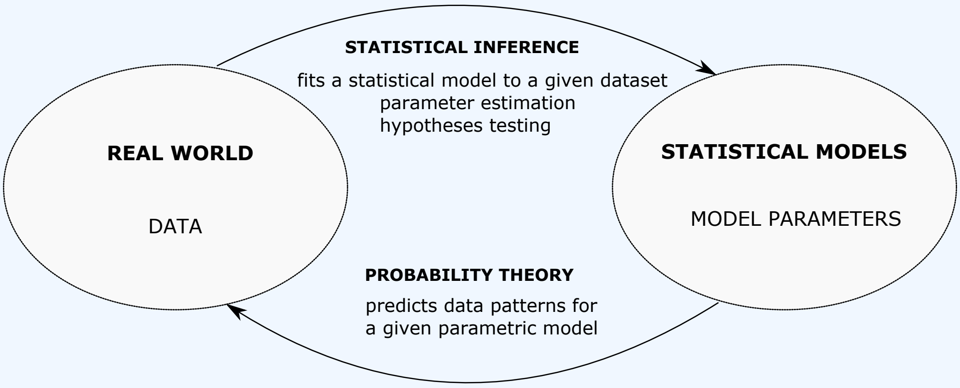
To Elin ♢ Camilla ♣ Diana ♡ Nelli ♠

Statistical Inference

Serik Sagitov

Chalmers University of Technology and the University of Gothenburg



Gothenburg, Sweden, March 7, 2023

Contents

Introduction

Statistical analysis consists of three parts: collection of data, summarising data, and making inferences. The main focus of this text is on the key tools of statistical inference: parameter estimation and hypothesis testing based upon properly collected, relatively small data sets. Special attention, therefore, is paid to the basic principles of experimental design: randomisation, blocking, and replication. The reader will get a deeper understanding of some traditional topics in mathematical statistics such as methods based on likelihood, aspects of experimental design, non-parametric testing, analysis of variance, introduction to Bayesian inference, chi-squared tests, and multiple regression.

It is expected that the reader is familiar with the basics of probability theory. The diagram on the front page illustrates the relationship between probability theory and statistical inference. An appreciation of these two different perspectives is provided by the following statements.

> PROBABILITY THEORY. Previous studies showed that the drug was 80% effective. We can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with probability 0.9999.
>
> STATISTICAL INFERENCE. It was observed that 78 out of 100 patients were cured. We are 95% confident that for other similar studies, the drug will be effective on between 69.9% and 86.1% of patients.

About the author

Serik Sagitov is a full professor of Mathematical Statistics at Chalmers University of Technology. His research papers are devoted to the theory and applications of branching and coalescent processes (click here). Please, use *serik@chalmers.se* if you wish to send to the author any comments or corrections concerning the text.

Acknowledgements

This text has grown from the lecture notes for the undergraduate course "Statistical Inference", given at the Chalmers University of Technology and the University of Gothenburg. The author is grateful to constructive feedback from the students who have taken this course. Special thanks go to professor Aila Särkkä for many insightful comments and corrections. The course material was originally based on the second edition of the book "Mathematical statistics and data analysis" by John Rice. A great number of examples and exercises included in this compendium are borrowed from Rice's textbook.

List of the course topics

Statistical inference and probability theory. Statistical models. Random variable, z-core. Correlation coefficient.
Population distribution. Population mean and standard deviation, population proportion.
Randomisation. Random sample, sampling with replacement.
Simple random sample, sampling without replacement.

Statistic, point estimate, sampling distribution.
Mean square error, systematic error and random (sampling) error.
Unbiased point estimate, consistent point estimate.
Sample mean, sample variance, sample standard deviation, sample proportion.
Finite population correction.
Standard error of the sample mean and sample proportion.
Approximate confidence intervals for the mean and the proportion.
Stratified random sampling. Optimal allocation of observations, proportional allocation.

Parametric models, population parameters.
Bernoulli, binomial, geometric, Poisson, and discrete uniform distributions.
Continuous uniform, exponential, gamma distributions.
Normal distribution, central limit theorem, continuity correction.
Method of moments for point estimation.
Maximum likelihood estimate (MLE). Likelihood function.
Normal approximation for the sampling distribution of MLE.

Sufficient statistics for population parameters.
Exact confidence intervals for the mean and variance. Chi-squared distribution, t-score and t-distribution.

Statistical hypotheses, simple and composite hypotheses, null hypothesis and alternative hypothesis.
Test statistic, rejection region. Two types of error.
Significance level, test power, effect size.
P-value of the test, one-sided and two-sided p-values.
Large-sample tests for the mean and for the proportion. One-sample t-test. The binomial test.
Nested hypotheses, generalised likelihood ratio test.
Chi-squared test of goodness of fit, its approximate nature. Multinomial distribution.

Bayes formulas for probabilities and densities. Prior and posterior distributions.
Conjugate priors. Normal-normal model.
Beta and Dirichlet distributions. Beta-binomial model and Dirichlet-multinomial model.
Bayesian estimation, zero-one loss function and squared error loss. Posterior risk.
Maximum aposteriori estimate (MAP) and posterior mean estimate (PME). Credibility interval.
Posterior odds. Bayesian hypotheses testing.

Empirical cumulative distribution function. Empirical variance.
Survival function and hazard function. Weibull distribution. Empirical survival function.
Histograms and kernel density estimates.
Population quantiles. Ordered sample and empirical quantiles.
QQ plots, normal QQ plot.
Coefficient of skewness and kurtosis. Light tails and heavy tails of probability distributions.
Leptokurtic and platykurtic distributions.

Population mean, mode, and median. Sample median, outliers.
Sign test and non-parametric confidence interval for the median.
Sample range, quartiles, interquartile range (IQR) and median of the absolute deviations (MAD). Boxplots.

Two independent samples versus paired samples.
Approximate confidence interval and large sample test for the mean difference.
Two-sample t-test, pooled sample variance.
Exact confidence interval for the mean difference. Transformation of variables.
Ranks versus exact measurements. Rank sum test. Signed rank test.
Approximate confidence interval for the difference between two proportions.
Large sample test for two proportions.
Fisher's exact test.
Simpson's paradox.

One-way ANOVA, sums of squares and mean squares.
Normal theory model, F-test, F-distribution, normal QQ plots for the residuals.
The problem of multiple comparison and multiple testing.
Simultaneous confidence intervals, Bonferroni's method and Tukey's method.
Two-way ANOVA, main effects and interaction. The noise and the residuals. Three F-tests.
Additive model. Randomised block design.
Non-parametric ANOVA: Kruskal-Wallis test, Friedman's test.

Categorical data.
Dichotomous data. Cross classification.
Chi-squared tests of homogeneity and independence.
Prospective and retrospective studies. Matched-pairs design, McNemar's test. Odds ratio.

Simple linear regression model. Normal equations. Least squares estimates.
Sample covariance, sample correlation coefficient.
Bias-corrected MLE of the noise variance. Coefficient of determination.
Confidence interval and hypotheses testing for the intercept and slope. Model utility test, t-value.
Prediction interval for a new observation.
Multiple regression. Design matrix.
Coefficient of multiple determination. Adjusted coefficient of multiple determination.
Collinearity problem.

Chapter 1

Parametric models

Notationally, we distinguish between random variables X, Y, Z and their realisations x, y, z by consistently using either capital or small letters. For a random variable X, we usually denote its mean value and variance by

$$\mu = \mathrm{E}(X), \quad \sigma^2 = \mathrm{Var}(X).$$

Recall that the mean (expected) value of X is computed as

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad \text{or} \quad \mathrm{E}(X) = \sum_{i=1}^{\infty} x_i p_i,$$

depending on whether the distribution of X is continuous, with the probability density function $f(x)$, or discrete, with the probability mass function $p_i = \mathrm{P}(X = x_i)$. The difference $(X - \mu)$ is called the deviation from the mean, and the variance of X is defined by

$$\mathrm{Var}(X) = \mathrm{E}((X - \mu)^2).$$

The square root of the variance, σ, is called the standard deviation of X.

By $X \sim \mathcal{F}(\mu, \sigma)$ we will mean that X has a distribution with mean μ and standard deviation σ. The symbol $\mathcal{F}(\mu, \sigma)$ will often denote the so-called population distribution.

The standardized version of X,

$$Z = \frac{X - \mu}{\sigma},$$

often called a *z-score*, is the result of a linear transformation of X such that $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$. Notice that an arbitrary linear transformation $Y = a + bX$, having mean $\mu_y = a + b\mu$ and standard deviation $\sigma_y = b\sigma$, brings the same z-score:

$$\frac{Y - \mu_y}{\sigma_y} = \frac{X - \mu}{\sigma} = Z.$$

It follows, that the z-score is a unit-less entity (for example, giving the same number of one's standardized height irrespectively of the length unit, be it centimeters or inches).

The covariance of two random variables (X_1, X_2), with means (μ_1, μ_2), is defined by

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}((X_1 - \mu_1)(X_2 - \mu_2)).$$

If (Z_1, Z_2) are standardised (X_1, X_2), then the correlation coefficient for (X_1, X_2) is given by $\rho = \mathrm{Cov}(Z_1, Z_2)$.

The correlation coefficient is used as a measure of linear dependence between a pair of random variables.

In general, the correlation coefficient belongs to the interval $-1 \leq \rho \leq 1$. In particular, if X_1 and X_2 are independent random variables, then $\rho = 0$, and if $X_2 = a + bX_1$, then $\rho = \pm 1$ depending on the sign of the non-zero slope b.

1.1 Normal distribution

A key parametric statistical model is the normal distribution which will be denoted by $\mathrm{N}(\mu, \sigma)$. A normally distributed random variable

$$X \sim \mathrm{N}(\mu, \sigma)$$

has mean $\mathrm{E}(X) = \mu$, variance $\mathrm{Var}(X) = \sigma^2$, and the probability density function

$$f(x) = \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

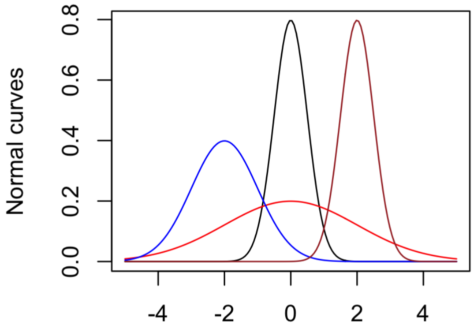
The corresponding z-score $Z = \frac{X-\mu}{\sigma}$ has the so-called standard normal distribution $\mathrm{N}(0, 1)$. The cumulative distribution function of $\mathrm{N}(0, 1)$ is assigned a special notation

$$\Phi(x) = \mathrm{P}(Z \le x), \quad -\infty < x < \infty.$$

The values of $\Phi(x)$ for $x \ge 0$, may be found using the table in Section 11.1. Importantly, if $X \sim \ \mathrm{N}(\mu, \sigma)$, then

$$\mathrm{P}(X \le x) = \mathrm{P}(\tfrac{X-\mu}{\sigma} \le \tfrac{x-\mu}{\sigma}) = \mathrm{P}(Z \le \tfrac{x-\mu}{\sigma}) = \Phi(\tfrac{x-\mu}{\sigma}).$$

The next figure presents four probability density functions $\mathrm{N}(0, 0.5)$, $\mathrm{N}(0, 2)$, $\mathrm{N}(-2, 1)$, and $\mathrm{N}(2, 0.5)$. All normal distribution curves have the same shape and differ only in their location parameter μ and their scale parameter σ.



The key role of the normal model in the statistical inference is due to the central limit theorem. Let $(X_1, \dots, X_n)$ be independent and identically distributed random variables with mean μ and variance σ^2. The arithmetical average

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

represents the sample mean. According to the central limit theorem, for a large sample size n,

$$\bar{X} \approx \mathrm{N}(\mu, \tfrac{\sigma}{\sqrt{n}}),$$

meaning that the random variable $\bar{X}$ is asymptotically normally distributed with mean μ and variance $\frac{\sigma^2}{n}$.

Signal and noise

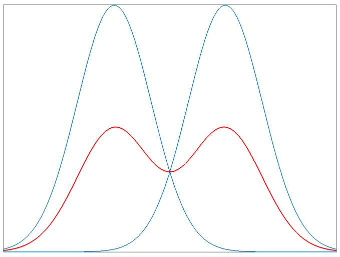
Suppose we want to measure the unknown signal value μ and each measurement Y has a random measurement error. Let

$$Y = \mu + \sigma Z, \quad Z \sim \ \mathrm{N}(0, 1),$$

where σZ is the measurement error or, in other words, the noise component. Then, the response variable $Y \sim \ \mathrm{N}(\mu, \sigma)$. In view of the central limit theorem, it is natural to model the random noise using the $\mathrm{N}(0, \sigma)$-model, since the noise is an accumulation of all external factors neither of which having a dominating effect on the response variable. In such a setting, we refer to σ as the *size of the noise.*

Mixtures of normal distribution

A motivating example for the mixture model is the hight of people in a population consisting of women and men. Let $\mathrm{N}(\mu_1, \sigma_1)$ be the distribution of women's height and $\mathrm{N}(\mu_2, \sigma_2)$ be the distribution of men's height. Then, the mixed population distribution describes the outcome of a two step random experiment: first toss a coin for choosing index i to be either 1 or 2, then generate a value using $\mathrm{N}(\mu_i, \sigma_i)$. The resulting density function has the shape of a "camel curve" as illustrated below (red line).



More generally, suppose that we are given $k \geq 2$ normally distributed random variables

$$X_1 \sim \mathrm{N}(\mu_1, \sigma_1), \ldots, X_k \sim \mathrm{N}(\mu_k, \sigma_k).$$

Define the response variable Y as X_i with a random index i taking one of the values $1, \ldots, k$ with probabilities $w_1, \ldots, w_k$, so that

$$w_1 + \ldots + w_k = 1.$$

This yields the following expressions for the mean $\mu = \mathrm{E}(Y)$ and variance $\sigma^2 = \mathrm{Var}(Y)$

$$\mu = w_1\mu_1 + \ldots + w_k\mu_k,$$
$$\sigma^2 = \sum_{j=1}^{k} w_j(\mu_j - \mu)^2 + \sum_{j=1}^{k} w_j\sigma_j^2.$$

The above expression for σ^2 is due to the *law of total variance*, see Wikipedia, which recognises two sources of variation

variation between the strata $\sum_{j=1}^k w_j(\mu_j - \mu)^2$,

variation within the strata $\sum_{j=1}^k w_j\sigma_i^2$.

1.2 One-way and two-way layout models

Suppose the expectation μ_i of the response variable is a function of the level i for a single main factor A having I different levels:

$$Y_i \sim \ \mathrm{N}(\mu_i, \sigma), \quad i = 1, \ldots, I.$$

It is helpful to represent the population means μ_i as the sum

$$\mu_i = \mu + \alpha_i, \quad \alpha_i = \mu_i - \mu,$$

of the so called *grand mean*

$$\mu = \frac{\mu_1 + \ldots + \mu_I}{I}$$

and the effect α_i of the main factor A at the level i. Observe that $\sum_{i=1}^I \alpha_i = 0$, meaning the total effect of the factor A is zero.

In the case of two main (categorical) factors, with factor A having I different levels, and factor B having J different levels, assume that the response variable depends on the combination of the levels of the two main factors in the following way

$$Y_{ij} \sim \ \mathrm{N}(\mu_{ij}, \sigma), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

where

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij},$$

is the sum of the grand mean μ, the main effect α_i of the factor A at the level i, the main effect β_j of the factor B at the level j, and the interaction δ_{ij} of the two main factors. Here it is assumed that

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0,$$

and

$$\sum_{i=1}^{I} \delta_{ij} = 0, \quad j = 1, \ldots, J, \qquad \sum_{j=1}^{J} \delta_{ij} = 0, \quad i = 1, \ldots, I.$$

In general, at different combinations of levels (i, j) of the two factors may interact either negatively, $\delta_{ij} < 0$, or positively, $\delta_{ij} > 0$.

Additive model

In the special case, with $\delta_{ij} = 0$ for all (i, j), the model claims that there is no interaction and the main factors contribute additively:

$$\mu_{ij} = \mu + \alpha_i + \beta_j.$$

Example: pay gap

Let the response variable be the salary of a person chosen from a large population. Factor A is person's sex having $I = 2$ levels: $i = 1$ for a female and $i = 2$ for a male. Factor B is person's profession having say $J = 20$ levels, where $j = 1$ is a farmer, $j = 2$ is a police officer, $j = 3$ is a doctor, and so on. In this example, the difference $\alpha_1 - \alpha_2$ represents the pay gap between women and men.

1.3 Sample mean, sample variance, and t-distributions

Suppose we are given a vector $(X_1, \ldots, X_n)$ of independent random variables having the same distribution $\mathcal{F}(\mu, \sigma)$. A realisation of this vector $(x_1, \ldots, x_n)$ will be called (with a slight abuse of the established terminology) a random sample drawn from the population distribution $\mathcal{F}(\mu, \sigma)$ with population mean μ and population standard deviation σ. For the given sample $(x_1, \ldots, x_n)$ define the sample mean, sample variance, and sample standard deviation by

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}.$$

An alternative formula for the sample variance

$$s^2 = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2), \quad \overline{x^2} = \frac{x_1^2 + \ldots + x_n^2}{n},$$

is often more convenient to use for pen and paper calculations. The sample mean $\bar{x}$ and sample variance s^2 are realisations of the random variables

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which have the following means and variances

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \tfrac{\sigma^2}{n}, \quad \mathrm{E}(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \tfrac{\sigma^4}{n}\Big(\mathrm{E}(\tfrac{X-\mu}{\sigma})^4 - \tfrac{n-3}{n-1}\Big).$$

If the population distribution is normal

$$\mathcal{F}(\mu, \sigma) = \mathrm{N}(\mu, \sigma),$$

then by *Cochran's theorem*, see Wikipedia, the so-called *t-score* of the random sample

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

has the t-distribution with $n-1$ degrees of freedom. The density function of the t-distribution with $k \geq 1$ degrees of freedom

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\Big(1 + \tfrac{x^2}{k}\Big)^{-\frac{k+1}{2}}, \quad -\infty < x < \infty$$

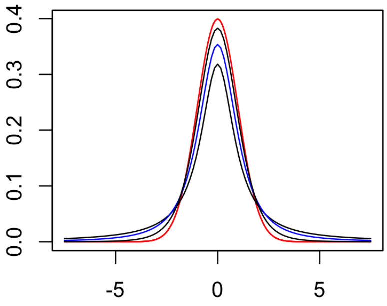
involves the gamma function

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$$

which is an extension of the factorial function to the non-integer positive numbers a, in that

$$\Gamma(k) = (k-1)!, \quad k = 1, 2, \ldots$$

The t-distribution curve with $k \geq 3$ of degrees of freedom looks similar to the $\mathrm{N}(0,1)$-curve, being symmetric around zero and having the standard deviations $\sqrt{\frac{k}{k-2}}$ which is larger than 1. The figure below depicts three t-distribution curves together with the $\mathrm{N}(0,1)$-curve (in red). The degrees of freedom used in the figure are $k = 1$, $k = 2$ in blue, and $k = 6$. The t-distribution with $k = 1$ degree of freedom has undefined mean value and infinite variance. The t-distribution with $k = 2$ degrees of freedom has zero mean and infinite variance.



The connection between the t-distribution and the standard normal distribution can be described in the following way: if $Z, Z_1, \ldots, Z_k$ are independent random variables with $\mathrm{N}(0,1)$-distribution, then

$$\frac{Z}{\sqrt{(Z_1^2 + \ldots + Z_k^2)/k}} \sim t_k.$$

1.4 Gamma, exponential, and chi-squared distributions

The gamma distribution $\mathrm{Gam}(\alpha, \lambda)$ is a continuous distribution described by two parameters: the shape parameter $\alpha > 0$ and the inverse scale (or the rate) parameter $\lambda > 0$. Its probability density function has the form

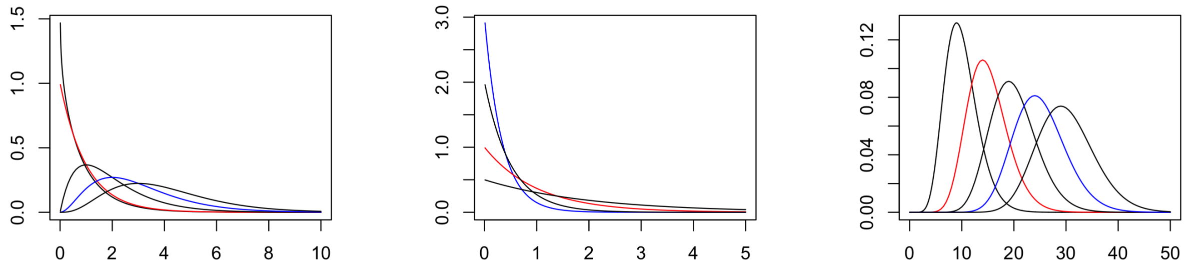
$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

The next figure depicts

on the left panel, the gamma densities with $\lambda = 1$ and $\alpha = 0.9, 1, 2, 3, 4$,

on the middle panel, the gamma densities with $\alpha = 1$ and $\lambda = 0.5, 1, 2, 3$,

on the right panel, the gamma densities with $\lambda = 1$ and $\alpha = 10, 15, 20, 25, 30$.



The gamma distribution model, despite being restricted to positive values, is more flexible than the normal distribution model since for the different values of α, the density curves have different shapes. In particular, if $\alpha = 1$, then we obtain the exponential distribution (see the middle panel above)

$$\mathrm{Gam}(1, \lambda) = \mathrm{Exp}(\lambda).$$

Moreover, if $X_i \sim \mathrm{Exp}(\lambda)$, $i = 1, \ldots, k$ are independent, then

$$X_1 + \ldots + X_k \sim \mathrm{Gam}(k, \lambda), \quad k = 1, 2, \ldots$$

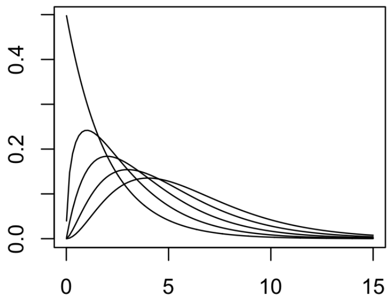
The mean and variance of the gamma distribution are

$$\mu = \tfrac{\alpha}{\lambda}, \quad \sigma^2 = \tfrac{\alpha}{\lambda^2}.$$

For large values of the shape parameter, there is a useful normal approximation for the gamma distribution:

$$\mathrm{Gam}(\alpha, \lambda) \approx \mathrm{N}(\tfrac{\alpha}{\lambda}, \tfrac{\sqrt{\alpha}}{\lambda}), \qquad \alpha \gg 1.$$

The chi-squared distribution with k degrees of freedom is the gamma distribution with $\alpha = \frac{k}{2}, \lambda = \frac{1}{2}$. The figure below depicts the chi-squared distribution densities with $k = 2, 3, 4, 5, 6$ degrees of freedom.



The chi-squared distribution is connected to the standard normal distribution as follows: if $Z_1, \ldots, Z_k$ are independent random variables with $\mathrm{N}(0,1)$-distribution, then

$$Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2.$$

Importantly, if $(X_1, \ldots, X_n)$ are independent and random variables each having the $\mathrm{N}(\mu, \sigma)$ distribution, then

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Here, the number of degrees of freedom is $n - 1$ instead of n, because one degree of freedom is consumed after μ being replaced by $\bar{X}$.

1.5 Bernoulli, binomial, and multinomial distributions

Let X be the outcome of a Bernoulli trial with probability of success p, meaning that

$$\mathrm{P}(X = 1) = p, \quad \mathrm{P}(X = 0) = 1 - p.$$

In this case, we write $X \sim \mathrm{Bin}(1, p)$ and say that X has the Bernoulli distribution with parameter $p \in [0, 1]$. The Bernoulli model is used for describing dichotomous data, when observations have two possible outcomes: female or male, heads or tails, passed or failed. Under such a dichotomy, one outcome is usually called a success and the other failure, and the value $x = 1$ is assigned to the successful outcome. The mean and variance of $X \sim \mathrm{Bin}(1, p)$ are

$$\mu = p, \quad \sigma^2 = p(1 - p).$$

If X is the sum of outcomes of n independent Bernoulli trials with probability p of success, then its distribution is called the binomial distribution $\mathrm{Bin}(n, p)$. This is a discrete distribution with the probability mass function

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, \ldots, n,$$

yielding the following formulas for the mean and variance

$$\mu = np, \quad \sigma^2 = np(1 - p).$$

The normal approximation for the binomial distribution

$$\mathrm{Bin}(n, p) \approx \mathrm{N}(np, \sqrt{np(1 - p)}).$$

is an instrumental example of the central limit theorem. The rule of thumb says that this normal approximation is good enough if both $np \geq 5$ and $n(1 - p) \geq 5$, so that

$$\mathrm{P}(X \leq x) \approx \Phi\Big(\tfrac{x - np}{\sqrt{np(1-p)}}\Big).$$

Continuity correction

For smaller values of n, this approximation is improved with help of the *continuity correction* trick, which in view of the equality

$$\mathrm{P}(X \leq x) = \mathrm{P}(X < x + 1),$$

suggests replacing x by $x + \frac{1}{2}$ on the right hand side of the approximation formula

$$\mathrm{P}(X \leq x) \approx \Phi\Big(\tfrac{x - np}{\sqrt{np(1-p)}}\Big),$$

resulting in

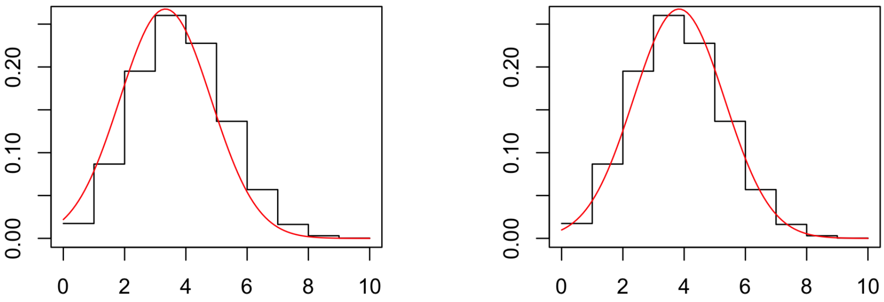
$$\mathrm{P}(X \leq x) \approx \Phi\Big(\tfrac{x + \frac{1}{2} - np}{\sqrt{np(1-p)}}\Big),$$

or similarly,

$$\mathrm{P}(X < x) \approx \Phi\Big(\tfrac{x - \frac{1}{2} - np}{\sqrt{np(1-p)}}\Big).$$

Example

To illustrate, we plot the $\mathrm{Bin}(10, 1/3)$ distribution together with its normal approximation: on the left panel without the continuity correction and on the right panel with the continuity correction. Observe that $n = 10$ is so small that with $p = 1/3$ we have $np = 3.33$ which smaller than the recommended lower bound 5. Still, the normal approximation with the continuity corrections is quite close.



The multinomial distribution $(X_1,\ldots,X_r) \sim \mathrm{Mn}(n;p_1,\ldots,p_r)$ is defined by

$$\mathrm{P}(X_1 = x_1,\ldots,X_r = x_r) = \binom{n}{x_1,\ldots,x_r} p_1^{x_1}\ldots p_r^{x_r},$$

where

$$x_i = 0,\ldots,n, \quad i = 1,\ldots,r,$$

and $(p_1,\ldots,p_r)$ is a vector of probabilities such that

$$p_1 + \ldots + p_r = 1.$$

This is an extension of the binomial distribution $\mathrm{Bin}(n,p) = \mathrm{Mn}(n;p,1-p)$. The $\mathrm{Mn}(n;p_1,\ldots,p_r)$ distribution describes the outcome of n independent trials with r possible outcomes labeled by $i = 1,\ldots,r$. If each trial outcome has distribution $(p_1,\ldots,p_r)$ over the set of possible labels $\{1,\ldots,r\}$, then X_i should be treated as the number of trials with the outcome labeled by i. We have

$$X_1 + \ldots + X_r = n,$$

the marginal distribution of X_i is binomial $X_i \sim \mathrm{Bin}(n,p_i)$, and the different counts (X_i,X_j) are negatively correlated:

$$\mathrm{Cov}(X_i,X_j) = -np_ip_j, \quad i \neq j.$$

1.6 Poisson, geometric, and hypergeometric distributions

The Poisson distribution $X \sim \mathrm{Pois}(\mu)$ is a discrete distribution with

$$\mathrm{P}(X = x) = \frac{\mu^x}{x!}e^{-\mu}, \quad x = 0,1,\ldots, \quad \mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \mu.$$

The Poisson distribution is obtained as an approximation for the $\mathrm{Bin}(n,p)$ distribution in the case

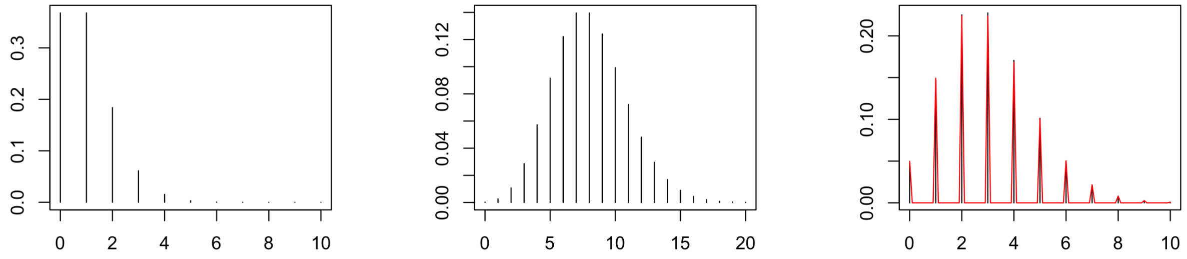
$$n \to \infty,\ p \to 0,\ \text{and } np \to \mu.$$

It is used to describe the number of rear events (like accidents) observed during a given time interval. The next figure depicts

on the left panel, the Poisson distribution with $\mu = 1$,

on the middle panel, the Poisson distribution with $\mu = 8$,

on the right panel, the Poisson distribution with $\mu = 3$ in red is compared to the $\mathrm{Bin}(100,0.03)$ distribution.



Geometric distribution

Consider a sequence of independent Bernoulli trials with probability p of success. The geometric distribution is either one of two discrete probability distributions:

the distribution of the number X of trials needed to get one success,

the distribution of the number $Y = X - 1$ of failures before the first success.

Which of these is called the geometric distribution is a matter of convention and convenience. Often, the name shifted geometric distribution is adopted for the distribution of X supported on the set $\{1,2,\ldots\}$. To avoid ambiguity, in this text, we say that X has the geometric distribution with parameter p and write $X \sim \mathrm{Geom}(p)$ if

$$\mathrm{P}(X = x) = (1-p)^{x-1}p, \quad x = 1,2,\ldots,$$

yielding the mean and variance formulas

$$\mu = \tfrac{1}{p}, \qquad \sigma^2 = \tfrac{1-p}{p^2}.$$

Like its continuous analogue, the exponential distribution, the geometric distribution is memoryless: the die one throws or the coin one tosses does not have a memory of how many failures have been observed so far.

Hypergeometric distribution

The hypergeometric distribution $X \sim \mathrm{Hg}(N, n, p)$ describes the number x of black balls among n balls drawn without replacement from a box with N balls, of which

$B = Np$ balls are black and

$W = N(1-p)$ balls are white.

In this case, X is the number of successes in n Bernoulli trials which depend on each other. The distribution of X is given by the formula

$$\mathrm{P}(X = x) = \frac{\binom{B}{x}\binom{W}{n-x}}{\binom{N}{n}},$$

for the integer numbers x satisfying

$$\max(0, n - W) \le x \le \min(n, B).$$

The mean and variance of X are

$$\mu = np, \quad \sigma^2 = np(1-p)\tfrac{N-n}{N-1}.$$

Compared to the variance of the $\mathrm{Bin}(n, p)$ distribution, the last formula contains the factor

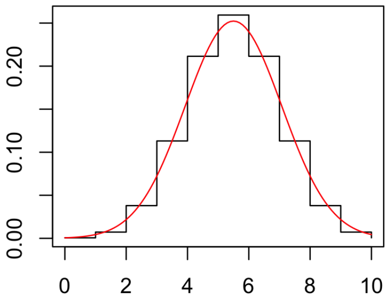
$$\tfrac{N-n}{N-1} = 1 - \tfrac{n-1}{N-1},$$

which is called the *finite population correction* factor. With a small fraction value n/N, the finite population correction is close to 1 and $\mathrm{Hg}(N, n, p) \approx \mathrm{Bin}(n, p)$. One may say that the binomial distribution is a version of the hypergeometric distribution with the infinite population size.

Despite the dependence between the drawings without replacement, there is a normal approximation also for the hypergeometric distribution:

$$\mathrm{Hg}(N, n, p) \approx \mathrm{N}(\mu, \sigma), \qquad \mu = np, \quad \sigma = \sqrt{np(1-p)}\sqrt{1 - \tfrac{n-1}{N-1}},$$

which is recommended to be applied provided $np \ge 5$ and $n(1-p) \ge 5$. On the figure below the $\mathrm{Hg}(100, 10, 0.5)$ distribution is compared to its normal approximation with the continuity correction. Again, with the continuity correction the normal approximation works well even for smaller values of np and $n(1-p)$.



1.7 Exercises

Problem 1

For any pair of random variables (X_1, X_2) with means (μ_1, μ_2), show that

$$\mathrm{Var}(X_i) = \mathrm{E}(X_i^2) - \mu_i^2, \quad \mathrm{Cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mu_1\mu_2.$$

Problem 2

Let

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1}, \quad Z_2 = \frac{X_2 - \mu_2}{\sigma_2},$$

be the standardised versions of X_1 and X_2. Verify that $\mathrm{E}(Z_i) = 0$ and $\mathrm{Var}(Z_i) = 1$. Show that the correlation coefficient for X_1 and X_2 is given by

$$\rho = \mathrm{E}\left(Z_1Z_2\right),$$

and explain in what sense it is a dimensionless quantity.

Problem 3

For $(X_1, \dots, X_r) \sim \text{Mn}(n; p_1, \dots, p_r)$, what is the distribution of the sum $X_i + X_j$ assuming $i \neq j$?

Problem 4

Let $X \sim \text{Gam}(\alpha, \lambda)$. To see that the parameter λ influences only the scale of the gamma distribution, show that the scaled random variable $Y = \lambda X$ has the gamma distribution $\text{Gam}(\alpha, 1)$.

Problem 5

Show that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

Problem 6

The average number of goals in a World Cup soccer match is approximately 2.5 and the Poisson model is appropriate. Compute the probabilities of k goals in a match for $k = 0, 1, 2$.

Problem 7

Show that for large N,

$$\text{Hg}(N, n, p) \approx \text{Bin}(n, p).$$

Problem 8

Consider a truncated version of the geometric distribution $\text{Geom}(n, p)$ such that for $X \sim \text{Geom}(n, p)$,

$$\text{P}(X = x) = (1 - p)^x p, \quad x = 0, 1, \dots, n - 1, \quad \text{P}(X = n) = 1 - \sum_{x=1}^{n-1} \text{P}(X = x),$$

and compute $\text{P}(X = n)$.

Chapter 2

Random sampling

Statistical inference is the use of data analysis for inferring relevant statistical patterns in a large population with help of a random sample drawn from the population in question.

> Picking at random one element from the population, produces a realisation x of a random variable $X \sim \mathcal{F}(\mu, \sigma)$ having the population distribution.

In many situations, studying the population distribution by enumeration is either very expensive or even impossible. Luckily, a good guess is available by studying a sample of n observations $(x_1, \ldots, x_n)$ drawn independently from the population distribution $\mathcal{F}(\mu, \sigma)$. Such a random sample is a single realisation of the vector $(X_1, \ldots, X_n)$ of independent and identically distributed random variables. If the sampling experiment is repeated, the new realization $(x'_1, \ldots, x'_n)$ will differ from $(x_1, \ldots, x_n)$.

> *Randomisation* in sampling protects against investigator's biases even unconscious.

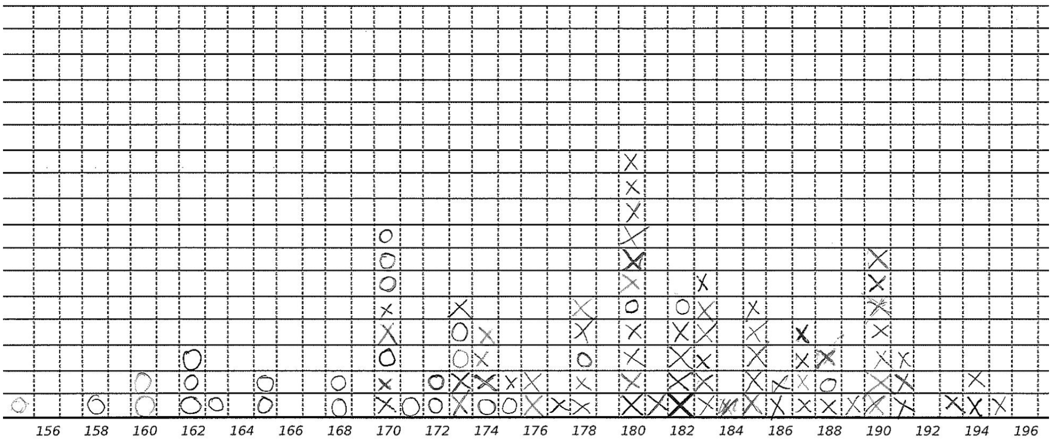
Any function $g(x_1, \ldots, x_n)$ of the sample data is called a *statistic*. The most important examples of statistics are the sample mean and sample variance

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad s^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}.$$

Example: in class experiment

The figure below presents the data of heights and gender for the students attending the course Statistical Inference at a certain year. The collected height values form a sample drawn from the population distribution of the heights of Gothenburg students for that year.

- Can this dataset be viewed as a random sample, in the sense that the students are drawn independently at random from the population of Gothenburg students?
- How would you estimate the population mean and variance of the heights using the collected data?
- How would you estimate the population proportion of women?



Report your hight in cm by adding **O** if you are female and **X** if you are male

2.1 Point estimation

Suppose the population distribution of interest is a gamma distribution

$$\mathcal{F}(\mu, \sigma) = \text{Gam}(\alpha, \lambda)$$

with unknown parameters. Given a random sample $(x_1, \ldots, x_n)$ drawn from $\mathrm{Gam}(\alpha, \lambda)$ distribution, one can try to estimate the shape parameter α using a relevant statistic $g(x_1, \ldots, x_n)$.

More generally, to estimate a population parameter θ based on a given random sample $(x_1, \ldots, x_n)$, we need a sensible point estimate $\hat{\theta} = g(x_1, \ldots, x_n)$. Observe, that in the same way as $(x_1, \ldots, x_n)$ is a realisation of a random vector $(X_1, \ldots, X_n)$, the point estimate $\hat{\theta}$ is a realisation of a random variable

$$\widehat{\Theta} = g(X_1, \ldots, X_n),$$

which we will call a point *estimator* of θ. The distribution of the random variable $\widehat{\Theta}$ is called the sampling distribution of the point estimator. The quality of the the point estimator $\widehat{\Theta}$ is measured by the mean square error

$$\mathrm{E}((\widehat{\Theta} - \theta)^2) = \mathrm{Var}(\widehat{\Theta}) + (\mathrm{E}(\widehat{\Theta}) - \theta)^2$$

which is the sum of two components involving

the bias size $\mathrm{E}(\widehat{\Theta}) - \theta$, measuring the lack of accuracy (systematic error),
$\mathrm{Var}(\widehat{\Theta})$, measuring the lack of precision (random error).

If the mean square error vanishes $\mathrm{E}((\widehat{\Theta} - \theta)^2) \to 0$ as $n \to \infty$, the point estimate $\hat{\theta}$ is called *consistent.* If $\mathrm{E}(\widehat{\Theta}) = \theta$, the estimate is called *unbiased.* The standard deviation $\sigma_{\widehat{\Theta}} = \sqrt{\mathrm{Var}(\widehat{\Theta})}$ of the estimator $\widehat{\Theta}$ is called the *standard error* of the point estimate $\hat{\theta}$.

The estimated standard error $s_{\hat{\theta}}$ of the point estimate $\hat{\theta}$
is a point estimate of $\sigma_{\widehat{\Theta}}$ computed from the data.

Remark on termilogy

Consistency is a feature of estimator. Sometimes, we write consistent estimate and mean that the estimate is a value of a consistent estimator. The same holds for unbiased estimators and estimates.

Sample mean, sample variance, and sample standard deviation

In view of

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \tfrac{\sigma^2}{n}, \quad \mathrm{E}(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \tfrac{\sigma^4}{n}\Big(\mathrm{E}(\tfrac{X-\mu}{\sigma})^4 - \tfrac{n-3}{n-1}\Big),$$

we conclude that the sample mean $\bar{x}$ is an unbiased and consistent estimate of the population mean $\theta = \mu$. Furthermore, the sample variance s^2 is an unbiased and consistent estimate for the population variance $\theta = \sigma^2$.

The estimated standard error for the sample mean $\bar{x}$ is given by $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

Notice that the sample standard deviation s systematically underestimates the population standard deviation σ since $\mathrm{E}(S) < \sigma$, provided $\mathrm{Var}(S) > 0$. To see this, observe that $(\mathrm{E}(S))^2 < \mathrm{E}(S^2)$, where $\mathrm{E}(S^2) = \sigma^2$. However, s is an asymptotically unbiased and consistent estimate of σ.

2.2 Approximate confidence intervals

By the central limit theorem, for the large sample sizes n,

$$\bar{X} \approx \mathrm{N}(\mu, \tfrac{\sigma}{\sqrt{n}}),$$

or in terms of the z-score

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx \mathrm{N}(0, 1).$$

Due to the consistency property $S \approx \sigma$ it follows that in terms of the t-score we have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx \mathrm{N}(0, 1),$$

This yields

$$\mathrm{P}(\bar{X} - zS/\sqrt{n} < \mu < \bar{X} + zS/\sqrt{n})) \approx \mathrm{P}(-z\sigma/\sqrt{n} < \bar{X} - \mu < z\sigma/\sqrt{n}) \approx 2(1 - \Phi(z)),$$

giving the following formula of an approximate $100(1–\alpha)\%$ two-sided confidence interval for the population mean μ:

$$I_\mu \approx \bar{x} \pm z(\tfrac{\alpha}{2}) \cdot s_{\bar{x}}$$

or equivalently,

$$I_\mu \approx \bar{x} \pm z(\tfrac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}.$$

Here $z(\alpha)$ is obtained from the normal distribution table using the relation

$$\Phi(z(\alpha)) = 1 - \alpha, \quad \alpha \in (0,1),$$

so that for example,

100(1–α)%	68%	80%	90%	95%	99%	99.7%
$z(\frac{\alpha}{2})$	1.00	1.28	1.64	1.96	2.58	3.00

The higher the confidence level 100(1–α)%, the wider is the confidence interval I_μ.
On the other hand, the larger the sample size n, the narrower is I_μ.

The exact meaning of the confidence level is a bit tricky. It is important to realise that the source of randomness in the expression $\bar{x} \pm z(\frac{\alpha}{2}) \cdot s_{\bar{x}}$ is in the sampling procedure. For a given sample $(x_1, \ldots, x_n)$, the concrete interval I_μ is a realisation of a random interval $\mathcal{I}_\mu$, such that

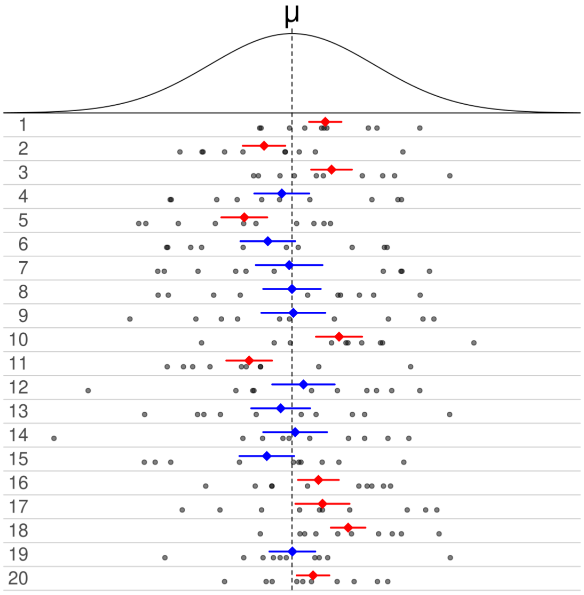
$$\mathrm{P}(\text{random interval } \mathcal{I}_\mu \text{ covers the point } \mu) \approx 1 - \alpha.$$

For example, out of a hundred 95% confidence intervals I_μ computed for 100 samples, on average 95 intervals are expected to cover the true value of μ. Notice that in this case, the random number of the confidence intervals covering μ has distribution $\mathrm{Bin}(100, 0.95)$ which is approximately normal $\mathrm{N}(95, 2.18)$.

Example: 50% confidence level

The figure below presents a case for the 50% confidence interval. Here 20 samples of size $n = 8$ were drawn from a population with mean μ. The 20 samples produce different sample means and sample variances. The resulting 20 confidence intervals for the mean μ have different middle points and widths:

some of the intervals cover the mean value (blue intervals),
some of the intervals fail to capture the mean value (red intervals).



2.3 Simple random sample

A finite population of size N can be viewed as a set of N elements characterised by numerical values $x \in \{a_1, a_2, \ldots, a_N\}$. The corresponding population distribution $\mathcal{F}(\mu, \sigma)$ is then given by the probability mass function

$$\mathrm{P}(X = x) = \frac{N_x}{N},$$

where N_x is the number of elements labeled by $a_i = x$. There are two basic ways of drawing a random sample of size n from a population of size N:

- *sampling with replacement* produces what we call a random sample consisting of independent and identically distributed observations,
- sampling without replacement produces a so called *simple random sample* having identically distributed but dependent observations.

Notice that if the ratio $\frac{n}{N}$ is small, then the two approaches are almost indistinguishable.

In this section we consider a simple random sample $(x_1, \ldots, x_n)$ with dependent observations. In this case the sample mean $\bar{x}$ is again an unbiased and consistent estimate for the population mean, such that

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \tfrac{\sigma^2}{n}(1 - \tfrac{n-1}{N-1}).$$

Here N is the finite population size and the finite population correction

$$1 - \tfrac{n-1}{N-1} = \tfrac{N-n}{N-1}$$

reflects the negative correlation between observations due to sampling without replacement.

Observe that the sample variance s^2 becomes a biased estimate of σ^2. Indeed, since

$$\begin{aligned}\mathrm{E}(S^2) &= \tfrac{n}{n-1}\mathrm{E}\Big(\tfrac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2\Big) = \tfrac{n}{n-1}(\mathrm{E}(X^2) - \mathrm{E}(\bar{X}^2))\\ &= \tfrac{n}{n-1}(\sigma^2 + \mu^2 - \tfrac{\sigma^2}{n}(1 - \tfrac{n-1}{N-1}) - \mu^2) = \tfrac{n}{n-1}(\sigma^2 - \tfrac{\sigma^2}{n}(1 - \tfrac{n-1}{N-1})) = \sigma^2 \tfrac{N}{N-1},\end{aligned}$$

we find that

$$\mathrm{E}(S^2) = \sigma^2 \tfrac{N}{N-1}.$$

Replacing σ^2 by $s^2\frac{N-1}{N}$ in the formula $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$, we obtain the following unbiased estimate of $\mathrm{Var}(\bar{X})$:

$$s_{\bar{x}}^2 = \tfrac{s^2}{n}\tfrac{N-1}{N}(1 - \tfrac{n-1}{N-1}) = \tfrac{s^2}{n}(1 - \tfrac{n}{N}).$$

Thus, for the sampling without replacement, the formula for the estimated standard error of the sample mean $\bar{x}$ takes the form

$$s_{\bar{x}} = \tfrac{s}{\sqrt{n}}\sqrt{1 - \tfrac{n}{N}}.$$

With this new formula for the estimated standard error $s_{\bar{x}}$, the formula of an approximate $100(1-\alpha)\%$ confidence interval

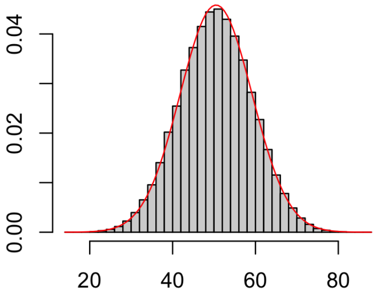
$$I_\mu \approx \bar{x} \pm z(\alpha/2)s_{\bar{x}},$$

remains to be valid even for the sampling without replacement, due to the central limit theorem under weak dependence.

The figure below plots the sampling distribution of the sample mean and its normal approximation. Here the sampling is performed without replacement from the population

$$\{a_1, a_2, \ldots, a_N\} = \{1, 2, \ldots, 100\}$$

using the sample size $n = 10$.



2.4 Dichotomous data

Consider the important case when the population distribution is a Bernoulli distribution

$$\mathcal{F}(\mu,\sigma) = \mathrm{Bin}(1,p)$$

with the unknown parameter p. A random sample drawn from the Bernoulli distribution consists of 0 or 1 values $x_i \in \{0,1\}$, therefore, the data $(x_1,\ldots,x_n)$ will be called *dichotomous*. This model can be used for categorical data, where the observed values are of non-numerical form, like male or female, after converting the non-numerical values to the 0-1 format: female = 1, male = 0.

The defining parameter of the Bernoulli model

$$p = \mathrm{P}(X=1),$$

will be called called the population proportion. The population proportion p defines both population mean and variance by

$$\mu = p, \quad \sigma^2 = p(1-p),$$

and the sample mean turns into a sample proportion $\hat{p} = \bar{x}$. The sample proportion is an unbiased and consistent estimate of p. Since $x_i^2 = x_i$ for $x_i \in \{0,1\}$, we find that the sample variance takes the form

$$\begin{aligned} s^2 &= \frac{(x_1-\hat{p})^2+\ldots+(x_n-\hat{p})^2}{n-1} = \frac{x_1^2-2x_1\hat{p}+\hat{p}^2+\ldots+x_n^2-2x_n\hat{p}+\hat{p}^2}{n-1} \\ &= \frac{x_1-2x_1\hat{p}+\hat{p}^2+\ldots+x_n-2x_n\hat{p}+\hat{p}^2}{n-1} = \frac{n\hat{p}^2-(x_1+\ldots+x_n)\hat{p}}{n-1} \\ &= \frac{n\hat{p}(1-\hat{p})}{n-1}. \end{aligned}$$

The estimated standard error for the sample proportion $\hat{p}$ is $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$.

Simple random sample

If sampling from a dichotomous population of size N is performed without replacement, the formula for the estimated standard error of $\hat{p}$ becomes

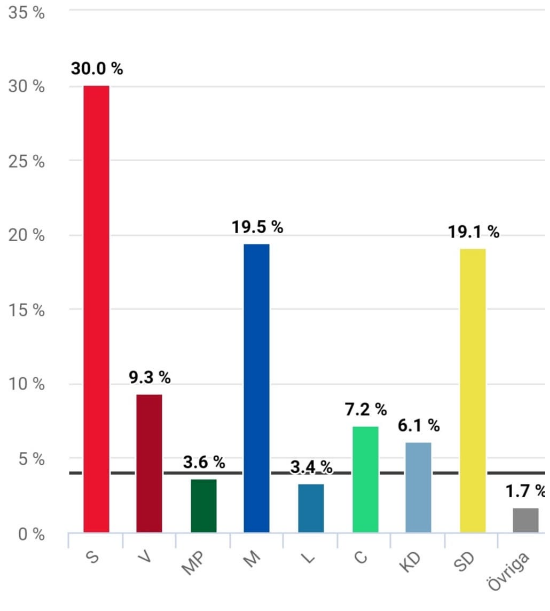
$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\tfrac{n}{N}}.$$

The following confidence interval formula is valid both for sampling with replacement and for sampling without replacement.

An approximate 100(1–α)% two-sided confidence interval for p is given by $I_p \approx \hat{p} \pm z(\frac{\alpha}{2}) \cdot s_{\hat{p}}$

Example: opinion polls

Swedish population of eligible voters has the size N of 7 millions. You are interested in the current attitude among voters towards a certain Swedish political party, quantified by a relevant population proportion p. How informative would be the result of a survey involving $n = 2000$ people? An intuitive response is that asking 2000 people could not accurately reflect the overall opinion of 7 000 000 people. However, if the sample is truly random, the error in $\hat{p}$ is quite small.



Indeed, the standard error of the sample proportion $\hat{p}$ with $n = 2000$ and $N = 7000000$ equals

$$s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1 - \tfrac{n}{N}} = 0.0224\sqrt{\hat{p}(1-\hat{p})},$$

which is less or equal than 1.12%, since $\hat{p}(1-\hat{p}) \le 0.25$.

2.5 Stratified random sampling

Given additional information on the population structure, one can reduce the sampling error using the method of stratified sampling. Assume that a population of size N consists of k strata of sizes $N_1, \ldots, N_k$, so that $N = N_1 + \ldots + N_k$. Suppose the strata fractions $w_j = N_j/N$ are known. A simple example of a stratified population is the Swedish population divided in two strata: the subpopulation of females and the subpopulation of males, so that $k = 2$ and $w_1 = w_2 = 0.5$.

In terms of the unknown strata means and standard deviations

$$(\mu_j, \sigma_j),\ j = 1, \ldots, k,$$

we get the following expressions for the population mean and variance

$$\mu = w_1\mu_1 + \ldots + w_k\mu_k,$$
$$\sigma^2 = \overline{\sigma^2} + \sum_{j=1}^{k} w_j(\mu_j - \mu)^2.$$

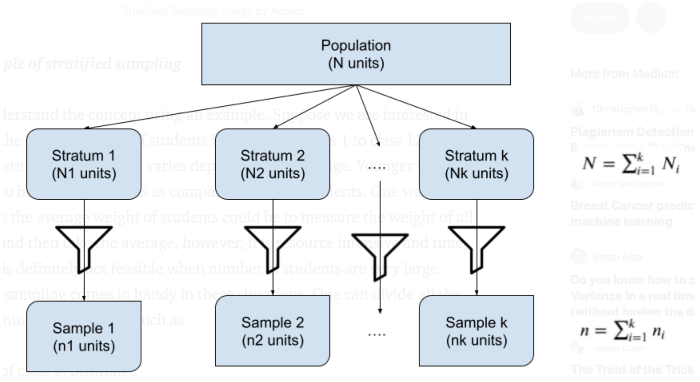
In view of

$$w_1 + \ldots + w_k = 1,$$

the above formulas for μ and σ^2 are obtained using the *law of total expectation* and *law of total variance*, see Wikipedia. The expression for σ^2 is due to the law of total variance, with

$$\overline{\sigma^2} = w_1\sigma_1^2 + \ldots + w_k\sigma_k^2$$

being the average variance, see the end of Section 1.1.



The stratified random sampling procedure consists of taking k independent random samples from each stratum with sample sizes $(n_1, \ldots, n_k)$ and sample means $\bar{x}_1, \ldots, \bar{x}_k$.

$$\boxed{\text{The stratified sample mean is the weighted average of } k \text{ sample means } \bar{x}_{\text{s}} = w_1\bar{x}_1 + \ldots + w_k\bar{x}_k.}$$

Observe that for any allocation $(n_1, \ldots, n_k)$ of

$$n = n_1 + \ldots + n_k$$

observations, the stratified sample mean is an unbiased estimate of μ since

$$\mathrm{E}(\bar{X}_{\text{s}}) = w_1\mathrm{E}(\bar{X}_1) + \ldots + w_k\mathrm{E}(\bar{X}_k) = w_1\mu_1 + \ldots + w_k\mu_k = \mu.$$

The variance of $\bar{X}_{\text{s}}$ is given by the formula

$$\mathrm{Var}(\bar{X}_{\text{s}}) = w_1^2\mathrm{Var}(\bar{X}_1) + \ldots + w_k^2\mathrm{Var}(\bar{X}_k) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k}.$$

It is estimated by

$$s_{\bar{x}_{\text{s}}}^2 = w_1^2 s_{\bar{x}_1}^2 + \ldots + w_k^2 s_{\bar{x}_k}^2 = \tfrac{w_1^2 s_1^2}{n_1} + \ldots + \tfrac{w_k^2 s_k^2}{n_k},$$

where s_j is the sample standard deviation corresponding to the sample mean $\bar{x}_j$.

A confidence interval for the mean based on a stratified sample: $I_\mu \approx \bar{x}_\mathrm{s} \pm z(\frac{\alpha}{2}) \cdot s_{\bar{x}_\mathrm{s}}$

Suppose we are allowed to collect n observations from the stratified population of size N and assume that $n \ll N$. What would be the optimal allocation $(n_1, \ldots, n_k)$ of n observations among different strata minimising the sampling error $s_{\bar{x}_\mathrm{s}}$ of $\bar{x}_\mathrm{s}$? The solution of this optimisation problem is given by the formula $n_j = n\frac{w_j \sigma_j}{\bar{\sigma}}$, $j = 1, \ldots, n$, where $\bar{\sigma}$ is the average standard deviation

$$\bar{\sigma} = w_1\sigma_1 + \ldots + w_k\sigma_k.$$

The stratified sample mean $\bar{X}_\mathrm{so}$ with the optimal allocation $n_j = n\frac{w_j\sigma_j}{\bar{\sigma}}$ has the smallest variance $\mathrm{Var}(\bar{X}_\mathrm{so}) = \frac{\bar{\sigma}^2}{n}$ among all allocations of n observations.

The optimal allocation assigns more observations to larger strata and strata with larger variation. The major drawback of the optimal allocation formula is that it requires the knowledge of the standard deviations σ_j.

If σ_j are unknown, which is often the case, then a sensible choice is to allocate the observations proportionally to the strata sizes, so that $n_i = nw_i$. Observe that with the proportional allocation, the stratified sample mean is equal to the usual sample mean

$$\bar{x}_\mathrm{sp} = w_1\bar{x}_1 + \ldots + w_n\bar{x}_k = \tfrac{n_1}{n}\bar{x}_1 + \ldots + \tfrac{n_k}{n}\bar{x}_k = \frac{x_1 + \ldots + x_n}{n} = \bar{x}.$$

However, this is not the mean of a random sample with independently allocated observations, since the n observations are forcefully allocated among the k strata proportionally to the strata sizes. For the truly random sample, the sample sizes $n_1, \ldots, n_k$ are the outcome of a random allocation of n observations among k strata following the multinomial distribution $\mathrm{Mn}(n, w_1, \ldots, w_k)$.

The stratified sample mean $\bar{X}_\mathrm{sp}$ for the proportional allocation $n_j = nw_j$ has the variance $\mathrm{Var}(\bar{X}_\mathrm{sp}) = \frac{\overline{\sigma^2}}{n}$.

Comparing the three unbiased estimates of the population mean $(\bar{x}_\mathrm{so}, \bar{x}_\mathrm{sp}, \bar{x})$, we find that their variances are ordered in the following way

$$\mathrm{Var}(\bar{X}_\mathrm{so}) \le \mathrm{Var}(\bar{X}_\mathrm{sp}) \le \mathrm{Var}(\bar{X}),$$

since

$$\frac{(\bar{\sigma})^2}{n} \le \frac{\overline{\sigma^2}}{n} \le \frac{\sigma^2}{n}.$$

Variability of σ_j across the different strata makes the optimal allocation more effective than proportional

$$\mathrm{Var}(\bar{X}_\mathrm{sp}) - \mathrm{Var}(\bar{X}_\mathrm{so}) = \tfrac{1}{n}(\overline{\sigma^2} - \bar{\sigma}^2) = \tfrac{1}{n}\sum w_j(\sigma_j - \bar{\sigma})^2.$$

Variability of μ_j across the strata makes the proportional allocation more effective than the sample mean produced by a random sample

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}(\bar{X}_\mathrm{sp}) = \tfrac{1}{n}(\sigma^2 - \overline{\sigma^2}) = \tfrac{1}{n}\sum w_j(\mu_j - \mu)^2.$$

2.6 Exact confidence intervals

Recall the approximate confidence interval formula for the mean

$$I_\mu \approx \bar{x} \pm z(\tfrac{\alpha}{2}) \cdot s_{\bar{x}},$$

which is based on the central limit theorem for the t-score of a random sample with a sufficiently large size n:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx \mathrm{N}(0, 1).$$

The approximate confidence interval formula does not require that the population distribution is normal, but it works only for large sample sizes. In this section, we state two confidence interval formulas based on the probability theory facts mentioned in Sections 1.3 and 1.4:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

These formulas are valid for small sample values under the assumption that the population distribution is normal.

Assume that a random sample $(x_1, \ldots, x_n)$ is taken from the normal distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{N}(\mu, \sigma)$$

with unspecified parameters μ and σ. Replacing the normal approximation by the exact distribution for the t-score

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

we arrive at the exact $100(1-\alpha)\%$ confidence interval formula for the mean

$$I_\mu = \bar{x} \pm t_{n-1}(\tfrac{\alpha}{2}) \cdot s_{\bar{x}},$$

which is valid even for the small values of $n \geq 2$. Remember that this formula requires that the population distribution is normal.

For example, with $\alpha = 0.05$ and $n = 10, 16, 25, 30$ we get the following four 95% confidence intervals for the mean

$$\begin{aligned} I_\mu &= \bar{x} \pm 2.26 \cdot s_{\bar{x}} \text{ for } n = 10, \\ I_\mu &= \bar{x} \pm 2.13 \cdot s_{\bar{x}} \text{ for } n = 16, \\ I_\mu &= \bar{x} \pm 2.06 \cdot s_{\bar{x}} \text{ for } n = 25, \\ I_\mu &= \bar{x} \pm 2.05 \cdot s_{\bar{x}} \text{ for } n = 30. \end{aligned}$$

Here the critical values for the t-distribution

$$t_9(0.025) = 2.26, \quad t_{15}(0.025) = 2.13, \quad t_{24}(0.025) = 2.06, \quad t_{29}(0.025) = 2.05$$

are obtained from the table in Section 11.2 using the column 0.025 and the rows $k = 9, 15, 24, 29$. These intervals are getting narrower for larger values of n asymptotically approaching the familiar approximate 95% confidence interval

$$I_\mu \approx \bar{x} \pm 1.96 \cdot s_{\bar{x}}.$$

The second exact confidence interval formula of this section is aimed at the population variance σ^2. We want to have a formula for a random interval $\mathcal{I}_{\sigma^2}$, such that

$$\mathrm{P}(\sigma^2 \in \mathcal{I}_{\sigma^2}) = 1 - \alpha.$$

Under the normality assumption we have $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$, so that an unbiased estimated of the standard error for the unbiased estimate s^2 of σ^2 is

$$s_{s^2} = \sqrt{\frac{2}{n-1}} s^2.$$

However, we can not use the same kind of formula $I_{\sigma^2} = s^2 \pm t_{n-1}(\frac{\alpha}{2}) \cdot s_{s^2}$ as for the mean.

The correct exact $100(1-\alpha)\%$ confidence interval formula for σ^2 is based on the non-symmetric chi-squared distribution in the above mentioned relation

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

and has the following non-symmetric form

$$I_{\sigma^2} = \Big(\frac{(n-1)s^2}{x_{n-1}(\frac{\alpha}{2})}; \frac{(n-1)s^2}{x_{n-1}(1-\frac{\alpha}{2})}\Big).$$

The numbers $x_k(\alpha)$ are obtained from the table of Section 11.3 giving the critical values of the chi-squared distributions.

Example

Examples of the 95% confidence intervals for σ^2 are

$$\begin{aligned} I_{\sigma^2} &= (0.47s^2, 3.33s^2) \text{ for } n = 10, \\ I_{\sigma^2} &= (0.55s^2, 2.40s^2) \text{ for } n = 16, \\ I_{\sigma^2} &= (0.61s^2, 1.94s^2) \text{ for } n = 25, \\ I_{\sigma^2} &= (0.63s^2, 1.81s^2) \text{ for } n = 30. \end{aligned}$$

To clarify, turn to the last formula dealing with $n = 30$. The χ^2_{29}-distribution table gives the critical values

$$x_{29}(0.025) = 45.722, \quad x_{29}(0.975) = 16.047,$$

yielding

$$\frac{n-1}{x_{n-1}(\frac{\alpha}{2})} = \frac{29}{45.722} = 0.63, \quad \frac{n-1}{x_{n-1}(1-\frac{\alpha}{2})} = \frac{29}{16.047} = 1.81.$$

We conclude that for $n = 30$,

$$I_{\sigma^2} = \Big(\frac{(n-1)s^2}{x_{n-1}(\frac{\alpha}{2})}; \frac{(n-1)s^2}{x_{n-1}(1-\frac{\alpha}{2})}\Big) = (0.63s^2, 1.81s^2).$$

2.7 Exercises

Problem 1

Consider a population consisting of five values

$$1, 2, 2, 4, 8.$$

Find the population mean and variance. Calculate the sampling distribution of the mean of a random sample of size 2 by generating all possible such samples (x_1, x_2). Then find the mean and variance of the sampling distribution, and verify the formulas given in this section

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Problem 2

In a simple random sample of 1500 voters, 55% said they planned to vote for a particular proposition, and 45% said they planned to vote against it. The estimated margin of victory for the proposition is thus 10%. What is the standard error of this estimated margin? What is an approximate 95% confidence interval for the margin of victory?

Problem 3

This problem introduces the concept of a one-sided confidence interval. Using the central limit theorem, how should the constant k_1 be chosen so that the interval

$$(-\infty, \bar{x} + k_1 s_{\bar{x}})$$

is an approximate 90% confidence interval for μ? How should k_2 be chosen so that

$$(\bar{x} - k_2 s_{\bar{x}}, \infty)$$

is an approximate 95% confidence interval for μ?

Problem 4

Verify the formula for the mean square error

$$\mathrm{E}((\widehat{\Theta} - \theta)^2) = \mathrm{Var}(\widehat{\Theta}) + (\mathrm{E}(\widehat{\Theta}) - \theta)^2.$$

Problem 5

A simple random sample of a population size 2000 yields 25 values with

104	109	11	109	87
86	80	119	88	122
91	103	99	108	96
104	98	98	83	107
79	87	94	92	97

(a) Calculate an unbiased estimate of the population mean.

(b) Calculate an unbiased estimates of the population variance and $\mathrm{Var}(\bar{X})$.

(c) Give an approximate 95% confidence interval for the population mean.

Problem 6

For a simple random sample, take $\bar{x}^2$ as a point estimate of μ^2. (This is an example of the method of moments estimate introduced in the next chapter.) Compute the bias of this point estimate, if any.

Problem 7

The following table (Cochran 1977) shows the stratification of all farms in a county by farm size and the mean and standard deviation of the number of acres of corn in each stratum.

Farm size	0-40	41-80	81-120	121-160	161-200	201-240	241+
Number of farms N_j	394	461	391	334	169	113	148
Stratum mean μ_j	5.4	16.3	24.3	34.5	42.1	50.1	63.8
Stratum standard deviation σ_j	8.3	13.3	15.1	19.8	24.5	26.0	35.2

(a) What are the population mean and variance?

(b) For a sample size of 100 farms, compute the sample sizes from each stratum for proportional and optimal allocation, and compare them.

(c) Calculate the variances of three sample means with different allocations of 100 observations: (1) proportional allocation, (2) optimal allocation, (3) random sampling.

(d) Suppose that ten farms are sampled per stratum. What is $\mathrm{Var}(\bar{X}_{\mathrm{s}})$? How large a simple random sample would have to be taken to attain the same variance? Ignore the finite population correction.

(e) Repeat part (d) using proportional allocation of $n = 70$ observations.

Problem 8

How might stratification be used in each of the following sampling problems?

(a) A survey of household expenses in a city.

(b) A survey to examine the lead concentration in the soil in a large plot of land.

(c) A survey to estimate the number of people who use elevators in a large building with a single bank of elevators.

Problem 9

Consider stratifying the population of Problem 1 into two strata (1,2,2) and (4,8). Assuming that one observation is taken from each stratum, find the sampling distribution of the estimate of the population mean and the mean and standard deviation of the sampling distribution. Check the formulas of Section 2.5.

Problem 10

The following 16 numbers were generated from a normal distribution $\mathrm{N}(\mu, \sigma)$

5.3299	4.2537	3.1502	3.7032
1.6070	6.3923	3.1181	6.5941
3.5281	4.7433	0.1077	1.5977
5.4920	1.7220	4.1547	2.2799

(a) Give unbiased estimates of μ and σ^2.

(b) Give 90%, 95%, and 99% confidence intervals for μ and σ^2.

(c) Give 90%, 95%, and 99% confidence intervals for σ.

(d) How much larger sample would you need to halve the length of the confidence interval for μ?

Chapter 3

Parameter estimation

Given a parametric model determined by a vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_k)$, we wish to estimate θ from a given random sample $(x_1, \ldots, x_n)$. There are two basic methods of finding good point estimates: (1) the method of moments and (2) the maximum likelihood method. The method of moments is based on k summary statistics called *sample moments*, while the maximum likelihood method uses the full information on the joint distribution of $(X_1, \ldots, X_n)$.

3.1 Method of moments

Consider the case of a parametric model characterised by a pair of parameters (θ_1, θ_2). Suppose we have explicit expressions for the first and second population moments $(\mathrm{E}(X), \mathrm{E}(X^2))$ in terms of (θ_1, θ_2):

$$\mathrm{E}(X) = f(\theta_1, \theta_2), \quad \mathrm{E}(X^2) = g(\theta_1, \theta_2).$$

Given a random sample $(x_1, \ldots, x_n)$ drawn from this parametric model, we define the first and second sample moments by

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad \overline{x^2} = \frac{x_1^2 + \ldots + x_n^2}{n}.$$

By the Law of Large Numbers saying that for any $j \geq 1$,

$$\frac{X_1^j + \ldots + X_n^j}{n} \to \mathrm{E}(X^j), \quad n \to \infty,$$

implying that $(\bar{x}, \overline{x^2})$ are consistent estimates of $(\mathrm{E}(X), \mathrm{E}(X^2))$. After replacing the population moments with the corresponding sample moments, we arrive at the equations

$$\bar{x} = f(\tilde{\theta}_1, \tilde{\theta}_2), \quad \overline{x^2} = g(\tilde{\theta}_1, \tilde{\theta}_2),$$

whose solution $(\tilde{\theta}_1, \tilde{\theta}_2)$ will be called the method of moments estimates of (θ_1, θ_2).

Geometric model

A researcher has observed $n = 130$ birds and counted the number of hops that each bird does between flights. As a result, she obtained a random sample $(x_1, \ldots, x_n)$, where

$$x_i = \text{number of hops that the } i\text{-th bird does between flights.}$$

The observed range of the number of hops j was between 1 and 12. The next table summarizes the dataset $(x_1, \ldots, x_n)$

number of hops x	1	2	3	4	5	6	7	8	9	10	11	12	total
observed counts o_x	48	31	20	9	6	5	4	2	1	1	2	1	130

in terms of the observed counts o_x, defined as the number of observed birds, who hopped x times. The observed counts are computed from $(x_1, \ldots, x_n)$ as

$$o_x = 1_{\{x_1 = x\}} + \ldots + 1_{\{x_n = x\}},$$

where 1_A is the indicator function of the relation A, which equals 1 if A is true and 0 otherwise.

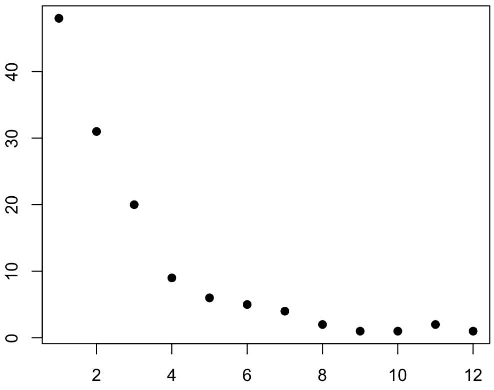
The data produces the following summary statistics

$$\begin{aligned}
\bar{x} &= \tfrac{\text{total number of hops}}{\text{number of birds}} = \tfrac{363}{130} = 2.79, \\
\overline{x^2} &= 1^2 \cdot \tfrac{48}{130} + 2^2 \cdot \tfrac{31}{130} + \ldots + 11^2 \cdot \tfrac{2}{130} + 12^2 \cdot \tfrac{1}{130} = 13.20, \\
s^2 &= \tfrac{130}{129}(\overline{x^2} - \bar{x}^2) = 5.47, \\
s_{\bar{x}} &= \sqrt{\tfrac{5.47}{130}} = 0.205.
\end{aligned}$$

An approximate 95% confidence interval for μ, the mean number of hops per bird is given by

$$I_\mu \approx \bar{x} \pm z(0.025) \cdot s_{\bar{x}} = 2.79 \pm 1.96 \cdot 0.205 = 2.79 \pm 0.40.$$

The data plot below exhibits the frequencies descending by a certain factor.



This suggests a geometric model for the number of jumps $X \sim \text{Geom}(p)$ for a random bird:

$$\mathrm{P}(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots$$

(such a geometric model implies that a bird "does not remember" the number of hops made so far, and the next move of the bird is to jump with probability $1-p$ or to fly away with probability p).

The method of moment estimate for the parameter $\theta = p$ of the geometric model requires a single equation arising from the expression for the first population moment

$$\mu = \tfrac{1}{p}.$$

This expression leads to the equation $\bar{x} = \frac{1}{\tilde{p}}$ which gives the method of moment estimate

$$\tilde{p} = 1/\bar{x} = 0.36.$$

In this case, we can even compute an approximate 95% confidence interval for p using the above mentioned I_μ:

$$I_p \approx (\tfrac{1}{2.79+0.40}, \tfrac{1}{2.79-0.40}) = (0.31, 0.42).$$

It is useful to compare the observed frequencies (counts) to the frequencies expected from the geometric distribution with parameter $\tilde{p}$:

x	1	2	3	4	5	6	7+
c_x	48	31	20	9	6	5	11
E_x	46.8	30.0	19.2	12.3	7.9	5.0	8.8

The expected counts E_x are computed in terms of independent geometric random variables $(X_1, \ldots, X_n)$

$$\begin{aligned} E_x &= \mathrm{E}(C_x) = \mathrm{E}(1_{\{X_1=x\}} + \ldots + 1_{\{X_n=x\}}) \\ &= n\mathrm{P}(X = x) = n(1-\tilde{p})^{x-1}\tilde{p} = 130 \cdot (0.64)^{x-1}(0.36), \quad x = 1, \ldots, 6, \\ E_7 &= n - E_1 - \ldots - E_6. \end{aligned}$$

An appropriate measure of discrepancy between the observed and expected counts is given by the following so-called chi-squared test statistic

$$\mathrm{x}^2 = \sum_{x=1}^{7} \tfrac{(c_x - E_x)^2}{E_x} = 1.86.$$

As it will be explained later on, the obtained small value 1.86 of the chi-squared test statistic allows us to conclude that the geometric model fits the bird data very well.

3.2 Maximum likelihood estimation

In a parametric setting with population density function $f(x|\theta)$, the observed sample $(x_1, \ldots, x_n)$ is a realization of the random vector $(X_1, \ldots, X_n)$ having the joint probability distribution

$$f(y_1, \ldots y_n|\theta) = f(y_1|\theta) \cdots f(y_n|\theta)$$

over the possible sample vectors $(y_1, \ldots, y_n)$, obtained as the product of the marginal densities due to the assumption of independence. Fixing the sample values $(y_1, \ldots, y_n) = (x_1, \ldots, x_n)$ and allowing the parameter value θ to vary, we obtain the so-called likelihood function

$$L(\theta) = f(x_1, \ldots x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta).$$

Observe thar the likelihood function is not a density function over θ.

Clearly, if $L(\theta_1) > L(\theta_2)$, then the parametric model $f(x|\theta_1)$ has a better support for the given data $(x_1, \ldots, x_n)$ compared to the model $f(x|\theta_2)$.

The maximum likelihood estimate $\hat{\theta}$ of θ is the value of θ that maximises the likelihood function $L(\theta)$.

Observe that it is often more convenient to find the maximum likelihood estimate $\hat{\theta}$ by maximizing the log-likelihood function

$$l(\theta) = \ln L(\theta) = \ln f(x_1|\theta) + \ldots + \ln f(x_n|\theta).$$

Sufficiency

Suppose there is a summary statistic $t = g(x_1, \ldots, x_n)$ such that

$$L(\theta) = f(x_1, \ldots, x_n|\theta) = h(t, \theta)c(x_1, \ldots, x_n) \propto h(t, \theta),$$

where the sign $\propto$ means "proportional to". Here the coefficient of proportionality $c(x_1, \ldots, x_n)$ does not explicitly depend on θ. In this case, the maximum likelihood estimate $\hat{\theta}$ depends on the data $(x_1, \ldots, x_n)$ only through the statistic t. Given such a factorisation property, we call t a sufficient statistic, as no other statistic that can be calculated from the same sample provides any additional information on the value of the maximum likelihood estimate $\hat{\theta}$.

Example: normal distribution model

The two-parameter normal distribution model

$$\mathcal{F}(\mu, \sigma) = \mathrm{N}(\mu, \sigma)$$

has a two-dimensional sufficient statistic $t = (t_1, t_2)$, where

$$t_1 = \sum_{i=1}^{n} x_i, \qquad t_2 = \sum_{i=1}^{n} x_i^2,$$

which follows from

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{t_2 - 2\mu t_1 + n\mu^2}{2\sigma^2}}.$$

Thus two samples having the same values for (t_1, t_2) will produce the same maximum likelihood estimates for (μ, σ). Notice the match between the number of sufficient statistics and the number of parameters of the model.

Example: Bernoulli distribution model

Consider the case of the Bernoulli model $X \sim \mathrm{Bin}(1, p)$ described by the probability mass function

$$f(x|p) = \mathrm{P}(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}.$$

Suppose we are given a random sample $(x_1, \ldots, x_n)$ of zeros and ones drawn from a population distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{Bin}(1, p)$$

with unknown p. Since $\mu = p$, the method of moment estimate of p is computed as the sample proportion $\tilde{p} = \bar{x}$.

The likelihood function,

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^t (1-p)^{n-t},$$

where $\prod$ stands for the product, is fully determined by the number of successes

$$t = x_1 + \ldots + x_n = n\bar{x},$$

providing with another example of a sufficient statistic. To maximise the log-likelihood function

$$l(p) = \ln L(p) = t \log p + (n - t) \log(1 - p),$$

take its derivative

$$l'(p) = \frac{t}{p} - \frac{n-t}{1-p}$$

and put it equal to zero. As a result, we find that even the maximum likelihood estimate of p is the sample proportion $\hat{p} = \frac{t}{n}$.

Large sample properties of the maximum likelihood estimates

For a random sample $(x_1, \ldots, x_n)$ drawn from a parametric population distribution $f(x|\theta)$, the log-likelihood function is the sum

$$l(\theta) = \ln f(x_1|\theta) + \ldots + \ln f(x_n|\theta)$$

due to independence between n observations. This implies that the log-likelihood function can be treated as a realisation of a sum of independent and identically distributed random variables $Y_i = \ln f(X_i|\theta)$. Using the central limit theorem argument one can derive the normal approximation

$$\hat{\Theta} \approx \mathrm{N}(\theta, \frac{\sigma_\theta}{\sqrt{n}})$$

for the maximum likelihood estimator. Here σ_θ^2 is the inverse of the so-called Fisher information in a single observation (if interested, see Wikipedia). It follows that the maximum likelihood estimators are asymptotically unbiased and consistent. Moreover, they are asymptotically efficient estimates in the sense of the following Cramer-Rao inequality.

Cramer-Rao inequality: if θ^* is an unbiased estimator of θ, then $\mathrm{Var}(\Theta^*) \geq \frac{\sigma_\theta^2}{n}$.

Example: exponential model

The lifetimes of five batteries measured in hours

$$x_1 = 0.5, \quad x_2 = 14.6, \quad x_3 = 5.0, \quad x_4 = 7.2, \quad x_5 = 1.2,$$

are assumed to be generated by the exponential population distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{Exp}(\theta).$$

In this case, the mean lifetime is $\mu = 1/\theta$, and θ can be viewed as the battery death rate per hour. The likelihood function

$$L(\theta) = \theta e^{-\theta x_1} \theta e^{-\theta x_2} \theta e^{-\theta x_3} \theta e^{-\theta x_4} \theta e^{-\theta x_5} = \theta^n e^{-\theta(x_1 + \ldots + x_n)} = \theta^5 e^{-\theta \cdot 28.5}$$

first grows from 0 to $2.2 \cdot 10^{-7}$ and then falls down towards zero. The likelihood maximum is reached at $\hat{\theta} = 0.175$.

For the exponential model, $t = x_1 + \ldots + x_n$ is a sufficient statistic, and the maximum likelihood estimate

$$\hat{\theta} = n/t = 1/\bar{x}$$

coincides with the method of moment estimate. It is a biased estimate of θ, since

$$\mathrm{E}(\widehat{\Theta}) = \mathrm{E}(1/\bar{X}) \neq 1/\mathrm{E}(\bar{X}) = 1/\mu = \theta,$$

but asymptotically unbiased due to the Law of Large Numbers saying that $\bar{X} \approx \mu$ for the large sample sizes.

3.3 Maximum likelihood estimation for the gamma distribution

Let $(x_1, \cdots, x_n)$ be a random sample from the gamma distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{Gam}(\alpha, \lambda)$$

with the two-dimensional unknown parameter $\theta = (\alpha, \lambda)$. Put

$$t_1 = x_1 + \ldots + x_n, \qquad t_2 = x_1 \cdots x_n,$$

and observe that the corresponding likelihood function takes the form

$$L(\alpha,\lambda)=\prod_{i=1}^{n}\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x_i^{\alpha-1}e^{-\lambda x_i}=\frac{\lambda^{n\alpha}}{\Gamma^n(\alpha)}(x_1\cdots x_n)^{\alpha-1}e^{-\lambda(x_1+\ldots+x_n)}=\frac{\lambda^{n\alpha}}{\Gamma^n(\alpha)}t_2^{\alpha-1}e^{-\lambda t_1}.$$

We see that (t_1,t_2) is a pair of sufficient statistics containing all information from the data needed to compute the likelihood function. To maximise the log-likelihood function

$$l(\alpha,\lambda)=\ln L(\alpha,\lambda),$$

set the two derivatives

$$\begin{aligned}\tfrac{\partial}{\partial\alpha}l(\alpha,\lambda)&=n\ln(\lambda)-n\tfrac{\Gamma'(\alpha)}{\Gamma(\alpha)}+\ln t_2,\\ \tfrac{\partial}{\partial\lambda}l(\alpha,\lambda)&=\tfrac{n\alpha}{\lambda}-t_1,\end{aligned}$$

equal to zero. The maximum likelihood estimates of the parameters (α,λ) can be obtained from the following two equations

$$\begin{aligned}\hat\lambda&=\hat\alpha/\bar x,\\ \ln(\hat\alpha/\bar x)&=-\tfrac{1}{n}\ln t_2+\Gamma'(\hat\alpha)/\Gamma(\hat\alpha).\end{aligned}$$

A numerical solution of the second equation would benefit of using the method of moment estimate $\tilde\alpha$ as the initial guess.

Example: male heights

Consider a random sample of $n=24$ male heights (cm) given in ascending order:

$$170,175,176,176,177,178,178,179,179,180,180,180,180,180,181,181,182,183,184,186,187,192,192,199.$$

Assuming that these numbers are generated by the $\mathrm{Gam}(\alpha,\lambda)$ distribution, we would like to estimate the parameters λ and α.

To apply the method of moments, we use the formulas for the first and second population moments

$$\mathrm{E}(X)=\tfrac{\alpha}{\lambda},\quad \mathrm{E}(X^2)=\mathrm{Var}(X)+(\mathrm{E}(X))^2=\frac{\alpha}{\lambda^2}+\frac{\alpha^2}{\lambda^2}=\frac{\alpha(1+\alpha)}{\lambda^2},$$

and the two sample moments computed from the data

$$\bar x=181.46,\quad \overline{x^2}=32964.2.$$

From the equations

$$\bar x=\frac{\tilde\alpha}{\tilde\lambda},\quad \overline{x^2}=\frac{\tilde\alpha(1+\tilde\alpha)}{(\tilde\lambda)^2}$$

we get

$$\frac{\tilde\alpha}{\tilde\lambda}=181.46,\quad \frac{1+\tilde\alpha}{\tilde\lambda}=\frac{32964.2}{181.46}=181.66,$$

yielding the method of moments estimates

$$\tilde\lambda=5.00,\quad \tilde\alpha=907.3.$$

The maximum likelihood estimate of the shape parameter α is obtained from the equation

$$\ln(\hat\alpha/\bar x)=-\tfrac{1}{n}\ln t_2+\Gamma'(\hat\alpha)/\Gamma(\hat\alpha)$$

using the method of moment estimate $\tilde\alpha=907.3$ as the initial guess. The Mathematica command

FindRoot[Log[a] == 0.00055+Gamma′[a]/Gamma[a], {a, 907.3}]

gives

$$\hat\alpha=908.76.$$

Then the maximum likelihood estimate of the scale parameter λ is obtained from the equation $\hat\lambda=\hat\alpha/\bar x$ yielding

$$\hat\lambda=5.01.$$

Notice that the obtained maximum likelihood estimates are close to the method of moment estimates.

Example: Gam$(\alpha, 1)$ model

A random sample $(1.23, 0.62, 2.22, 2.55, 1.42)$ was generated by the gamma distribution with the shape parameter $\alpha = 2$ and the scale parameter $\lambda = 1$. Treating $\theta = \alpha$ as the unknown parameter, and assuming that $\lambda = 1$ is known, we arrive at the Gam$(\alpha, 1)$ model and wish to estimate α using the given sample of size $n = 5$.

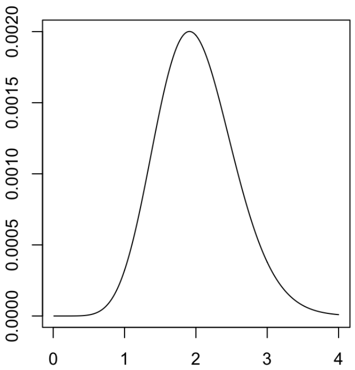
Since, the population mean is $\mu = \alpha/\lambda = \alpha$, the method of moments estimate is $\tilde{\alpha} = \bar{x} = 1.61$. This estimate should be compared to the true value $\alpha = 2$. The corresponding likelihood function has the form

$$L(\alpha) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i} = \Gamma^{-n}(\alpha) t^{\alpha-1} e^{-8.04},$$

where the constant 8.04 is obtained as $x_1 + \ldots + x_n$ and

$$t = x_1 \cdots x_n = 6.13$$

is a sufficient statistic in this case. The graph of the likelihood function $L(\alpha)$ on the figure below, shows that the area under the curve is much smaller than 1. A closer look at the figure reveals that the maximum likelihood estimate is slightly smaller than the true value $\alpha = 2$.



To maximise the log-likelihood function

$$l(\alpha) = -n \ln(\Gamma(\alpha)) + (\alpha - 1) \ln t - 8.04,$$

take its derivative and put it equal to zero. This results in the equation

$$0 = \ln t - n\Gamma'(\alpha)/\Gamma(\alpha).$$

The Mathematica command

FindRoot[Log[a] == Log[6.14] - 5*Gamma$'[a]/$Gamma[a], {a, 1.61}]

gives a numerical solution $\hat{\alpha} = 1.90989$, to be compared with the true value $\alpha = 2$. Here the method of moments estimate 1.61 is used as the initial guess for the algorithm solving the equation. In this example, the maximum likelihood estimate brings a drastic improvement compared to the method of moments estimate.

3.4 Exercises

Problem 1

The Poisson distribution has been used by traffic engineers as a model for light traffic. The following table shows the number of right turns during 300 three-min intervals at a specific intersection.

x	frequency
0	14
1	30
2	36
3	68
4	43
5	43
6	30
7	14
8	10
9	6
10	4
11	1
12	1
13+	0

Fit a Poisson distribution. Comment on the fit by comparing observed and expected counts. It is useful to know that the 300 intervals were distributed over various hours of the day and various days of the week.

Problem 2

Let $(x_1, \ldots, x_n)$ be a random sample from a geometric distribution

$$\mathrm{P}(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots$$

(a) Write down the likelihood function based on this random sample and suggest a simple sufficient statistic.

(b) Find the maximum likelihood estimate of p.

(c) Is this estimate consistent? Explain.

Problem 3

Suppose that X is a discrete random variable with

$$\begin{aligned}\mathrm{P}(X=0) &= \tfrac{2}{3}\theta,\\ \mathrm{P}(X=1) &= \tfrac{1}{3}\theta,\\ \mathrm{P}(X=2) &= \tfrac{2}{3}(1-\theta),\\ \mathrm{P}(X=3) &= \tfrac{1}{3}(1-\theta),\end{aligned}$$

where the parameter $\theta \in [0,1]$ is unknown. The following 10 independent observations were drawn from this distribution:

$$(3, 0, 2, 1, 3, 2, 1, 0, 2, 1).$$

(a) Find the method of moments estimate $\tilde{\theta}$ of θ.

(b) Estimate the standard error of $\tilde{\theta}$.

(c) What is the maximum likelihood estimate $\hat{\theta}$ of θ?

(d) Estimate the standard error of $\hat{\theta}$.

Problem 4

Suppose that x is generated by a $\mathrm{Bin}(n, p)$ distribution with unknown p.

(a) Show that the maximum likelihood estimate of p is $\hat{p} = \frac{x}{n}$.

(b) Given $n = 10$ and $x = 5$, sketch the graph of the likelihood function.

Problem 5

A company has manufactured certain objects and has printed a serial number on each object. The serial numbers start at 1 and end at N, where N is the number of objects that have been manufactured. One of these objects is selected at random, and the serial number of that object is 888.

(a) What is the method of moments estimate of N?

(b) What is the maximum likelihood estimate of N?

Problem 6

To estimate the number N of fish living in a lake, a master degree student has applied the capture-recapture method according to the two-step procedure:

1. capture and tag $n = 100$ fish, then release them in the lake,
2. capture and release $k = 50$ fish one by one, and count the number of the tagged fish among those captured on the second stage.

Suppose $x = 20$ fish were tagged among the $k = 50$ fish captured on the second stage. Find a maximum likelihood estimate of N after suggesting a simple parametric model.

Problem 7

The following 16 numbers came from the normal random number generator on a computer:

5.33	4.25	3.15	3.70
1.61	6.39	3.12	6.59
3.53	4.74	0.11	1.60
5.49	1.72	4.15	2.28

(a) Write down the likelihood function based on this sample. (Hint: to avoid tedious calculations on your calculator use the numbers in the next subquestion.)

(b) In what sense the sum of the sample values (which is close to 58), and the sum of their squares (which is close to 260) are sufficient statistics in this case?

(c) Turning to the log-likelihood function compute the maximum likelihood estimates for the mean and variance. Is the obtained variance estimate unbiased?

Problem 8

Let $(x_1, \ldots, x_n)$ be a random sample generated by the continuous uniform distribution over the interval $[0, \theta]$ with unknown θ.

(a) Find the method of moments estimate of θ and its mean and variance.

(b) Find the maximum likelihood estimate of θ.

(c) Find the probability density of the maximum likelihood estimator and calculate its mean and variance. Compare the variance, the bias, and the mean square error to those of the method of moments estimate.

(d) Find a modification of the maximum likelihood estimate that renders it unbiased.

Problem 9

For two factors, starchy-or-sugary and green-or-white base leaf, the following counts for the progeny of self-fertilized heterozygotes were observed (Fisher 1958)

Type	Count
Starchy green	$c_1 = 1997$
Starchy white	$c_2 = 906$
Sugary green	$c_3 = 904$
Sugary white	$c_4 = 32$

According to the genetic theory the cell probabilities are

$$p_1 = \frac{2+\theta}{4}, \quad p_2 = \frac{1-\theta}{4}, \quad p_3 = \frac{1-\theta}{4}, \quad p_4 = \frac{\theta}{4},$$

where $0 < \theta < 1$. In particular, if $\theta = 0.25$, then the genes are unlinked and the genotype frequencies are

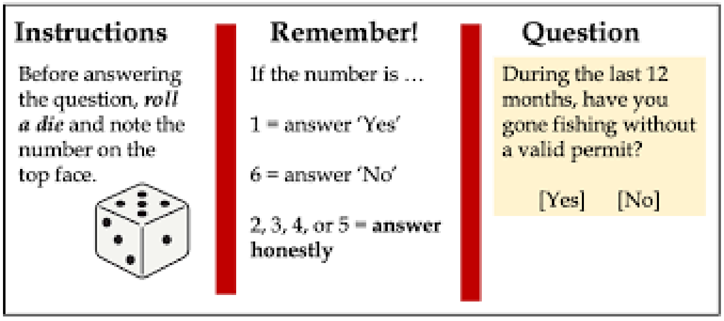
	Green	White	Total
Starchy	9/16	3/16	3/4
Sugary	3/16	1/16	1/4
Total	3/4	1/4	1

(a) Find the likelihood function of θ. Specify a sufficient statistic.

(b) Find the maximum likelihood estimate of θ.

Problem 10

The method of randomised response deals with surveys asking sensitive questions. Suppose we want to estimate the proportion q of the fishermen who during the last 12 months have gone fishing without a valid permit. We are interested in the population as a whole - not in punishing particular individuals. Suppose randomly chosen n fishermen have responded yes/no to a randomised statement according to the instructions in the figure below. Suggest a probability model for this experiment, and find a method of moments estimate for q. What is the standard error of the estimated proportion?



Instructions

Before answering the question, ***roll a die*** and note the number on the top face.

Remember!

If the number is ...

1 = answer 'Yes'

6 = answer 'No'

2, 3, 4, or 5 = **answer honestly**

Question

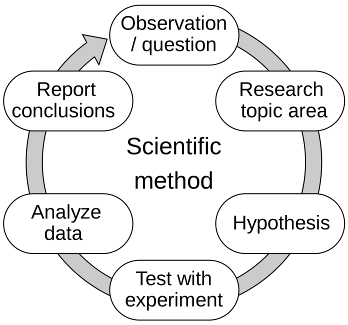
During the last 12 months, have you gone fishing without a valid permit?

[Yes] [No]

Chapter 4

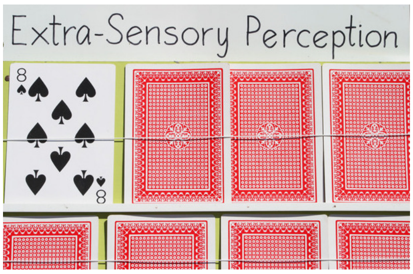
Hypothesis testing

Hypothesis testing is a crucial part of the scientific method. In this chapter we learn about the statistical hypothesis testing.



Example: extrasensory perception

Your friend claims that he has extrasensory perception ability. To test this claim you arrange an experiment where he asked to guess the suits of $n = 100$ playing cards. As a result the friend guessed correctly the suits of 30 cards out of 100. What would be your conclusion?



One particular way of reasoning in such a case is the topic of this chapter. Under the hypothesis of pure guessing the number of correct answers is $X \sim \text{Bin}(100, 1/4)$. Then the observed outcome $x = 30$ deviates from the mean $100 \cdot 1/4 = 25$ for about one standard deviation. Since such an outcome is not unusual, a sensible conclusion is that the data does not contradict the hypothesis of pure guessing.

4.1 Statistical significance

Often we need a rule based on data for choosing between two mutually exclusive hypotheses

H_0: the effect of interest is zero,

H_1: the effect of interest is not zero.

Here the null hypothesis H_0 represents an established theory that must be discredited in order to demonstrate a phenomenon contradicting the established theory stated in the form of the alternative hypothesis H_1. The decision rule for hypothesis testing is based on a test statistic $t = t(x_1, \ldots, x_n)$, a function of the data with distinct typical values under H_0 and H_1. The task is to find an appropriately chosen rejection region $\mathcal{R}$ so that

$$\text{we reject } H_0 \text{ in favour of } H_1 \text{ if and only if } t \in \mathcal{R}.$$

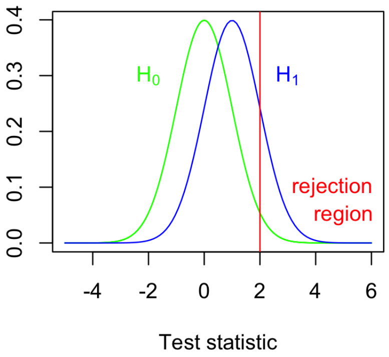
Making a such decision we are facing four possible outcomes:

	Negative decision: do not reject H_0	Positive decision: reject H_0 in favour of H_1
If H_0 is true	True negative outcome	False positive outcome, type I error
If H_1 is true	False negative outcome, type II error	True positive outcome

The figure below illustrates the case of two simple hypotheses

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1,$$

where $\theta_1 > \theta_0$, so that the alternative hypothesis claims a positive effect size $\theta_1 - \theta_0$. The figure depicts two sampling distributions of the test statistic T: the null distribution in green and the alternative distribution in blue. The rejection region consists on the values of t to the right of the red line.



This figure illustrates the following four important conditional probabilities:

$\alpha = \mathrm{P}(T \in \mathcal{R} \vert H_0)$	the conditional probability of type I error, called the significance level of the test, is given by the area below the green line to the right of the red line
$1 - \alpha = \mathrm{P}(T \notin \mathcal{R} \vert H_0)$	the specificity of the test, is given by the area below the green line to the left of the red line
$\beta = \mathrm{P}(T \notin \mathcal{R} \vert H_1)$	the conditional probability of type II error, is given by the area below the blue line to the left of the red line
$1 - \beta = \mathrm{P}(T \in \mathcal{R} \vert H_1)$	the sensitivity of the test or the power of the test is given by the area below the blue line to the right of the red line

The larger the power of the test $1 - \beta$, the better the ability of the test to recognise the effect of size $\theta_1 - \theta_0$. It is desirable to place the red line in such a way that both α and β are minimised, however, according to the figure, moving the red line to the right or to the left would decrease one of the error sizes and increase the other.

For a given test statistic, if the sample size is fixed, one can not make both α and β smaller by changing $\mathcal{R}$.

A *significance test* resolves the conflict between the two types of errors, by controlling the significance level α. Given the value of α, say 5%, the rejection region $\mathcal{R}$ is found from the equation

$$\alpha = \mathrm{P}(T \in \mathcal{R} | H_0)$$

using the null distribution of the test statistic T. After the rejection region is determined, the size of type II error is computed by

$$\beta = \mathrm{P}(T \notin \mathcal{R} | H_1).$$

To summarise, the significance testing calculations follow the next flow chart

$$\text{choose } \alpha \to \text{find } \mathcal{R} \to \text{compute } \beta \to \text{compute the power of the test } 1 - \beta.$$

Example: extrasensory perception

Your friend guessed correctly the suits of $y = 30$ cards out of $n = 100$. To analyse this data we use a binomial model, assuming that the number of cards guessed correctly is generated by the $Y \sim \mathrm{Bin}(n, p)$ distribution, where the unknown p is the probability of your friend guessing correctly the suit of a playing card. The null hypothesis, $H_0 : p = 0.25$, posits pure guessing, and the alternative hypothesis of interest, $H_1 : p > 0.25$, suggests the presence of the extrasensory perception ability. Taking y as a test statistic and putting $\alpha = 0.05$, define the rejection region by

$$\mathcal{R} = \{y : y \geq y(0.05)\},$$

where the critical value $y(0.05)$ is determined by the equation

$$0.05 = \mathrm{P}(Y \geq y(0.05)|p = 0.25).$$

Using the normal approximation for the z-score

$$Z_0 = \frac{Y - 25}{4.33} \overset{H_0}{\approx} \mathrm{N}(0,1),$$

we find that

$$y(0.05) \approx 25 + 4.33 \cdot 1.645 \approx 32.$$

Since the observed test statistic $y_{\text{obs}} = 30$ is below 32, we conclude that the experimental result is not significant, and we do not reject H_0 at the 5% significance level.

4.2 Large-sample test for the proportion

The last example is an illustration of the large-sample test for the proportion described next. Consider a random sample of size n drawn from a Bernoulli distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{Bin}(1, p)$$

with the unknown population proportion p. Recall that the corresponding sample proportion $\hat{p}$ gives the maximum likelihood estimate of p.

$$\boxed{\text{Given the null hypothesis } H_0\text{: } p = p_0, \text{ use the test statistic } z_0 = \tfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}.}$$

By the central limit theorem, the null distribution of the z-score is approximately normal: $Z_0 \overset{H_0}{\approx}$ N(0,1). Depending on the problem in hand, there might arise three different alternative hypotheses:

$$H_1\text{: } p > p_0, \quad H_1\text{: } p < p_0, \quad H_1\text{: } p \neq p_0.$$

The corresponding rejection region depends on the exact form of the alternative hypothesis

	H_1	Rejection region
one-sided	$p > p_0$	$\mathcal{R} = \{z \geq z(\alpha)\}$
one-sided	$p < p_0$	$\mathcal{R} = \{z \leq -z(\alpha)\}$
two-sided	$p \neq p_0$	$\mathcal{R} = \{z \leq -z(\frac{\alpha}{2})\} \cup \{z \geq z(\frac{\alpha}{2})\}$

where $z(\alpha)$ is found from $\Phi(z(\alpha)) = 1 - \alpha$ using the normal distribution table.

Confidence interval method of hypotheses testing

Consider testing H_0: $p = p_0$ against the two-sided alternative H_1: $p \neq p_0$. Observe that at the significance level α, the rejection rule can be expressed as

$$\mathcal{R} = \{p_0 \notin I_p\},$$

in terms of a 100(1-α)% confidence interval for the proportion. The corresponding decision rule is simple: reject the null hypothesis stating $p = p_0$, if the confidence interval I_p does not cover the value p_0.

Power of the test

Consider the one-sided setting

$$H_0\text{: } p = p_0 \text{ against } H_1\text{: } p > p_0.$$

The *power function* Pw(p) of the one-sided test can be computed using the normal approximation $Z \approx \mathrm{N}(0,1)$ under H_1, for the sampling distribution of the z-score

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}.$$

To this end observe that

$$z\sqrt{p(1-p)} = z_0\sqrt{p_0(1-p_0)} + p_0 - p.$$

We have

$$\begin{aligned}\mathrm{Pw}(p) = \mathrm{P}(Z_0 \geq z(\alpha)|H_1) &= \mathrm{P}\Big(Z \geq \frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p)}{\sqrt{p(1-p)}}|H_1\Big) \\ &\approx 1 - \Phi\big(\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p)}{\sqrt{p(1-p)}}\big), \quad p > p_0.\end{aligned}$$

Observe that with $p = p_0$, we get Pw(p) = α. The larger is the effect size $p - p_0$, the larger is the power Pw(p) of the test.

Planning the sample size

In the setting

$$H_0\colon\ p = p_0 \text{ against } H_1\colon\ p = p_1,$$

the formula

$$\sqrt{n} \approx \frac{z(\alpha)\sqrt{p_0(1-p_0)} + z(\beta)\sqrt{p_1(1-p_1)}}{|p_1 - p_0|}$$

gives the sample size n corresponding to given values of α and β. If the alternatives are very close to each other, the denominator goes to zero and the required sample size becomes very large. This is very intuitive as it becomes more difficult to distinguish between two close parameter values. On the other hand, if we decrease the levels α and β, the values $z(\alpha)$ and $z(\beta)$ from the normal distribution table become larger and the corresponding sample size $n = n(\alpha, \beta)$ will be larger as well, meaning that for both types of errors to be small, you have to collect more data.

Next, we derive this formula for $p_1 > p_0$ leaving the other case $p_1 < p_0$ as an exercise. As shown before, for $p_1 > p_0$,

$$1 - \beta = \mathrm{Pw}(p_1) \approx 1 - \Phi\Big(\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}}\Big)$$

yielding the equation

$$\beta \approx \Phi\Big(\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}}\Big).$$

Combining this with

$$\beta = \Phi(-z(\beta)),$$

we arrive at the relation

$$\frac{z(\alpha)\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}} \approx -z(\beta)$$

which brings the desired formula.

Binomial test

The binomial test is a test for proportion for a small sample size n. Consider a random sample $(x_1, \ldots, x_n)$ drawn from a Bernoulli distribution $\mathrm{Bin}(1, p)$ with the unknown population proportion p. Let $c = x_1 + \ldots + x_n$ count the number of successes, so that the corresponding random variable has a binomial distribution

$$C \sim \mathrm{Bin}(n, p).$$

Under $H_0 : p = p_0$, we have $C \sim \mathrm{Bin}(n, p_0)$. Using this distribution define b_α and c_α by

$$\mathrm{P}(C \le b_\alpha | H_0) = \alpha, \quad \mathrm{P}(C \ge c_\alpha | H_0) = \alpha.$$

Then the rejection region of the binomial test is determined according to the next table

	H_1	Rejection region
one-sided	$p > p_0$	$\mathcal{R} = \{c \ge c_\alpha\}$
one-sided	$p < p_0$	$\mathcal{R} = \{c \le b_\alpha\}$
two-sided	$p \ne p_0$	$\mathcal{R} = \{c \le b_{\alpha/2}\} \cup \{c \ge c_{\alpha/2}\}$

To illustrate the binomial test, consider the extrasensory perception test, where the subject is asked to guess the suits of $n = 20$ cards. The number of cards guessed correctly is $C \sim \mathrm{Bin}(20, p)$ and the null hypothesis of interest is $H_0 : p = 0.25$. The null distribution $\mathrm{Bin}(20, 0.25)$ of the test statistic c gives the following probabilities

c	8	9	10	11
$\mathrm{P}(C \ge c)$	0.101	0.041	0.014	0.004

We conclude that in particular, for the one-sided alternative $H_1 : p > 0.25$ and $\alpha = 4.1\%$, the rejection region of the binomial test is $\mathcal{R} = \{c \ge 9\}$. The corresponding power function

$$\mathrm{Pw}(p) = \mathrm{P}(C \ge 9 | C \sim \mathrm{Bin}(20, p))$$

takes the following values

p	0.27	0.30	0.40	0.50	0.60	0.70
$\mathrm{Pw}(p)$	0.064	0.113	0.404	0.748	0.943	0.995

4.3 P-value of the test

A p-value of the test base on the test statistic t is the probability p of obtaining a test statistic value as extreme or more extreme than the observed value t_{obs}, given that H_0 is true.

$$\boxed{\text{For a given significance level } \alpha, \text{ we reject } H_0, \text{ if p} \le \alpha, \text{ and do not reject } H_0 \text{ otherwise.}}$$

The p-value depends on the observed data t_{obs} and therefore, is a realisation of a random variable P. The source of randomness is in the sampling procedure: if you take another sample, you obtain a different p-value. To illustrate, suppose we are testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ with help of a test statistic Z whose null distribution is N(0,1). In this case, the p-value is computed as

$$\text{p} = \text{P}(Z > z_{\text{obs}}|H_0) = 1 - \Phi(z_{\text{obs}}),$$

and in terms of the underlying random variables

$$\text{P} = \text{P}(Z > Z_{\text{obs}}|H_0) = 1 - \Phi(Z_{\text{obs}}).$$

Since $Z_{\text{obs}} \overset{H_0}{\sim} \text{N}(0,1)$, we conclude that the p-value has a uniform null distribution:

$$\text{P}(\text{P} \le p|H_0) = \text{P}(1-\Phi(Z_{\text{obs}}) \le p|H_0) = \text{P}(\Phi(Z_{\text{obs}}) \ge 1-p|H_0) = \text{P}(Z_{\text{obs}} \ge \Phi_{-1}(1-p)|H_0) = 1-\Phi(\Phi_{-1}(1-p)) = p.$$

Example: extrasensory perception

A subject is asked to guess the suits of $n = 100$ cards, and we want to test

$$H_0 : p = 0.25 \text{ (pure guessing), against } H_1 : p > 0.25 \text{ (extrasensory perception ability),}$$

Applying the large sample test for proportion we find the rejection rule at 5% significance level to be

$$\mathcal{R} = \{\tfrac{\hat{p}-0.25}{0.0433} \ge 1.645\} = \{\hat{p} \ge 0.32\} = \{y \ge 32\},$$

where y is the number of suits guessed correctly. With a simple alternative $H_1 : p = 0.30$ the power of the test is

$$1 - \Phi(\tfrac{1.645 \cdot 0.433 - 0.5}{0.458}) = 32\%.$$

The sample size required for the 90% power is

$$n = (\tfrac{1.645 \cdot 0.433 + 1.28 \cdot 0.458}{0.05})^2 = 675.$$

If the observed sample count is $y_{\text{obs}} = 30$, then the observed z-score $z_0 = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$ takes the value

$$z_{\text{obs}} = \frac{0.3 - 0.25}{0.0433} = 1.15,$$

and the one-sided p-value is computed as

$$\text{P}(Z_0 \ge 1.15|H_0) = 12.5\%.$$

Since the p-value is larger that 10%, the result is not significant, and we do not reject H_0.

Large-sample test for mean

The large-sample test for the mean deals with H_0: $\mu = \mu_0$ against either the two-sided or a one-sided alternative for continuous or discrete data. The corresponding test statistic is the t-score

$$t_0 = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}.$$

This test works well even if the population distribution $\mathcal{F}(\mu, \sigma)$ is not necessarily normal, provided the sample size n is sufficiently large. The rejection region is computed using the normal approximation of the null distribution

$$T_0 \overset{H_0}{\approx} \text{N}(0,1).$$

One-sample t-test

For small n, under the assumption that the population distribution is normal

$$\mathcal{F}(\mu, \sigma) = \text{N}(\mu, \sigma),$$

the t-test of H_0: $\mu = \mu_0$ is based on the exact null distribution

$$T_0 \overset{H_0}{\sim} t_{n-1}.$$

4.4 Likelihood-ratio test

A general method of finding asymptotically optimal tests (having the largest power for a given α) uses the likelihood ratio as the test statistic. Consider first the case of two simple hypotheses. For testing

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1,$$

use the likelihood-ratio $\frac{L(\theta_0)}{L(\theta_1)}$ as the test statistic. A larger value of the likelihood-ratio would suggest that the parameter value θ_0 explains the data set better than θ_1, while a smaller value $\frac{L(\theta_0)}{L(\theta_1)}$ would indicate that the alternative hypothesis H_1 explains the data set better than H_0.

$$\boxed{\text{The likelihood-ratio test rejects } H_0 \text{ for small values of the likelihood-ratio.}}$$

By the Neyman-Pearson lemma, see Wikipedia, the likelihood-ratio test is the optimal test in the case of two simple hypotheses.

Nested hypotheses

The general case of two composite alternatives can be stated in term of a pair of nested parameter sets $\Omega_0 \subset \Omega$

$$H_0 : \theta \in \Omega_0 \text{ against } H_1 : \theta \in \Omega \setminus \Omega_0.$$

It will be more convenient to recast this setting in terms of two nested hypotheses

$$H_0 : \theta \in \Omega_0, \quad H : \theta \in \Omega,$$

leading to two maximum likelihood estimates

$\hat{\theta}_0$ = maximises the likelihood function $L(\theta)$ over $\theta \in \Omega_0$,
$\hat{\theta}$ = maximises the likelihood function $L(\theta)$ over $\theta \in \Omega$.

In this case, the likelihood-ratio is defined by

$$w = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})},$$

and the likelihood-ratio test rejects H_0 for smaller values of w, or equivalently for larger values of

$$-\ln w = \ln L(\hat{\theta}) - \ln L(\hat{\theta}_0).$$

Observe that $0 < w \leq 1$, so that $-\ln w \geq 0$. By Wilks' theorem, see Wikipedia, the test statistic $(-2\ln w)$ has a nice approximation for its null distribution

$$-2\ln W \overset{H_0}{\approx} \chi^2_{\text{df}}, \quad \text{where df} = \dim(\Omega) - \dim(\Omega_0).$$

The approximation is valid for sufficiently large sample sizes n.

4.5 Chi-squared test of goodness of fit

Suppose that the random sample consists of n independent observations, with each observation belonging to one of J classes with probabilities $(p_1, \ldots, p_J)$. Such data are summarised as the vector of observed counts whose joint distribution is multinomial

$$(C_1, \ldots, C_J) \sim \text{Mn}(n; p_1, \ldots, p_J), \qquad \text{P}(C_1 = c_1, \ldots, C_J = c_J) = \frac{n!}{k_1! \cdots k_J!} p_1^{c_1} \cdots p_J^{c_J}.$$

The general parameter space

$$\Omega = \{(p_1, \ldots, p_J) : p_1 + \ldots + p_J = 1, p_1 \geq 0, \ldots, p_J \geq 0\}$$

has dimension

$$\dim(\Omega) = J - 1.$$

Consider a parametric model for the data

$$H_0 : (p_1, \ldots, p_J) \in \Omega_0,$$

where

$$\Omega_0 = \{(p_1, \ldots, p_J) \in \Omega : (p_1, \ldots, p_J) = (p_1(\lambda), \ldots, p_J(\lambda)), \quad \lambda \in \Lambda\},$$

with

$$\dim(\Omega_0) = r, \quad 0 \leq r < J - 1.$$

To see if the proposed model fits the data, compute $\hat{\lambda}$, the maximum likelihood estimate of λ, and then the expected cell counts

$$e_j = n \cdot p_j(\hat{\lambda}),$$

where "expected" means expected under the null hypothesis model. In this special setting, the above mentioned likelihood-ratio test statistic

$$-2\ln w \approx \mathrm{x}^2$$

is approximated by the so-called chi-squared test statistic

$$\mathrm{x}^2 = \sum_{j=1}^{J} \frac{(c_j - e_j)^2}{e_j},$$

and the approximate null distribution of the chi-squared test statistic is

$$\mathrm{X}^2 \overset{H_0}{\approx} \chi^2_{J-1-r},$$

where the number of degrees of freedom is computed as the difference

$$\mathrm{df} = \dim(\Omega) - \dim(\Omega_0) = (J-1) - r = J - 1 - r.$$

The chi-squared test is approximate in that all *expected* counts are recommended to be at least 5. If not, then you should combine small cells in larger cells and recalculate the number of degrees of freedom df.

Example: geometric model

Returning to the data on the number of hops for birds, consider

$$H_0: \text{ number of hops that a bird does between flights has a geometric distribution Geom}(p).$$

Using $\hat{p} = 0.358$ and $J = 7$ we obtain $\mathrm{x}^2 = 1.86$. Using the chi-squared distribution table with df $= 7 - 1 - 1 = 5$, we find that the p-value is close to 90%. This implies that the geometric distribution model fits very well to the data.

Case study: sex ratio in german families

A 1889 study made in Germany recorded the numbers of boys $(x_1, \ldots, x_n)$ for $n = 6115$ families with 12 children each. Each x_i is an independent realisation of the random variable X having a discrete distribution

$$p_x = \mathrm{P}(X = x), \quad x = 0, 1, \ldots, 12.$$

The corresponding parameter space Ω has dimension $\dim(\Omega) = 12$. The data is given in the table below in the form of thirteen observed counts.

Number of boys x	Observed count c_x	Model 1: e_x	and $\frac{(c_x-e_x)^2}{e_x}$	Model 2: e_x	and $\frac{(c_x-e_x)^2}{e_x}$
0	7	1.5	20.2	2.3	9.6
1	45	17.9	41.0	26.1	13.7
2	181	98.5	69.1	132.8	17.5
3	478	328.4	68.1	410.0	11.3
4	829	739.0	11.0	854.2	0.7
5	1112	1182.4	4.2	1265.6	18.6
6	1343	1379.5	1.0	1367.3	0.4
7	1033	1182.4	18.9	1085.2	2.5
8	670	739.0	6.4	628.1	2.8
9	286	328.4	5.5	258.5	2.9
10	104	98.5	0.3	71.8	14.4
11	24	17.9	2.1	12.1	11.7
12	3	1.5	1.5	0.9	4.9
Total	6115	6115	$\mathrm{x}^2 = 249.2$	6115	$\mathrm{x}^2 = 110.5$

Model 1

A simple way for describing the vector of parameters $(p_0, \ldots, p_{12})$ is to use the symmetric binomial distribution $X \sim \mathrm{Bin}(12, 0.5)$. This leads to the setting

$$H_0: (p_0, \ldots, p_{12}) = (p_0^{(0)}, \ldots, p_{12}^{(0)}), \quad \text{where } p_x^{(0)} = \binom{12}{x} \cdot 2^{-12}, \quad x = 0, 1, \ldots, 12,$$

with the set Ω_0 consisting of a single point

$$\Omega_0 = \{(p_0^{(0)}, \ldots, p_{12}^{(0)})\},$$

so that $\dim(\Omega_0) = 0$. Under this H_0, the expected counts are computed as

$$e_x = np_x^{(0)} = 6115 \cdot \binom{12}{x} \cdot 2^{-12}, \quad x = 0, 1, \ldots, 12,$$

see the table below. The observed chi-squared test statistic is $\text{x}^2 = 249.2$, df $= 12$. Since $x_{12}(0.005) = 28.3$, we can reject H_0 at 0.5% level.

Model 2

Consider a more flexible model $X \sim \text{Bin}(12, \lambda)$ with an unspecified probability of a boy λ. The corresponding null hypothesis takes the form

$$H_0 : (p_0, \ldots, p_{12}) = (p_1(\lambda), \ldots, p_{12}(\lambda)), \text{ where } p_x(\lambda) = \binom{12}{x} \cdot \lambda^j (1-\lambda)^{12-x}, \quad x = 0, \ldots, 12, \quad 0 \leq \lambda \leq 1.$$

Clearly, the corresponding parameter space Ω_0 has dimension $\dim(\Omega_0) = 1$. The expected cell counts

$$e_x = 6115 \cdot \binom{12}{x} \cdot \hat{\lambda}^x \cdot (1 - \hat{\lambda})^{12-x}$$

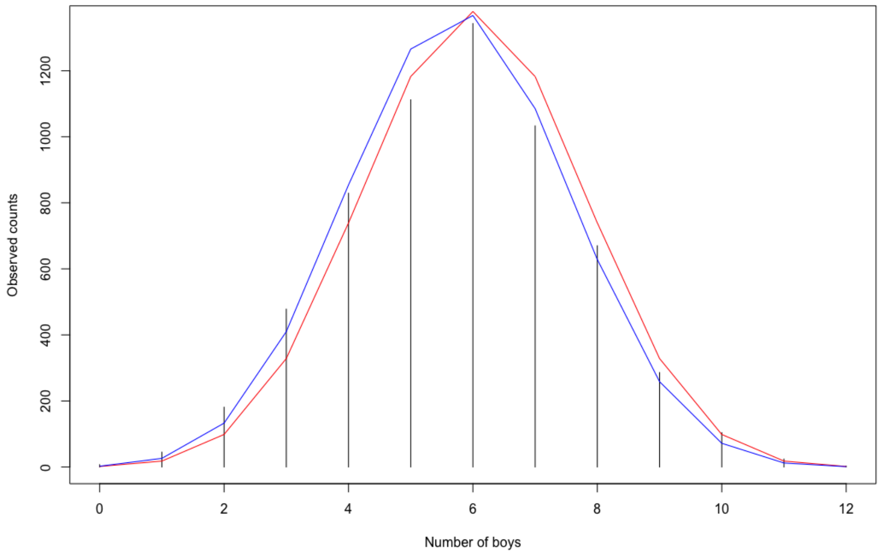
are computed using the maximum likelihood estimate of the proportion of boys λ

$$\hat{\lambda} = \frac{\text{total number of boys}}{\text{total number of children}} = \frac{1 \cdot 45 + 2 \cdot 181 + \ldots + 12 \cdot 3}{6115 \cdot 12} = 0.481$$

The observed chi-squared test statistic $\text{x}^2 = 110.5$ is much smaller than the one for the Model 2. However, since $x_{11}(0.005) = 26.76$, even the Model 2 should be rejected at 0.5% significance level.

Case summary

The figure below compares the observed counts (black histogram) with the Model 1 expected counts (red line) and the Model 2 expected counts (blue line). The red line has a better fit to the data, however it underestimates the variation of the observed cell counts. To make the model even more flexible model, one should allow the probability of a boy λ to differ from family to family.



4.6 Exercises

Problem 1

Suppose that $X \sim \text{Bin}(100, p)$. Consider a test

$$H_0 : p = 1/2, \quad H_1 : p \neq 1/2.$$

that rejects H_0 in favour of H_1 for $|x - 50| > 10$. Use the normal approximation to the binomial distribution to respond to the following items:

(a) What is α?

(b) Graph the power as a function of p.

Problem 2

This problem introduces the case of one-sided null hypothesis. A random sample $(x_1, \ldots, x_n)$ is drawn from a normal population distribution $\mathrm{N}(\mu, 1)$. Consider two alternative composite hypotheses

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0.$$

(a) Rewrite H_0 and H_1 as a pair of nested hypotheses H_0 and H.

(b) Demonstrate that the likelihood function satisfies

$$L(\mu) \propto \exp\{-\tfrac{n}{2}(\mu - \bar{x})^2\}.$$

(c) Show that the corresponding likelihood ratio has the form

$$w = \begin{cases} 1 & \text{if } \bar{x} < \mu_0, \\ e^{-\frac{n}{2}(\bar{x}-\mu_0)^2} & \text{if } \bar{x} \geq \mu_0 \end{cases}$$

(d) Explain why the rejection region of the likelihood ratio test can be expressed as

$$\mathcal{R} = \{\bar{x} > \mu_0 + c_\alpha\},$$

where c_α is determined by the equation

$$\alpha = \max_{\mu \in H_0} \mathrm{P}(\bar{X} > \mu_0 + c_\alpha | \mu).$$

In particular, for $\alpha = 0.05$ and $n = 25$, show that

$$\mathcal{R} = \{\bar{x} > \mu_0 + 0.33\}.$$

Problem 3

Let $(x_1, \ldots, x_n)$ be a sample from a Poisson distribution $\mathrm{Pois}(\mu)$. Find the likelihood ratio for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Use the fact that the sum of independent Poisson random variables follows a Poisson distribution to explain how to determine a rejection region for this test at the significance level α.

Problem 4

Let $(x_1, \ldots, x_{25})$ be a sample from a normal distribution having a variance of 100.

(a) Find the rejection region for a test at level $\alpha = 0.1$ of $H_0 : \mu = 0$ versus $H_1 : \mu = 1.5$.

(b) What is the power of the test?

(c) Repeat (a) and (b) for $\alpha = 0.01$.

Problem 5

Under H_0, a random variable has a cumulative distribution function

$$F(x) = x^2, \quad 0 \leq x \leq 1,$$

and under H_1, it has a cumulative distribution function

$$F(x) = x^3, \quad 0 \leq x \leq 1.$$

(a) What is the form of the likelihood ratio test of H_0 versus H_1?

(b) What is the rejection region of this test for the level α?

(c) What is the power of the test?

Problem 6

A random sample from $\mathrm{N}(\mu, \sigma)$ results in $I_\mu = (-2, 3)$ as a 99% confidence interval for μ. Test

$$H_0 : \mu = -3 \quad \text{against} \quad H_1 : \mu \neq -3$$

at $\alpha = 0.01$.

Problem 7

Let $(x_1, \dots, x_{15})$ be a random sample from a normal distribution $\mathrm{N}(\mu, \sigma)$. The sample standard deviation is $s = 0.7$. Test $H_0 : \sigma = 1$ versus $H_1 : \sigma < 1$ at the significance level $\alpha = 0.05$.

Problem 8

Consider the binomial model for the data value x:

$$X \sim \mathrm{Bin}(n, p).$$

(a) What is the likelihood ratio for testing $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$?

(b) Show that the corresponding likelihood ratio test should reject for larger values of $|x - \frac{n}{2}|$.

(c) For the rejection region

$$\mathcal{R} = \{|x - \tfrac{n}{2}| > k\}$$

determine the significance level α as a function of k.

(d) If $n = 10$ and $k = 2$, what is the significance level of the test?

(e) Use the normal approximation to the binomial distribution to find the significance level given $n = 100$ and $k = 10$.

Problem 9

Suppose that a test statistic Z has a standard normal distribution under the null hypothesis.

(a) If the test rejects for larger values of $|z|$, what is the p-value corresponding to the observed value 1.5 of the test statistic z?

(b) Answer the same question if the test rejects only for larger values of z.

Problem 10

It has been suggested that H_1 : dying people may be able to postpone their death until after an important occasion, such as a wedding or birthday. Phillips and King (1988) studied the patterns of death surrounding Passover, an important Jewish holiday.

(a) California data (1966–1984). They compared the number of deaths during the week before Passover to the number of deaths during the week after Passover for 1919 people who had Jewish surnames. Of these, 922 occurred in the week before and 997 in the week after Passover. Apply a statistical test to see if there is evidence supporting the claim H_1.

(b) For 852 males of Chinese and Japanese ancestry, 418 died in the week before and 434 died in the week after Passover. Can we reject H_0 : death cannot be postponed, using these numbers?

Problem 11

If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities

$$p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2,$$

for some $0 \leq \theta \leq 1$. Plato et al. (1964) published the following data on haptoglobin type in a sample of 190 people

Genotype	AA	Aa	aa
Observed count	10	68	112

Test the goodness of fit of the data to the equilibrium model.

Problem 12

Check for the seasonal variation in the following data on the US suicides in 1970.

Month	Number of suicides
Jan	1867
Feb	1789
Mar	1944
Apr	2094
May	2097
Jun	1981
Jul	1887
Aug	2024
Sep	1928
Oct	2032
Nov	1978
Dec	1859

Problem 13

In 1965, a newspaper carried a story about a high school student who reported getting 9207 heads and 8743 tails in 17950 coin tosses.

(a) Is this a significant discrepancy from the null hypothesis $H_0 : p = \frac{1}{2}$, where p is the probability of heads?

(b) A statistician contacted the student and asked him exactly how he had performed the experiment (Youden 1974). To save time the student had tossed groups of five coins at a time, and a younger brother had recorded the results, shown in the table:

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590

Are the data consistent with the hypothesis that all the coins were fair ($p = \frac{1}{2}$)?

(c) Are the data consistent with the hypothesis that all five coins had the same probability of heads but this probability was not necessarily $\frac{1}{2}$?

Chapter 5

Bayesian inference

The statistical tools introduced in this course so far are based on the so called *frequentist approach.* In the parametric case, the frequentist treats the data x as randomly generated by a distribution $f(x|\theta)$ involving the unknown true population parameter value θ, which may be estimated using the method of maximum likelihood. This section presents basic concepts of the *Bayesian approach* relying on the following model for the observed data x:

$$\text{Apriori distribution } \xrightarrow{g(\theta)} \text{ generates a value } \theta \xrightarrow{f(x|\theta)} \text{ data } x.$$

The model assumes that before the data is collected the parameter of interest θ is randomly generated by a prior distribution $g(\theta)$. The computational power of the Bayesian approach stems from the possibility to treat θ as a realisation of a random variable Θ.

The prior distribution $g(\theta)$ brings into the statistical model our knowledge (or lack of knowledge) on θ before the data x is generated using a conditional distribution $f(x|\theta)$, which in this section is called the likelihood function. After the data x is generated by such a two-step procedure involving the pair $g(\theta)$ and $f(x|\theta)$, we may update our knowledge on θ and compute a posterior distribution $h(\theta|x)$ using the Bayes formula

$$h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\phi(x)}.$$

The denominator, depending on whether the distribution is continuous or discrete,

$$\phi(x) = \int f(x|\theta)g(\theta)d\theta \quad \text{or} \quad \phi(x) = \sum_\theta f(x|\theta)g(\theta)$$

gives the marginal distribution of the random data X. For a fixed realization x, treating the denominator $\phi(x)$ as a constant which does not explicitly involve θ, the Bayes formula can be summarized as

$$\boxed{\text{posterior} \propto \text{likelihood} \times \text{prior}}$$

where the sign $\propto$ means proportional.

If we have no prior knowledge on θ, the prior distribution is often modelled by the uniform distribution. In this case of uninformative prior, with $g(\theta)$ being a constant over a certain interval, we have $h(\theta|x) \propto f(x|\theta)$, implying that the posterior knowledge comes solely from the likelihood function.

Example: IQ measurement

A randomly chosen individual has an unknown true intelligence quotient value θ. Suppose the IQ test is calibrated in such a way that θ can be viewed as a realisation of a random variable Θ having the normal prior distribution $\mathrm{N}(100, 15)$. This normal distribution describes the population distribution of people's IQ with population mean $\mu_0 = 100$ and population standard deviation $\sigma_0 = 15$. For a person with an IQ value θ, the result x of an IQ measurement is generated by another normal distribution $\mathrm{N}(\theta,\, 10)$, with no systematic error and a random error $\sigma = 10$. Since

$$g(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}, \quad f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}},$$

we find that likelihood times prior equals

$$g(\theta)f(x|\theta) = \frac{1}{2\pi\sigma_0\sigma} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2} - \frac{(x-\theta)^2}{2\sigma^2}},$$

which is proportional to

$$\exp\Big\{ -\frac{\theta^2 - 2\mu_1\theta}{2\sigma_1^2} \Big\} \propto \exp\Big\{ -\frac{(\theta-\mu_1)^2}{2\sigma_1^2} \Big\},$$

where

$$\mu_1 = \gamma\mu_0 + (1-\gamma)x, \quad \sigma_1^2 = \gamma\sigma_0^2, \quad \gamma = \frac{\sigma^2}{\sigma^2+\sigma_0^2}.$$

It follows that the posterior distribution is also normal $\mathrm{N}(\mu_1, \sigma_1)$.

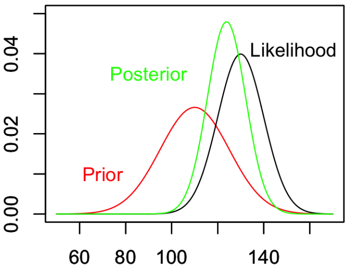
The parameter

$$\gamma = \frac{\sigma_1^2}{\sigma_0^2}$$

is called a shrinkage factor. Being $\gamma \in (0,1)$ it measures the reduction in variance when the posterior distribution is compared to the prior distribution. The smaller is γ the larger is the gain from the data.

In particular, if the observed IQ result is $x = 130$, then the posterior distribution becomes $\mathrm{N}(120.7, 8.3)$. The posterior mean $\mu_1 = 120.7$ is obtained as a down-corrected measurement result $x = 130$ in view of the lower prior expectation $\mu_0 = 100$. The posterior variance $\sigma_1^2 = 69.2$ is smaller than that of the prior distribution $\sigma_0 = 225$ by the shrinkage factor $\gamma = 0.308$, reflecting the fact that the updated knowledge brings much more certainty about the true IQ value compared to the available prior knowledge.

The figure on the right depicts three probability density curves over the values of the parameter θ representing possible IQ values.
Observe that the posterior curve is close to zero
not only for those θ where either the prior curve is close to zero
but also where the likelihood curve is close to zero.
As a result, the posterior curve becomes narrower than the prior curve.



5.1 Conjugate priors

Suppose the data x is generated by a parametric model having the likelihood function $f(x|\theta)$. Consider a parametric family of the prior distributions $\mathcal{G}$.

> $\mathcal{G}$ is called a family of conjugate priors for the likelihood function $f(x|\theta)$
> if for any prior $g(\theta) \in \mathcal{G}$, the corresponding posterior distribution $h(\theta|x) \in \mathcal{G}$

The next table presents five Bayesian models involving conjugate priors. The details of the first three models come next. Notice that the posterior variance is always smaller than the prior variance. This list also illustrates that the contribution of the prior distribution to the posterior distribution becomes smaller as the sample size n increases.

Parametric model for the data	Unknown θ	Prior	Posterior distribution
$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma)$	$\theta = \mu$	$\mathrm{N}(\mu_0, \sigma_0)$	$\mathrm{N}(\gamma_n\mu_0 + (1-\gamma_n)\bar{x}; \sigma_0\sqrt{\gamma_n})$
$X \sim \mathrm{Bin}(n, p)$	$\theta = p$	$\mathrm{Beta}(a, b)$	$\mathrm{Beta}(a + x, b + n - x)$
$(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$	$\theta = (p_1, \ldots, p_r)$	$\mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$	$\mathrm{Dir}(\alpha_1 + x_1, \ldots, \alpha_r + x_r)$
$X_1, \ldots, X_n \sim \mathrm{Geom}(p)$	$\theta = p$	$\mathrm{Beta}(a, b)$	$\mathrm{Beta}(a + n, b + n\bar{x} - n)$
$X_1, \ldots, X_n \sim \mathrm{Pois}(\mu)$	$\theta = \mu$	$\mathrm{Gam}(\alpha_0, \lambda_0)$	$\mathrm{Gam}(\alpha_0 + n\bar{x}, \lambda_0 + n)$
$X_1, \ldots, X_n \sim \mathrm{Gam}(\alpha, \lambda)$	$\theta = \lambda$	$\mathrm{Gam}(\alpha_0, \lambda_0)$	$\mathrm{Gam}(\alpha_0 + \alpha n, \lambda_0 + n\bar{x})$

Normal-normal model

Suppose a random sample $(x_1, \ldots, x_n)$ is drawn from the normal distribution $\mathrm{N}(\mu, \sigma)$ with a known standard deviation σ and the unknown mean $\theta = \mu$. Taking the normal prior $\Theta \sim \mathrm{N}(\mu_0, \sigma_0)$ with known (μ_0, σ_0) results in the normal posterior $\mathrm{N}(\mu_1, \sigma_1)$ with

$$\mu_1 = \gamma_n\mu_0 + (1-\gamma_n)\bar{x}, \quad \sigma_1^2 = \sigma_0^2\gamma_n,$$

where

$$\gamma_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \sigma_0^2}$$

is the shrinkage factor which becomes smaller for the larger sample sizes n. As a result for the large samples, the posterior mean μ_1 gets close to the maximum likelihood estimate and the input $\gamma_n\mu_0$ involving the prior mean becomes negligible.

Binomial-beta model

Next, we introduce the beta distribution which serves as a convenient family of conjugate priors for Bayesian inference for p, in the case when the data x is generated by the $\mathrm{Bin}(n, p)$.

Beta distribution

Beta distribution $\text{Beta}(a, b)$ is determined by two parameters $a > 0$, $b > 0$ which are called pseudo-counts. It is defined by the probability density function

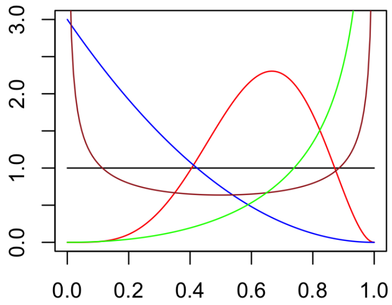
$$g(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1,$$

its mean and variance are given by

$$\mu = \frac{a}{a+b}, \quad \sigma^2 = \frac{\mu(1-\mu)}{a+b+1}.$$

The figure below depicts five different beta-distribution curves.

- flat black $\text{Beta}(1, 1)$,
- U-shaped brown $\text{Beta}(0.5, 0.5)$,
- bell-shaped red $\text{Beta}(5, 3)$,
- L-shaped blue $\text{Beta}(1, 3)$,
- J-shaped green $\text{Beta}(3, 0.5)$.



Proof of the conjugacy property

To demonstrate that the beta distribution is a conjugate prior for the binomial likelihood, observe that

$$\text{prior} \ \propto p^{a-1}(1-p)^{b-1},$$

and

$$\text{likelihood} \ \propto p^{x}(1-p)^{n-x},$$

imply

$$\text{posterior} \ \propto \text{prior} \ \times \text{likelihood} \ \propto p^{a+x-1}(1-p)^{b+n-x-1}.$$

This entails that the posterior is also a beta distribution $\text{Beta}(a_1, b_1)$ with the updated parameters

$$a_1 = a + x, \quad b_1 = b + n - x.$$

Example: thumbtack landing on its base

Suppose we are interested in the probability p of a thumbtack landing on its base. Two experiments are performed. An experiment consists of n tosses of the thumbtack with the number of base landings $X \sim \text{Bin}(n, p)$ being counted.

Experiment 1: after $n_1 = 10$ tosses, the observed count of the base landings is $x_1 = 2$. We apply the uninformative prior distribution $\text{Beta}(1, 1)$ with the mean $\mu_0 = 0.50$ and standard deviation $\sigma_0 = 0.29$. The resulting posterior distribution is the Beta(3, 9) distribution with the posterior mean $\mu_1 = \frac{3}{12} = 0.25$ and standard deviation $\sigma_1 = 0.12$.

Experiment 2: after $n_2 = 40$ tosses, the observed count of the base landings is $x_2 = 9$. As a new prior distribution we use the posterior distribution obtained from the first experiment Beta(3, 9). The new posterior distribution becomes Beta(12, 40) with the mean $\mu_2 = \frac{12}{52} = 0.23$ and standard deviation $\sigma_2 = 0.06$.

Multinomial-Dirichlet model

The multinomial-Dirichlet model is a multivariate version of the binomial-beta model. For both the binomial-beta and multinomial-Dirichlet models, the updating rule has the form

$$\boxed{\text{the posterior pseudo-counts} = \text{the prior pseudo-counts plus the sample counts}}$$

Dirichlet distribution

The Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_r)$ is a multivariate extension of the beta distribution. It is a probability distribution over the vectors $(p_1, \ldots, p_r)$ with non-negative components such that

$$p_1 + \ldots + p_r = 1.$$

The positive parameters $\alpha_1, \ldots, \alpha_r$ of the Dirichlet distribution are often called the pseudo-counts. The probability density function of $\text{Dir}(\alpha_1, \ldots, \alpha_r)$ is given by

$$g(p_1, \ldots, p_r) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_r)} p_1^{\alpha_1 - 1} \ldots p_r^{\alpha_r - 1}, \quad \alpha_0 = \alpha_1 + \ldots + \alpha_r.$$

The marginal distributions of the random vector $(X_1, \ldots, X_r) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$ are the beta distributions

$$X_j \sim \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j), \quad j = 1, \ldots, r.$$

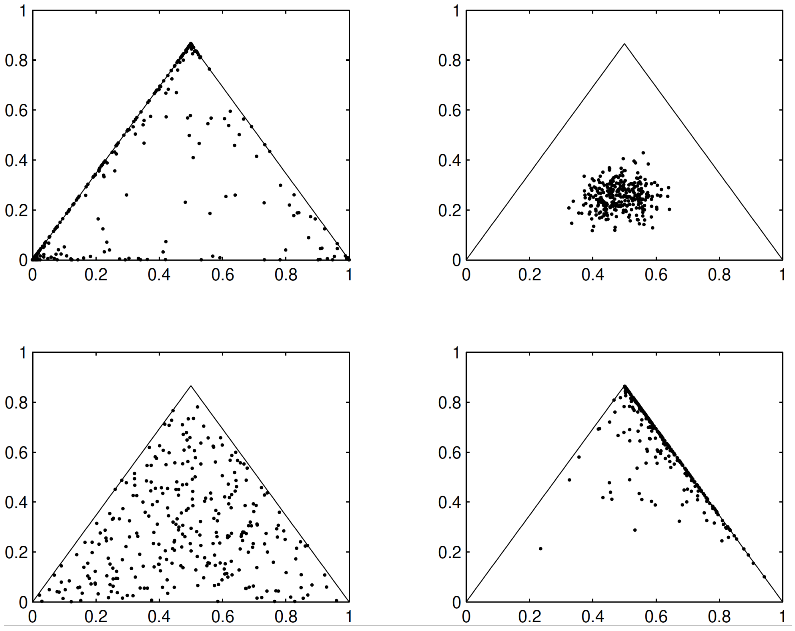
Different components of the vector have negative covariances

$$\mathrm{Cov}(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \text{ for } i \neq j.$$

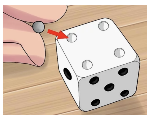
The figure below illustrates four examples of $\mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3)$ distribution. Each triangle contains $n = 300$ points generated using different sets of parameters $(\alpha_1, \alpha_2, \alpha_3)$:

upper left $(0.3, 0.3, 0.1)$, upper right $(13, 16, 15)$, lower left $(1, 1, 1)$, lower right $(3, 0.1, 1)$.

A dot in a triangle gives a realisation (x_1, x_2, x_3) of the vector $(X_1, X_2, X_3) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \alpha_3)$ as the distances to the bottom edge of the triangle (x_1), to the right edge of the triangle (x_2), and to the left edge of the triangle (x_3).



Example: loaded die experiment



A possibly loaded die is rolled 18 times, giving 4 ones, 3 twos, 4 threes, 4 fours, 3 fives, and 0 sixes:

$$2, 1, 1, 4, 5, 3, 3, 2, 4, 1, 4, 2, 3, 4, 3, 5, 1, 5.$$

The parameter of interest is the vector of six probabilities $\theta = (p_1, \ldots, p_6)$. The data can be viewed as generated by a multinomial distribution $\mathrm{Mn}(18; p_1, \ldots, p_6)$. Since we have no idea about the values of the probabilities $(p_1, \ldots, p_6)$ we will use the uninformative prior distribution $\mathrm{Dir}(1, 1, 1, 1, 1, 1)$. Due to the conjugacy property we obtain the posterior distribution to be $\mathrm{Dir}(5, 4, 5, 5, 4, 1)$, where the posterior pseudo-counts are computed as

$$(5, 4, 5, 5, 4, 1) = (1, 1, 1, 1, 1, 1) + (4, 3, 4, 4, 3, 0).$$

5.2 Bayesian estimation

In the language of decision theory, finding a point estimate a for the unknown population parameter θ is an action of assigning the value a to the unknown parameter θ. In the frequentist setting, the optimal a is found by maximising the likelihood function. In the Bayesian setting, the optimal choice of a is determined by the so-called loss function $l(\theta, a)$. The so-called Bayes action minimises the posterior risk

$$R(a|x) = \mathrm{E}\big(l(\Theta, a)|x\big),$$

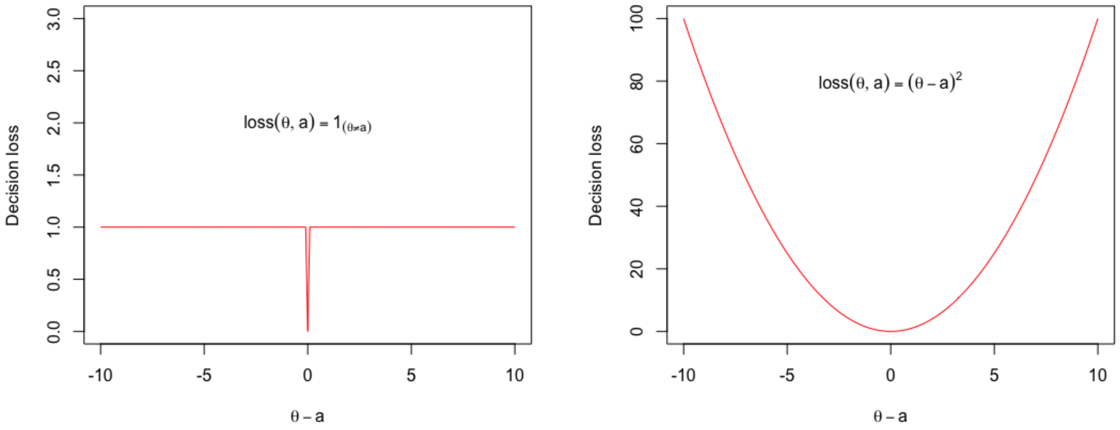
computed using the posterior distribution

$$R(a|x) = \int l(\theta, a) h(\theta|x) d\theta \quad \text{or} \quad R(a|x) = \sum_\theta l(\theta, a) h(\theta|x).$$

We consider two loss functions leading to two different Bayesian estimators. These two loss functions called the zero-one loss and the squared error loss

$\boxed{\text{Zero-one loss function: } l(\theta, a) = 1_{\{\theta \neq a\}}}$ $\boxed{\text{Squared error loss: } l(\theta, a) = (\theta - a)^2}$

are schematically depicted on the figure below.



Zero-one loss function and maximum a posteriori probability

With the zero-one loss function, the posterior risk is equal to the probability of misclassification

$$R(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x).$$

In this case, to minimise the risk we have to maximise the posterior probability $h(a|x)$. We define $\hat{\theta}_{\text{map}}$ as the value of θ that maximises $h(\theta|x)$. Observe that with the uninformative prior, $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$.

Squared error loss function and posterior mean estimate

Using the squared error loss function we find that the posterior risk is a sum of two components

$$R(a|x) = \mathrm{E}((\Theta - a)^2|x) = \mathrm{Var}(\Theta|x) + (\mathrm{E}(\Theta|x) - a)^2.$$

Since the first component is independent of a, we minimise the posterior risk by putting

$$\hat{\theta}_{\text{pme}} = \mathrm{E}(\Theta|x),$$

resulting in the posterior mean value as the Bayesian point estimate of θ.

Example: loaded die experiment

Turning to the loaded die experiment, observe that the maximum likelihood estimate based on the sample counts is given by the vector of sample proportions

$$\hat{\theta}_{\text{mle}} = (\tfrac{4}{18}, \tfrac{3}{18}, \tfrac{4}{18}, \tfrac{4}{18}, \tfrac{3}{18}, 0).$$

Notably, the maximum likelihood estimate assigns value zero to p_6, thereby predicting that in the future observations there will be no sixes.

Considering the two alternative Bayesian estimates based on the posterior distribution $\mathrm{Dir}(5, 4, 5, 5, 4, 1)$

$$\hat{\theta}_{\text{map}} = (\tfrac{4}{18}, \tfrac{3}{18}, \tfrac{4}{18}, \tfrac{4}{18}, \tfrac{3}{18}, 0), \quad \hat{\theta}_{\text{pme}} = (\tfrac{5}{24}, \tfrac{4}{24}, \tfrac{5}{24}, \tfrac{5}{24}, \tfrac{4}{24}, \tfrac{1}{24}),$$

we see that the former coincides with $\hat{\theta}_{\text{mle}}$, while the latter estimate has the advantage of assigning a positive value to p_6.

5.3 Credibility interval

Given data x coming from a parametric model with the likelihood function $f(x|\theta)$, a $100(1-\alpha)\%$ confidence interval for the parameter θ,

$$I_\theta = (a_1(x), a_2(x)),$$

is viewed as a realisation of a random interval $(a_1(X), a_2(X))$ such that

$$\mathrm{P}(a_1(X) < \theta < a_2(X)) = 1 - \alpha.$$

This frequentist interpretation of the confidence level $100(1-\alpha)\%$ is rather cumbersome as it requires mentioning other samples and potential confidence intervals which as a group cover the true unknown value of θ with probability $1-\alpha$.

In the framework of Bayesian inference we can refer to θ as a realisation of a random variable Θ with a certain posterior distribution $h(\theta|x)$. This allows us to define a $100(1-\alpha)\%$ credibility interval (or credible interval)

$$J_\theta = (b_1(x), b_2(x))$$

by the relation based on the posterior distribution

$$\mathrm{P}(b_1(x) < \Theta < b_2(x)|x) = 1 - \alpha.$$

The interpretation of the credibility interval is more intuitive as it does not refer to some potential, never observed data values.

Example: IQ measurement

Given a single IQ value $x = 130$, we have $\bar{X} \sim \mathrm{N}(\mu; 10)$ and an exact 95% confidence interval for the true IQ value θ takes the form

$$I_\theta = 130 \pm 1.96 \cdot 10 = 130 \pm 19.6.$$

The interval 130 ± 19.6 has 95% confidence level in the sense that if we repeat the IQ measurement for the same person, then the new IQ result X will produce a random confidence interval which will cover the true IQ of the subject with probability 0.95.

With the posterior distribution $\Theta \sim \mathrm{N}(120.7; 8.3)$ in hand, a 95% credibility interval for θ is computed as

$$J_\mu = 120.7 \pm 1.96 \cdot 8.3 = 120.7 \pm 16.3.$$

The proper understanding of the obtained interval 120.7 ± 16.3 is the following. If we choose another person from the general population and the person's IQ turns out to be $x = 130$, then the new person's IQ Θ belongs to the interval 120.7 ± 16.3 with probability 0.95. On the other hand, if we get a second IQ test result x_2 for the same person, then we can compute a new credibility interval based on x_2 using the posterior distribution after the first IQ test result as the new prior.

5.4 Bayesian hypotheses testing

Considering the case of two simple hypotheses

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta = \theta_1$$

we wish to choose between H_0 and H_1 using not only the two likelihood functions $f(x|\theta_0)$, $f(x|\theta_1)$ but also the prior probabilities of the two optional values

$$\mathrm{P}(H_0) = \pi_0, \quad \mathrm{P}(H_1) = \pi_1, \quad \pi_0 + \pi_1 = 1.$$

In terms of an appropriate rejection region $\mathcal{R}$ for the available data x, the Bayesian decision should be taken depending of a cost function having the following four cost values

	Decision	H_0 true	H_1 true
$x \notin \mathcal{R}$	Do not reject H_0	0	cost_1
$x \in \mathcal{R}$	Reject H_0	cost_0	0

where cost_0 is the error type I cost and cost_1 is the error type II cost. For a given set $\mathcal{R}$, the average cost is the weighted mean of two values cost_0 and cost_1

$$\mathrm{cost}_0\pi_0\mathrm{P}(X \in \mathcal{R}|H_0) + \mathrm{cost}_1\pi_1\mathrm{P}(X \notin \mathcal{R}|H_1) = \mathrm{cost}_1\pi_1 + \int_{\mathcal{R}} \Big(\mathrm{cost}_0\pi_0 f(x|\theta_0) - \mathrm{cost}_1\pi_1 f(x|\theta_1)\Big)dx.$$

Now observe that

$$\int_{\mathcal{R}} \Big(\mathrm{cost}_0\pi_0 f(x|\theta_0) - \mathrm{cost}_1\pi_1 f(x|\theta_1)\Big)dx \geq \int_{\mathcal{R}^*} \Big(\mathrm{cost}_0\pi_0 f(x|\theta_0) - \mathrm{cost}_1\pi_1 f(x|\theta_1)\Big)dx,$$

where

$$\mathcal{R}^* = \{x : \mathrm{cost}_0\pi_0 f(x|\theta_0) < \mathrm{cost}_1\pi_1 f(x|\theta_1)\}.$$

It follows that the rejection region minimising the average cost is $\mathcal{R} = \mathcal{R}^*$. Taking $\mathcal{R}^*$ as the rejection region, we should reject H_0 for values of the likelihood ratio which are smaller than a certain critical value:

$$\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{\text{cost}_1\pi_1}{\text{cost}_0\pi_0},$$

determined by the prior odds π_0/π_1 and the cost ratio $\text{cost}_1/\text{cost}_0$. In other terms, we reject H_0 for the values of the posterior odds smaller than the cost ratio

$$\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{\text{cost}_1}{\text{cost}_0}.$$

Example of Bayesian hypothesis testing

The defendant N charged with rape, is a male of age 37 living in the area not very far from the crime place. The jury have to choose between two alternative hypotheses H_0: N is innocent, H_1: N is guilty. There are three conditionally independent pieces of evidence

E_1: a DNA match,

E_2: defendant N is not recognised by the victim,

E_3: an alibi supported by the N's girlfriend.

The reliability of these pieces of evidence was quantified as

$$\begin{array}{lll}
\mathrm{P}(E_1|H_0) = \frac{1}{200{,}000{,}000}, & \mathrm{P}(E_1|H_1)=1, & \text{very strong evidence in favour of } H_1 \text{ with } \frac{\mathrm{P}(E_1|H_0)}{\mathrm{P}(E_1|H_1)} = \frac{1}{200{,}000{,}000} \\
\mathrm{P}(E_2|H_1) = 0.1, & \mathrm{P}(E_2|H_0) = 0.9, & \text{strong evidence in favour of } H_0 \text{ with } \frac{\mathrm{P}(E_2|H_0)}{\mathrm{P}(E_2|H_1)} = 9 \\
\mathrm{P}(E_3|H_1) = 0.25, & \mathrm{P}(E_3|H_0) = 0.5, & \text{evidence in favour of } H_0 \text{ with } \frac{\mathrm{P}(E_3|H_0)}{\mathrm{P}(E_3|H_1)} = 2
\end{array}$$

For the sake of Bayesian inference the non-informative prior probability

$$\pi_1 = \mathrm{P}(H_1) = \tfrac{1}{200000},$$

is suggested, taking into account the number of males who theoretically could have committed the crime without any evidence taken into account. This yields the prior odds for H_0 to be very high

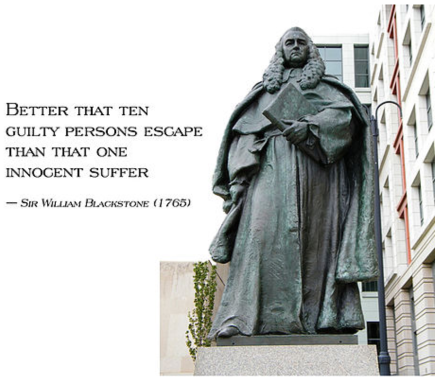
$$\frac{\pi_0}{\pi_1} = 200000.$$

The resulting posterior odds is

$$\frac{\mathrm{P}(H_0|E_1,E_2,E_3)}{\mathrm{P}(H_1|E_1,E_2,E_3)} = \frac{\pi_0\mathrm{P}(E_1,E_2,E_3|H_0)}{\pi_1\mathrm{P}(E_1,E_2,E_3|H_1)} = \frac{\pi_0}{\pi_1}\cdot\frac{\mathrm{P}(E_1|H_0)}{\mathrm{P}(E_1|H_1)}\cdot\frac{\mathrm{P}(E_2|H_0)}{\mathrm{P}(E_2|H_1)}\cdot\frac{\mathrm{P}(E_3|H_0)}{\mathrm{P}(E_3|H_1)} = 0.018.$$

Conclusion: the defendant N would deemed to be guilty if the cost values assigned by the jury are such that

$$\frac{\text{cost}_1}{\text{cost}_0} = \frac{\text{cost for unpunished crime}}{\text{cost for punishing an innocent}} > 0.018.$$



5.5 Exercises

Problem 1

This is a continuation of the Problem 3 (a-d) from Section 3.4.

(e) Assume the uniform prior for the parameter θ and find the posterior density. Sketch the posterior curve. Find the MAP estimate of θ.

Problem 2

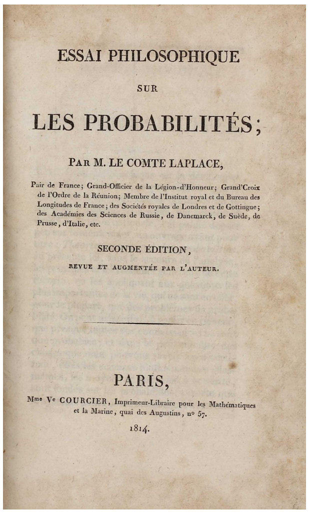
In an ecological study of the feeding behaviour of birds, the number of hops between flights was counted for several birds.

Number of hops j	1	2	3	4	5	6	7	8	9	10	11	12	Tot
Observed count c_j	48	31	20	9	6	5	4	2	1	1	2	1	130

Assume that the data were generated by a Geom(p) model and take the uniform prior for p. What is the posterior distribution and what are the posterior mean and standard deviation?

Problem 3

On the Laplace rule of succession.
Laplace claimed that when an event happens n times in a row and never fails to happen, the probability that the event will occur the next time is $\frac{n+1}{n+2}$.
Can you suggest a rationale for this claim?



ESSAI PHILOSOPHIQUE

SUR

LES PROBABILITÉS;

PAR M. LE COMTE LAPLACE,

Pair de France; Grand-Officier de la Légion-d'Honneur; Grand'Croix de l'Ordre de la Réunion; Membre de l'Institut royal et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, d'Italie, etc.

SECONDE ÉDITION,
REVUE ET AUGMENTÉE PAR L'AUTEUR.

PARIS,
Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques
et la Marine, quai des Augustins, nº 57.
1814.

Problem 4

It is known that the random variable X has one of the following two distributions

X-values	x_1	x_2	x_3	x_4
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.3	0.2
$\mathrm{P}(x\|H_1)$	0.1	0.4	0.1	0.4

(a) Compare the likelihood ratio, $\frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)}$, for each x_i and order the x_i accordingly.

(b) What is the likelihood ratio test of H_0 versus H_1 at level $\alpha = 0.2$? What is the test at level $\alpha = 0.5$?

(c) If the prior probabilities are $\mathrm{P}(H_0) = \mathrm{P}(H_1) = \frac{1}{2}$, which outcomes favour H_0?

(d) What prior probabilities correspond to the decision rules with $\alpha = 0.2$ and $\alpha = 0.5$?

Problem 5

Suppose that under H_0, a measurement X is $\mathrm{N}(0, \sigma)$, and under H_1, the measurement X is $\mathrm{N}(1, \sigma)$. Assume that the prior probabilities satisfy

$$\mathrm{P}(H_0) = 2\mathrm{P}(H_1).$$

The hypothesis H_0 will be chosen if $\mathrm{P}(H_0|x) > \mathrm{P}(H_1|x)$. Answer the following questions referring to the different choices of $\sigma^2 = 0.1, 0.5, 1.0, 5.0$.

(a) For what values of $X = x$ will H_0 be chosen?

(b) In the long run, what proportion of the time will H_0 be chosen if H_0 is true $\frac{2}{3}$ of the time?

Problem 6

Under H_0, a random variable X has a cumulative distribution function $F(x) = x^2, \quad 0 \leq x \leq 1$, and under H_1, it has a cumulative distribution function $F(x) = x^3, \quad 0 \leq x \leq 1$.

If the two hypotheses have equal prior probabilities, for what values of x is the posterior probability of H_0 greater than that of H_1?

Problem 7

Suppose your prior beliefs about the probability p of success have mean 1/3 and variance 1/32. What is the posterior mean after having observed 8 successes in 20 trials?

Problem 8

Mice were injected with a bacterial solution, some of the mice were also given penicillin. The results were

	Without penicillin	With penicillin
Survived	8	12
Died	48	62

(a) Find a 95% confidence interval for the difference between two probabilities of survival.

(b) Assume that both groups have the probability of survival p. How would you compute an exact credibility interval for the population proportion p, if you could use a computer? Compute an approximate 95% credibility interval using a normal approximation.

Problem 9

The gamma distribution $\mathrm{Gam}(\alpha, \lambda)$ is a conjugate prior for the Poisson likelihood with mean θ. If x is a single observed value randomly sampled from the $\mathrm{Pois}(\theta)$ distribution, then the parameters (α_1, λ_1) for the posterior gamma distribution of Θ are found by the following updating rule:
- the shape parameter $\alpha_1 = \alpha + x$,
- the inverse scale parameter $\lambda_1 = \lambda + 1$.

(a) Find $\hat{\theta}_{\mathrm{PME}}$, the posteriori mean estimate for the θ, under the exponential prior $\mathrm{Exp}(1)$, given the following random sample values from the $\mathrm{Pois}(\theta)$ population distribution

$$x_1 = 2, \quad x_2 = 0, \quad x_3 = 2, \quad x_4 = 5.$$

(b) What is the updating rule for an arbitrary sample size n? Compare the value of $\hat{\theta}_{\mathrm{PME}}$ with the maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ as $n \to \infty$. Your conclusions?

Problem 10

Consider the $\mathrm{Beta}(a, b)$ distribution. Verify that given $a > 1$ and $b > 1$, the maximum of the probability density function $g(p)$ is attained at

$$p^* = \frac{a-1}{a+b-2}.$$

Problem 11

Credible intervals are not unique on a posterior distribution. Methods for defining a suitable credible interval include:

1. choosing the narrowest interval, which for a unimodal distribution will involve choosing those values of highest probability density including the mode (the maximum a posteriori). This is sometimes called the highest posterior density interval (HPDI),

2. choosing the interval where the probability of being below the interval is as likely as being above it, this is sometimes called the equal-tailed interval,

3. the interval for which the posterior mean is the central point.

Respond to the following:

(a) demonstrate that the equal-tailed interval includes the posterior median,

(b) compare these three options for the beta-posterior $\mathrm{Beta}(5, 5)$,

(c) compute the 95% HPDI for the normal-posterior $\mathrm{N}(3, 0.2)$.

Chapter 6

Summarising data

Consider a random sample $(x_1, \ldots, x_n)$ taken from the unknown population distribution $\mathcal{F}(\mu, \sigma)$ characterized by the cumulative distribution function

$$F(x) = \mathrm{P}(X \le x).$$

Even for moderate sample sizes n, a proper preprocessing of the dataset is crucial. A very useful first rearrangement is to sort the random sample $(x_1, \ldots, x_n)$ in the ascending order. As a result, we arrive at the ordered sample values

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)},$$

where in particular,

$$x_{(1)} = \min\{x_1, \ldots, x_n\}, \quad x_{(n)} = \max\{x_1, \ldots, x_n\}.$$

Doing this we do not lose any essential information about the data making it much more tractable. For example, the sample mean and sample standard deviations can be computed from the ordered sample $(x_{(1)}, \ldots, x_{(n)})$. In this chapter we present several other basic tools for summarising the data in hand.

6.1 Empirical probability distribution

Given a random sample $(x_1, \ldots, x_n)$, the unknown population distribution function $F(x)$ can be estimated by the empirical distribution function defined as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \le x\}}.$$

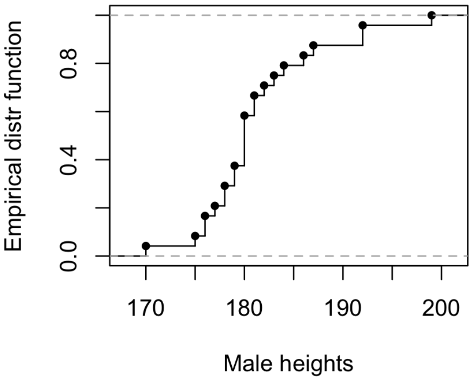
Observe that for a fixed x,

$$\hat{F}(x) = \hat{p}$$

is the sample proportion estimating the population proportion $p = F(x)$. This implies that $\hat{F}(x)$ is an unbiased consistent estimate of $F(x)$. The next figure illustrates this definition in terms of the earlier mentioned $n = 24$ male heights

$$(x_{(1)}, \ldots, x_{(24)}) =$$
$$(170, 175, 176, 176, 177, 178, 178, 179, 179, 180, 180, 180, 180, 180, 181, 181, 182, 183, 184, 186, 187, 192, 192, 199).$$

The graph of the empirical distribution function $\hat{F}(x)$ gives an idea about the shape of the population distribution function $F(x)$.



As a function of x, $\hat{F}(x)$ is a cumulative distribution function for a random variable Y with the discrete uniform distribution

$$\mathrm{P}(Y = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n,$$

assuming that all sample values x_i are pairwise different. Clearly,

$$\mathrm{E}(Y) = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x},$$

and since

$$\mathrm{E}(Y^2) = \sum_{i=1}^{n} \frac{x_i^2}{n} = \overline{x^2},$$

we get

$$\mathrm{Var}(Y) = \overline{x^2} - (\bar{x})^2 = \frac{n-1}{n} s^2.$$

It is easy to verify that even if some of x_i coincide, $\hat{F}(\cdot)$ is a cumulative distribution function with mean $\bar{x}$ and variance $\frac{n-1}{n}s^2$. We will call

$$\hat{\sigma}^2 = \tfrac{n-1}{n} s^2 = \tfrac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

the *empirical variance.*

Empirical survival function

If the data describes life lengths, then it is more convenient to use the empirical survival function

$$\hat{S}(x) = 1 - \hat{F}(x),$$

the proportion of the data greater than x. If the life length L has distribution function $F(x) = \mathrm{P}(L \leq x)$, then its survival function is

$$S(x) = \mathrm{P}(L > x) = 1 - F(x), \quad x \geq 0.$$

If $f(x) = F'(x)$ is the probability density function for the life length L, then the corresponding hazard function is defined by

$$h(x) = \frac{f(x)}{S(x)}.$$

The hazard $h(x)$ is the mortality rate at age x in that

$$\frac{\mathrm{P}(x < L \leq x + \delta | L \geq x)}{\delta} = \frac{\mathrm{P}(x < L \leq x + \delta)}{\delta \mathrm{P}(L \geq x)} = \frac{F(x+\delta) - F(x)}{\delta S(x)} \to h(x), \quad \delta \to 0.$$

The hazard function can be viewed as the negative slope of the log survival function:

$$h(x) = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = -\tfrac{d}{dx} \ln S(x).$$

$$\boxed{\text{If } L \text{ has the exponential distribution } \mathrm{Exp}(\lambda)\text{, then the hazard rate } h(x) \equiv \lambda \text{ is constant over the ages.}}$$

Example: guinea pigs

Guinea pigs were randomly divided in 5 treatment groups of 72 animals each and one control group of 107 animals. The guinea pigs in the treatment groups were infected with increasing doses of tubercle bacilli (Bjerkdal, 1960). The survival times were recorded in days (note that not all the animals in the lower-dosage regimens died).

Control lifetimes

18 36 50 52 86 87 89 91 102 105 114 114 115 118 119 120 149 160 165 166 167 167 173 178 189 209 212 216 273 278 279 292 341 355 367 380 382 421 421 432 446 455 463 474 506 515 546 559 576 590 603 607 608 621 634 634 637 638 641 650 663 665 688 725 735

Dose I lifetimes

76 93 97 107 108 113 114 119 136 137 138 139 152 154 154 160 164 164 166 168 178 179 181 181 183 185 194 198 212 213 216 220 225 225 244 253 256 259 265 268 268 270 283 289 291 311 315 326 326 361 373 373 376 397 398

406 452 466 592 598

Dose II lifetimes

72 72 78 83 85 99 99 110 113 113 114 114 118 119 123 124 131 133 135 137 140 142 144 145 154 156 157 162 162 164 165 167 171 176 177 181 182 187 192 196 211 214 216 216 218 228 238 242 248 256 257 262 264 267 267 270 286 303 309 324 326 334 335 358 409 473 550

Dose III lifetimes

10 33 44 56 59 72 74 77 92 93 96 100 100 102 105 107 107 108 108 108 109 112 113 115 116 120 121 122 122 124 130 134 136 139 144 146 153 159 160 163 163 168 171 172 176 183 195 196 197 202 213 215 216 222 230 231 240 245 251 253 254 254 278 293 327 342 347 361 402 432 458 555

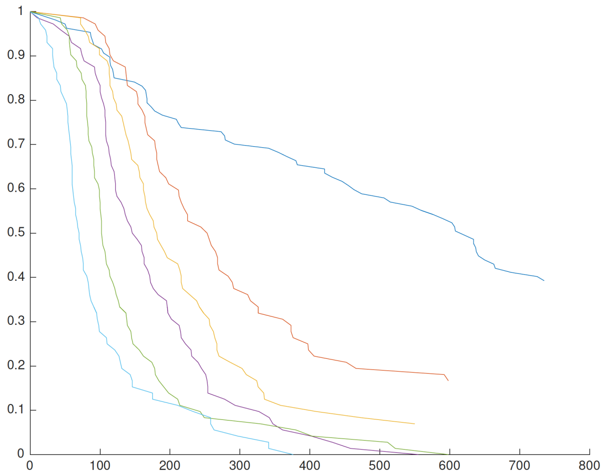
Dose IV lifetimes

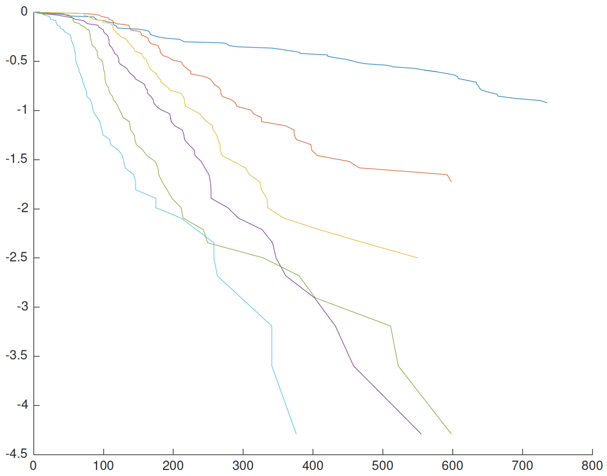
43 45 53 56 56 57 58 66 67 73 74 79 80 80 81 81 81 82 83 83 84 88 89 91 91 92 92 97 99 99 100 100 101 102 102 102 103 104 107 108 109 113 114 118 121 123 126 128 137 138 139 144 145 147 156 162 174 178 179 184 191 198 211 214 243 249 329 380 403 511 522 598

Dose V lifetimes

12 15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56 57 58 58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75 76 76 81 83 84 85 87 91 95 96 98 99 109 110 121 127 129 131 143 146 146 175 175 211 233 258 258 263 297 341 341 376

It is difficult to compare the six groups just looking at the numbers. The data is illuminated by two graphs: the left one for the survival functions and the right one for the log-survival functions.





The negative slopes of the curves to the right illustrate the hazard rates for different groups. Notice that the top line is almost linear. This suggests an exponential model for the life lengths of the guinea pigs in the control group, with the constant hazard rate $\lambda \approx 0.001$ deaths per day.

6.2 Quantiles and QQ-plots

The inverse of the cumulative distribution function $F(x)$ is called the quantile function

$$Q(p) = F^{-1}(p), \quad 0 < p < 1.$$

For a given population distribution function F and $0 < p < 1$, the population p-quantile is defined by

$$x_p = Q(p).$$

The following three quantiles are of special importance:

median	$m = x_{0.5} = Q(0.5)$,
lower quartile	$x_{0.25} = Q(0.25)$,
upper quartile	$x_{0.75} = Q(0.75)$.

The p-quantile x_p cuts off the proportion p of smallest values for the random variable X with $\mathrm{P}(X \le x) = F(x)$:

$$\mathrm{P}(X \le x_p) = F(x_p) = F(Q(p)) = p.$$

In the continuous distribution case, the ordered random sample

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)},$$

gives the strictly ordered n jump points for the empirical distribution function $\hat{F}$, so that

$$\hat{F}(x_{(k)}) = \tfrac{k}{n}, \qquad \hat{F}(x_{(k)} - \epsilon) = \tfrac{k-1}{n}.$$

This observation leads to the following definition of empirical quantiles

$$\boxed{x_{(k)} \text{ is called the empirical } (\tfrac{k-0.5}{n})\text{-quantile}}$$

QQ-plots and normal QQ-plots

Suppose we have two independent samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ of equal size n which are taken from two population distributions with distribution functions F_1 and F_2. A relevant null hypothesis H_0: $F_1 \equiv F_2$ has an equivalent expression in terms of the quantile functions H_0: $Q_1 \equiv Q_2$. The latter hypothesis can be tested graphically using a QQ-plot.

$$\boxed{\text{QQ-plot is a scatter plot of } n \text{ dots with coordinates } (x_{(k)}, y_{(k)}).}$$

If such a QQ-plot closely follows the 45 degree line, that is when we observe almost equal quantiles, we may conclude that the data supports the null hypothesis of equality H_0: $F_1 \equiv F_2$.

More generally, if the QQ-plot approximates a straight line $y = a + bx$, then we take this as evidence for the linear relation

$$Y = a + bX \text{ in distribution.}$$

Indeed, the latter claim means that for all x,

$$F_1(x) = F_2(a + bx),$$

so that putting $Q_1(p) = x$, we get $Q_2(p) = a + bx$, which yields the linear relationship for the QQ-plot

$$Q_2(p) = a + bQ_1(p), \quad 0 < p < 1.$$

The normality hypothesis H_0 states that the random sample $(x_1, \ldots, x_n)$ is drawn from the normal distribution

$$\mathcal{F}(\mu, \sigma) = \mathrm{N}(\mu, \sigma),$$

with unspecified parameter values. A QQ-plot used for testing this hypothesis is called a normal QQ-plot or normal probability plot. To define the normal QQ-plot, let

$$y_k = \Phi^{-1}(\tfrac{k-0.5}{n}),$$

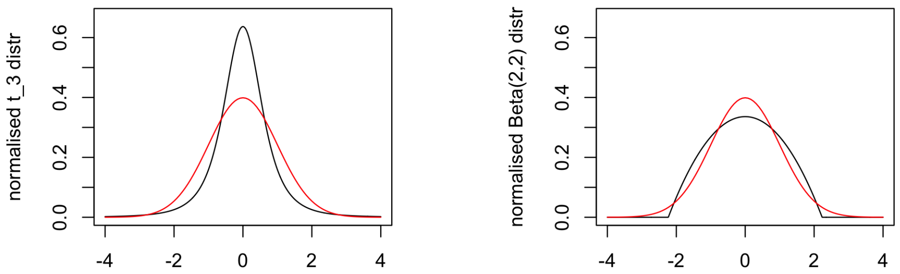
where Φ^{-1} is the quantile function for the $\mathrm{N}(0,1)$ distribution function Φ. The normal QQ-plot is the scatter plot for the sample quantiles matched with the theoretical quantiles

$$(x_{(1)}, y_1), \ldots, (x_{(n)}, y_n).$$

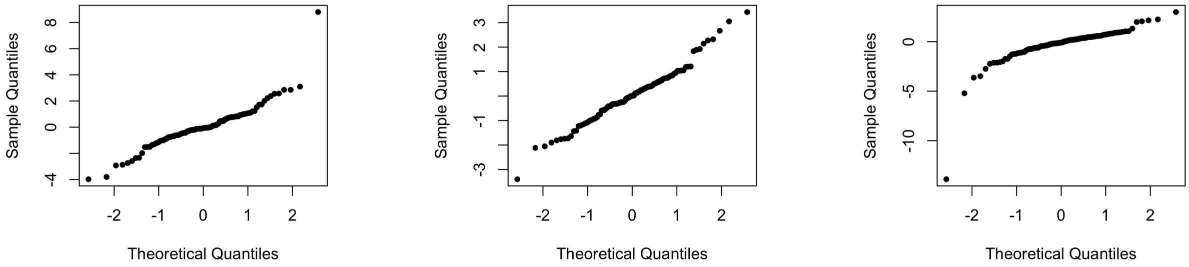
If the normal QQ-plot is close to a straight line $y = a + bx$, then we take it as an evidence supporting the hypothesis of normality, and we may use the point estimates $\hat{\mu} = -\frac{a}{b}$, $\hat{\sigma} = \frac{1}{b}$ for the population mean and standard deviation.

Example: $\mathbf{t}_3$ and $\mathbf{Beta}(2,2)$

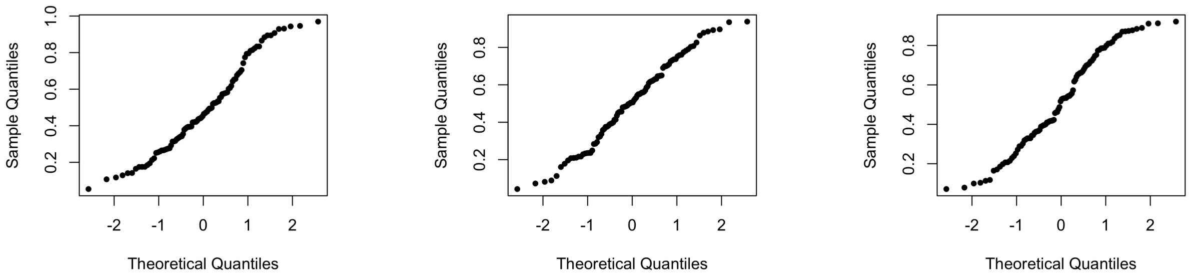
Consider two distributions t_3 and $\mathrm{Beta}(2,2)$, both being symmetric around 0 and 0.5 respectively. On the figure below each of them is plotted after the linear transformation $\frac{X-\mu}{\sigma}$ together with the $\mathrm{N}(0,1)$ curve. This figure is meant to demonstrate the differences in the tails of these two distributions: the t-distribution has heavy tails and the beta distribution has light tails.



The panels below show the normal QQ-plots for three samples of size $n = 100$ drawn from the t-distribution. Observe the deviations from the straight lines characteristic for the heavy tails: the sample values in the tails are more extreme than those projected by the straight lines going through the middle part of the QQ-plot.



The next panels show the normal QQ-plots for three samples of size $n = 100$ drawn from the beta-distribution. The deviations from the straight lines are typical for the light tails: the sample values in the tails are less extreme than those projected by the straight lines going through the middle part of the QQ-plot.



6.3 Density estimation

Estimating the probability density function $f(x) = F'(x)$ for the population distribution is trickier than estimating $F(x)$. For example, taking the derivative of the empirical distribution function $\hat{F}(x)$ doesn't work. One can use a solution based on the histogram displaying the observed counts

$$c_j = \sum_{i=1}^{n} 1_{\{x_i \in \text{cell}_j\}}$$

over the adjacent cells of a width h. By scaling the observed counts

$$f_h(x) = \tfrac{c_j}{nh}, \quad \text{for } x \in \text{cell}_j,$$

we arrive at the scaled histogram giving us a density estimate since

$$\int f_h(x)dx = \tfrac{h}{nh} \sum_j c_j = 1.$$

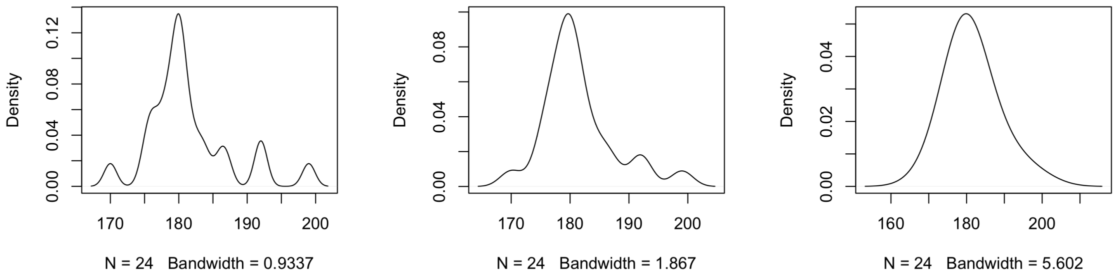
The choice of a balanced width h is important: smaller h give ragged profiles, larger h give obscured profiles.

To produce a smooth version of the scaled histogram, one can use the *kernel density estimate* with bandwidth h defined by

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} \phi(\tfrac{x-x_i}{h}), \text{ where } \phi(x) = \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Example: male heights

Turning to the example of Section 3.3 with 24 male heights, we get the following plots for three kernel density estimates with different bandwidths h. With smaller h, the curve is oversensitive to the individual sample values. With larger h, the curve is almost fully determined by its sample mean and standard deviation.



6.4 Skewness and kurtosis

Another simple way of testing normality relies on the third and forth central moments of the data distribution: coefficient of skewness and kurtosis. Let $X \sim \mathcal{F}(\mu, \sigma)$. In terms of its normalized version $Z = \frac{X-\mu}{\sigma}$, having

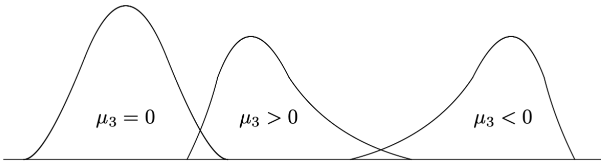
$$\mu_1 = \mathrm{E}(Z) = 0, \quad \mu_2 = \mathrm{E}(Z^2) = 1,$$

the population coefficient of skewness and population kurtosis are defined by

$$\mu_3 = \mathrm{E}(Z^3), \qquad \mu_4 = \mathrm{E}(Z^4).$$

Importantly, for the normal distribution, $\mu_3 = 0$ and $\mu_4 = 3$.

Depending on the sign of the coefficient of skewness, we distinguish between symmetric $\mu_3 = 0$, skewed to the right $\mu_3 > 0$, and skewed to the left $\mu_3 < 0$ distributions.



With μ_3 close to zero, kurtosis being close to 3, can be used as an indication of the curve profile to be close to that of the normal distribution. Otherwise, we distinguish between the leptokurtic distributions with $\mu_4 > 3$ (heavy tails), and platykurtic distributions with $\mu_4 < 3$ (light tails). Given a random sample $(x_1, \ldots, x_n)$ with the sample mean $\bar{x}$ and sample variance s^2, the sample skewness and sample kurtosis are computed by

$$m_3 = \frac{1}{s^3 n}\sum_{i=1}^{n}(x_i - \bar{x})^3, \qquad m_4 = \frac{1}{s^4 n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

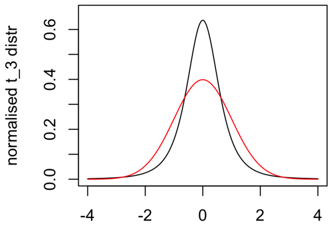
The next table gives a list of parametric distributions with their skewness and kurtosis coefficients.

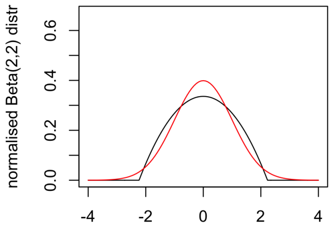
Parametric distribution	**Skewness**	**Kurtosis**
$\mathrm{N}(\mu, \sigma)$ normal distribution	$\mu_3 = 0$ symmetric	$\mu_4 = 3$
$\mathrm{Gam}(\alpha, \lambda)$ gamma distribution	$\mu_3 = \frac{2}{\sqrt{\alpha}}$ skewed to the right	$\mu_4 = 3 + \frac{6}{\alpha}$ leptokurtic
$\mathrm{Beta}(a, a)$ beta distribution	$\mu_3 = 0$ symmetric	$\mu_4 = 3 - \frac{1}{2a+3}$ platykurtic
t-distribution with df $= k$ and $k \geq 5$	$\mu_3 = 0$ symmetric	$\mu_4 = 3 + \frac{6}{k-4}$ leptokurtic

For example, we see that the gamma distribution $\mathrm{Gam}(\alpha, \lambda)$ is positively skewed having $\mu_3 = \frac{2}{\sqrt{\alpha}}$. As the shape parameter α gets larger, the skewness gets smaller and kurtosis becomes close to 3.

Example: t_3 and Beta$(2, 2)$

Returning to the example of the t_3 and Beta$(2, 2)$ distributions, we find that the heavy-tailed t-distribution has infinite kurtosis $\mu_4 = \infty$, while the light-tailed beta-distribution has kurtosis $\mu_4 = 2.143$ which is smaller than 3. Recalling the two standardised curves,





observe that the heavy-tailed curve is indeed leptokurtic (from Greek lepto 'narrow' + kurtos 'bulging') and the light-tailed curve is platykurtic (from platy 'broad, flat').

Example: male heights

For the random sample of $n = 24$ male heights given in ascending order

$$170, 175, 176, 176, 177, 178, 178, 179, 179, 180, 180, 180, 180, 180, 181, 181, 182, 183, 184, 186, 187, 192, 192, 199,$$

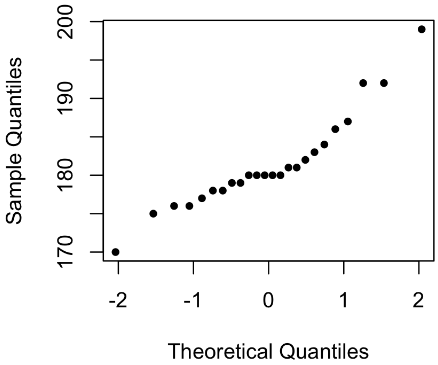
we compute the following summary statistics:

$$\bar{x} = 181.46, \quad \hat{m} = 180, \quad m_3 = 1.05, \quad m_4 = 4.31,$$

saying that the distribution is skewed to the right.

Good to know: the distribution of the heights of adult males is positively skewed, implying $m < \mu$, so that more than half of heights are below the average.

The normal QQ-plot of the data reveals a deviation from normality.



6.5 Inference on the population median

The central point of a distribution can be defined in terms of various measures of location, for example, as the population mean μ or the median m. The population mean μ is estimated by the sample mean $\bar{x}$, and the population median m is estimated by the sample median $\hat{m}$. Given the ordered sample $(x_{(1)}, \ldots, x_{(n)})$, the sample median is defined as

$$\hat{m} = x_{(k)} \quad \text{or} \quad \hat{m} = \frac{x_{(k)} + x_{(k+1)}}{2},$$

depending on whether $n = 2k - 1$ is an odd number or an even number $n = 2k$. The sample mean $\bar{x}$ is sensitive to outliers, while the sample median $\hat{m}$ is not. Therefore, we say that $\hat{m}$ is an estimator robust to outliers. This explains why in many cases the data (like personal income in the United States, see Wikipedia) is summarised by the sample median rather than by the sample mean.

Confidence interval for the median

Consider a random sample $(x_1, \ldots, x_n)$ without assuming any parametric model for the unknown population distribution $\mathcal{F}(\mu, \sigma)$. Suppose that the population distribution $\mathcal{F}(\mu, \sigma)$ is continuous so that all sample values $(x_1, \ldots, x_n)$ are different (no ties). Observe that the number of observations below the true population median m,

$$y = \sum_{i=1}^{n} 1_{\{x_i \le m\}}$$

is a realisation of a random variable Y having the symmetric binomial distribution $Y \sim \text{Bin}(n, 0.5)$. Denote

$$p_k = \text{P}(Y < k) = \sum_{i=0}^{k-1} \binom{n}{i} 2^{-n}, \quad k = 1, \ldots, n.$$

Since

$$\{Y \ge k\} = \{X_{(k)} < m\}, \qquad \{Y \le n-k\} = \{X_{(n-k+1)} > m\},$$

we obtain for $k < n/2$,

$$\text{P}(X_{(k)} < m < X_{(n-k+1)}) = \text{P}(k \le Y \le n-k) = 1 - 2p_k.$$

This yields the following non-parametric formula for an exact confidence interval for the median.

$$\boxed{I_m = (x_{(k)}, x_{(n-k+1)}) \text{ is a } 100 \cdot (1 - 2p_k)\% \text{ confidence interval for the population median } m}$$

For example, if $n = 25$, then from the table below we find that $(X_{(8)}, X_{(18)})$ gives a 95.7% confidence interval for the median.

k	6	7	8	9	10	11	12
$100 \cdot (1 - 2p_k)$	99.6	98.6	95.7	89.2	77.0	57.6	31.0

Sign test

The sign test is a non-parametric test of H_0: $m = m_0$ against the two-sided alternative H_1: $m \ne m_0$. The sign test statistic

$$y_0 = \sum_{i=1}^{n} 1_{\{x_i \le m_0\}}$$

counts the number of observations below the null hypothesis value. This test rejects the null hypothesis for the larger or smaller observed values y_0 with the reference to the null distribution $Y_0 \overset{H_0}{\sim} \text{Bin}(n, 0.5)$. There is a simple connection between the sign test and the confidence interval for the median: we reject H_0 if m_0 falls outside the confidence interval

$$I_m = (x_{(k)}, x_{(n-k+1)}),$$

where k is computed from the given significance level of the test.

6.6 Measures of dispersion

So far we used the sample variance s^2 as the measure of dispersion in the data. A more straightforward measure of dispersion is the sample range $x_{(n)} - x_{(1)}$. Both the sample variance and the sample range are sensitive to outliers. Consider two robust measures of dispersion:

the $\underline{\text{i}}$nter$\underline{\text{q}}$uartile $\underline{\text{r}}$ange, $\text{IQR} = x_{0.75} - x_{0.25}$, is the difference between the upper and lower quartiles,

the $\underline{\text{m}}$edian of $\underline{\text{a}}$bsolute values of $\underline{\text{d}}$eviations, MAD, is defined as the sample median of $\{|x_i - \hat{m}|, i = 1, \ldots, n\}$.

Turning to the standard normal distribution, we get

$$\Phi(0.675) = 0.75, \quad \Phi^{-1}(0.75) = 0.675,$$

so that for the general normal distribution $\text{N}(\mu, \sigma)$, the theoretical lower and upper quartiles are $\mu \pm 0.675 \cdot \sigma$ yielding

$$\text{IQR} = (\mu + 0.675 \cdot \sigma) - (\mu - 0.675 \cdot \sigma) = 1.35 \cdot \sigma.$$

On the other hand, since

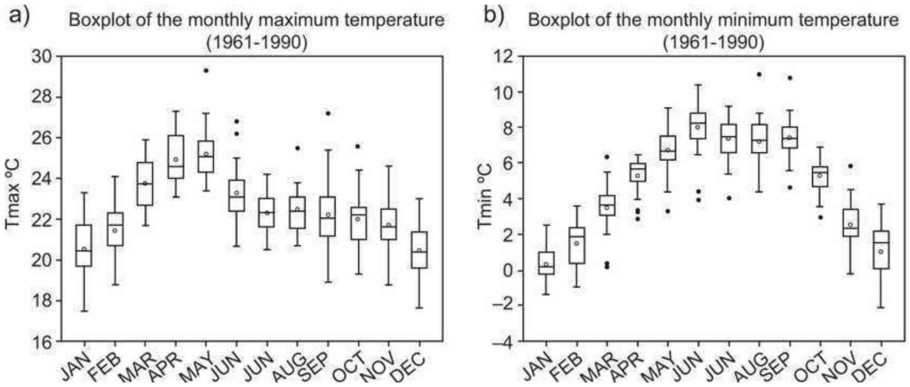
$$\text{P}(|X - \mu| \le 0.675 \cdot \sigma) = (\Phi(0.675) - 0.5) \cdot 2 = 0.5,$$

we obtain $\text{MAD} = 0.675 \cdot \sigma$. To summarise, for the normal distribution $\mathcal{F}(\mu, \sigma) = \text{N}(\mu, \sigma)$ model, we have three estimates of σ:

$$s, \quad \frac{\text{IQR}}{1.35}, \quad \frac{\text{MAD}}{0.675}.$$

Boxplots

The boxplot is a convenient graphical tool for comparing the distributions of different samples in terms of the measures of location and dispersion.



The boxplot is built of the following components: box, whiskers and outliers.

Box

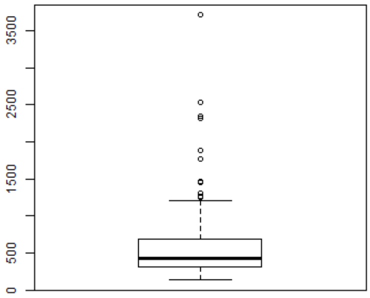
- upper edge of the box = upper quartile (UQ)
- box center = median
- lower edge of the box = lower quartile (LQ)
- box height = interquartile range (IQR)

Wiskers

- upper whisker end = {largest data point ≤ UQ + 1.5 × IQR}
- lower whisker end = {smallest data point ≥ LQ – 1.5 × IQR}

Outliers

- upper dots = {data points > UQ + 1.5 × IQR}
- lower dots = {data points < LQ – 1.5 × IQR}



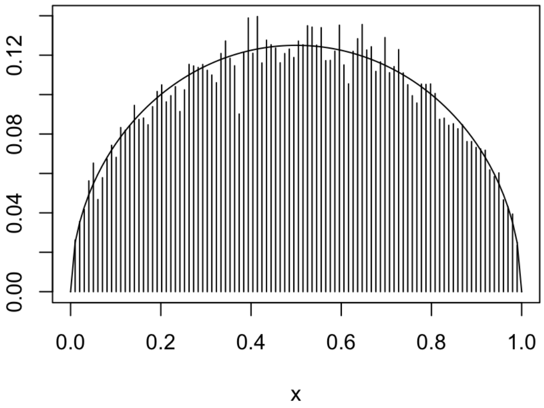
6.7 Exercises

Problem 1

Suppose that $(X_1,\dots,X_n)$ are independent uniform U$(0,1)$ random variables. The corresponding empirical distribution function

$$\hat{F}(x)=\frac{1}{n}\sum_{i=1}^{n}1_{\{X_i\le x\}}$$

is a random variable with the mean x. One hundred such samples of size $n=16$ were generated, and from these samples the standard errors of $\hat{F}(x)$ were computed for different values of x. In the following figure the estimated standard errors of $\hat{F}(x)$ are depicted as the vertical bars.



The elliptic line on the graph giving the theoretical prediction of the observed profile of the vertical bars follows the formula

$$(x-0.5)^2+16y^2=0.25.$$

Explain this formula by computing a certain variance.

Problem 2

Using the setting of the previous problem, show for $x<y$, that

$$\mathrm{Cov}(\hat{F}(x),\hat{F}(y))=\frac{x(1-y)}{n}.$$

Observe that this positive correlation is quite intuitive: if the empirical distribution function $\hat{F}(x)$ overshoots its expected value x, then we would expect that the other value $\hat{F}(y)$ would also overshoot the respective value y.

Problem 3

The data below are the percentages of hydrocarbons in $n=59$ samples of beeswax.

14.27 14.80 12.28 17.09 15.10 12.92 15.56 15.38 15.15 13.98
14.90 15.91 14.52 15.63 13.83 13.66 13.98 14.47 14.65 14.73
15.18 14.49 14.56 15.03 15.40 14.68 13.33 14.41 14.19 15.21
14.75 14.41 14.04 13.68 15.31 14.32 13.64 14.77 14.30 14.62
14.10 15.47 13.73 13.65 15.02 14.01 14.92 15.47 13.75 14.87
15.28 14.43 13.96 14.57 15.49 15.13 14.23 14.44 14.57

(a) For this data, plot the empirical distribution function, a histogram, and the normal QQ-plot. Find the 0.9, 0.75, 0.5, 0.25, and 0.1 quantiles. Does the distribution appear Gaussian?

(b) The average percentage of hydrocarbons in a synthetic wax is 85%. Suppose that beeswax was diluted with 1% synthetic wax. Could this be detected? What about 3% and 5% dilution?

Problem 4

Calculate the hazard function for the Weibull distribution having the distribution function

$$F(x)=1-e^{-\alpha x^{\beta}},\quad x\geq 0,$$

where α and β are two positive parameters. (Waloddi Weibull was a Swedish engineer, scientist, and mathematician.)

Problem 5

Give an example of a distribution with an increasing failure rate (hazard function). Give an example of a distribution with a decreasing failure rate.

Problem 6

Olson, Simpson, and Eden (1975) discuss the analysis of data obtained from a cloud seeding experiment. The following data present the rainfall from 26 seeded and 26 control clouds.

Seeded clouds
129.6, 31.4, 2745.6, 489.1, 430, 302.8, 119, 4.1, 92.4, 17.5,
200.7, 274.7, 274.7, 7.7, 1656, 978, 198.6, 703.4, 1697.8, 334.1,
118.3, 255, 115.3, 242.5, 32.7, 40.6

Control clouds
26.1, 26.3, 87, 95, 372.4, 0.01, 17.3, 24.4, 11.5, 321.2,
68.5, 81.5, 47.3, 28.6, 830.1, 345.5, 1202.6, 36.6, 4.9, 4.9,
41.1, 29, 163, 244.3, 147.8, 21.7

Make a QQ-plot for seeded rainfall versus control rainfall and log seeded rainfall versus log control rainfall. What do these plots suggest about the effect, if any, of seeding?

Problem 7

The Laplace distribution with a positive parameter λ is a two-sided exponential distribution. Its density function is $f(x) = \frac{\lambda}{2}e^{-\lambda|x|}$ for $x \in (-\infty, \infty)$. The variance of this distribution is $2\lambda^{-2}$ and kurtosis is 6.

(a) Take $\lambda = \sqrt{2}$. Plot carefully the density $f(x)$ together with the standard normal distribution density.

(b) Use the drawn picture to explain the exact meaning of the following citation. "Kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable, although some sources are insistent that heavy tails, and not peakedness, is what is really being measured by kurtosis".

Problem 8

Given a random sample $(x_1, \ldots, x_n)$ with $n = 25$, we are interested in testing $H_0 : m = 20$ against the two-sided alternative $H_1 : m \neq 20$ concerning the population median m. No parametric model is assumed. As a test statistic we take $y = \sum_{i=1}^{n} 1_{\{x_i \leq 20\}}$, the number of observations below the null hypothesis value.

(a) Find the exact null distribution of the test statistic. What are your assumptions?

(b) Suggest an approximate confidence interval formula for m.

Problem 9

Miscellaneous questions.

(a) Describe a situation when a stratified sampling is more effective than a simple random sampling for estimating the population mean. Which characteristics of the strata will influence your sample allocation choice?

(b) Given a dataset how do you compute kurtosis? What is the purpose of this summary statistic? Why is it important to compute the coefficient of skewness for a proper interpretation of the kurtosis value?

(c) Suppose we are interested in the average height for a population of size 2,000,000. To what extend can a sample of 200 individuals be representative for the whole population?

Problem 10

Let $x_1, \ldots, x_{25}$ be a random sample drawn from $\mathrm{N}(0.3, 1)$. Consider testing

$$H_0 : m = 0 \quad \text{vs} \quad H_1 : m > 0$$

at $\alpha = 0.05$. Compare the power of two tests:

(a) the power of the sign test,

(b) the power of the test based on the normal theory assuming that $\sigma = 1$ is known.

Chapter 7

Comparing two samples

Suppose we wish to compare two population distributions $\mathcal{F}(\mu_1, \sigma_1)$ and $\mathcal{F}(\mu_2, \sigma_2)$ based on two random samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ from these two populations. Consider two sample means

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \ldots + y_m}{m},$$

two sample variances

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2,$$

and the standard errors of $\bar{x}$ and $\bar{y}$

$$s_{\bar{x}} = \frac{s_1}{\sqrt{n}}, \quad s_{\bar{y}} = \frac{s_2}{\sqrt{m}}.$$

The difference $(\bar{x} - \bar{y})$ is an unbiased estimate of $\mu_1 - \mu_2$. We are interested in finding the standard error of $\bar{x} - \bar{y}$, a confidence interval for the difference $\mu_1 - \mu_2$, as well as testing the null hypothesis of equality

$$H_0 : \mu_1 = \mu_2.$$

Two main settings will be addressed: (1) two independent samples and (2) paired samples for sampling the differences.

7.1 Two independent samples

In the case of two independent samples, due to independence between $\bar{X}$ and $\bar{Y}$, we have

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

implying that

$$s_{\bar{x}-\bar{y}}^2 = s_{\bar{x}}^2 + s_{\bar{y}}^2 = \frac{s_1^2}{n} + \frac{s_2^2}{m}$$

gives an unbiased estimate of $\mathrm{Var}(\bar{X} - \bar{Y})$. Taking the square root we obtain the following formula for the standard error of the point estimate $\bar{x} - \bar{y}$

$$s_{\bar{x}-\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2} = \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}.$$

Large sample test for the difference between two means

If n and m are large, we can use the normal approximation

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_{\bar{X}}^2 + S_{\bar{Y}}^2}} \approx \mathrm{N}(0, 1).$$

The hypothesis of equality

$$H_0 : \mu_1 = \mu_2$$

is tested using the test statistic

$$z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}}$$

whose null distribution is approximated by the standard normal $\mathrm{N}(0, 1)$.

$$\boxed{\text{Approximate } 100(1-\alpha)\% \text{ confidence interval } I_{\mu_1-\mu_2} \approx \bar{x} - \bar{y} \pm z(\tfrac{\alpha}{2}) \cdot \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}.}$$

Two-sample t-test

The two–sample t-test is based on the following three assumptions on the population distributions:

1. two random samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ are independent of each other,
2. both underlying population distributions are normal, $\mathcal{F}(\mu_1, \sigma_1) = \mathrm{N}(\mu_1, \sigma_1)$ and $\mathcal{F}(\mu_2, \sigma_2) = \mathrm{N}(\mu_2, \sigma_2)$,
3. these two normal distributions have equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

To estimate the common variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we can use

$$s_{\mathrm{p}}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2}{n + m - 2},$$

which is called the *pooled sample variance*. Using the equality

$$s_{\mathrm{p}}^2 = \frac{n-1}{n+m-2} \cdot s_1^2 + \frac{m-1}{n+m-2} \cdot s_2^2,$$

it is easy to see that the pooled sample variance is an unbiased estimate of σ^2:

$$\mathrm{E}(S_{\mathrm{p}}^2) = \frac{n-1}{n+m-2}\mathrm{E}(S_1^2) + \frac{m-1}{n+m-2}\mathrm{E}(S_2^2) = \sigma^2.$$

In the case of equal variances and independent samples, the variance

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma^2 \cdot \frac{n+m}{nm},$$

has the following unbiased estimate

$$s_{\bar{x}-\bar{y}}^2 = s_p^2 \cdot \frac{n+m}{nm}.$$

Under the normality assumption with equal variances, the sampling distribution of the two-sample t-score

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}}$$

is the t-distribution with $(n + m - 2)$ degrees of freedom.

$$\boxed{\text{Exact confidence interval } I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm t_{n+m-2}(\tfrac{\alpha}{2}) \cdot s_{\mathrm{p}} \cdot \sqrt{\tfrac{n+m}{nm}}}$$

The two sample t-test for testing H_0: $\mu_1 = \mu_2$ is based on the test statistic

$$t_0 = \frac{\bar{x} - \bar{y}}{s_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}}$$

obtained from the two-sample t-score t by putting $\mu_1 - \mu_2 = 0$. The null distribution of the test statistic is $T_0 \overset{H_0}{\sim} t_{n+m-2}$.

Case study: iron retention

Fe^{3+} (10.2)	Fe^{3+} (1.2)	Fe^{3+} (0.3)	Fe^{2+} (10.2)	Fe^{2+} (1.2)	Fe^{2+} (0.3)
0.71	2.20	2.25	2.20	4.04	2.71
1.66	2.93	3.93	2.69	4.16	5.43
2.01	3.08	5.08	3.54	4.42	6.38
2.16	3.49	5.82	3.75	4.93	6.38
2.42	4.11	5.84	3.83	5.49	8.32
2.42	4.95	6.89	4.08	5.77	9.04
2.56	5.16	8.50	4.27	5.86	9.56
2.60	5.54	8.56	4.53	6.28	10.01
3.31	5.68	9.44	5.32	6.97	10.08
3.64	6.25	10.52	6.18	7.06	10.62
3.74	7.25	13.46	6.22	7.78	13.80
3.74	7.90	13.57	6.33	9.23	15.99
4.39	8.85	14.76	6.97	9.34	17.90
4.50	11.96	16.41	6.97	9.91	18.25
5.07	15.54	16.96	7.52	13.46	19.32
5.26	15.89	17.56	8.36	18.40	19.87
8.15	18.3	22.82	11.65	23.89	21.60
8.24	18.59	29.13	12.45	26.39	22.25

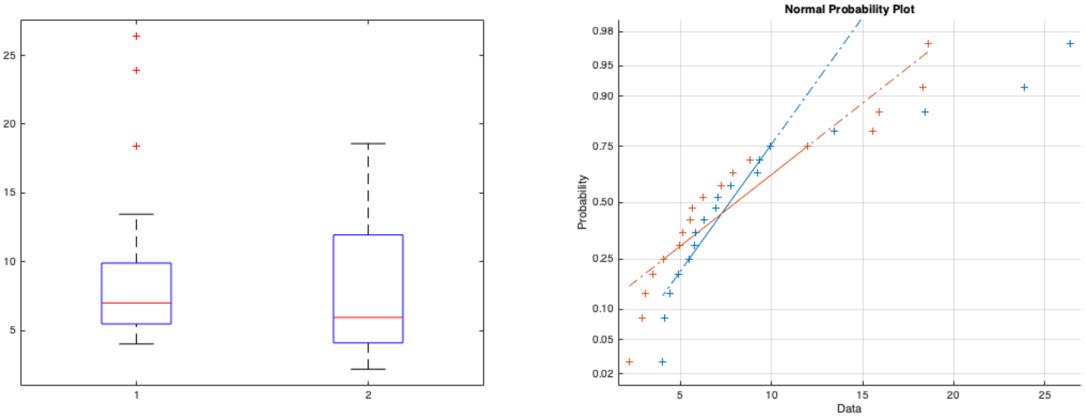
Two forms of iron Fe^{2+} and Fe^{3+} were compared in a randomised study on mice. The obtained data is the percentage of iron retained in mice under three different iron dosage concentrations: 10.2, 1.2, and 0.3 millimolar. The six samples were collected using $n = 18$ mice randomly assigned at each combination of the iron form and concentration. Turning to the two independent samples at the medium concentration, we get the following summary of the data:

Fe^{2+} percentage retained: $n = 18$, $\bar{x} = 9.63$, $s_1 = 6.69$, $s_{\bar{x}} = 1.58$
Fe^{3+} percentage retained: $m = 18$, $\bar{y} = 8.20$, $s_2 = 5.45$, $s_{\bar{y}} = 1.28$

The graphs below show that the population distributions are not normal. Therefore, to test H_0: $\mu_1 = \mu_2$ we turn to the large sample test. Using the observed value

$$z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}} = 0.7,$$

and applying the normal distribution table we find an approximate two-sided p-value = 0.48.



Left panel: boxplots for percentages of Fe^{2+} (left) and Fe^{3+} (right). Right panel: two normal QQ plots.

After the log transformation of the data

$$x_i^\circ = \ln x_i, \quad y_i^\circ = \ln y_i,$$

the normality assumption becomes more acceptable, as seen from the graphs below. For the transformed data, we get

Fe^{2+} log-percentage retained: $n = 18$, $\bar{x}^\circ = 2.09$, $s_1^\circ = 0.659$, $s_{\bar{x}^\circ} = 0.155$,
Fe^{3+} log-percentage retained: $m = 18$, $\bar{y}^\circ = 1.90$, $s_2^\circ = 0.574$, $s_{\bar{y}^\circ} = 0.135$.

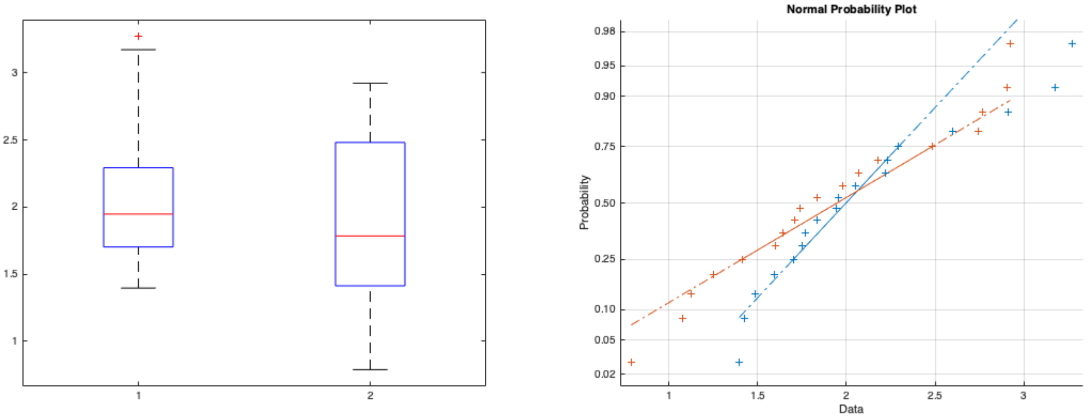
Assuming that two population variances are equal to the unknown σ^2, we estimate σ^2 by the pooled sample variance formula

$$s_{\text{p}}^2 = \frac{n-1}{n+m-2} \cdot (s_1^\circ)^2 + \frac{m-1}{n+m-2} \cdot (s_2^\circ)^2 = 0.5(0.659^2 + 0.574^2) = 0.382.$$

The two-sample t-score for the transformed data is

$$t_0 = \frac{\bar{x}^\circ - \bar{y}^\circ}{s_{\text{p}}\sqrt{2/n}} = \frac{2.09 - 1.90}{0.618/3} = 0.922.$$

Applying the two-sample t-test with df = 34, we find the two-sided p-value = 0.366 also resulting in non-significant difference. In the next chapter we will analyse the full dataset of six samples using the two-way layout model of the analysis of variance.



Left panel: boxplots for log-percentages of Fe^{2+} (left) and Fe^{3+} (right). Right panel: two normal QQ plots.

Rank sum test

The rank sum test is a nonparametric test for two independent samples, which does not assume normality of population distributions. Suppose we have two independent samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ drawn from two continuous population distributions F_1 and F_2, and we would like to test

$$H_0\colon\ F_1 = F_2 \text{ against } H_1\colon\ F_1 \neq F_2$$

in the case when either one of the distributions F_1 and F_2 is not normal, or their variances are not equal. The rank sum test procedure consists of the following steps:

> pool the samples and replace the data values by their ranks $1, 2, \ldots, n+m$, starting from the smallest sample value to the largest, and then compute two test statistics r_1 = sum of the ranks of x-observations, and r_2 = sum of y-ranks.

Clearly, these two test statistics have a fixed sum

$$r_1 + r_2 = 1 + 2 + \ldots + (n+m) = \tfrac{(n+m)(n+m+1)}{2}.$$

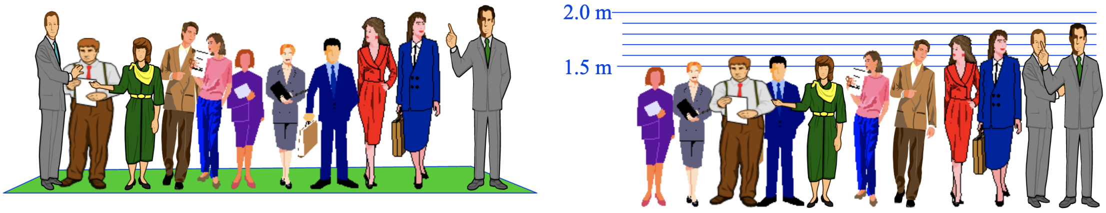
The exact null distributions for R_1 and R_2 depend only on the sample sizes n and m, and the rank sum test is included in all statistical packages. For $n \geq 10$, $m \geq 10$, we may apply the normal approximation for the null distributions of R_1 and R_2 with

$$\mathrm{E}(R_1) = \frac{n(n+m+1)}{2},\ \mathrm{E}(R_2) = \frac{m(n+m+1)}{2},\ \mathrm{Var}(R_1) = \mathrm{Var}(R_2) = \frac{mn(n+m+1)}{12}.$$

Note that the variances of R_1 and R_2 must be equal since their sum is a constant.

Example: comparing heights

The picture below illustrates a sampling experiment for testing the equality $H_0 : F_1 = F_2$ of the height distributions for females F_1 and males F_2, against $H_1\colon\ F_1 \neq F_2$. According to the left panel we can measure the heights of $n = 6$ females and $m = 5$ males.



From the right panel we find the two rank sums

$$r_1 = 1 + 2 + 5 + 6 + 8 + 9 = 31, \quad r_2 = 3 + 4 + 7 + 10 + 11 = 35.$$

Despite the sample sizes are small, we will use the normal approximation for the null distribution of R_2, which has the mean and variance

$$\mu = \frac{m(n+m+1)}{2} = 30, \quad \sigma^2 = \frac{mn(n+m+1)}{12} = 30.$$

The resulting one-sided p-value is larger than 10%,

$$1 - \Phi(\tfrac{35.5-30}{\sqrt{30}}) = 1 - \Phi(1.00) = 0.16,$$

implying that we can not reject the null hypothesis of equal means judging from these small samples.

7.2 Two independent samples: comparing population proportions

Consider the special case of two independent samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ drawn from two Bernoulli distributions $X \sim \mathrm{Bin}(1, p_1)$ and $Y \sim \mathrm{Bin}(1, p_2)$. The null hypothesis of equality takes the form

$$H_0 : p_1 = p_2,$$

the two-sample t-test is not valid, and we would like to modify the large sample test for the mean difference. For the Bernoulli model, the sample means turn into the sample proportions

$$\bar{x} = \hat{p}_1, \qquad \bar{y} = \hat{p}_2,$$

giving unbiased estimates of p_1, p_2 with the standard errors

$$s_{\hat{p}_1} = \sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n-1}}, \quad s_{\hat{p}_2} = \sqrt{\tfrac{\hat{p}_2(1-\hat{p}_2)}{m-1}}.$$

Large sample test for two proportions

If the samples sizes m and n are large, then an approximate confidence interval for the difference $p_1 - p_2$ is given by

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z(\tfrac{\alpha}{2}) \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{m-1}}.$$

If this interval does not cover 0, we reject the null hypothesis of equality $H_0 : p_1 = p_2$ in favour of the two-sided alternative $H_1 : p_1 \neq p_2$ at the significance level α.

Example: opinion polls

Consider two consecutive monthly poll results $\hat{p}_1$ and $\hat{p}_2$ with $n \approx m \approx 5000$ interviews. A change in support to a major political party from $\hat{p}_1$ to $\hat{p}_2$, with both numbers being close to 40%, is deemed significant if the difference between these two sample proportions satisfies

$$|\hat{p}_1 - \hat{p}_2| > 1.96 \cdot \sqrt{2 \cdot \tfrac{0.4 \cdot 0.6}{5000}} = 0.019.$$

This result may be compared with the one-sample hypothesis testing

$$H_0 : p = 0.4 \text{ vs } H_0 : p \neq 0.4.$$

The approximate 95% confidence interval for p is

$$I_p \approx \hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}},$$

and if $\hat{p}$ is not far from 0.4, then the result of the opinion poll is significant if

$$|\hat{p} - 0.4| > 1.96 \cdot \sqrt{\tfrac{0.4 \cdot 0.6}{5000}} = 0.013.$$

Fisher's exact test

> *A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. [...] [It] consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of that the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind [...]. — Fisher, 1935.*

Fisher's exact test deals with the null hypothesis

$$H_0 : p_1 = p_2,$$

when the sample sizes m and n are not sufficiently large for applying normal approximations for the distributions of the sample means. We summarise the data of two independent samples as a 2×2 table of sample counts

	Sample 1	Sample 2	Total
Number of successes	c_{11}	c_{12}	$c_{11} + c_{12}$
Number of failures	c_{01}	c_{02}	$c_{01} + c_{02}$
Sample sizes	n	m	$N = n + m$

where

$$c_{11} = x_1 + \ldots + x_n, \quad c_{01} = n - c_{11}, \quad c_{12} = y_1 + \ldots + y_m, \quad c_{02} = m - c_{12}.$$

The key of the Fisher's idea is to treat the observed count c_{11} as a random outcome of drawing without replacement of n balls from an urn with N balls. Assuming that the urn contains $c_{11} + c_{12}$ black balls and $c_{01} + c_{02}$ white balls, the c_{11} is the number of black balls among n sampled balls. Putting $N = n + m$, and freezing the proportion

$$p = \frac{c_{11} + c_{12}}{N},$$

we find that under the null hypothesis of equality, the sample count c_{11} is generated by the hypergeometric distribution

$$C_{11} \overset{H_0}{\sim} \text{Hg}(N, n, p).$$

This null distribution should be used for determining the rejection rule of the exact Fisher test in terms of the sample counts (c_{11}, c_{12}) and the sample sizes (n, m).

Example: gender bias

The following data were collected after 48 copies of the same file with 24 files labeled as “male” and the other 24 labeled as “female” were sent to 48 experts. Each expert decision had two possible outcomes: promote or hold file. We wish to test

$$H_0\colon\ p_1 = p_2 \text{ no gender bias,}$$

against

$$H_1\colon\ p_1 > p_2 \text{ males are favoured.}$$

Given the data

	Male	Female	Total
Promote	21	14	35
Hold file	3	10	13
Total	24	24	48

Fisher’s test would reject H_0 in favour of the one-sided alternative H_1 if the observed value $c_{11} = 21$ will be deemed significantly large under the hypergeometric distribution $\text{Hg}(N, n, p)$ with $N = 48$, $n = 24$, and $p = 35/48$:

$$\text{P}(C_{11} = k) = \frac{\binom{35}{k}\binom{13}{24-k}}{\binom{48}{24}} = \frac{\binom{35}{35-k}\binom{13}{k-11}}{\binom{48}{24}}, \quad 11 \le k \le 24.$$

This distribution is symmetric around the mean $\mu = 17.5$ with

$$\text{P}(C_{11} \le 14) = \text{P}(C_{11} \ge 21) = 0.0245.$$

We conclude that the one-sided p-value = 0.0245, and a two-sided p-value = 0.049. According to this dataset, there is a significant evidence of sex bias in favor of male candidates.

7.3 Paired samples

Paired difference tests discussed in this section are based on a specific type of *blocking.* Examples of paired observations:

- two different drugs applied to two patients matched by age and sex, with measured effects (x, y),
- a fruit’s weight x before shipment and y after shipment,
- two types of tires tested on the same car, with measured effects (x, y).

A paired samples is a vector of independent and identically distributed two-dimensional random variables

$$(x_1, y_1), \ldots, (x_n, y_n).$$

As before, we have two samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ taken from two possibly different population means and variances

$$\text{E}(X) = \mu_1, \quad \text{E}(Y) = \mu_2, \quad \text{Var}(X) = \sigma_1^2, \quad \text{Var}(Y) = \sigma_2^2,$$

and our main question is again whether the difference $\mu_1 - \mu_2$ is statistically significant.

In contrast to the two independent samples case, in the paired sampling setting, there is a positive dependence in the joint distribution of (X, Y):

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_1\sigma_2} > 0.$$

Even with paired samples, the difference $\bar{x} - \bar{y}$ gives an unbiased estimate of the population mean difference $\mu_1 - \mu_2$. This is an unbiased estimate whose variance value takes into account dependence between X and Y. Observe that

$$\begin{aligned}\text{Var}(\bar{X} - \bar{Y}) &= \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2\text{Cov}(\bar{X}, \bar{Y}) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2}{n^2}\text{Cov}(X_1 + \ldots + X_n, Y_1 + \ldots + Y_n).\end{aligned}$$

Since X_i and Y_j are independent for $i \ne j$, we get

$$\text{Cov}(X_1 + \ldots + X_n, Y_1 + \ldots + Y_n) = \text{Cov}(X_1, Y_1) + \ldots + \text{Cov}(X_n, Y_n) = n\text{Cov}(X, Y) = n\sigma_1\sigma_2\rho,$$

implying

$$\text{Var}(\bar{X} - \bar{Y}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho).$$

In particular, if the two samples are independent and have equal sizes, then $\rho = 0$ and

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2).$$

It follows that with $\rho > 0$, the paired sampling ensures a smaller standard error for the estimate $\bar{x} - \bar{y}$ compared to the independent X and Y.

The key idea in handling the paired samples is to reduce the two-sample case to the single random sample case by taking the differences

$$(d_1, \ldots, d_n), \quad d_i = x_i - y_i.$$

For the sample of the differences, we have

$$\bar{d} = \bar{x} - \bar{y},$$

with

$$\mathrm{E}(\bar{D}) = \mu_1 - \mu_2, \quad \mathrm{E}(\bar{D}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho).$$

Smoking and platelet aggregation

To study the effect of cigarette smoking on platelet aggregation, Levine (1973) drew blood samples from 11 individuals before and after they smoked a cigarette and measured the extend to which the blood platelets aggregated. Platelets are involved in the formation of blod clots, and it is known that smokers suffer more often from disorders involving blood clots than do nonsmokers. The obtained data are shown in the following table, which gives the maximum percentage of all the platelets that aggregated after being exposed to a stimulus.

Before smoking y_i	After smoking x_i	$d_i = x_i - y_i$
25	27	2
25	29	4
27	37	10
28	43	15
30	46	16
44	56	12
52	61	9
53	57	4
53	80	27
60	59	–1
67	82	15

The first two columns show the paired measurements (x_i, y_i) for $n = 11$ individuals. Using the data we estimate the correlation coefficient ρ for the measurement results before and after smoking by computing the sample correlation coefficient

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_n - \bar{x})(y_n - \bar{y})}{(n-1)s_1 s_2} = 0.90.$$

The obtained high value indicates a strong positive correlation between the first two columns, and the third column reduces the two sample data to the single random sample of the differences $d_i = x_i - y_i$.

Assuming that the population distribution for differences $D \sim \mathrm{N}(\mu, \sigma)$ is normal with $\mu = \mu_1 - \mu_2$, we may apply the one-sample t-test for

$$H_0\colon \mu = 0 \text{ against } H_1\colon \mu \neq 0.$$

The observed test statistic value

$$t_0 = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10.27}{2.40} = 4.28$$

according to t-distribution with 10 degrees of freedom produces the two-sided p-value 0.0016. We conclude that smoking has a significant health effect.

Without the assumption of normality on the distribution of differences, we can apply the non-parametric sign test for a pair of hypotheses concerning the population median m of the difference $D = X - Y$:

$$H_0\colon m = 0 \text{ against } H_1\colon m \neq 0.$$

We may use

$$y_0 = 1_{\{d_1 < 0\}} + \ldots + 1_{\{d_n < 0\}} = 1$$

as the test statistic which has the null distribution

$$Y_0 \overset{H_0}{\sim} \mathrm{Bin}(n, 0.5).$$

The two-sided p-value of the sign test is

$$2\mathrm{P}(Y_0 \leq 1) = 2[(0.5)^{11} + 11 \cdot (0.5)^{11}] = 0.012.$$

Thus we reject the null hypothesis of no difference also with the help of the non-parametric sign test.

Signed rank test

The sign test disregards a lot of information in the data taking into account only the sign of the differences. The signed rank test pays attention to the sizes of positive and negative differences. This non-parametric test requires the following assumption: the population distribution of $D = X - Y$ is symmetric about its median m. The test statistic for

$$H_0\colon\ m = 0 \text{ against } H_1\colon\ m \neq 0.$$

is computed in terms the ranks of the absolute values of the differences

$$r_i = \text{rank}(|d_i|), \quad i = 1, \ldots, n.$$

Here $r_i = 1$ if d_i has the smallest absolute value and $r_i = n$ if d_i has the largest absolute value among the differences irrespective of the sign of the difference d_i. The signed ranks used for computing the signed rank test statistic w are defined as follows: if $d_i > 0$, then r_i is called a positive rank, and if $d_i < 0$, then r_i is called a negative rank. There are two possible ways to compute the signed rank test statistic w: either as the sum of positive ranks

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i>0\}}$$

or the sum of the negative ranks

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i<0\}}.$$

Irrespective of the sign of the ranks used for computing the test statistic its null distributions is the same. The exact null distribution of W is included in all statistical packages. For $n \geq 20$, one can use the normal approximation $\mathrm{N}(\mu, \sigma)$ of the null distribution of W, with

$$\mu = \frac{n(n+1)}{4}, \qquad \sigma^2 = \frac{n(n+1)(2n+1)}{24}.$$

The signed rank test uses more data information than the sign test but requires symmetric distribution of differences.

Example: platelet aggregation

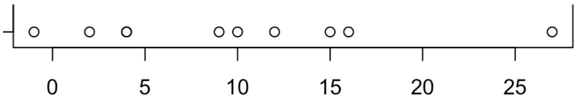
The table below illustrates how the signed ranks are computed for the data on platelet aggregation. Notice how the problem of ties (equal differences) is resolved for assigning the ranks.

Before smoking y_i	After smoking x_i	$d_i = x_i - y_i$	$\lvert d_i\rvert$	Rank of $\lvert d_i\rvert$	Signed rank
25	27	2	2	2	+2
25	29	4	4	3.5	+3.5
27	37	10	10	6	+6
28	43	15	15	8.5	+8.5
30	46	16	16	10	+10
44	56	12	12	7	+7
52	61	9	9	5	+5
53	57	4	4	3.5	+3.5
53	80	27	27	11	+11
60	59	–1	1	1	–1
67	82	15	15	8.5	+8.5

The observed value of the signed rank test statistic is

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i<0\}} = 1.$$

It gives the two-sided p-value = 0.002. Given the small p-value, before we conclude that there is a significant effect, we have to verify the symmetry assumption on the distribution of differences D. Judging from the plot below, the vector of observed differences $(2, 4, 10, 15, 16, 12, 9, 4, 27, -1)$ is symmetric enough around its median.



7.4 Paired samples: comparing population proportions

Suppose we have two dependent Bernoulli random variables $X \sim \text{Bin}(1, p_1)$ and $Y \sim \text{Bin}(1, p_2)$. The vector (X, Y) have four possible values $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ with probabilities denoted as $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$. With n independent paired observations of the vector (X, Y), we get four counts $(c_{00}, c_{01}, c_{10}, c_{11})$ for the different outcomes. The corresponding joint distribution is multinomial

$$(C_{00}, C_{01}, C_{10}, C_{11}) \sim \text{Mn}(n, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}).$$

An unbiased estimate for the difference

$$p_1 - p_2 = \pi_{10} - \pi_{01},$$

in the case of the paired dichotomous observations is given by

$$\hat{p}_1 - \hat{p}_2 = \hat{\pi}_{10} - \hat{\pi}_{01}, \quad \hat{\pi}_{10} = \frac{c_{10}}{n}, \quad \hat{\pi}_{01} = \frac{c_{01}}{n}.$$

The variance of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is estimated by

$$s^2_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1},$$

in view of

$$\text{Var}(C_{10} - C_{01}) = n\pi_{10}(1 - \pi_{10}) + n\pi_{01}(1 - \pi_{01}) + 2n\pi_{10}\pi_{01} = n(\pi_{10} + \pi_{01} - (\pi_{10} - \pi_{01})^2).$$

Referring to the central limit theorem we arrive at the following approximate $100(1-\alpha)\%$ confidence interval formula for the difference $p_1 - p_2$

$$I_{p_1 - p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z(\tfrac{\alpha}{2}) \cdot s_{\hat{p}_1 - \hat{p}_2}.$$

McNemar's test

A significant difference between p_1 and p_2 is established when the confidence interval $I_{p_1 - p_2}$ does not cover zero, that is when

$$|\hat{p}_1 - \hat{p}_2| > z(\tfrac{\alpha}{2}) \cdot s_{\hat{p}_1 - \hat{p}_2},$$

or in other words, the rejection region for

$$H_0 : p_1 = p_2 \text{ against } H_1 : p_1 \neq p_2$$

och equivalently

$$H_0 : \pi_{10} = \pi_{01} \text{ against } H_1 : \pi_{10} \neq \pi_{01}$$

has the form

$$\mathcal{R} = \left\{ \frac{|\hat{\pi}_{10} - \hat{\pi}_{01}|}{\sqrt{\frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1}}} > z(\tfrac{\alpha}{2}) \right\}.$$

Now notice that for the large sample sizes, the squared left hand side approximately equals

$$\frac{n(\hat{\pi}_{10} - \hat{\pi}_{01})^2}{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2} \approx \frac{(c_{10} - c_{01})^2}{c_{10} + c_{01}}.$$

This observation leads to the McNemar test statistic

$$\mathrm{x}^2 = \frac{(c_{10} - c_{01})^2}{c_{10} + c_{01}}.$$

Its null distribution is approximated by the χ^2_1-distribution. Observe that out of the four sample counts $(c_{00}, c_{01}, c_{10}, c_{11})$ the McNemar test statistic depends only on two of them (c_{01}, c_{10}). Thus the number of informative counts is just $c_{01} + c_{10}$ out of total n. We will return to the McNemar test in Section 9.3.

7.5 List of statistical tests

At this point, the reader might find it useful to get a look at the next list mentioning all the frequentist tests, parametric and non-parametric, covered in this textbook (so far and later on).

One-sample tests

- One sample t-test: normal population distribution
- Large sample test for mean

- Large sample test for proportion: categorical data
- Small sample test for proportion: categorical data
- Chi-squared test of goodness of fit: categorical data, large sample
- Chi-squared test of independence: categorical data, large sample
- Model utility test: linear model, several explanatory variables, normal noise, homoscedasticity

Two-sample tests

- Two sample t-test: normal populations, equal variances, independent samples
- Fisher's exact test: categorical data, independent samples
- McNemar: categorical data, matched samples, large samples

Several samples

- ANOVA 1: normal population distributions, equal variances, independent samples
- ANOVA 2: normal population distributions, equal variances, matched samples
- Chi-squared test of homogeneity: categorical data, independent samples, large samples

Non-parametric tests

- Sign test: one sample
- Signed rank test: two matched samples, symmetric distribution of differences
- Rank sum test: two independent samples
- Kruskal-Wallis: several independent samples
- Fridman: several matched samples

7.6 Exercises

Problem 1

Four random numbers were generated from a normal distribution $\mathrm{N}(\mu_1, \sigma^2)$

$$x_1 = 1.1650, \quad x_2 = 0.6268, \quad x_3 = 0.0751, \quad x_4 = 0.3516,$$

along with five random numbers with the same variance σ^2 but perhaps a different mean μ_2

$$y_1 = 0.3035, \quad y_2 = 2.6961, \quad y_3 = 1.0591, \quad y_4 = 2.7971, \quad y_5 = 1.2641.$$

(a) What do you think the means of the random normal number generators were? What do you think the difference of the means was?

(b) Estimate σ^2.

(c) What is the estimated standard error of your estimate of the difference of the means?

(d) Form a 90% confidence interval for the difference of the means.

(e) In this situation, is it more appropriate to use a one-sided test or a two-sided test of the equality of the means?

(f) What is the p-value of a two-sided test of the null hypothesis of equal means?

(g) Would the hypothesis that the means were the same versus a two-sided alternative be rejected at the significance level $\alpha = 0.1$?

(h) Suppose you know that the variance of the normal distribution was $\sigma^2 = 1$. How would your answers to the preceding questions change?

Problem 2

In the "two independent samples" setting we have two ways of estimating the variance of $\bar{X} - \bar{Y}$:

(a) $s_{\mathrm{p}}^2(\frac{1}{n} + \frac{1}{m})$, if $\sigma_1 = \sigma_2 = \sigma$,

(b) $\frac{s_1^2}{n} + \frac{s_2^2}{m}$ without the assumption of equal variances.

Show that if $m = n$, then these two estimates are identical.

Problem 3

An experiment of the efficacy of a drug for reducing high blood pressure is performed using four subjects in the following way:

> two of the subjects are chosen at random for the control group and two for the treatment group.

During the course of a treatment with the drug, the blood pressure of each of the subjects in the teatment group is measured for ten consecutive days as is the blood pressure of each of the subjects in the control group.

(a) In order to test whether the treatment has an effect, do you think it is appropriate to use the two-sample t-test with $n = m = 20$?

(b) Do you think it is appropriate to use the rank sum test?

Problem 4

This is an example of the so-called *Simpson's paradox.*
Hospital A has higher overall death rate than hospital B:

Hospital:	A	B
Died	63	16
Survived	2037	784
Total	2100	800
Death Rate	0.03	0.02

However, if we split the data in two parts, patients in good (+) and bad (−) conditions, for both parts hospital A performs better:

Hospital:	A^+	A^-	B^+	B^-
Died	6	57	8	8
Survived	594	1443	592	192
Total	600	1500	600	200
Death Rate	0.010	0.038	0.013	0.040

Resolve this paradox.

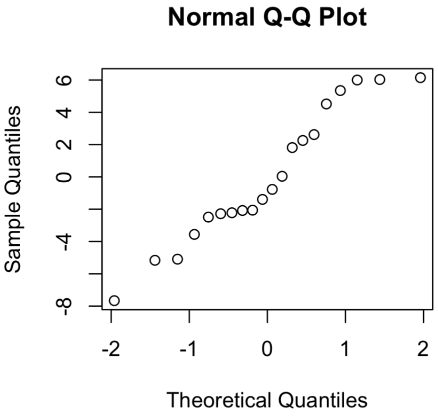
Problem 5

Suppose that n measurements are to be taken under a treatment condition and another n measurements are to be taken independently under a control condition. It is thought that the standard deviation of a single observation is about 10 under both conditions. How large should n be so that a 95% confidence interval for the mean difference has a width of 2? Use the normal distribution rather than the t-distribution, since n will turn out to be quite large.

Problem 6

Two types of engine bearings are compared.
The data in the table give the millions of cycles until failure.

Type I	Type II
3.03	3.19
5.53	4.26
5.60	4.47
9.30	4.53
9.92	4.67
12.51	4.69
12.95	6.79
15.21	9.37
16.04	12.75
16.84	12.78



(a) Use a normal theory test for the null hypothesis of no difference against the two-sided alternative

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 \neq \mu_2.$$

(b) Test the hypothesis of no difference between the two types of bearing using a non-parametric method.

(c) Which of the methods (a) or (b) do you think is better in this case? (Hint: refer to the normal QQ-plot constructed for the deviations of the 20 observations from their sample means.)

(d) Estimate π, the probability that a type I bearing will outlast a type II bearing.

Problem 7

Find the exact null distribution for the test statistic of the signed rank test with $n = 4$.

Problem 8

Turn to the signed rank test and denote by W a random variable having the null distribution of its test statistic. For $n = 10, 20, 25$ and $\alpha = 0.05, 0.01$, the next table gives the critical values w_α such that $2\mathrm{P}(W \le w_\alpha)$ is closest to α.

	$n = 10$	$n = 20$	$n = 25$
$\alpha = 0.05$	8	52	89
$\alpha = 0.01$	3	38	68

Compare these critical values with their counterparts w_α° obtained using the normal approximation for W.

Problem 9

From two population distributions with the same $\sigma = 10$, two samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ of sizes $n = 25$ can be taken in two ways

(a) paired sample with $\mathrm{Cov}(X_i, Y_i) = 50$, $i = 1, \ldots, 25$,
(b) two independent random samples.

Compare the power curves for testing

$$H_0 : \mu_1 = \mu_2, \qquad H_1 : \mu_1 > \mu_2, \qquad \alpha = 0.05,$$

for the settings (a) and (b).

Problem 10

The table on the right reports fifteen pairs (x_i, y_i) of measurements. What are $\bar{x} - \bar{y}$ and $s_{\bar{x}-\bar{y}}$?

If the pairing had been erroneously ignored and it had been assumed that the two samples were independent, what would have been the estimate of the standard deviation of $\bar{X} - \bar{Y}$?

Analyse the data to determine if there is a systematic difference between μ_1 and μ_2.

Sample 1	Sample 2
97.2	97.2
105.8	97.8
99.5	96.2
100	101.8
93.8	88
79.2	74
72	75
72	67.5
69.5	65.8
20.5	21.2
95.2	94.8
90.8	95.8
96.2	98
96.2	99
91	100.2

Problem 11

The media often present short reports of the results of experiments. To the critical reader, such reports often raise more questions than they answer. Comment on the following pitfalls in the interpretation of each of the following.

(a) It is reported that patients whose hospital rooms have a window recover faster than those whose rooms do not.

(b) Nonsmoking wives whose husbands smoke have a cancer rate twice that of wives whose husbands do not smoke.

(c) A two-year study in North Carolina found that 75% of all industrial accidents in the state happened to workers who had skipped breakfast.

(d) A 15-year study of more than 45 000 Swedish soldiers revealed that heavy users of marijuana were six times more likely than nonusers to develop schizophrenia.

(e) A study of nearly 4000 elderly North Carolinians has found that those who attended religious services every week were 46% less likely to die over a six-year period than people who attended less often or not at all.

Problem 12

The following table shows admission rates for the six most popular majors at the graduate school at the University of California at Berkeley. The numbers in the table are the number of applicants and the percentage admitted.

	Men	Women
Major A	825 (62%)	108 (82%)
Major B	560 (63%)	25 (68%)
Major C	325 (37%)	593 (34%)
Major D	417 (33%)	375 (35%)
Major E	191 (28%)	393 (34%)
Major F	373 (6%)	341 (7%)

(a) If the percentage admitted are compared, women do not seem to be unfavourably treated. But when the combined admission rates for all six majors are calculated, it is found that 44% of the men and only 30% of the women were admitted. How this paradox is resolved?

(b) This is an example of an observational study. Suggest a controlled experiment testing relevant statistical hypotheses.

Problem 13

You have got a grant to measure the average weight of the hippopotamus at birth. You have seen in a previous publication by Stanley and Livingstone that for male calves the distribution of weights has a mean of roughly 70 kg and a standard deviation of 10 kg, while these numbers are 60 kg and 5 kg for females, but you are interested in a better remeasurement of the overall average.

The experimental procedure is simple: you wait for the herd of hippopotami to be sleeping, you approach a newborn, you put it quickly on the scales, and you pray for the mother not to wake up. You managed to weigh 13 female and 23 male newborns with the following results:

	Female	Male
Sample mean	62.8	69.7
Sample standard deviation	6.8	11.7

(a) Test the null hypothesis of the equal sex ratio for the newborn hippopotami (meaning that the ratio of males to females at birth is 1 to 1).

(b) Assuming the ratio of males to females at birth is 1 to 1, suggest two different unbiased point estimates for the overall average weight of the hippopotamus at birth: the random sample mean and the stratified sample mean.

(c) Compute the standard errors for the stratified sample mean.

Problem 14

For each of nine horses, a veterinary anatomist measured the density of nerve cells at specified sites in the intestine:

Animal	Site I	Site II
1	50.6	38.0
2	39.2	18.6
3	35.2	23.2
4	17.0	19.0
5	11.2	6.6
6	14.2	16.4
7	24.2	14.4
8	37.4	37.6
9	35.2	24.4

The null hypothesis of interest is that in the population of all horses there is no difference between the two sites.

(a) Which of the two non-parametric tests is appropriate here: the rank sum test or the signed rank test? Explain your choice.

(b) On the basis of the data, would you reject the null hypothesis? Use one of the tests named in the item (a).

(c) Explain the following extract from the course text book:

> More precisely, with the signed rank test, H_0 states that the distribution of the differences is symmetric about zero. This will be true if the members of pairs of experimental units are assigned randomly to treatment and control conditions, and the treatment has no effect at all.

Problem 15

A study is conducted of the association between the rate at which words are spoken and the ability of a “talking computer” to recognise commands that it is programmed to accept. A random sample of 50 commands is spoken first at a rate under 60 words per minute, and then the same commands are repeated at a rate over 60 words per minute. In the first case the computer recognised 42 out of 50 commands while in the second case it recognised only 35 commands. Is the observed difference statistically significant? Assume that if the "talking computer" made a correct answer in the fast regime, then it recognised correctly the same command in the slow regime.

Chapter 8

Analysis of variance

The two sample setting from the previous chapter is the case, where the response is explained by a single main factor having two levels. In this chapter devoted to the analysis of variance or in short ANOVA, we extend the settings with one or two main factors having arbitrary many levels. The one-way layout setting extends the two independent samples case, while the two-way layout generalises the paired sample setting of the previous section.

8.1 One-way layout ANOVA

Consider the one-way layout model from Section 1.2. Suppose that for each of the I levels of the main factor A, we have independently collected a random sample $(y_{i1}, \ldots, y_{in})$ of size n. With I independent samples of size n in hand, we want to test

$$H_0 : \mu_1 = \ldots = \mu_I, \text{ against } H_1 : \mu_i \neq \mu_j \text{ for some pair of indices } (i, j).$$

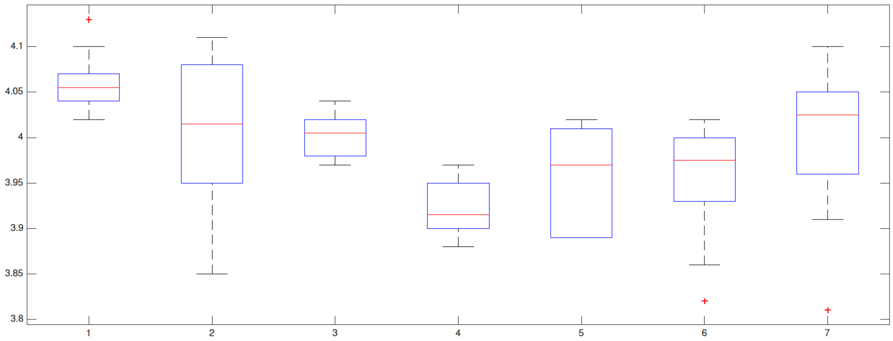
If the levels of the factor A are I different treatments in a comparison study, then the above null hypothesis claims that the compared treatments have the same effect, implying that the factor A has no influence on the measured response and the suggested one-way layout model is not useful.

Example: seven labs

The table below lists 70 measurements of chlorpheniramine maleate in tablets with a nominal dosage of 4 mg made by seven different labs. This is an example of a one-way layout data consisting of $I = 7$ independent samples each of size $n = 10$.

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
4.13	3.86	4.00	3.88	4.02	4.02	4.00
4.07	3.85	4.02	3.88	3.95	3.86	4.02
4.04	4.08	4.01	3.91	4.02	3.96	4.03
4.07	4.11	4.01	3.95	3.89	3.97	4.04
4.05	4.08	4.04	3.92	3.91	4.00	4.10
4.04	4.01	3.99	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.92	3.89	3.98	3.91
4.06	4.04	3.97	3.9	3.89	3.99	3.96
4.10	3.97	3.98	3.97	3.99	4.02	4.05
4.04	3.95	3.98	3.90	4.00	3.93	4.06

The data is graphically presented in the form of seven boxplots.



The null hypothesis of interest H_0 states that there is no significant difference between the output of the seven laboratories. We would reject H_0 if the discrepancy between the ordered sample means

Lab i	1	3	7	2	5	6	4
Mean μ_i	4.062	4.003	3.998	3.997	3.957	3.955	3.920

can not be explained by the noise factors.

Normal theory model

Assume that $N = I \cdot n$ response variables have the form

$$Y_{ik} = \mu + \alpha_i + \sigma Z_{ik}, \quad Z_{ik} \sim \mathrm{N}(0,1), \quad i = 1, \ldots, I, \quad k = 1, \ldots, n,$$

where $\alpha_i = \mu_i - \mu$ are the main effects such that

$$\alpha_1 + \ldots + \alpha_I = 0.$$

Here the noise components σZ_{ik} are assumed to have the same standard deviation σ. The corresponding maximum likelihood estimates of μ, $\mu_i = \mu + \alpha_i$, and α_i,

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\mu}_i = \bar{y}_{i.}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..},$$

can be expressed in terms of the sample means

$$\bar{y}_{i.} = \frac{1}{n}\sum_{k=1}^{n} y_{ik}, \quad \bar{y}_{..} = \frac{\bar{y}_{1.} + \ldots + \bar{y}_{I.}}{I} = \frac{1}{N}\sum_{i=1}^{I}\sum_{k=1}^{n} y_{ik}.$$

It is easy to verify that

$$\hat{\alpha}_1 + \ldots + \hat{\alpha}_I = 0.$$

The observed response values can be represented as

$$y_{ik} = \hat{\mu} + \hat{\alpha}_i + \hat{e}_{ik}, \quad \hat{e}_{ik} = y_{ik} - \bar{y}_{i.},$$

where $\hat{e}_{ik}$ are the so-called residuals giving the discrepancy between the observed response and the corresponding sample mean.

The ANOVA test is built around the following decomposition

$$\mathrm{ss_T} = \mathrm{ss_A} + \mathrm{ss_E},$$

where

$\mathrm{ss_T} = \sum_i \sum_k (y_{ik} - \bar{y}_{..})^2$ is the total sum of squares for the pooled sample,

$\mathrm{ss_A} = n\sum_i \hat{\alpha}_i^2$ is the factor A sum of squares,

$\mathrm{ss_E} = \sum_i \sum_k \hat{e}_{ik}^2$ is the error sum of squares.

Each of these sums of squares $(\mathrm{ss_T}, \mathrm{ss_A}, \mathrm{ss_E})$ is characterised by its number of degrees of freedom:

$$\mathrm{df_T} = N - 1, \quad \mathrm{df_A} = I - 1, \quad \mathrm{df_E} = I \cdot (n-1).$$

Notice that

$$\mathrm{df_T} = \mathrm{df_A} + \mathrm{df_E}.$$

This decomposition says that the total variation in the response values is the sum of the between-group variation and the within-group variation, making clear why the data analysis based on this decomposition is called the analysis of variance.

By normalising the sums of squares with help of the numbers of degrees of freedom, we obtain the so-called mean sums of squares

$$\mathrm{ms_A} = \frac{\mathrm{ss_A}}{\mathrm{df_A}}, \quad \mathrm{ms_E} = \frac{\mathrm{ss_E}}{\mathrm{df_E}}.$$

The underlying random variables have the mean values

$$\mathrm{E}(\mathrm{MS_A}) = \sigma^2 + \frac{n}{I-1}\sum_i \alpha_i^2, \qquad \mathrm{E}(\mathrm{MS_E}) = \sigma^2.$$

Notice that $\mathrm{ms_E}$ gives an unbiased estimate of the noise variance σ^2. It can be viewed as the pooled sample variance

$$s_\mathrm{p}^2 = \mathrm{ms_E} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{n}(y_{ik} - \bar{y}_{i.})^2}{I(n-1)}.$$

One-way F-test

Under the null hypothesis $H_0 : \alpha_1 = \ldots = \alpha_I = 0$, we expect

$$\mathrm{E}(\mathrm{MS_A}) = \mathrm{E}(\mathrm{MS_E}) = \sigma^2.$$

On the other hand, if H_0 is violated, then $\mathrm{E}(MS_{\mathrm{A}}) > \mathrm{E}(MS_{\mathrm{E}})$, as given some $\alpha_i \neq 0$, we have

$$\sigma^2 + \frac{n}{I-1}\sum_i \alpha_i^2 > \sigma^2.$$

Thus taking the ratio between the two mean squares

$$f = \frac{\mathrm{ms_A}}{\mathrm{ms_E}}$$

as the test statistic, we should reject H_0 using the rejection region

$$\mathcal{R} = \{f \geq f_\alpha\}.$$

The critical value f_α should be determined by the null distribution

$$F \overset{H_0}{\sim} F_{k_1,k_2}, \qquad k_1 = I-1,\ k_2 = I(n-1),$$

where F_{k_1,k_2} is the so-called F-distribution with degrees of freedom (k_1, k_2). The F-distribution is the distribution for the ratio

$$\frac{X_1/k_1}{X_2/k_2} \sim F_{k_1,k_2},$$

where $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ are two independent random variables having chi-squared distributions. The critical values of the F-distributions for different significance levels α are given in Section 11.4.

Example: seven labs

The one-way ANOVA table below

Source	df	ss	ms	f	p
Labs	6	0.125	0.0210	5.66	0.0001
Error	63	0.231	0.0037		
Total	69	0.356			

summarises the F-test applied to the seven labs data. The p-value 0.01% of the test is computed from the observed test statistic $f = 5.66$ using the $F_{6,63}$-distribution built in the R-program. The use of the table in Section 11.4 allows us to conclude that the p-value must be smaller than 0.1% since 5.66 is larger than the closest available table value $F_{6,63}(0.001) = 4.3395$.

The normal QQ-plot of residuals $\hat{e}_{ik}$ supports the normality assumption. The noise size σ is estimated by $s_{\mathrm{p}} = \sqrt{0.0037} = 0.061$.

8.2 Simultaneous confidence interval

The analysis of the last example leads to the conclusion that at least one of the $c = \binom{7}{2} = 21$ pairwise differences is significant. The table below lists the 10 largest pairwise difference between the sample means.

(i,j)	$(1,4)$	$(1,6)$	$(1,5)$	$(3,4)$	$(7,4)$	$(2,4)$	$(1,2)$	$(1,7)$	$(1,3)$	$(5,4)$
$\bar{y}_{i.} - \bar{y}_{j.}$	0.142	0.107	0.105	0.083	0.078	0.077	0.065	0.064	0.059	0.047

Clearly, the difference $(\mu_1 - \mu_4)$ corresponding to to the largest sample mean difference

$$\bar{y}_{1.} - \bar{y}_{4.} = 0.142$$

must be significant at the 5% significance level. The important follow up question is: which of the other 21 pairwise differences are significant and which are not at the 5% significance level?

A naive answer to this question would rely on the 95% confidence interval for a single pair of independent samples $(\mu_i - \mu_j)$ from the previous chapter

$$I_{\mu_i - \mu_j} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{63}(0.025) \cdot \frac{s_{\mathrm{p}}}{\sqrt{5}} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm 0.055,$$

where $t_{63}(0.025) = 2.00$. Notice that here we use the t-distribution with 63 degrees of freedom, as the corresponding pooled sample variance $s_{\mathrm{p}}^2 = 0.0037$ is based on $I = 7$ samples each of size $n = 10$. According to this confidence interval formula we have 9 significant differences: $(1,4),(1,6),(1,5),(3,4),(7,4),(2,4),(1,2),(1,7),(1,3)$, since the next ordered difference 0.047 for the pair $(5,4)$ is smaller than 0.055. The inequality $0.047 < 0.055$ ensures that the 95% confidence interval

$$I_{\mu_5-\mu_4} = 0.047 \pm 0.055$$

contains 0, implying that the difference $(\mu_5 - \mu_4)$ is not significant at the 5% significance level.

The problem with the suggested solution is that the use of the confidence interval formula aimed at a single difference may produce false discoveries. This is an example of the famous *multiple comparison problem* nicely illustrated by the illuminating cartoon "Jelly beans cause acne", see https://xkcd.com/882/. What we need here is a simultaneous confidence interval formula addressing all $c = 21$ pairwise comparisons. Next, we introduce two formulas for the simultaneous confidence interval: the Bonferrroni formula $B_{\mu_i-\mu_j}$, and the Tukey formula $T_{\mu_i-\mu_j}$.

Bonferroni method

Think of a statistical test involving a null hypothesis H_0 at the significance level α_c is repeatedly applied to c independently sampled datasets. The overall result is deemed to be positive if we get at least one positive result among these c tests. Then the overall significance level α is computed as

$$\alpha = \mathrm{P}(X_c \geq 1|H_0),$$

where X_c is the number of positive results among these c tests. Due to the independence of the tests, we have

$$X_c \overset{H_0}{\sim} \mathrm{Bin}(c, \alpha_c),$$

yielding

$$\mathrm{P}(X \geq 1|H_0) = 1 - (1-\alpha_c)^c.$$

Assuming α_c is small, we find that

$$\alpha = 1 - (1-\alpha_c)^c \approx c\alpha_c.$$

This reasoning leads to the following important conclusion. To obtain the overall significance level α for the multiple testing procedure involving c independent tests, each individual test should be performed at the level α/c. Applying this idea to the current setting of $c = \binom{I}{2}$ pairwise differences $(\mu_i - \mu_j)$, we arrive at the the Bonferroni's formula of the $100(1-\alpha)\%$ simultaneous confidence interval:

$$B_{\mu_i-\mu_j} = \bar{y}_{i.} - \bar{y}_{j.} \pm t_{I(n-1)}(\tfrac{\alpha}{I(I-1)}) \cdot s_{\mathrm{p}}\sqrt{\tfrac{2}{n}}, \quad 1 \leq i,j \leq I.$$

This formula is obtained from the confidence interval for a single difference

$$I_{\mu_i-\mu_j} = \bar{y}_{i.} - \bar{y}_{j.} \pm t_{I(n-1)}(\tfrac{\alpha}{2}) \cdot s_{\mathrm{p}}\sqrt{\tfrac{2}{n}}, \quad 1 \leq i,j \leq I,$$

by applying the Bonferroni correction of the significance level

$$\alpha \to \tfrac{\alpha}{c} = \tfrac{2\alpha}{I(I-1)}.$$

Observe that the pairwise differences $\mu_i - \mu_j$ are not independent as required by Bonferroni method: for example, knowing the differences $(\mu_1 - \mu_2)$ and $(\mu_2 - \mu_3)$ we may exactly predict the third difference $\mu_1 - \mu_3$ using the equality

$$\mu_1 - \mu_2 + \mu_2 - \mu_3 = \mu_1 - \mu_3.$$

As a result the Bonferroni method gives slightly wider intervals compared to the Tukey method introduced below.

Example: seven labs

For the seven labs example, the 95% Bonferroni interval takes the form

$$B_{\mu_i-\mu_j} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{63}(\tfrac{0.025}{21}) \cdot \tfrac{s_{\mathrm{p}}}{\sqrt{5}} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm 0.086,$$

where $t_{63}(0.0012) = 3.17$. Compared to the naive interval

$$I_{\mu_i-\mu_j} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm 0.055,$$

the Bonferroni interval

$$B_{\mu_i-\mu_j} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm 0.086$$

is much wider, resulting in only 3 significant differences: $(1,4),(1,5),(1,6)$.

Tukey method

Under the normality assumption with equal variances, the centred sample means

$$X_i = \bar{Y}_{i.} - \mu_i \sim \mathrm{N}(0, \tfrac{\sigma}{\sqrt{n}}), \quad i = 1, \ldots, I,$$

are independent normal variables. Consider the range of all paiwise differences $X_i - X_j$:

$$\max\{X_1, \ldots, X_I\} - \min\{X_1, \ldots, X_I\}.$$

giving the largest pairwise difference between the components of the vector $(Z_1, \ldots, Z_I)$. The corresponding normalised range has a distribution that is free from the parameter σ

$$\frac{\max\{X_1, \ldots, X_I\} - \min\{X_1, \ldots, X_I\}}{S_{\mathrm{p}}/\sqrt{n}} \sim \mathrm{SR}_{k_1,k_2}, \quad k_1 = I, \ k_2 = I(n-1).$$

Here, the so-called studentised range distribution SR_{k_1,k_2} is controlled by two parameters: the number of independent samples k_1 and the number of degrees of freedom k_2 used in the variance estimate s_{p}^2.

$$\boxed{\text{Tukey's } 100(1-\alpha)\% \text{ simultaneous confidence interval } T_{\mu_i-\mu_j} = \bar{y}_{i.} - \bar{y}_{j.} \pm q_{I,I(n-1)}(\alpha) \cdot \tfrac{s_{\mathrm{p}}}{\sqrt{n}}}$$

The Tukey interval is built using the factor $q_{k_1,k_2}(\alpha)$ determined by

$$\mathrm{P}(Q > q_{k_1,k_2}(\alpha)) = \alpha, \text{ where } Q \sim \mathrm{SR}_{k_1,k_2}.$$

In contrast to Bonferroni, Tukey takes into account the dependences between the differences $(\mu_i - \mu_j)$.

Example: seven labs

Using the R-command `qtukey(0.95,7,63)` we find $q_{7,63}(0.05) = 4.307$ so that

$$T_{\mu_i-\mu_j} = \bar{y}_{i.} - \bar{y}_{j.} \pm q_{7,63}(0.05) \cdot \tfrac{0.061}{\sqrt{10}} = \bar{y}_{i.} - \bar{y}_{j.} \pm 0.083,$$

recognising four significant pairwise differences: $(1,4), (1,5), (1,6), (3,4)$.

8.3 Kruskal-Wallis test

The F-test requires the residual analysis where for example with help of the normal QQ-plot of the residuals one should evaluate whether the assumption of normality and equal variances is reasonable. If these assumptions are not valid then one can use the non-parametric Kruskal-Wallis test described next.

Assume that we have I independent random samples and consider the null hypothesis of no effect

$$H_0: \text{ the underlying } I \text{ population distributions are equal.}$$

Extending the idea of the rank-sum test, turn to the pooled sample of size $N = I \cdot n$. Let r_{ik} be the pooled ranks of the sample values y_{ik}, so that

$$\sum_i \sum_k r_{ik} = 1 + 2 + \ldots + N = \frac{N(N+1)}{2},$$

implying that the overall mean rank is

$$\bar{r}_{..} = \frac{N(N+1)}{2N} = \frac{N+1}{2}.$$

The Kruskal-Wallis test statistic

$$w = \frac{12n}{N(N+1)} \sum_{i=1}^{I} (\bar{r}_{i.} - \tfrac{N+1}{2})^2$$

measures the discrepancy between the sample means of the ranks

$$\bar{r}_{i.} = \frac{r_{i1} + \ldots + r_{in}}{n}, \quad i = 1, \ldots, I.$$

A large value of w would indicate a deviation from the null distribution. Thus the Kruskall-Wallis test rejects H_0 for larger values of w. For $I = 3$, $n \geq 5$ or $I \geq 4$, $n \geq 4$, one can use the approximate null distribution

$$W \overset{H_0}{\approx} \chi^2_{I-1}.$$

Example: seven labs

In the table below the actual measurements are replaced by their ranks $1 \div 70$. For the observed Kruskal-Wallis test statistic $w = 28.17$, using the χ^2_6-distribution table, we see that the p-value of the Kruskal-Wallis test is smaller than 0.005.

Labs	1	2	3	4	5	6	7
	70	4	35	6	46	48	38
	63	3	45	7	21	5	50
	53	65	40	13	47	22	52
	64	69	41	20	8	28	58
	59	66	57	16	14	37	68
	54	39	32	26	42	2	1
	43	44	51	17	9	31	15
	61	56	25	11	10	34	23
	67	24	29	27	33	49	60
	55	19	30	12	36	18	62
Means	58.9	38.9	38.5	15.5	26.6	27.4	42.7

8.4 Two-way layout ANOVA

Referring to the two-way layout model from Section 1.2, assume that the dataset in hand

$$\{y_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, n\}$$

is generated in the following way

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \sigma Z_{ijk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n,$$

where Z_{ijk} are independent random variables having the standard normal distribution $\mathrm{N}(0, 1)$. We start by assuming that $n \geq 2$, so that for each combination of levels (i, j) of the main factors we have at least two replications, which would allow us to measure the variation in the response variable within each cell (i, j). The important case $n = 1$ will be addressed separately.

The maximum likelihood estimates under the assumption of normality take the form

$$\begin{aligned}
\hat\mu &= \bar y_{...} = \tfrac{1}{IJn}\sum_i\sum_j\sum_k y_{ijk}, \\
\hat\alpha_i &= \bar y_{i..} - \bar y_{...}, \quad \bar y_{i..} = \tfrac{1}{Jn}\sum_j\sum_k y_{ijk}, \\
\hat\beta_j &= \bar y_{.j.} - \bar y_{...}, \quad \bar y_{.j.} = \tfrac{1}{In}\sum_i\sum_k y_{ijk}, \\
\hat\delta_{ij} &= \bar y_{ij.} - \bar y_{...} - \hat\alpha_i - \hat\beta_j = \bar y_{ij.} - \bar y_{i..} - \bar y_{.j.} + \bar y_{...}, \quad \bar y_{ij.} = \tfrac{1}{n}\sum_k y_{ijk}.
\end{aligned}$$

Given $n \geq 2$, after computing the residuals

$$\hat e_{ijk} = y_{ijk} - \bar y_{ij.}, \quad k = 1, \ldots, n,$$

we arrive at the crucial decomposition

$$y_{ijk} = \hat\mu + \hat\alpha_i + \hat\beta_j + \hat\delta_{ij} + \hat e_{ijk}.$$

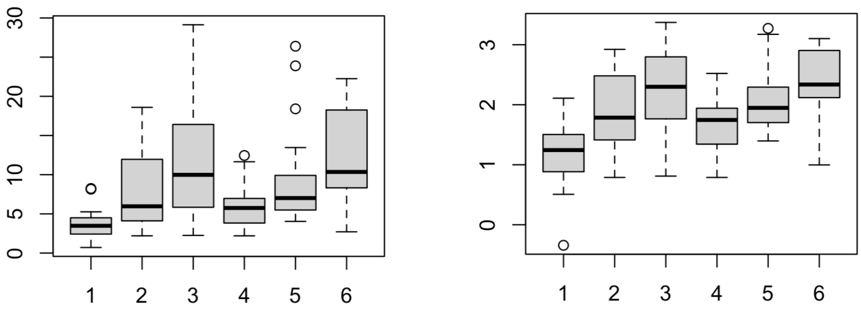
Case study: iron retention

The table below gives the percentages of iron retained in 108 mice randomly allocated into six groups. Here the factor A has two levels $I = 2$ representing two iron forms, and the factor B has three levels $J = 3$ representing the different dosage concentrations. Each sample has size $n = 18$.

Fe^{3+} (10.2)	Fe^{3+} (1.2)	Fe^{3+} (0.3)	Fe^{2+} (10.2)	Fe^{2+} (1.2)	Fe^{2+} (0.3)
0.71	2.20	2.25	2.20	4.04	2.71
1.66	2.93	3.93	2.69	4.16	5.43
2.01	3.08	5.08	3.54	4.42	6.38
2.16	3.49	5.82	3.75	4.93	6.38
2.42	4.11	5.84	3.83	5.49	8.32
2.42	4.95	6.89	4.08	5.77	9.04
2.56	5.16	8.50	4.27	5.86	9.56
2.60	5.54	8.56	4.53	6.28	10.01
3.31	5.68	9.44	5.32	6.97	10.08
3.64	6.25	10.52	6.18	7.06	10.62
3.74	7.25	13.46	6.22	7.78	13.80
3.74	7.90	13.57	6.33	9.23	15.99
4.39	8.85	14.76	6.97	9.34	17.90
4.50	11.96	16.41	6.97	9.91	18.25
5.07	15.54	16.96	7.52	13.46	19.32
5.26	15.89	17.56	8.36	18.40	19.87
8.15	18.3	22.82	11.65	23.89	21.60
8.24	18.59	29.13	12.45	26.39	22.25

The six boxplots for the retention data (see the next figure, left panel) show that all samples come from population distributions skewed to the right, and that there is a clear discrepancy among different sample standard deviations.

With the data skewed to the right, taking the logarithms might serve as a simple remedy. For this reason, we will focus on the transformed data (y_{ijk}) obtained by taking the natural logarithms of the percentage of iron retention. The right panel of the next figure shows that the boxplots for the transformed data (y_{ijk}) look more symmetric and exhibit less discrepancy between the within sample variation.



The next table lists the sample means $(\bar{y}_{ij\cdot}, \bar{y}_{i\cdot\cdot}, \bar{y}_{\cdot j\cdot}, \bar{y}_{\cdot\cdot\cdot})$ for the transformed data (y_{ijk}):

	10.2	1.2	0.3	Level mean
Fe^{3+}	1.16	1.90	2.28	1.78
Fe^{2+}	1.68	2.09	2.40	2.06
Level mean	1.42	2.00	2.34	1.92

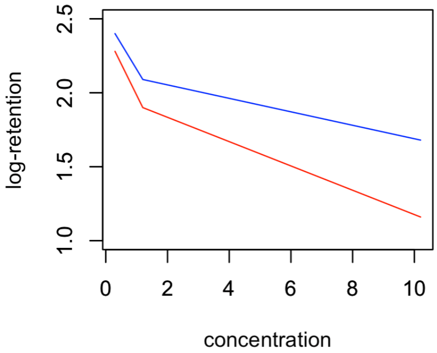
The resulting maximum likelihood estimates are

$$\hat{\mu} = 1.92, \qquad \hat{\alpha}_1 = -0.14, \qquad \hat{\alpha}_2 = 0.14, \qquad \hat{\beta}_1 = -0.50, \qquad \hat{\beta}_2 = 0.08, \qquad \hat{\beta}_3 = 0.42,$$

and

$$(\hat{\delta}_{ij}) = \begin{pmatrix} -0.12 & 0.04 & 0.08 \\ 0.12 & -0.04 & -0.08 \end{pmatrix}.$$

It is useful to graphically compare the two profiles shown on the figure below, where the red line connects the three sample means for the iron form Fe^{3+}, see the first row of the previous table, and the blue line does the same for for Fe^{2+}. In particular, the fact that the two rows are not parallel may indicate a possible interaction between the two main factors.



Observe that the estimated effect size for the log-retention is

$$\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{y}_{2..} - \bar{y}_{1..} = 0.28.$$

This yields the multiplicative effect of $e^{0.28} = 1.32$ on the original scale, implying that the retention percentage of Fe^{2+} is higher than that of Fe^{3+} by factor 1.32.

Three F-tests for the two-way ANOVA

The two-way ANOVA version of the decomposition of the sums of squares has the form

$$\mathrm{ss_T} = \mathrm{ss_A} + \mathrm{ss_B} + \mathrm{ss_{AB}} + \mathrm{ss_E},$$

where

$$\begin{aligned}
\mathrm{ss_T} &= \textstyle\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2, & \mathrm{df_T} &= IJn - 1,\\
\mathrm{ss_A} &= Jn \textstyle\sum_i \hat{\alpha}_i^2, & \mathrm{df_A} &= I - 1,\\
\mathrm{ss_B} &= In \textstyle\sum_j \hat{\beta}_j^2, & \mathrm{df_B} &= J - 1,\\
\mathrm{ss_{AB}} &= n \textstyle\sum_i \sum_j \hat{\delta}_{ij}^2, & \mathrm{df_{AB}} &= (I-1)(J-1),\\
\mathrm{ss_E} &= \textstyle\sum_i \sum_j \sum_k \hat{e}_{ijk}^2, & \mathrm{df_E} &= IJ(n-1).
\end{aligned}$$

The corresponding mean sums of squares and their expected values are

$$\begin{aligned}
\mathrm{ms_A} &= \tfrac{\mathrm{ss}_A}{\mathrm{df_A}}, & \mathrm{E(MS_A)} &= \sigma^2 + \tfrac{Jn}{I-1} \textstyle\sum_i \alpha_i^2,\\
\mathrm{ms_B} &= \tfrac{\mathrm{ss_B}}{\mathrm{df_B}}, & \mathrm{E(MS_B)} &= \sigma^2 + \tfrac{In}{J-1} \textstyle\sum_j \beta_j^2,\\
\mathrm{ms_{AB}} &= \tfrac{\mathrm{ss_{AB}}}{\mathrm{df_{AB}}}, & \mathrm{E(MS_{AB})} &= \sigma^2 + \tfrac{n}{(I-1)(J-1)} \textstyle\sum_i \sum_j \delta_{ij}^2,\\
\mathrm{ms_E} &= \tfrac{\mathrm{ss_E}}{\mathrm{df_E}}, & \mathrm{E(MS_E)} &= \sigma^2.
\end{aligned}$$

The error mean sum of squares $\mathrm{ms_E}$, an unbiased estimate of σ^2, is in fact the pooled sample variance

$$s_\mathrm{p}^2 = \mathrm{ms_E} = \frac{\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2}{IJ(n-1)}.$$

In the two-way ANOVA setting, there are three different null hypotheses of interest

H_A: $\alpha_1 = \ldots = \alpha_I = 0$, there is no systematic difference between different levels of the factor A,

H_B: $\beta_1 = \ldots = \beta_J = 0$, there is no systematic difference between different levels of the factor B,

H_AB: all $\delta_{ij} = 0$, there is no interaction between the factors A and B.

Following the idea of the F-test in the one-way setting, we arrive at three F-tests addressing the three null hypotheses using three different F-test statistics.

Null hypothesis	No effect property	Test statistics and its null distribution
H_A	$\mathrm{E(MS_A)} = \mathrm{E(MS_E)} = \sigma^2$	$F_\mathrm{A} = \frac{\mathrm{MS_A}}{\mathrm{MS_E}} \sim F_{\mathrm{df_A},\mathrm{df_E}}$
H_B	$\mathrm{E(MS_B)} = \mathrm{E(MS_E)} = \sigma^2$	$F_\mathrm{B} = \frac{\mathrm{MS_B}}{\mathrm{MS_E}} \sim F_{\mathrm{df_B},\mathrm{df_E}}$
H_AB	$\mathrm{E(MS_{AB})} = \mathrm{E(MS_E)} = \sigma^2$	$F_\mathrm{AB} = \frac{\mathrm{MS_{AB}}}{\mathrm{MS_E}} \sim F_{\mathrm{df_{AB}},\mathrm{df_E}}$

Again, we reject the null hypothesis for the values of the F-test statistic larger than the critical value determined by the corresponding F-distribution.

Example: iron retention

The two-way ANOVA table for the transformed iron retention data takes the form

Source	df	ss	ms	f	p
Iron form	1	2.074	2.074	5.99	0.017
Dosage	2	15.588	7.794	22.53	0.000
Interaction	2	0.810	0.405	1.17	0.315
Error	102	35.296	0.346		
Total	107	53.768			

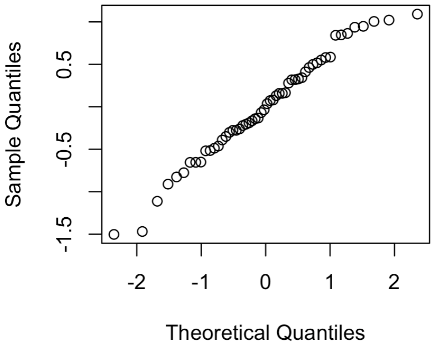
Referring to the three p-values in the rightmost column, obtained using the three null distributions

$$(F_{1,102}, F_{2,102}, F_{2,102}),$$

we draw the following three conclusions

- there is a significant effect due to iron form,
- the dosage effect is undoubtably significant, however, this is something expected,
- the interaction effect between the iron form and the dosage is not statistically significant.

Concerning the main conclusion, recall that our previous analysis of two samples at the intermediate dosage has produced a non-significant result.



Finally, after inspecting the normal QQ-plot for the residuals $(\hat{e}_{ijk})$ we approve the assumptions of normality and equal variance for the transformed data.

Randomised block design

The famous data collecting principle suggests:

block what you can, randomise what you cannot.

Blocking is used to remove the effects of the most important nuisance variable. Randomisation is then used to reduce the contaminating effects of the remaining nuisance variables. Applied to the ANOVA two-way layout setting, this principle leads to the following experimental design: randomly assign I treatments within each of J blocks. The first of the next three examples is illustrated by the figure above.

Block	Treatments	Observation
A homogeneous plot of land divided into I subplots	I fertilisers each applied to a randomly chosen subplot	The yield on the subplot (i, j)
A four-wheel car	4 types of tires tested on the same car	tire's life-length
A litter of I animals	I diets randomly assigned to I siblings	the weight gain

8.5 Additive model

In this section we assume that $n = 1$. With only one replication per cell (i, j), we cannot estimate the interaction effect between two main effects A and B. This restricts us to the additive model without interaction

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sigma Z_{ij}, \quad Z_{ij} \sim \mathrm{N}(0, 1).$$

For the given data (y_{ij}), we find the maximum likelihood estimates and the residuals using

$$\hat\mu = \bar y_{..}, \qquad \hat\alpha_i = \bar y_{i.} - \bar y_{..}, \qquad \hat\beta_i = \bar y_{.j} - \bar y_{..},$$
$$\hat e_{ij} = y_{ij} - \bar y_{..} - \hat\alpha_i - \hat\beta_i = y_{ij} - \bar y_{i.} - \bar y_{.j} + \bar y_{..}.$$

It is easy to check that

$$y_{ij} = \hat\mu + \hat\alpha_i + \hat\beta_j + \hat e_{ij}.$$

In this case, the decomposition of the sums of squares takes a reduced form

$$\mathrm{ss_T} = \mathrm{ss_A} + \mathrm{ss_B} + \mathrm{ss_E},$$

with

$$\begin{aligned}
\mathrm{ss_T} &= \textstyle\sum_i \sum_j (\bar y_{ij} - \bar y_{..})^2, & \mathrm{df_T} &= IJ - 1,\\
\mathrm{ss_A} &= J \textstyle\sum_i \hat\alpha_i^2, & \mathrm{df_A} &= I - 1,\\
\mathrm{ss_B} &= I \textstyle\sum_j \hat\beta_j^2, & \mathrm{df_B} &= J - 1,\\
\mathrm{ss_E} &= \textstyle\sum_i \sum_j \hat e_{ij}^2, & \mathrm{df_E} &= (I - 1)(J - 1).
\end{aligned}$$

Using

$$\begin{aligned}
\mathrm{ms_A} &= \tfrac{\mathrm{ss}_A}{\mathrm{df_A}}, & \mathrm{E(MS_A)} &= \sigma^2 + \tfrac{J}{I-1} \textstyle\sum_i \alpha_i^2\\
\mathrm{ms_B} &= \tfrac{\mathrm{ss_B}}{\mathrm{df_B}}, & \mathrm{E(MS_B)} &= \sigma^2 + \tfrac{I}{J-1} \textstyle\sum_j \beta_j^2\\
\mathrm{ms_E} &= \tfrac{\mathrm{ss_E}}{\mathrm{df_E}}, & \mathrm{E(MS_E)} &= \sigma^2.
\end{aligned}$$

we can apply two F-tests for two different null hypotheses

$$\begin{aligned}
H_\mathrm{A} &: \alpha_1 = \ldots = \alpha_I = 0, & F_\mathrm{A} &= \frac{\mathrm{MS_A}}{\mathrm{MS_E}} \overset{H_\mathrm{A}}{\sim} F_{\mathrm{df_A},\mathrm{df_E}},\\
H_\mathrm{B} &: \beta_1 = \ldots = \beta_J = 0, & F_\mathrm{B} &= \frac{\mathrm{MS_B}}{\mathrm{MS_E}} \overset{H_\mathrm{B}}{\sim} F_{\mathrm{df_B},\mathrm{df_E}}.
\end{aligned}$$

Example: itching treatments

In a study the following $I = 7$ treatments

Treat 1: no drug, Treat 2: placebo, Treat 3: papaverine, Treat 4: morphine,

Treat 5: aminophylline, Treat 6: pentabarbital, Treat 7: tripelennamine.

were compared in their effect to relieve itching. Each treatment was applied to $J = 10$ male volunteers aged 20 - 30 years after an itching condition had been initiated by a certain injection. The data in the table give the durations of the itching in seconds (y_{ij}).

Subject	Treat 1	Treat 2	Treat 3	Treat 4	Treat 5	Treat 6	Treat 7
BG	174	263	105	199	141	108	141
JF	224	213	103	143	168	341	184
BS	260	231	145	113	78	159	125
SI	225	291	103	225	164	135	227
BW	165	168	144	176	127	239	194
TS	237	121	94	144	114	136	155
GM	191	137	35	87	96	140	121
SS	100	102	133	120	222	134	129
MU	115	89	83	100	165	185	79
OS	189	433	237	173	168	188	317

The resulting two-way ANOVA table

Source	df	ss	ms	f	p
Drugs	6	53013	8835	2.85	0.018
Subjects	9	103280	11476	3.71	0.001
Error	54	167130	3096		
Total	69	323422			

shows that the treatment effect is significant. The differences between the subjects are significant, as expected. The two p-values are computed from the $F_{6,54}$ and $F_{9,54}$-distributions.

Friedman test

Here we introduce a non-parametric test for the two-way layout of the ANOVA setting with $n = 1$, which does not require that ϵ_{ij} are normally distributed. The Friedman tests null hypothesis H_0: of no treatment effect based on within block ranking. Let the ranks within j-th block be:

$$(r_{1j}, \dots, r_{Ij}) = \text{ranks of } (y_{1j}, \dots, y_{Ij}),$$

so that

$$r_{1j} + \ldots + r_{Ij} = 1 + 2 + \ldots + I = \frac{I(I+1)}{2},$$

and the average ranks are

$$\bar{r}_{i.} = \frac{r_{i1} + \ldots + r_{iJ}}{J}, \quad \bar{r}_{..} = \frac{\bar{r}_{1.} + \ldots + \bar{r}_{I.}}{I} = \frac{I+1}{2}.$$

$$\boxed{\text{Friedman test statistic } q = \tfrac{12J}{I(I+1)} \textstyle\sum_{i=1}^{I} (\bar{r}_{i.} - \tfrac{I+1}{2})^2 \text{ has an approximate null distribution } Q \overset{H_0}{\approx} \chi^2_{I-1}.}$$

The Friedman test statistic q is a measure of agreement between J rankings, so we reject H_0 for large values of q.

Example: itching treatments

The boxplots for the seven treatments indicate violations of the assumptions of normality. Notably, the placebo results exhibit larger variance for the itching duration in seconds. We complement the previous ANOVA results with the Friedman non-parametric test based the following rank values (r_{ij}) and their treatment means $(\bar{r}_{1.}, \dots, \bar{r}_{7.})$.

Subject	Treat 1	Treat 2	Treat 3	Treat 4	Treat 5	Treat 6	Treat 7
BG	5	7	1	6	3.5	2	3.5
JF	6	5	1	2	3	7	4
BS	7	6	4	2	1	5	3
SI	6	7	1	4	3	2	5
BW	3	4	2	5	1	7	6
TS	7	3	1	5	2	4	6
GM	7	5	1	2	3	6	4
SS	1	2	5	3	7	6	4
MU	5	3	2	4	6	7	1
OS	4	7	5	2	1	3	6
$\bar{r}_{i.}$	5.10	4.90	2.30	3.50	3.05	4.90	4.25

Each subject acts as a jury member by ranking the seven treatment effects. Notice how the subject BG resolves the tie between treatments 5 and 7. The Friedman test statistic q is a measure of agreement among the jury members: if q is larger than the critical value, then the observed disagreement will allow us to reject the null hypothesis of no difference. With $\frac{I+1}{2} = 4$, we find the Friedman test statistic value to be

$$q = 2.14 \cdot \Big((5.10-4)^2 + (4.90-4)^2 + (2.30-4)^2 + (3.50-4)^2 + (3.05-4)^2 + (4.90-4)^2 + (4.25-4)^2\Big) = 14.28.$$

Using the chi-squared distribution table with $k = 6$ degrees of freedom we find that the p-value of the test is approximately 3%. The Friedman test also rejects the null hypothesis and suggests the following top 3 list in the effectiveness against itching:

(1) papaverine, (2) aminophylline, (3) morphine.

8.6 Exercises

Problem 1

For the one-way analysis of variance with two treatment groups, show that the F-test statistic equals t^2, where t is the test statistic for the two-sample t-test.

Problem 2

A study on the tensile strength of aluminium rods is conducted. Forty identical rods are randomly divided into four groups, each of size 10. Each group is subjected to a different heat treatment, and the tensile strength, in thousands of pounds per square inch, of each rod is determined. The following table presents the results of the measurements. Consider the null hypothesis of equality between the four treatment means of tensile strength

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Treatment	**1**	**2**	**3**	**4**	Combined data
	18.9	18.3	21.3	15.9	18.9 18.3 21.3 15.9
	20.0	19.2	21.5	16.0	20.0 19.2 21.5 16.0
	20.5	17.8	19.9	17.2	20.5 17.8 19.9 17.2
	20.6	18.4	20.2	17.5	20.6 18.4 20.2 17.5
	19.3	18.8	21.9	17.9	19.3 18.8 21.9 17.9
	19.5	18.6	21.8	16.8	19.5 18.6 21.8 16.8
	21.0	19.9	23.0	17.7	21.0 19.9 23.0 17.7
	22.1	17.5	22.5	18.1	22.1 17.5 22.5 18.1
	20.8	16.9	21.7	17.4	20.8 16.9 21.7 17.4
	20.7	18.0	21.9	19.0	20.7 18.0 21.9 19.0
mean	20.34	18.34	21.57	17.35	19.40
variance	0.88	0.74	0.88	0.89	3.58
skewness	0.16	0.14	-0.49	-0.08	0.05
kurtosis	2.51	2.59	2.58	2.46	1.98

(a) Test the null hypothesis applying an ANOVA test. Show clearly how all the sums of squares are computed using the sample means and variances given in the table.

(b) What are the assumptions of the ANOVA model you used? Do they appear fulfilled?

(c) The Bonferroni method gives the following formula for computing simultaneous 95% confidence intervals for six pairwise differences between four treatment means

$$B_{\mu_i-\mu_j} = (\bar{y}_{i.} - \bar{y}_{j.}) \pm t_{36}(\tfrac{0.025}{6}) \cdot 0.4472 \cdot s_{\mathrm{p}}.$$

Explain this formula and using it check which of the pairs of treatments have significantly different means.

Problem 3

Consider the likelihood ratio test for the null hypothesis of the one-way layout under the normality assumption, and show that it is equivalent to the F-test.

Problem 4

Suppose in a one-way layout there are 10 treatments and seven observations under each treatment. What is the ratio of the length of a simultaneous confidence interval for the difference of two means formed by Tukey's method to that of one formed by the Bonferroni method? How do both of these compare in length to an interval based on the t-distribution that does not take account of multiple comparisons?

Problem 5

During each of four experiments on the use of carbon tetrachloride as a worm killer, ten rats were infested with larvae (Armitage 1983). Eight days later, five rats were treated with carbon tetrachloride; the other five were kept as controls. After two more days, all the rats were killed and the numbers of worms were counted. The table below gives the counts of worms for the four control groups.

Group I	Group II	Group III	Group IV
279	378	172	381
338	275	335	346
334	412	335	340
198	265	282	471
303	286	250	318

Significant differences among the control groups, although not expected, might be attributable to changes in the experimental conditions. A finding of significant differences could result in more carefully controlled experimentation and thus greater precision in later work.

Use both graphical techniques and the F-test to test whether there are significant differences among the four groups. Use a nonparametric technique as well.

Problem 6

The concentrations (in nanogram per millimiter) of plasma epinephrine were measured for 10 dogs under isofluorane, halothane, and cyclopropane anesthesia. The measurements are given in the following table (Perry et al. 1974).

	Dog 1	Dog 2	Dog 3	Dog 4	Dog 5	Dog 6	Dog 7	Dog 8	Dog 9	Dog 10
Isofluorane	0.28	0.51	1.00	0.39	0.29	0.36	0.32	0.69	0.17	0.33
Halothane	0.30	0.39	0.63	0.68	0.38	0.21	0.88	0.39	0.51	0.32
Cyclopropane	1.07	1.35	0.69	0.28	1.24	1.53	0.49	0.56	1.02	0.30

Is there a difference in treatment effects? Use a parametric and a nonparametric analysis.

Problem 7

The following table gives the survival times (in hours) for animals in an experiment whose design consisted of three poisons, four treatments, and four observations per cell.

	Treatment A	Treatment B	Treatment C	Treatment D
Poison I	3.1 4.5	8.2 11.0	4.3 4.5	4.5 7.1
	4.6 4.3	8.8 7.2	6.3 7.6	6.6 6.2
Poison II	3.6 2.9	9.2 6.1	4.4 3.5	5.6 10.0
	4.0 2.3	4.9 12.4	3.1 4.0	7.1 3.8
Poison III	2.2 2.1	3.0 3.7	2.3 2.5	3.0 3.6
	1.8 2.3	3.8 2.9	2.4 2.2	3.1 3.3

(a) Conduct a two-way analysis of variance to test the effects of the two main factors and their interaction.

(b) Box and Cox (1964) analysed the reciprocals of the data, pointing out that the reciprocal of a survival time can be interpreted as the rate of death. Conduct a two-way analysis of variance, and compare to the results of part (a). Comment on how well the standard two-way ANOVA model fits and on the interaction in both analyses.

Problem 8

Officials of a small transit system with only five buses want to evaluate four types of tires with respect to wear. Applying a randomized block design, they decided to put one tire of each type on each of the five buses. The tires are run for 15,000 miles, after which the tread wear, in millimeters, is measured.

Bus	Tire 1	Tire 2	Tire 3	Tire 4	Mean
1	9.1	17.1	20.8	11.8	14.7
2	13.4	20.3	28.3	16.0	19.5
3	15.6	24.6	23.7	16.2	20.0
4	11.0	18.2	21.4	14.1	16.2
5	12.7	19.8	25.1	15.8	18.4
Mean	12.4	20.0	23.9	14.8	17.8

(a) State the most appropriate null hypothesis by referring to a suitable parametric model. What are the main assumptions of the parametric model?

(b) Using a non-parametric procedure test the null hypothesis of no difference between the four types of tires.

(c) What kind of external effects are controlled by the suggested randomised block design? How the wheel positions for different tire types should be assigned for each of the five buses?

Problem 9

The accompanying data resulted from an experiment carried out to investigate whether yield from a certain chemical process depended either on the formulation of a particular input or on mixer speed.

		Speed			
		60	70	80	Means
	1	189.7	185.1	189.0	
	1	188.6	179.4	193.0	187.03
	1	190.1	177.3	191.1	
Formulation					
	2	165.1	161.7	163.3	
	2	165.9	159.8	166.6	164.66
	2	167.6	161.6	170.3	
	Means	177.83	170.82	178.88	175.84

A statistical computer package gave

$$\text{ss}_{\text{A}} = 2253.44, \quad \text{ss}_{\text{B}} = 230.81, \quad \text{ss}_{\text{AB}} = 18.58, \quad \text{ss}_{\text{E}} = 71.87.$$

(a) Calculate estimates of the main effects.

(b) Does there appear to be interaction between Formulation and Speed? In which various ways interaction between these two factors could manifest itself? Illustrate with graphs.

(c) Does yield appear to depend either on formulation or speed.

(d) Why is it important to inspect the normal QQ-plot of the 18 residuals?

Problem 10

Three different varieties of tomato (Harvester, Pusa Early Dwarf, and Ife No. 1) and four different plant densities (10, 20, 30, and 40 thousands plants per hectare) are being considered for planting in a particular region. To see whether either variety or plant density affects yield, each combination of variety and plant density is used in three different plots, resulting in the following data on yields:

Variety	Density 10,000	Density 20,000	Density 30,000	Density 40,000	mean
Har	10.5, 9.2, 7.9	12.8, 11.2, 13.3	12.1, 12.6, 14.0	10.8, 9.1, 12.5	11.33
Ife	8.1, 8.6, 10.1	12.7, 13.7, 11.5	14.4, 15.4, 13.7	11.3, 12.5, 14.5	12.21
PED	16.1, 15.3, 17.5	16.6, 19.2, 18.5	20.8, 18.0, 21.0	18.4, 18.9, 17.2	18.13
mean	11.48	14.39	15.78	13.91	13.89

(a) Fill in the ANOVA table for the missing numbers

Source of variation	ss	df	ms	f
Varieties				
Density				
Interaction	8.03			
Errors	38.04			

(b) Clearly state the three pairs of hypotheses of interest. Test them using the normal theory approach.

(c) Estimate the noise size σ.

Problem 11

A population with mean μ consists of three subpopulations with means μ_1, μ_2, μ_3 and the same variance σ^2. Three independent random samples, each of size $n = 13$, from the three subpopulation distributions gave the following sample means and standard deviations:

	Sample 1	Sample 2	Sample 3
Mean	6.3	5.6	6.0
SD	2.14	2.47	3.27

(a) Compute a stratified sample mean, assuming that the three subpopulation sizes have the ratios $N_1 : N_2 : N_3 = 0.3 : 0.2 : 0.5$. Prove that this is an unbiased estimate for the population mean μ.

(b) Assume that all three subpopulation distributions are normal. Write down a simultaneous confidence interval for the differences $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$.

(c) Would you reject the null hypothesis of equality $\mu_1 = \mu_2 = \mu_3$ in this case?

Problem 12

In an experimental study two volunteer male subjects aged 23 and 25 underwent three treatments to compare a new drug against no drug and placebo. Each volunteer had one treatment per day and the time order of these three treatments was randomised.

(a) Comment on the details of the experimental design.

(b) Find the exact null distribution for the test statistic of an appropriate non-parametric test.

Chapter 9

Categorical data analysis

Categorical data appear in the form of a contingency table containing the sample counts for several, say k, mutually exclusive categories. The corresponding population distribution can be modelled as $\mathrm{Mn}(1, \pi_1, \ldots, \pi_k)$, an extension of the Bernoulli distribution with k possible outcomes having probabilities $(\pi_1, \ldots, \pi_k)$ such that

$$\pi_1 + \ldots + \pi_k = 1.$$

In this case, the sample counts for a random sample of size n have the multinomial distribution

$$(C_1, \ldots, C_k) \sim \mathrm{Mn}(n, \pi_1, \ldots, \pi_k).$$

Consider a cross-classification for a pair of categorical factors A and B. If factor A has I levels and factor B has J levels, then the population distribution of a single cross classification event has the form

	b_1	b_2	$\ldots$	b_J	total
a_1	π_{11}	π_{12}	$\ldots$	π_{1J}	$\pi_{1\cdot}$
a_2	π_{21}	π_{22}	$\ldots$	π_{2J}	$\pi_{2\cdot}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_I	π_{I1}	π_{I2}	$\ldots$	π_{IJ}	$\pi_{I\cdot}$
total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	$\ldots$	$\pi_{\cdot J}$	1

Here

$$\pi_{ij} = \mathrm{P}(A = a_i, B = b_j)$$

are the joint probabilities, and

$$\pi_{i\cdot} = \mathrm{P}(A = a_i), \qquad \pi_{\cdot j} = \mathrm{P}(B = b_j)$$

are the marginal probabilities. The null hypothesis of independence claims that there is no relationship between factors A and B

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \text{for all pairs } (i, j).$$

The conditional probabilities

$$\pi_{i|j} = \mathrm{P}(A = a_i | B = b_j) = \frac{\pi_{ij}}{\pi_{\cdot j}}$$

are summarised in the next table

	b_1	b_2	$\ldots$	b_J
a_1	$\pi_{1\|1}$	$\pi_{1\|2}$	$\ldots$	$\pi_{1\|J}$
a_2	$\pi_{2\|1}$	$\pi_{2\|2}$	$\ldots$	$\pi_{2\|J}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_I	$\pi_{I\|1}$	$\pi_{I\|2}$	$\ldots$	$\pi_{I\|J}$
total	1	1	$\ldots$	1

The null hypothesis of homogeneity states the equality of J population distributions

$$H_0 : \pi_{i|j} = \pi_{i\cdot} \quad \text{for all pairs } (i, j).$$

9.1 Chi-squared test of homogeneity

Consider a table of $I \times J$ observed counts obtained from J independent random samples

	sample 1	sample 2	$\ldots$	sample J	total counts
category a_1	c_{11}	c_{12}	$\ldots$	c_{1J}	c_1
category a_2	c_{21}	c_{22}	$\ldots$	c_{2J}	c_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
category a_I	c_{I1}	c_{I2}	$\ldots$	c_{IJ}	c_I
sample sizes	n_1	n_2	$\ldots$	n_J	n

This model is described by J multinomial distributions

$$(C_{1j},\dots,C_{Ij})\sim \mathrm{Mn}(n_j;\pi_{1|j},\dots,\pi_{I|j}),\quad j=1,\dots,J.$$

Notice that the total number of degrees of freedom for J independent samples from I-dimensional multinomial distributions is $J(I-1)$.

Under the hypothesis of homogeneity

$$H_0:\pi_{i|j}=\pi_i \text{ for all } (i,j),$$

the maximum likelihood estimates of π_i are the pooled sample proportions

$$\hat\pi_i=c_i/n,\quad i=1,\dots,I.$$

These estimates consumes $(I-1)$ degrees of freedom, since their sum is 1. Using these maximum likelihood estimates we compute the expected cell counts

$$e_{ij}=n_j\cdot\hat\pi_i=c_in_j/n$$

and the chi-squared test statistic takes the form

$$\mathrm{x}^2=\sum_{i=1}^I\sum_{j=1}^J\frac{(c_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i=1}^I\sum_{j=1}^J\frac{(c_{ij}-c_in_j/n)^2}{c_in_j/n}.$$

We reject H_0 for larger values of x^2 using the approximate null distribution $\mathrm{X}^2\approx\chi^2_{\mathrm{df}}$ with

$$\mathrm{df}\ =J(I-1)-(I-1)=(I-1)(J-1).$$

Example: small cars and personality

A car company studies how customers' attitude toward small cars relates to different personality types. The next table summarises the observed counts:

	cautious	middle-of-the-road	explorer	total
favourable	79	58	49	186
neutral	10	8	9	27
unfavourable	10	34	42	86
total	99	100	100	299

After putting in the brackets the corresponding expected counts

	cautious	middle-of-the-road	explorer	total
favourable	79(61.6)	58(62.2)	49(62.2)	186
neutral	10(8.9)	8(9.0)	9(9.0)	27
unfavourable	10(28.5)	34(28.8)	42(28.8)	86
total	99	100	100	299

we find the chi-squared test statistic to be

$$\mathrm{x}^2=27.24\ \text{ with df}=(3-1)\cdot(3-1)=4.$$

Since 27.24 is larger than the table value $x_4(0.005)=14.86$, we reject the hypothesis of homogeneity at 0.5% significance level. Conclusion: the persons who saw themselves as cautious are more likely to express a favourable opinion of small cars.

9.2 Chi-squared test of independence

Suppose we have single random sample of size n and the sample counts for a cross-classification are summarised in the following way.

	b_1	b_2	$\dots$	b_J	total
a_1	c_{11}	c_{12}	$\dots$	c_{1J}	c_1
a_2	c_{21}	c_{22}	$\dots$	c_{2J}	c_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
a_I	c_{I1}	c_{I2}	$\dots$	c_{IJ}	c_I
total	n_1	n_2	$\dots$	n_J	n

In contrast to the previous setting of J independent samples, the total counts $(n_1, n_2, \dots, n_J)$ are random outcomes cross-classification of the single sample of size n. This data is modelled by the multinomial joint distribution

$$(C_{11}, \dots, C_{IJ}) \sim \text{ Mn}(n_{\cdot\cdot}; \pi_{11}, \dots, \pi_{IJ}),$$

characterised by $IJ - 1$ degrees of freedom. Using such data we would like to test the null hypothesis of independence

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \text{for all pairs } (i, j).$$

The maximum likelihood estimates of $\pi_{i\cdot}$ and $\pi_{\cdot j}$ are

$$\hat{\pi}_{i\cdot} = \frac{c_i}{n}, \qquad \hat{\pi}_{\cdot j} = \frac{n_j}{n}.$$

Under the null hypothesis of independence the expected cell counts

$$e_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{c_i n_j}{n}$$

are exactly the same as for the homogeneity test, with the same number of degrees of freedom

$$\text{df} = (IJ - 1) - (I - 1 + J - 1) = (I - 1)(J - 1).$$

The chi-squared tests of homogeneity and independence have the same test rejection rule.

Example: marital status and educational level

A sample is drawn from a population of married women. Each observation is placed in a 2 × 2 contingency table depending on woman's educational level and her marital status.

	Married only once	Married more than once	Total
College	550 (523.8)	61(87.2)	611
No college	681(707.2)	144(117.8)	825
Total	1231	205	1436

We test H_0: no relationship between the marital status and the education level, by applying the chi-squared test of independence. Using the expected counts given in the brackets, we find the chi-squared test statistic to be $\mathrm{x}^2 = 16.01$. With df = 1 we can use the normal distribution table, since $Z \sim \text{N}(0, 1)$ is equivalent to $Z^2 \sim \chi^2_1$. Thus

$$\text{P}(\text{X}^2 > 16.01|H_0) \approx \text{P}(|Z| > 4.001) = 2(1 - \Phi(4.001)).$$

It follows that the p-value of the test is less that 0.1%, and we reject the null hypothesis of independence. Conclusion: the college educated women, once they marry, are less likely to divorce.

9.3 Matched-pairs designs

We start with an illuminating example concerning Hodgkin disease which has very low incidence of 2 in 10 000.

Case study: Hodgkin's disease and tonsillectomy

To test a possible influence of tonsillectomy on the onset of Hodgkin's disease, researchers use cross-classification data of the form

	A	$\bar{\text{A}}$	total
E	c_{11}	c_{12}	c_1
$\bar{\text{E}}$	c_{21}	c_{22}	c_2
total	n_1	n_2	n

where the four counts distinguish among sampled individuals who are

either affected or unaffected (A or $\bar{\text{A}}$),
and either exposed to tonsillectomy or non-exposed (E or $\bar{\text{E}}$).

There are three possible sampling designs to choose:

- a single random sample would give counts like $\begin{pmatrix} 0 & 0 \\ 0 & n \end{pmatrix}$, as the majority of people are unaffected and non-exposed,

- a prospective study: take an E-sample of size c_1 and a control Ē-sample of size c_2, then watch who gets affected, would give $\begin{pmatrix} 0 & c_1 \\ 0 & c_2 \end{pmatrix}$, because of the very low incidence of the Hodgkin disease,
- a retrospective study: take an A-sample of size n_1 and a control Ā-sample of size n_2, then find out who of them in the past had been exposed to tonsillectomy, would give informative counts.

Two retrospective case-control studies produced opposite results. Study (a) by Vianna, Greenwald, and Davis (1971) gave the following counts.

	A	Ā	total
E	67	43	110
Ē	34	64	98
total	101	107	208

The chi-squared test of homogeneity was applied, which resulted in the test statistic $\mathrm{x}_a^2 = 14.29$. With df = 1, the p-value was found to be very small

$$\mathrm{P}(\mathrm{X}_a^2 \geq 14.29|H_0) \approx 2(1 - \Phi(\sqrt{14.29})) = 0.0002.$$

Study (b) by Johnson and Johnson (1972) was summarised by the table

	A	Ā	total
E	41	33	74
Ē	44	52	96
total	85	85	170

and the chi-squared tests of homogeneity was applied. With $\mathrm{x}_b^2 = 1.53$ and df = 1, the p-value was found to be strikingly different

$$\mathrm{P}(\mathrm{X}_b^2 \geq 1.53|H_0) \approx 2(1 - \Phi(\sqrt{1.53})) = 0.215.$$

It turned out that the study (b) was based on a matched-pairs design violating the assumptions of the chi-squared test of homogeneity. The sample consisted of 85 sibling pairs having same sex and close age: one of the siblings was affected the other not. A proper summary of the study (b) sample distinguishes among four groups of sibling pairs:

$$(\mathrm{AE}, \bar{\mathrm{A}}\mathrm{E}), (\mathrm{AE}, \bar{\mathrm{A}}\bar{\mathrm{E}}), (\mathrm{A}\bar{\mathrm{E}}, \bar{\mathrm{A}}\mathrm{E}), (\mathrm{A}\bar{\mathrm{E}}, \bar{\mathrm{A}}\bar{\mathrm{E}}).$$

The next table lists the corresponding sample counts.

	ĀE	ĀĒ	total
AE	26	15	41
AĒ	7	37	44
total	33	52	85

Notice that this contingency table contains the information of the previous one in terms of the marginal counts.

An appropriate test in this setting is McNemar's test introduced in Section 7.4. For the data of study (b), the McNemar test statistic is

$$\mathrm{x}^2 = \frac{(7 - 15)^2}{7 + 15} = 2.91,$$

giving the p-value of

$$\mathrm{P}(\mathrm{X}^2 \geq 2.91|H_0) \approx 2(1 - \Phi(\sqrt{2.91})) = 0.09.$$

The correct p-value of 0.09 is much smaller than that of 0.215 computed using the test of homogeneity. Since there are only 7 + 15 = 22 informative observations, more data is required.

9.4 Odds ratios

The probability of a random event A is a number $\mathrm{P}(A)$ between 0 and 1. The odds of the event A are defined as

$$\mathrm{odds}(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(A^c)} = \frac{\mathrm{P}(A)}{1 - \mathrm{P}(A)},$$

taking non-negative values. Clearly,

$$\mathrm{P}(A) = \frac{\mathrm{odds}(A)}{1 + \mathrm{odds}(A)},$$

and for small $\mathrm{P}(A)$, we have $\mathrm{odds}(A)$ and $\mathrm{P}(A)$ are close to each other. Conditional odds for A given B are defined similarly,

$$\mathrm{odds}(A|B) = \frac{\mathrm{P}(A|B)}{\mathrm{P}(A^c|B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A^cB)}.$$

Defininng the odds ratio for a pair of events (A, B) by

$$\Delta_{AB} = \frac{\text{odds}(A|B)}{\text{odds}(A|B^c)},$$

we find that

$$\Delta_{AB} = \frac{\text{P}(AB)\text{P}(A^cB^c)}{\text{P}(A^cB)\text{P}(AB^c)},$$

implying

$$\Delta_{AB} = \Delta_{BA}, \quad \Delta_{AB^c} = \frac{1}{\Delta_{AB}}.$$

The odds ratio is a measure of dependence between a pair of random events. It has the following properties

if $\Delta_{AB} = 1$, then events A and B are independent,
if $\Delta_{AB} > 1$, then $\text{P}(A|B) > \text{P}(A|B^c)$ so that B increases the probability of A,
if $\Delta_{AB} < 1$, then $\text{P}(A|B) < \text{P}(A|B^c)$ so that B decreases the probability of A.

Odds ratios for case-control studies

To illustrate the use of odds ratios in statistical inference, return to the case-control study by Vianna, Greenwald, and Davis (1971) on the Hodgkin disease. Based on the counts for two independent samples

	A	Ā	total
E	67	43	110
Ē	34	64	98
total	101	107	208

the chi-squared test of homogeneity rejected the null hypothesis of no relationship between tonsillectomy and the disease. However, it would be useful to have a measure of the strength of this relationship. Such a measure is provided by the odds ratio

$$\Delta_{\text{AE}} = \frac{\text{odds(A|E)}}{\text{odds(A}|\bar{\text{E}})} = \frac{\text{odds(E|A)}}{\text{odds(E}|\bar{\text{A}})}$$

if we were able to estimate it from the data. As we show next, we can estimate the odds ratio by the ratio of the products of the counts on two diagonals

$$\widehat{\Delta}_{\text{AE}} = \frac{67 \cdot 64}{34 \cdot 43} = 2.93,$$

and conclude that tonsillectomy increases the odds for the onset of Hodgkin's disease by factor 2.93.

To this end, we refer to two population distributions underlying the two retrospective samples

	A	Ā
E	P(E|A)	P(E|Ā)
Ē	P(Ē|A)	P(Ē|Ā)
total	1	1

estimated by

	A	Ā
E	67/101	43/107
Ē	34/101	64/107
total	1	1

Thus the odds ratio

$$\Delta_{\text{AE}} = \frac{\text{P(E|A)}}{\text{P}(\bar{\text{E}}|\text{A})} \cdot \frac{\text{P}(\bar{\text{E}}|\bar{\text{A}})}{\text{P}(\text{E}|\bar{\text{A}})}$$

is consistently estimated by

$$\widehat{\Delta}_{\text{AE}} = \frac{67/101}{34/101} \cdot \frac{64/107}{43/107} = \frac{67 \cdot 64}{34 \cdot 43} = 2.93.$$

9.5 Exercises

Problem 1

Adult-onset diabetes is known to be highly genetically determined. A study was done comparing frequencies of a particular allele in a sample of such diabetics and a sample of non-diabetics.

	diabetic	normal	total
Bb or *bb*	12	4	16
BB	39	49	88
total	51	53	104

Is there a relationship between the allele frequencies and the adult-onset diabetes?

Problem 2

Overfield and Klauber (1980) published the following data on the incidence of tuberculosis in relation to blood groups in a sample of Eskimos. Is there any association of the disease and blood group within the ABO system or within the MN system?

	O	A	AB	B
moderate	7	5	3	13
minimal	27	32	8	18
not present	55	50	7	24

	MM	MN	NN
moderate	21	6	1
minimal	54	27	5
not present	74	51	11

Problem 3

It is conventional wisdom in military squadron that pilots tend to father more girls than boys. Snyder (1961) gathered data for military fighter pilots. The sex of the pilots' offspring were tabulated for three kinds of flight duty during the month of conception, as shown in the following table.

	girl	boy
flying fighter	51	38
flying transport	14	16
not flying	38	46

(a) Is there any significant difference between the three groups?

(b) In the United States in 1950, 105.37 males were born for every 100 females. Are the data consistent with this sex ratio?

Problem 4

A randomised double-blind experiment compared the effectiveness of several drugs in ameliorating postoperative nausea. All patients were anesthetized with nitrous oxide and ether. The following table shows the incidence of nausea during the first four hours for each of several drugs and a placebo (Beecher 1959).

	number of patients	incidence of nausea
placebo	165	95
chlorpromazine	152	52
dimenhydrinate	85	52
pentobarbital (100 mg)	67	35
pentobarbital (150 mg)	85	37

Compare the drugs to each other and to the placebo.

Problem 5

In a study of the relation of blood type to various diseases, the following data were gathered in London and Manchester (Woolf 1955).

London	Control	Peptic Ulcer
Group A	4219	579
Group O	4578	911

Manchester	Control	Peptic Ulcer
Group A	3775	246
Group O	4532	361

First, consider the two tables separately. Is there a relationship between blood type and propensity to peptic ulcer? If so, evaluate the strength of the relationship. Are the data from London and Manchester comparable?

Problem 6

Record of 317 patients at least 48 years old who were diagnosed as having endometrial carcinoma were obtained from two hospitals (Smith et al. 1975). Matched controls for each case were obtained from the two institutions: the controls had cervical cancer, ovarian cancer, or carcinoma of the vulva. Each control was matched by age at diagnosis (within four years) and year of diagnosis (within two years) to a corresponding case of endometrial carcinoma.

The following table gives the number of cases and controls who had taken estrogen for at least 6 months prior to the diagnosis of cancer.

	Controls: estrogen used	Controls: estrogen not used	Total
Cases: estrogen used	39	113	152
Cases: estrogen not used	15	150	165
Total	54	263	317

(a) Is there a significant relationship between estrogen use and endometrial cancer?

(b) This sort of design, called a retrospective case-control study, is frequently used in medical investigations where a randomised experiment is not possible. Do you see any possible weak points in a retrospective case-control design?

Problem 7

A psychological experiment was done to investigate the effect of anxiety on a person's desire to be alone or in company (Lehman 1975). A group of 30 subjects was randomly divided into two groups of sizes 13 and 17. The subjects were told that they would be subjected to some electric shocks, but one group (high-anxiety) was told that the shocks would be quite painful and the other group (low-anxiety) was told that they would be mild and painless. Both groups were told that there would be a 10-min wait before the experiment began, and each subject was given the choice of waiting alone or with the other subjects. The following are the results:

	Wait Together	Wait Alone	Total
High-Anxiety	12	5	17
Low-Anxiety	4	9	13
Total	16	14	30

Use Fisher's exact test to test whether there is a significant difference between the high- and low-anxiety groups. What is a reasonable one-sided alternative, if any?

Problem 8

Hill and Barton (2005) asked the question: red against blue outfits - does it matter in combat sports? Although other colors are also present in animal displays, it is specifically the presence and intensity of red coloration that correlates with male dominance and testosterone levels. Increased redness during aggressive interactions may reflect relative dominance.

In the 2004 Olympic Games, contestants in four combat sports were randomly assigned red and blue outfits. The winner counts in different sports

	Red	Biue	Total
Boxing	148	120	268
Freestyle wrestling	27	24	51
Greco-Roman wrestling	25	23	48
Tae Kwon Do	45	35	80
Total	245	202	447

Is there any evidence that wearing red is more favourable in some of the sports than others?

Problem 9

Suppose that a company wishes to examine the relationship of gender to job satisfaction, grouping job satisfaction into four categories: very satisfied, somewhat satisfied, somewhat dissatisfied, and very dissatisfied. The company plans to ask the opinion of 100 employees. Should you, the company's statistician, carry out a chi-square test of independence or a test of homogeneity?

Problem 10

Questions concerning hypotheses testing methodology. Try to give detailed answers.

(a) Consider a hypothetical study of the effects of birth control pills. In such a case, it would be impossible to assign women to a treatment or a placebo at random. However, a non-randomized study might be conducted by carefully matching control to treatments on such factors as age and medical history.

The two groups might be followed up on for some time, with several variables being recorded for each subject such as blood pressure, psychological measures, and incidences of various problems. After termination of the study, the two groups might be compared on each of these many variables, and it might be found, say, that there was a "significant difference" in the incidence of melanoma.

What is a common problem with such "significant findings"?

(b) You analyse cross-classification data summarized in a two by two contingency table. You wanted to apply the chi-square test but it showed that one of the expected counts was below 5. What alternative statistical test you may try applying?

(c) Why tests like rank sum test, Friedman test, and Kruskal-Wallis tests are often called distribution-free tests?

Problem 11

A public policy polling group is investigating whether people living in the same household tend to make independent political choices. They select 200 homes where exactly three voters live. The residents are asked separately for their opinion ("yes" or "no") on a city charter amendment. The results of the survey are summarized in the table:

Number of saying "yes"	0	1	2	3
Frequency	30	56	73	41

Based on these data can we claim that opinions are formed independently?

Problem 12

Verify that the hypothesis of homogeneity

$$H_0 : \pi_{i|j} = \pi_{i\cdot} \quad \text{for all pairs } (i,j),$$

is mathematically equivalent to the hypothesis of independence

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \text{for all pairs } (i,j).$$

Problem 13

A study of the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline is based on $n = 441$ stations.

	Pricing policy			
	Aggressive	Neutral	Nonaggressive	Total
Substandard condition	24	15	17	56
Standard condition	52	73	80	205
Modern condition	58	86	36	180
Total	134	174	133	441

(a) Suggest a parametric model for the data and write down the corresponding likelihood function.

(b) What is a relevant null hypothesis for the data?

(c) Properly analyse the data and draw your conclusions.

Problem 14

In a controlled clinical trial which began in 1982 and ended in 1987, more than 22000 physicians participated. The participants were randomly assigned in two groups: Aspirin and Placebo. The aspirin group have been taking 325 mg aspirin every second day. At the end of trial, the number of participants who suffered from myocardial infarctions was assessed.

	MyoInf	No MyoInf	Total
Aspirin	104	10933	11037
Placebo	189	10845	11034

The popular measure in assessing the results in clinical trials is the Risk Ratio

$$RR = \frac{\text{Risk of Aspirin}}{\text{Risk of Placebo}} = \frac{104/11037}{189/11034} = 0.55.$$

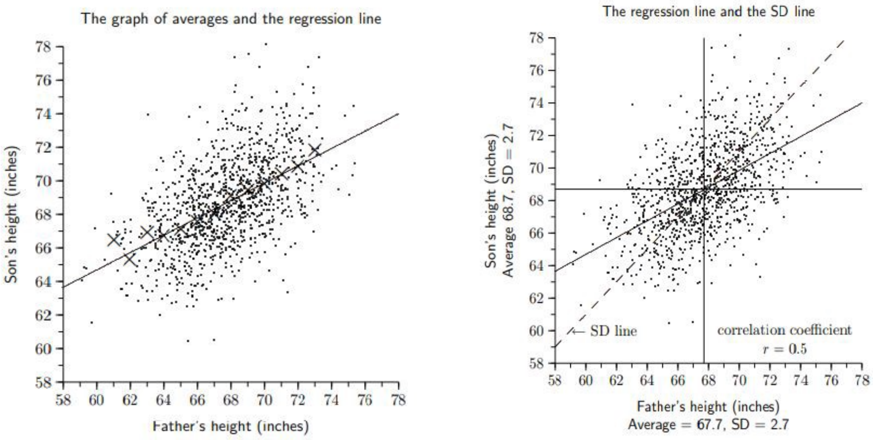
(a) How would you interpret the obtained value of the risk ratio? What ratio of conditional probabilities is estimated by RR?

(b) Is the observed value of RR significantly different from 1?

Chapter 10

Multiple regression

Karl Pearson collected the heights of 1,078 fathers and their full-grown sons in England, circa 1900. The following scatter diagram illustrates the data, with one dot for each father-son pair. The straight line on the left panel, is the so called regression line which indicates the average son-height for the various father-heights. Using this line one can predict the average son-height for a given father-height. In this sense the father-height x serves as a predictor variable for the response variable y, the son-height.



The right panel adds several important insights on this dataset. It shows that the regression line goes through the middle point of the scatter plot $(\bar{x}, \bar{y})$, where $\bar{x}$ is the sample mean of the father-heights and $\bar{y}$ is the sample mean of the son-heights. The cross going through the central point $(\bar{x}, \bar{y})$ divides the scatter plot in four quadrants. We see that the majority of the dots lies in the first and the third quadrants reflecting the fact that the heights of the father and the son are positively correlated.

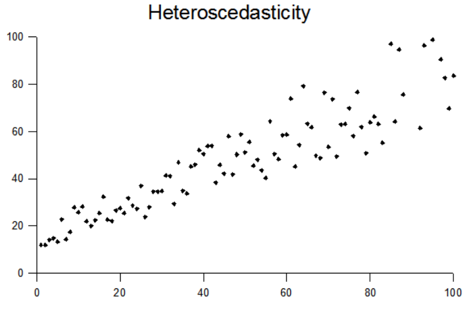
Consider now only the dots to the right of the vertical line, that is when the fathers are taller than average. For this part of the dots, we see that on average the son is shorter than his father. To the left of the vertical line the relation is opposite: on average the son is taller than his father. Francis Galton called this remarkable phenomenon *regression to mediocrity.*

10.1 Simple linear regression model

According to the simple linear regression model the response

$$Y = \beta_0 + \beta_1 x + \sigma Z, \qquad Z \sim \mathrm{N}(0, 1),$$

is a linear function of the predictor variable x plus a normally distributed noise of size σ. The key assumption of *homoscedasticity* requires that the noise size is independent of the x-value. Whenever this assumption is violated, the situation is described by the term *heteroscedasticity.*



Under this model, the data consist of n pairs

$$(x_1, y_1), \dots, (x_n, y_n),$$

where

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, n,$$

involves a vector $(e_1, \dots, e_n)$ of independent realisations of a random variable with distribution $\mathrm{N}(0, \sigma)$.

Maximum likelihood estimates

The corresponding likelihood is a function of the three-dimensional parameter $\theta = (\beta_0, \beta_1, \sigma^2)$

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \Big\} = C\sigma^{-n} \exp\Big\{ - \frac{1}{2\sigma^2} \sum_{i=1}^{n} e_i^2 \Big\},$$

where $C = (2\pi)^{-n/2}$. This implies the following expression for the log-likelihood function $l(\theta) = \ln L(\theta)$

$$l(\theta) = \ln C - (n/2) \ln \sigma^2 - \frac{\sum_{i=1}^{n} e_i^2}{2\sigma^2}.$$

Observe that

$$e_i = y_i - \beta_0 - \beta_1 x_i,$$

implying

$$n^{-1} \sum_{i=1}^{n} e_i^2 = \beta_0^2 + 2\beta_0\beta_1\bar{x} - 2\beta_0\bar{y} - 2\beta_1\overline{xy} + \beta_1^2\overline{x^2} + \overline{y^2},$$

where the five summary statistics

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad \bar{y} = \frac{y_1 + \ldots + y_n}{n}, \quad \overline{x^2} = \frac{x_1^2 + \ldots + x_n^2}{n}, \quad \overline{y^2} = \frac{y_1^2 + \ldots + y_n^2}{n}, \quad \overline{xy} = \frac{x_1y_1 + \ldots + x_ny_n}{n}$$

form the set of sufficient statistics.

To obtain the maximum likelihood estimates $\hat{\theta} = (b_0, b_1, \hat{\sigma}^2)$ of $\theta = (\beta_0, \beta_1, \sigma^2)$ compute the derivatives

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0} &= -\frac{n}{\sigma^2}(\beta_0 + \beta_1\bar{x} - \bar{y}), \\
\frac{\partial l}{\partial \beta_1} &= -\frac{n}{\sigma^2}(\beta_0\bar{x} - \overline{xy} + \beta_1\overline{x^2}), \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} e_i^2,
\end{aligned}$$

and set them all equal to zero. As a result we get a system of three equations

$$b_0 + b_1\bar{x} = \bar{y}, \quad b_0\bar{x} + b_1\overline{x^2} = \overline{xy}, \quad \hat{\sigma}^2 = \frac{\mathrm{ss_E}}{n},$$

where

$$\mathrm{ss_E} = \hat{e}_1^2 + \ldots + \hat{e}_n^2,$$

is the error sum of squares defined in terms of the residuals

$$\hat{e}_i = y_i - b_0 - b_1 x_i.$$

Solving the first two equations we obtain the maximum likelihood estimates for the slope and intercept

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad b_0 = \bar{y} - b_1\bar{x}.$$

Notice that the estimates (b_0, b_1) of parameters (β_0, β_1) are obtained by minimising the sum of squares

$$\mathrm{ss_E} = \min_{\beta_0, \beta_1} \Big\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Big\},$$

and for this reason, (b_0, b_1) are called the least squares estimates of (β_0, β_1).

The maximum likelihood estimate of $\hat{\sigma}^2 = \frac{\mathrm{ss_E}}{n}$ is a biased but asymptotically unbiased estimate of σ^2. An unbiased estimate of σ^2 is given by

$$s^2 = \frac{\mathrm{ss_E}}{n-2}.$$

Residuals

For each predictor value x_i define the predicted response by

$$\hat{y}_i = b_0 + b_1 x_i,$$

so that the corresponding residual is given by the difference $\hat{e}_i = y_i - \hat{y}_i$. The residuals $(\hat{e}_1, \ldots, \hat{e}_n)$ are linearly connected via

$$\hat{e}_1 + \ldots + \hat{e}_n = 0, \qquad x_1\hat{e}_1 + \ldots + x_n\hat{e}_n = 0, \qquad \hat{y}_1\hat{e}_1 + \ldots + \hat{y}_n\hat{e}_n = 0.$$

Thus under the simple linear regression model, the scatter plot of the residuals $\hat{e}_i$ versus x_i should look like a horizontal blur. If the linear model is not valid, it will show up in a somewhat bowed shape of the $(x_i, \hat{e}_i)$ scatter plot. In some cases, such a non-linearity problem can be fixed by some kind of a log-transformation of the data.

The residuals $\hat{e}_i$ are realisations of random variables $\hat{E}_i$ having normal distributions with zero means, whose variances and covariances are given by

$$\mathrm{Var}(\hat{E}_i) = \sigma^2\Big(1 - \frac{c_{ii}}{n-1}\Big), \qquad \mathrm{Cov}(\hat{E}_i, \hat{E}_j) = -\sigma^2\frac{c_{ij}}{n-1},$$

where

$$c_{ij} = \frac{\sum_{k=1}^n (x_i - x_k)(x_j - x_k)}{ns_x^2}.$$

For larger n, the random variables $\hat{E}_i$ can be viewed as independent having the same $\mathrm{N}(0, \sigma)$ distribution.

Sample correlation coefficient

Define the sample variances and the sample covariance by

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2, \qquad s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Then the sample correlation coefficient between the predictor and response variables is defined as

$$r = \frac{s_{xy}}{s_x s_y}.$$

Noticing that

$$b_1 = \frac{rs_y}{s_x},$$

we find that the fitted regression line $y = b_0 + b_1 x$ can be written in the form

$$y = \bar{y} + r\tfrac{s_y}{s_x}(x - \bar{x}),$$

or equivalently

$$\Big(\frac{y - \bar{y}}{s_y}\Big) = r\Big(\frac{x - \bar{x}}{s_x}\Big).$$

The last formula is most appealing to the intuition: it claims that the standardised versions of the predictor and the response are connected through the regression factor.

Coefficient of determination

Using the formula for the predicted responses

$$\hat{y}_i = \bar{y} + b_1(x_i - \bar{x}), \quad b_1 = r\tfrac{s_y}{s_x},$$

and the relation

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i,$$

we obtain a decomposition of the sum of squares

$$\mathrm{ss}_\mathrm{T} = \mathrm{ss}_\mathrm{R} + \mathrm{ss}_\mathrm{E},$$

where

$$\mathrm{ss}_\mathrm{T} = \sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$$

is the total sum of squares, and

$$\mathrm{ss}_\mathrm{R} = \sum_i (\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2 = (n-1)r^2 s_y^2$$

is the regression sum of squares. The obtained relations yield

$$\frac{\text{ss}_\text{R}}{\text{ss}_\text{T}} = r^2, \qquad \frac{\text{ss}_\text{E}}{\text{ss}_\text{T}} = 1 - r^2.$$

These equalities explain why the squared sample correlation coefficient r^2 is called the coefficient of determination. Coefficient of determination r^2 is the proportion of variation in the response variable explained by the variation of the predictor variable.

Coefficient of determination r^2 has a more intuitive meaning than the sample correlation coefficient r

On the other hand, we have

$$\text{ss}_\text{E} = \text{ss}_\text{T}(1 - r^2) = (n-1)s_y^2(1 - r^2),$$

implying the following useful expression for the unbiased estimate of σ^2:

$$s^2 = \frac{n-1}{n-2}\, s_y^2(1 - r^2).$$

10.2 Confidence intervals and hypothesis testing

The least squares estimators (b_0, b_1) are unbiased and consistent. Due to the normality assumption we have the following exact distributions for the underlying random variables (B_0, B_1)

$$B_0 \sim \text{N}(\beta_0, \sigma_0), \qquad \sigma_0^2 = \frac{\sigma^2 \sum x_i^2}{n(n-1)s_x^2}, \qquad s_{b_0}^2 = \frac{s^2 \sum x_i^2}{n(n-1)s_x^2}, \qquad \frac{B_0 - \beta_0}{S_{B_0}} \sim t_{n-2},$$

$$B_1 \sim \text{N}(\beta_1, \sigma_1), \qquad \sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}, \qquad s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}, \qquad \frac{B_1 - \beta_1}{S_{B_1}} \sim t_{n-2}.$$

There is a weak correlation between the two estimators:

$$\text{Cov}(B_0, B_1) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2}$$

which is negative, if $\bar{x} > 0$, and positive, if $\bar{x} < 0$.

For β_0 and β_1, we get two exact $100(1-\alpha)\%$ confidence intervals $I_{\beta_i} = b_i \pm t_{n-2}(\frac{\alpha}{2}) \cdot s_{b_i}$

For $i = 0$ or $i = 1$ and a given value β^*, one could be interested in testing the null hypothesis H_0: $\beta_i = \beta^*$. Use the test statistic

$$t = \frac{b_i - \beta^*}{s_{b_i}},$$

which is a realisation of a random variable T that has the exact null distribution

$$T \sim t_{n-2}.$$

Two important examples of hypotheses testing.

1. The model utility test is built around the null hypothesis

   $$H_0 : \beta_1 = 0$$

   stating that there is no relationship between the predictor variable x and the response y. The corresponding test statistic, often called t-value,

   $$t = \frac{b_1}{s_{b_1}}$$

   has t_{n-2} as the null distribution.

2. The zero-intercept test aims at

   $$H_0 : \beta_0 = 0.$$

   Its t-value $t = \frac{b_0}{s_{b_0}}$ has the same null distribution t_{n-2}.

Intervals for individual observations

Suppose a sample of size n resulted in the least squares estimates (b_0, b_1), the noise size estimate s, the sample mean $\bar{x}$ and the sample standard deviation s_x for the predictor values in the sample. We are going to make a new measurement with the predictor value $x = x_p$ and wish to say something on the response value

$$y_p = \beta_0 + \beta_1 x_p + \sigma z_p,$$

where z_p is generated by the $\mathrm{N}(0, 1)$ distribution independently of the available sample $(x_1, y_1), \ldots, (x_n, y_n)$. The expected value of Y_p,

$$\mu_p = \beta_0 + \beta_1 x_p$$

is estimated by

$$\hat{\mu}_p = b_0 + b_1 x_p.$$

The standard error σ_p of $\hat{\mu}_p$ is computed as the square root of

$$\mathrm{Var}(B_0 + B_1 x_p) = \mathrm{Var}(B_0) + x_p^2 \mathrm{Var}(B_1) + 2x_p \mathrm{Cov}(B_0, B_1) = \tfrac{\sigma^2}{n} + \tfrac{\sigma^2}{n-1} \cdot (\tfrac{x_p - \bar{x}}{s_x})^2.$$

This leads to the following exact $100(1-\alpha)\%$ confidence interval for μ_p

$$I_{\mu_p} = b_0 + b_1 x_p \pm t_{n-2}(\tfrac{\alpha}{2}) \cdot s\sqrt{\tfrac{1}{n} + \tfrac{1}{n-1}(\tfrac{x_p - \bar{x}}{s_x})^2}.$$

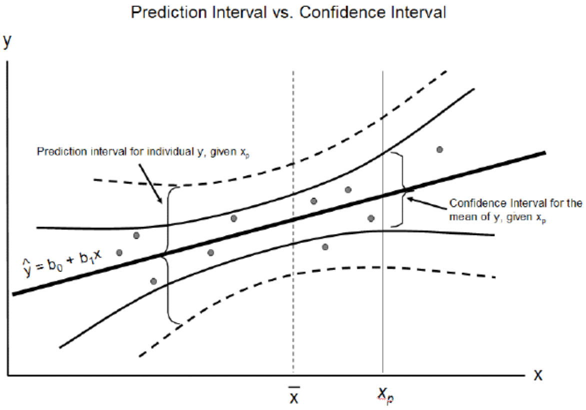
The new feature of this section is the so called prediction interval of the response value y_p with takes into account the noise factors accompanying the new measurement.

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ prediction interval } I_{y_p} = b_0 + b_1 x_p \pm t_{n-2}(\tfrac{\alpha}{2}) \cdot s\sqrt{1 + \tfrac{1}{n} + \tfrac{1}{n-1}(\tfrac{x_p - \bar{x}}{s_x})^2}}$$

This prediction interval has wider limits since it reflects the uncertainty due the noise factors. The larger factor appearing under the square root comes from the variance formula

$$\mathrm{Var}(Y_p - \hat{\mu}_p) = \mathrm{Var}(\mu_p + \sigma Z_p - \hat{\mu}_p) = \sigma^2 + \mathrm{Var}(\hat{\mu}_p) = \sigma^2(1 + \tfrac{1}{n} + \tfrac{1}{n-1} \cdot (\tfrac{x_p - \bar{x}}{s_x})^2).$$

The next figure illustrates the relationship between the 95% confidence and prediction intervals (I_{μ_p}, I_{y_p}). Two features are important to be aware of. Firstly, the prediction interval is broader. Observe that as $n \to \infty$, the width of the confidence interval goes to zero, while the width of the 95% prediction interval converges to 1.96σ. Secondly, both intervals gets wider as the value x_p is placed further from the sample mean $\bar{x}$ of the predictor values.


Prediction Interval vs. Confidence Interval
y
Prediction interval for individual y, given x_p
Confidence Interval for the mean of y, given x_p
$\hat{y} = b_0 + b_1 x$
x
$\bar{x}$
x_p

10.3 Multiple linear regression

An important extension of the simple linear regression model involving a single predictor x is the multiple regression model involving $p-1$ predictors with an arbitrary $p \geq 2$. The corresponding data set consists of n vectors $(x_{i,1}, \ldots, x_{i,p-1}, y_i)$ with $n > p$, such that

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1} x_{1,p-1} + e_1, \\ &\ldots \\ y_n &= \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1} x_{n,p-1} + e_n, \end{aligned}$$

where $e_1, \ldots, e_n$ are independently generated by the $\mathrm{N}(0, \sigma)$ distribution. In terms of the column vectors

$$\mathbf{y} = (y_1, \ldots, y_n)^\intercal, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\intercal, \quad \boldsymbol{e} = (e_1, \ldots, e_n)^\intercal,$$

the multiple regression model is compactly described by the relation

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\mathbb{X}$ is the so called design matrix assumed to have rank p:

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}.$$

The machinery developed for the simple linear regression model works well for the multiple regression. The least squares estimates $\boldsymbol{b} = (b_0, \ldots, b_{p-1})^\intercal$ are obtained as

$$\boldsymbol{b} = (\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal\mathbf{y},$$

and yield the next formula for the predicted responses $\hat{\mathbf{y}} = \mathbb{X}\boldsymbol{b}$:

$$\hat{\boldsymbol{y}} = \mathbb{P}\boldsymbol{y}, \text{ where } \mathbb{P} = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal.$$

Turning to the random vector $\boldsymbol{B}$ behind the the least squares estimates $\boldsymbol{b}$, we find that

$$\mathrm{E}(\boldsymbol{B}) = \boldsymbol{\beta}.$$

Furthermore, the covariance matrix, the $p \times p$ matrix with elements $\mathrm{Cov}(B_i, B_j)$, is given by

$$\mathrm{E}\{(\boldsymbol{B} - \boldsymbol{\beta})(\boldsymbol{B} - \boldsymbol{\beta})^\intercal\} = \sigma^2(\mathbb{X}^\intercal\mathbb{X})^{-1}.$$

The vector of residuals

$$\hat{\boldsymbol{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbb{I} - \mathbb{P})\mathbf{y}$$

has a zero mean vector and a covariance matrix $\sigma^2(\mathbb{I} - \mathbb{P})$.

$$\boxed{\text{An unbiased estimate of } \sigma^2 \text{ is given by } s^2 = \tfrac{\text{ss}_\text{E}}{n-p}, \quad \text{where } \text{ss}_\text{E} = \hat{e}_1^2 + \ldots + \hat{e}_n^2.}$$

Denote by $d_0^2, \ldots, d_{p-1}^2$ the diagonal elements of the matrix $(\mathbb{X}^\intercal\mathbb{X})^{-1}$. Then the standard error of b_j is computed as $s_{b_j} = sd_j$, entailing the exact sampling distributions for the least squares estimates

$$\frac{B_j - \beta_j}{S_{B_j}} \sim t_{n-p}, \quad j = 0, 1, \ldots, p-1.$$

Special case: $p = 2$

For the simple linear regression case, the design matrix $\mathbb{X}$ has the dimensions $n \times 2$, so that

$$\mathbb{X}^\intercal = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix}, \qquad \mathbb{X}^\intercal\mathbb{X} = \begin{pmatrix} n & x_1 + \ldots + x_n \\ x_1 + \ldots + x_n & x_1^2 + \ldots + x_n^2 \end{pmatrix} = n\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}.$$

It follows

$$(\mathbb{X}^\intercal\mathbb{X})^{-1} = \frac{1}{n(\overline{x^2} - (\bar{x})^2)}\begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}, \quad \mathbb{X}^\intercal\boldsymbol{y} = n\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix},$$

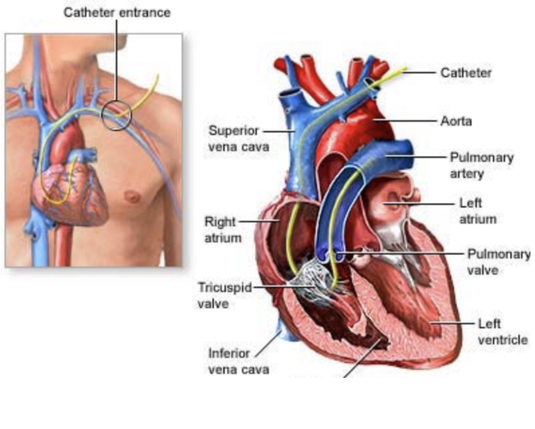
and the formulas for b_0 and b_1 are recovered in the matrix form

$$\boldsymbol{b} = (\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal\boldsymbol{y} = \frac{1}{\overline{x^2} - (\bar{x})^2}\begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}.$$

Case study: catheter length

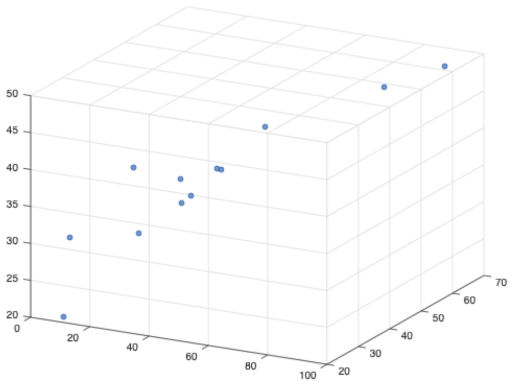
Doctors want predictions on heart catheter length depending on child's height and weight. The data consist of $n = 12$ observations for the distance to pulmonary artery:

Height (in)	Weight (lb)	Length (cm)
42.8	40.0	37.0
63.5	93.5	49.5
37.5	35.5	34.5
39.5	30.0	36.0
45.5	52.0	43.0
38.5	17.0	28.0
43.0	38.5	37.0
22.5	8.5	20.0
37.0	33.0	33.5
23.5	9.5	30.5
33.0	21.0	38.5
58.0	79.0	47.0



For this example the response variable y is the catheter length, and the two predictors are x_1 is the height of a patient and x_2 is the same patient's weight. In this case $p = 3$, with the response vector and the design matrix are given by

$$\boldsymbol{y} = \begin{pmatrix} 20.0 \\ 30.5 \\ 38.5 \\ 33.5 \\ 34.5 \\ 28.0 \\ 36.0 \\ 37.0 \\ 37.0 \\ 43.0 \\ 47.0 \\ 49.5 \end{pmatrix}, \qquad \mathbb{X} = \begin{pmatrix} 1 & 22.5 & 8.5 \\ 1 & 23.5 & 9.5 \\ 1 & 33.0 & 21.0 \\ 1 & 37.0 & 33.0 \\ 1 & 37.5 & 35.5 \\ 1 & 38.5 & 17.0 \\ 1 & 39.5 & 30.0 \\ 1 & 42.8 & 40.0 \\ 1 & 43.0 & 38.5 \\ 1 & 45.5 & 52.0 \\ 1 & 58.0 & 79.0 \\ 1 & 63.5 & 93.5 \end{pmatrix}.$$



Geometrically, the task is to fit a plain

$$y = b_0 + b_1 x_1 + b_2 x_2$$

to the three-dimensional scatter plot for $n = 12$ points. Applying the formulas given above, we compute four point estimates $b_0 = 21$, $b_1 = 0.20$, $b_2 = 0.19$, $s = 3.9$, and three standard errors $s_{b_0} = 8.8$ and $s_{b_1} = 0.36$, $s_{b_2} = 0.17$. Using t-distribution with $n - p = 9$ degrees of freedom, we compute three 95% confidence intervals

$$\begin{aligned} I_{\beta_0} &= 21 \pm 2.2622 \cdot 8.8 = 21 \pm 19.91, \\ I_{\beta_1} &= 0.20 \pm 2.2622 \cdot 0.36 = 0.20 \pm 0.81, \\ I_{\beta_2} &= 0.19 \pm 2.2622 \cdot 0.17 = 0.19 \pm 0.38. \end{aligned}$$

We see that the last two intervals cover zero.

10.4 Coefficients of multiple determination and model utility tests

The coefficient of multiple determination can be computed similarly to the coefficient determination R^2 for the simple linear regression model as

$$R^2 = 1 - \frac{\text{ss}_\text{E}}{\text{ss}_\text{T}}, \qquad \text{ss}_\text{T} = (n-1)s_y^2.$$

The problem with R^2 is that it increases with p even if totally irrelevant predictor variables are added to the model. To punish for irrelevant variables it is better to use the adjusted coefficient of multiple determination

$$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{\text{ss}_\text{E}}{\text{ss}_\text{T}}.$$

The adjustment factor $\frac{n-1}{n-p}$ reduces the coefficient of determination if p gets larger. Representing

$$R_a^2 = 1 - \frac{s^2}{s_y^2},$$

we arrive at the following transparent interpretation of the adjusted coefficient of determination. Since $\frac{s^2}{s_y^2}$ is the proportion of the variance in y explained by the noise, the adjusted R^2 is the proportion of the variance in y explained by the regression model.

Example: flow rate vs stream depth

The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan et al, 1976).

Depth x	0.34	0.29	0.28	0.42	0.29	0.41	0.76	0.73	0.46	0.40
Flow rate y	0.64	0.32	0.73	1.33	0.49	0.92	7.35	5.89	1.98	1.12

A bowed shape of the plot of the residuals versus depth suggests that the relation between x and y is not linear. The multiple linear regression framework can by applied to the quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

with $x_1 = x$ and $x_2 = x^2$. The results of such analysis are presented in the table:

Coefficient	Estimate	Standard Error	t value
β_0	1.68	1.06	1.52
β_1	-10.86	4.52	-2.40
β_2	23.54	4.27	5.51

The residuals show no sign of systematic misfit. The test statistic $t = 5.51$ of the utility test for the crucial null hypothesis $H_0 : \beta_2 = 0$, is larger than the critical value $t_7(0.0005) = 5.4079$. We conclude that the quadratic term in the model is statistically significant at the 0.1% significance level.

Case study: catheter length

Doctors want predictions distance to pulmonary artery (y) depending on child's height (h) and weight (w). The data consist of $n = 12$ of the triplets (h_i, w_i, y_i). We will compare three linear regressions

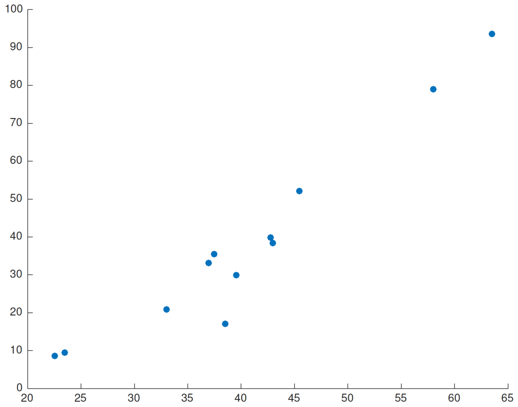
$$\text{model 1: } y = \beta_0 + \beta_1 h + \sigma z, \quad \text{model 2: } y = \beta_0 + \beta_1 w + \sigma z, \quad \text{model 3: } y = \beta_0 + \beta_1 h + \beta_2 w + \sigma z.$$

The analysis of these regression models is summarised in the next table

estimates	model 1	t-value	model 2	t-value	model 3	t-value
$b_0(s_{b_0})$	12.1 (4.3)	2.8	25.6 (2.0)	12.8	21 (8.8)	2.39
$b_1(s_{b_1})$	0.60 (0.10)	6.0	0.28 (0.04)	7.0	0.20 (0.36)	0.56
$b_2(s_{b_2})$	–	–	–	–	0.19 (0.17)	1.12
s	4.0		3.8		3.9.	

In contrast to the two simple linear regression models, for the multiple regression model we can not reject either $H_1 : \beta_1 = 0$ or $H_2 : \beta_2 = 0$ because the t-values 0.59 and 1.12 are not significant according to the t-distribution table with df $= 9$. This paradox is explained by the different meaning of the slope parameters in the simple and multiple regression models. In the multiple model β_1 is the expected change in L when H increased by one unit and W held constant.

The comparison of the coefficient of determination R^2 for these three models, $R^2 = 0.78, 0.80, 0.81$ for the models 1, 2, 3 respectively, leaves the false impression that the multiple regression model 3 is better than the simple linear regressions models 1 and 2. Turning to the adjusted coefficient of determination $R^2_a = 0.76, 0.78, 0.77$, we conclude that the model 2 is the best among the three regressions. This is an example of the collinearity problem: height and weight have a strong linear relationship, see the next figure, therefore, there is little or no gain from adding the predictor h to the regression model 2 with the single predictor w.



10.5 Exercises

Problem 1

Suppose we are given a two-dimensional random sample

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Verify that the sample covariance is an unbiased estimate of the population covariance.

Problem 2

Draw by hand a scatter plot for ten pairs of measurements

x	0.34	1.38	-0.65	0.68	1.40	-0.88	-0.30	-1.18	0.50	-1.75
y	0.27	1.34	-0.53	0.35	1.28	-0.98	-0.72	-0.81	0.64	-1.59

(a) Fit a straight line $y = a + bx$ by the method of least squares, and sketch it on the plot.

(b) Fit a straight line $x = c + dy$ by the method of least squares, and sketch it on the plot.

(c) Why the lines on (a) and (b) are not the same?

Problem 3

Each student receives two consecutive grade point averages

$x =$ the high school GPA and $y =$ the freshman GPA.

Starting from two coupled models for female students

$$y = \beta_0 + \beta_1 x + \sigma z, \quad Z \sim \mathrm{N}(0, \sigma),$$

and for male students

$$y = \beta_0'' + \beta_1 x + \sigma z, \quad Z \sim \mathrm{N}(0, \sigma),$$

suggest a joint multiple regression model involving an extra predictor f which equal 1 if the student is female and 0 otherwise. Give the form of the design matrix for such a model.

Problem 4

Check that $\mathbb{P} = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal$ is a projection matrix such that $\mathbb{P}^2 = \mathbb{P}$.

Problem 5

The sample $(x_1, y_1), \ldots, (x_n, y_n)$ was collected for n students who took a statistical course, with

x giving the midterm grade and y giving the final grade.

The data produced the following five summary statistics

$$r = 0.5, \quad \bar{x} = \bar{y} = 75, \quad s_x = s_y = 10.$$

(a) Given a student's midterm score $x = 95$, predict the final score of the student.

(b) Given the final score $y = 85$ and not knowing the midterm score of a student, predict the student's midterm score.

Problem 6

Let $X \sim \mathrm{N}(0, 1)$ and $Z \sim \mathrm{N}(0, 1)$ be two independent random variables and consider a third one

$$Y = X + \beta Z.$$

(a) Show that the correlation coefficient for X and Y is

$$\rho = \tfrac{1}{\sqrt{1+\beta^2}}.$$

(b) Using the result of part (a) suggest a way of generating bivariate vectors (x, y) with a specified positive population correlation coefficient.

Problem 7

The stopping distance of an automobile on a certain road was studied as a function of velocity (Brownee, 1960)

velocity of a car x (mi/h)	20.5	20.5	30.5	40.5	48.8	57.8
stopping distance y (ft)	15.4	13.3	33.9	73.1	113.0	142.6

Fit y and $\sqrt{y}$ as linear functions of the velocity, and examine the residuals in each case. Which fit is better? Can you suggest any physical reason that explains why?

Problem 8

An excerpt rom Wikipedia:

> "The American Psychological Association's 1995 report "Intelligence: Knowns and Unknowns" stated that the correlation between IQ and crime was -0.2. It was -0.19 between IQ scores and number of juvenile offences in a large Danish sample; with social class controlled, the correlation dropped to -0.17. A correlation of 0.20 means that the explained variance is less than 4%."

Explain how the 4% appearing in the last sentence was computed.

Problem 9

Verify that the test statistic for the model utility test can be expressed in terms of the sample correlation r as follows

$$t = \frac{b_1}{s_{b_1}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Problem 10

The article "Effects of gamma radiation on juvenile and mature cuttings of quaking aspen" (Forest science, 1967) reports the following data on exposure time to radiation x and dry weight of roots y:

x	0	2	4	6	8
y	110	123	119	86	62

The estimated quadratic regression function is $y = 111.8857 + 8.0643x - 1.8393x^2$.

(a) What is the underlying multiple regression model? Write down the corresponding design matrix.

(b) Compute the predicted responses. Find an unbiased estimate s^2 of the noise variance σ^2.

(c) Compute the coefficient of multiple determination.

Problem 11

A sports statistician studied the relation between the time (y in seconds) for a particular competitive swimming event and the swimmer's age (x in years) for 20 swimmers with age ranging from 8 to 18. She employed quadratic regression model and obtained the following result

$$\hat{y} = 147 - 11.11x + 0.2730x^2.$$

The standard error for the curvature effect coefficient was estimated as $s_{b_2} = 0.1157$.

(a) Plot the estimated regression function. Would it be reasonable to use this regression function when the swimmer's age is 40?

(b) Construct a 99 percent confidence interval for the curvature effect coefficient. Interpret your interval estimate.

(c) Test whether or not the curvature effect can be dropped from the quadratic regression model, controlling the α risk at 0.01. State the alternatives, the decision rule, the value of the test statistic, and the conclusion. What is the p-value of the test?

Problem 12

Suppose that grades of 10 students on a midterm and a final exams have a correlation coefficient of 0.5 and both exams have an average score of 75 and a standard deviation of 10.

(a) Sketch a scatterplot illustrating performance on two exams for this group of 10 students.

(b) If Carl's score on the midterm is 90, what would you predict his score on the final to be? How uncertain is this prediction?

(c) If Maria scored 80 on the final, what would you guess that her score on the midterm was?

(d) Exactly what assumptions do you make to make your calculations in (b) and (c)?

Chapter 11

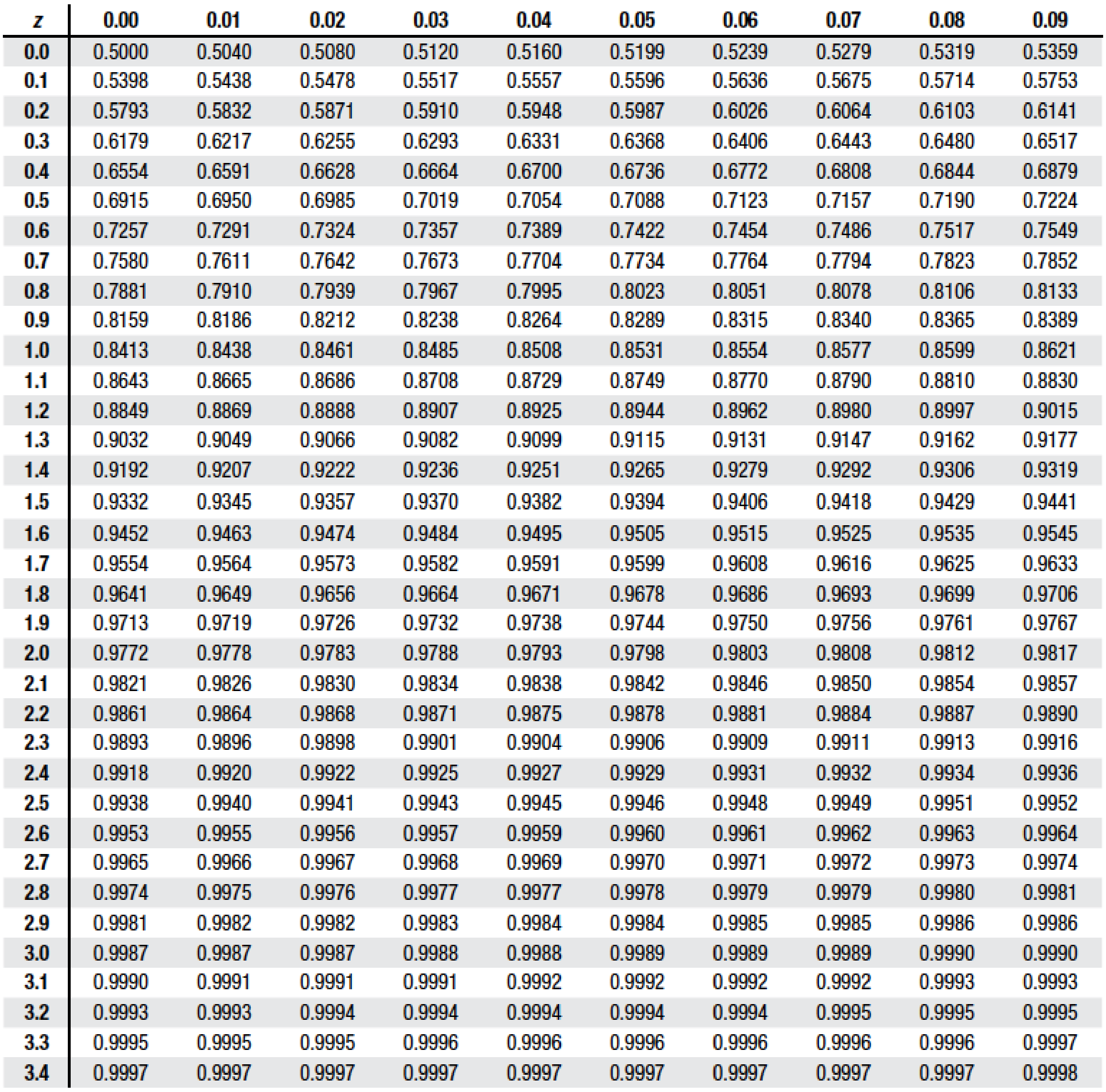
Statistical distribution tables

11.1 Normal distribution table

If Z has a standard normal distribution N(0,1), then the following table gives the probabilities

$$\Phi(z) = \mathrm{P}(Z \leq z).$$

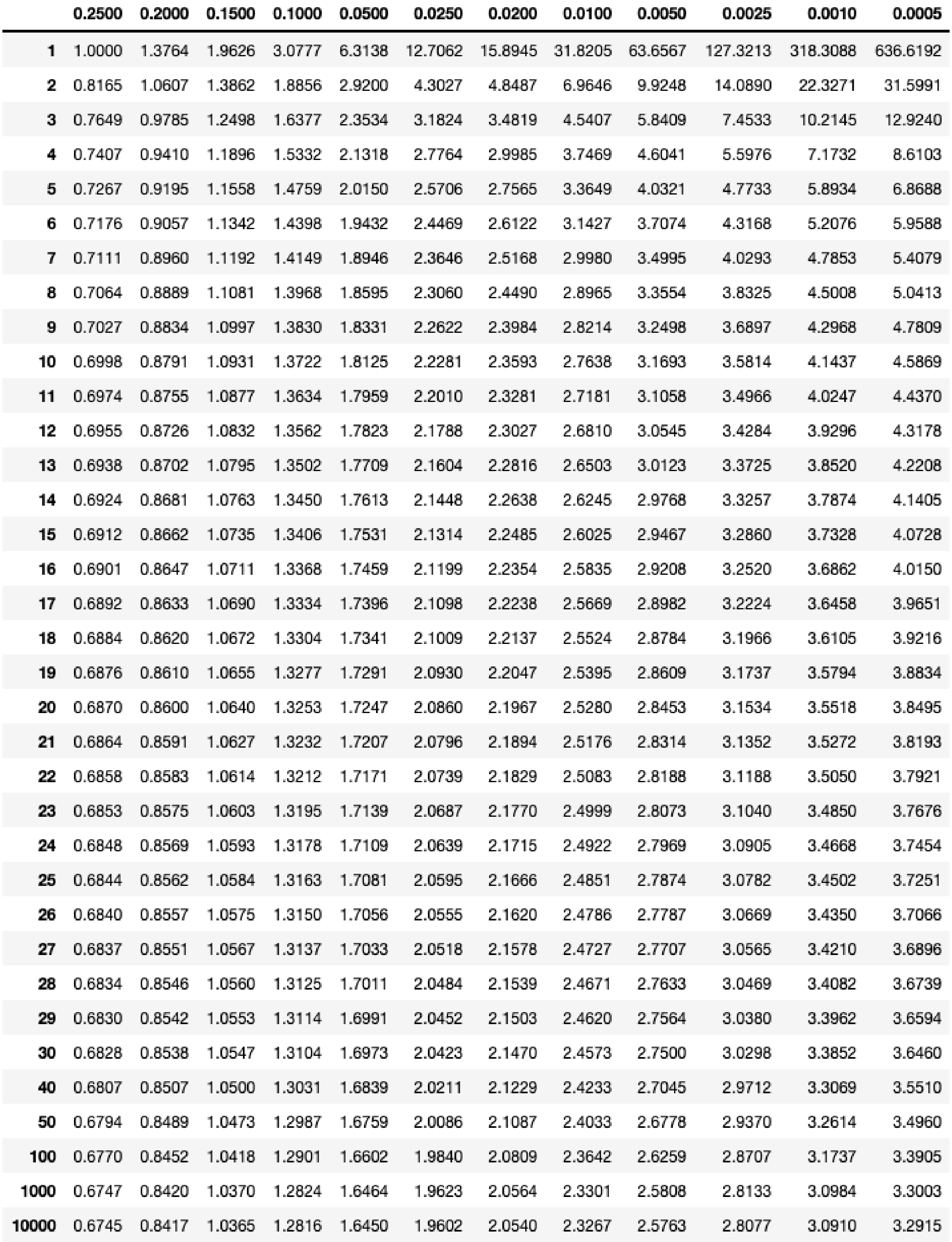
For example, the number on the row 1.9 and column 0.06 gives $\Phi(1.96) = 0.9750$.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

11.2 Critical values of the t-distribution

If T has a t_k-distribution with k degrees of freedom, then for a chosen α, the following table gives the value of $t_k(\alpha)$ such that $\mathrm{P}(T > t_k(\alpha)) = \alpha$. For example, if $k = 5$ and $\alpha = 0.05$, then $t_5(0.05) = 2.015$.



	0.2500	0.2000	0.1500	0.1000	0.0500	0.0250	0.0200	0.0100	0.0050	0.0025	0.0010	0.0005
1	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	15.8945	31.8205	63.6567	127.3213	318.3088	636.6192
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	4.8487	6.9646	9.9248	14.0890	22.3271	31.5991
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	3.4819	4.5407	5.8409	7.4533	10.2145	12.9240
4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	2.9985	3.7469	4.6041	5.5976	7.1732	8.6103
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	2.7565	3.3649	4.0321	4.7733	5.8934	6.8688
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	2.6122	3.1427	3.7074	4.3168	5.2076	5.9588
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.5168	2.9980	3.4995	4.0293	4.7853	5.4079
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.4490	2.8965	3.3554	3.8325	4.5008	5.0413
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.3984	2.8214	3.2498	3.6897	4.2968	4.7809
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.3593	2.7638	3.1693	3.5814	4.1437	4.5869
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.3281	2.7181	3.1058	3.4966	4.0247	4.4370
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.3027	2.6810	3.0545	3.4284	3.9296	4.3178
13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.2816	2.6503	3.0123	3.3725	3.8520	4.2208
14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.2638	2.6245	2.9768	3.3257	3.7874	4.1405
15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.2485	2.6025	2.9467	3.2860	3.7328	4.0728
16	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.2354	2.5835	2.9208	3.2520	3.6862	4.0150
17	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.2238	2.5669	2.8982	3.2224	3.6458	3.9651
18	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.2137	2.5524	2.8784	3.1966	3.6105	3.9216
19	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.2047	2.5395	2.8609	3.1737	3.5794	3.8834
20	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.1967	2.5280	2.8453	3.1534	3.5518	3.8495
21	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.1894	2.5176	2.8314	3.1352	3.5272	3.8193
22	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.1829	2.5083	2.8188	3.1188	3.5050	3.7921
23	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.1770	2.4999	2.8073	3.1040	3.4850	3.7676
24	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.1715	2.4922	2.7969	3.0905	3.4668	3.7454
25	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.1666	2.4851	2.7874	3.0782	3.4502	3.7251
26	0.6840	0.8557	1.0575	1.3150	1.7056	2.0555	2.1620	2.4786	2.7787	3.0669	3.4350	3.7066
27	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.1578	2.4727	2.7707	3.0565	3.4210	3.6896
28	0.6834	0.8546	1.0560	1.3125	1.7011	2.0484	2.1539	2.4671	2.7633	3.0469	3.4082	3.6739
29	0.6830	0.8542	1.0553	1.3114	1.6991	2.0452	2.1503	2.4620	2.7564	3.0380	3.3962	3.6594
30	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.1470	2.4573	2.7500	3.0298	3.3852	3.6460
40	0.6807	0.8507	1.0500	1.3031	1.6839	2.0211	2.1229	2.4233	2.7045	2.9712	3.3069	3.5510
50	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.1087	2.4033	2.6778	2.9370	3.2614	3.4960
100	0.6770	0.8452	1.0418	1.2901	1.6602	1.9840	2.0809	2.3642	2.6259	2.8707	3.1737	3.3905
1000	0.6747	0.8420	1.0370	1.2824	1.6464	1.9623	2.0564	2.3301	2.5808	2.8133	3.0984	3.3003
10000	0.6745	0.8417	1.0365	1.2816	1.6450	1.9602	2.0540	2.3267	2.5763	2.8077	3.0910	3.2915

11.3 Critical values of the chi-squared distribution

If X has a χ_k^2-distribution with k degrees of freedom, then for a chosen α, the following table gives the value of $x_k(\alpha)$ such that $\mathrm{P}(X > x_k(\alpha)) = \alpha$. For example, if $k = 5$ and $\alpha = 0.05$, then $x_5(0.05) = 11.070$.

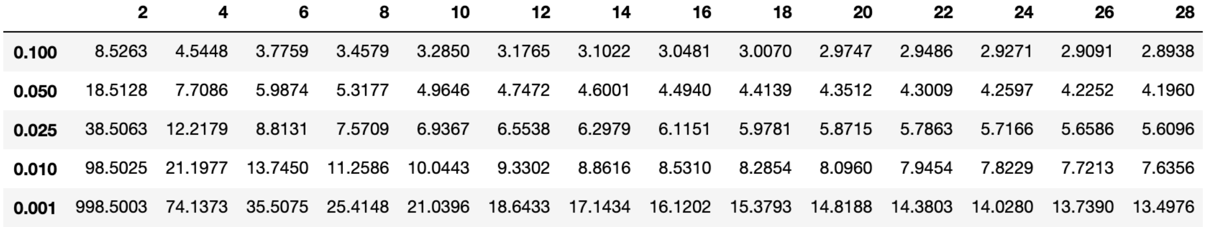
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

11.4 Critical values of the F-distribution

If F has a $F_{\mathrm{df},k}$-distribution, then the tables below give the critical value $F_{\mathrm{df},k}(\alpha)$ such that $\mathrm{P}(F > F_{\mathrm{df},k}(\alpha)) = \alpha$. Different k are shown as columns and different α are shown as rows.

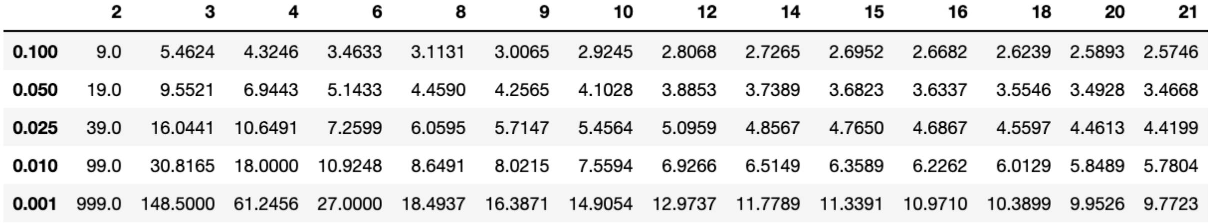
For example, if $\mathrm{df} = 2$, $k = 6$, and $\alpha = 0.025$, then $F_{2,6}(0.025) = 7.2599$.

$F_{1,k}$-distribution table

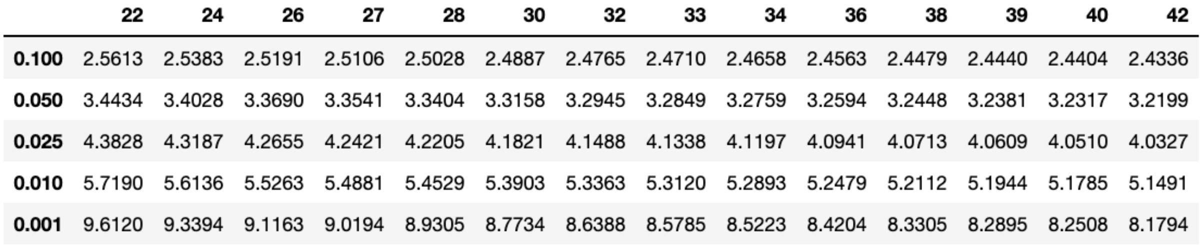


	2	4	6	8	10	12	14	16	18	20	22	24	26	28
0.100	8.5263	4.5448	3.7759	3.4579	3.2850	3.1765	3.1022	3.0481	3.0070	2.9747	2.9486	2.9271	2.9091	2.8938
0.050	18.5128	7.7086	5.9874	5.3177	4.9646	4.7472	4.6001	4.4940	4.4139	4.3512	4.3009	4.2597	4.2252	4.1960
0.025	38.5063	12.2179	8.8131	7.5709	6.9367	6.5538	6.2979	6.1151	5.9781	5.8715	5.7863	5.7166	5.6586	5.6096
0.010	98.5025	21.1977	13.7450	11.2586	10.0443	9.3302	8.8616	8.5310	8.2854	8.0960	7.9454	7.8229	7.7213	7.6356
0.001	998.5003	74.1373	35.5075	25.4148	21.0396	18.6433	17.1434	16.1202	15.3793	14.8188	14.3803	14.0280	13.7390	13.4976

$F_{2,k}$-distribution table

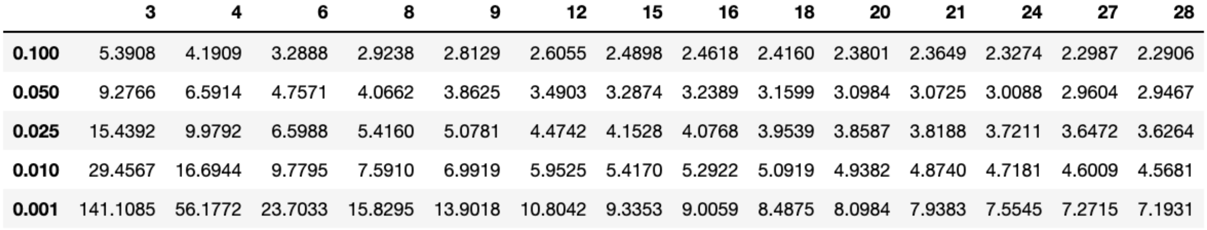


	2	3	4	6	8	9	10	12	14	15	16	18	20	21
0.100	9.0	5.4624	4.3246	3.4633	3.1131	3.0065	2.9245	2.8068	2.7265	2.6952	2.6682	2.6239	2.5893	2.5746
0.050	19.0	9.5521	6.9443	5.1433	4.4590	4.2565	4.1028	3.8853	3.7389	3.6823	3.6337	3.5546	3.4928	3.4668
0.025	39.0	16.0441	10.6491	7.2599	6.0595	5.7147	5.4564	5.0959	4.8567	4.7650	4.6867	4.5597	4.4613	4.4199
0.010	99.0	30.8165	18.0000	10.9248	8.6491	8.0215	7.5594	6.9266	6.5149	6.3589	6.2262	6.0129	5.8489	5.7804
0.001	999.0	148.5000	61.2456	27.0000	18.4937	16.3871	14.9054	12.9737	11.7789	11.3391	10.9710	10.3899	9.9526	9.7723

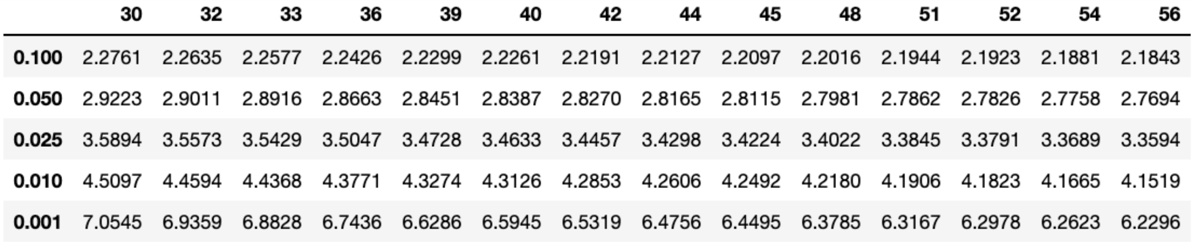


	22	24	26	27	28	30	32	33	34	36	38	39	40	42
0.100	2.5613	2.5383	2.5191	2.5106	2.5028	2.4887	2.4765	2.4710	2.4658	2.4563	2.4479	2.4440	2.4404	2.4336
0.050	3.4434	3.4028	3.3690	3.3541	3.3404	3.3158	3.2945	3.2849	3.2759	3.2594	3.2448	3.2381	3.2317	3.2199
0.025	4.3828	4.3187	4.2655	4.2421	4.2205	4.1821	4.1488	4.1338	4.1197	4.0941	4.0713	4.0609	4.0510	4.0327
0.010	5.7190	5.6136	5.5263	5.4881	5.4529	5.3903	5.3363	5.3120	5.2893	5.2479	5.2112	5.1944	5.1785	5.1491
0.001	9.6120	9.3394	9.1163	9.0194	8.9305	8.7734	8.6388	8.5785	8.5223	8.4204	8.3305	8.2895	8.2508	8.1794

$F_{3,k}$-distribution table

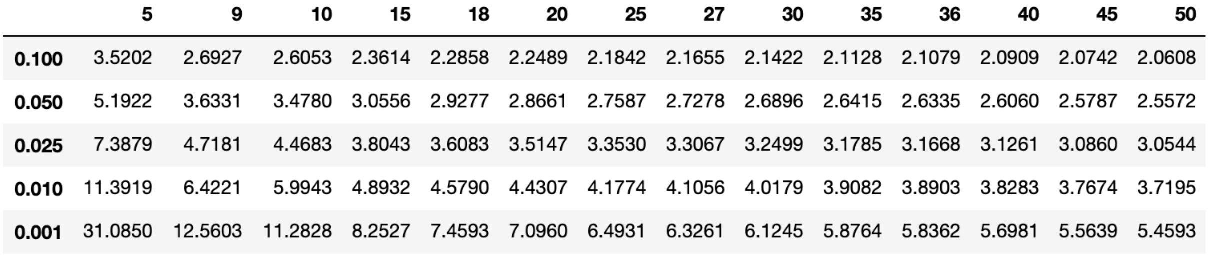


	3	4	6	8	9	12	15	16	18	20	21	24	27	28
0.100	5.3908	4.1909	3.2888	2.9238	2.8129	2.6055	2.4898	2.4618	2.4160	2.3801	2.3649	2.3274	2.2987	2.2906
0.050	9.2766	6.5914	4.7571	4.0662	3.8625	3.4903	3.2874	3.2389	3.1599	3.0984	3.0725	3.0088	2.9604	2.9467
0.025	15.4392	9.9792	6.5988	5.4160	5.0781	4.4742	4.1528	4.0768	3.9539	3.8587	3.8188	3.7211	3.6472	3.6264
0.010	29.4567	16.6944	9.7795	7.5910	6.9919	5.9525	5.4170	5.2922	5.0919	4.9382	4.8740	4.7181	4.6009	4.5681
0.001	141.1085	56.1772	23.7033	15.8295	13.9018	10.8042	9.3353	9.0059	8.4875	8.0984	7.9383	7.5545	7.2715	7.1931



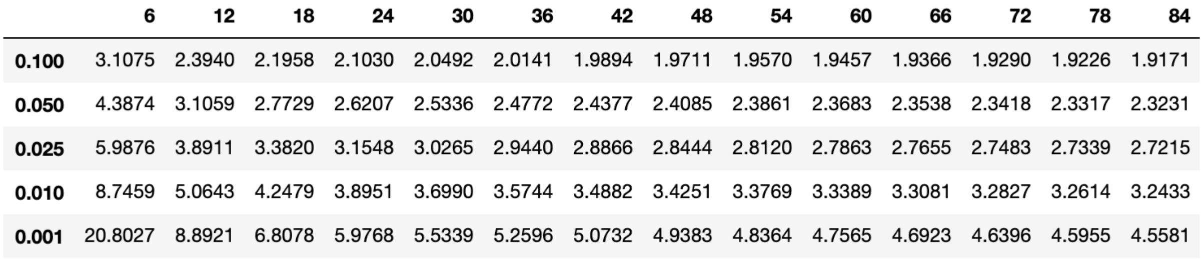
	30	32	33	36	39	40	42	44	45	48	51	52	54	56
0.100	2.2761	2.2635	2.2577	2.2426	2.2299	2.2261	2.2191	2.2127	2.2097	2.2016	2.1944	2.1923	2.1881	2.1843
0.050	2.9223	2.9011	2.8916	2.8663	2.8451	2.8387	2.8270	2.8165	2.8115	2.7981	2.7862	2.7826	2.7758	2.7694
0.025	3.5894	3.5573	3.5429	3.5047	3.4728	3.4633	3.4457	3.4298	3.4224	3.4022	3.3845	3.3791	3.3689	3.3594
0.010	4.5097	4.4594	4.4368	4.3771	4.3274	4.3126	4.2853	4.2606	4.2492	4.2180	4.1906	4.1823	4.1665	4.1519
0.001	7.0545	6.9359	6.8828	6.7436	6.6286	6.5945	6.5319	6.4756	6.4495	6.3785	6.3167	6.2978	6.2623	6.2296

$F_{4,k}$-distribution table



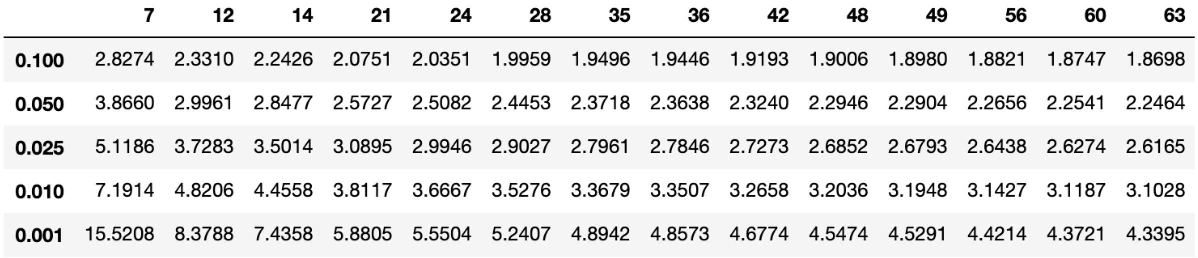
	5	9	10	15	18	20	25	27	30	35	36	40	45	50
0.100	3.5202	2.6927	2.6053	2.3614	2.2858	2.2489	2.1842	2.1655	2.1422	2.1128	2.1079	2.0909	2.0742	2.0608
0.050	5.1922	3.6331	3.4780	3.0556	2.9277	2.8661	2.7587	2.7278	2.6896	2.6415	2.6335	2.6060	2.5787	2.5572
0.025	7.3879	4.7181	4.4683	3.8043	3.6083	3.5147	3.3530	3.3067	3.2499	3.1785	3.1668	3.1261	3.0860	3.0544
0.010	11.3919	6.4221	5.9943	4.8932	4.5790	4.4307	4.1774	4.1056	4.0179	3.9082	3.8903	3.8283	3.7674	3.7195
0.001	31.0850	12.5603	11.2828	8.2527	7.4593	7.0960	6.4931	6.3261	6.1245	5.8764	5.8362	5.6981	5.5639	5.4593

$F_{5,k}$-distribution table



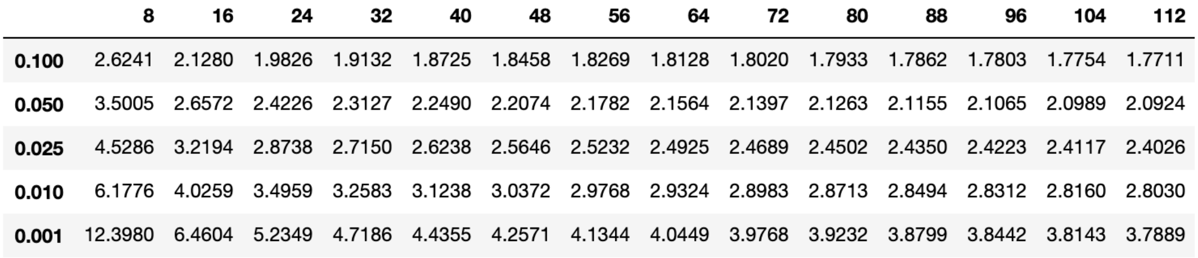
	6	12	18	24	30	36	42	48	54	60	66	72	78	84
0.100	3.1075	2.3940	2.1958	2.1030	2.0492	2.0141	1.9894	1.9711	1.9570	1.9457	1.9366	1.9290	1.9226	1.9171
0.050	4.3874	3.1059	2.7729	2.6207	2.5336	2.4772	2.4377	2.4085	2.3861	2.3683	2.3538	2.3418	2.3317	2.3231
0.025	5.9876	3.8911	3.3820	3.1548	3.0265	2.9440	2.8866	2.8444	2.8120	2.7863	2.7655	2.7483	2.7339	2.7215
0.010	8.7459	5.0643	4.2479	3.8951	3.6990	3.5744	3.4882	3.4251	3.3769	3.3389	3.3081	3.2827	3.2614	3.2433
0.001	20.8027	8.8921	6.8078	5.9768	5.5339	5.2596	5.0732	4.9383	4.8364	4.7565	4.6923	4.6396	4.5955	4.5581

$F_{6,k}$-distribution table



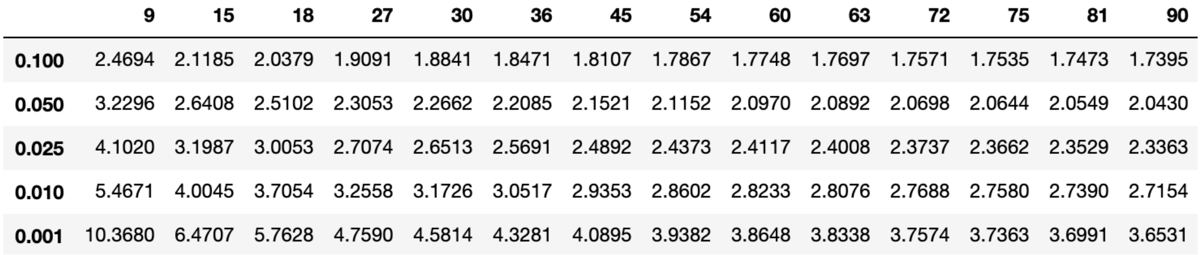
	7	12	14	21	24	28	35	36	42	48	49	56	60	63
0.100	2.8274	2.3310	2.2426	2.0751	2.0351	1.9959	1.9496	1.9446	1.9193	1.9006	1.8980	1.8821	1.8747	1.8698
0.050	3.8660	2.9961	2.8477	2.5727	2.5082	2.4453	2.3718	2.3638	2.3240	2.2946	2.2904	2.2656	2.2541	2.2464
0.025	5.1186	3.7283	3.5014	3.0895	2.9946	2.9027	2.7961	2.7846	2.7273	2.6852	2.6793	2.6438	2.6274	2.6165
0.010	7.1914	4.8206	4.4558	3.8117	3.6667	3.5276	3.3679	3.3507	3.2658	3.2036	3.1948	3.1427	3.1187	3.1028
0.001	15.5208	8.3788	7.4358	5.8805	5.5504	5.2407	4.8942	4.8573	4.6774	4.5474	4.5291	4.4214	4.3721	4.3395

$F_{7,k}$-distribution table



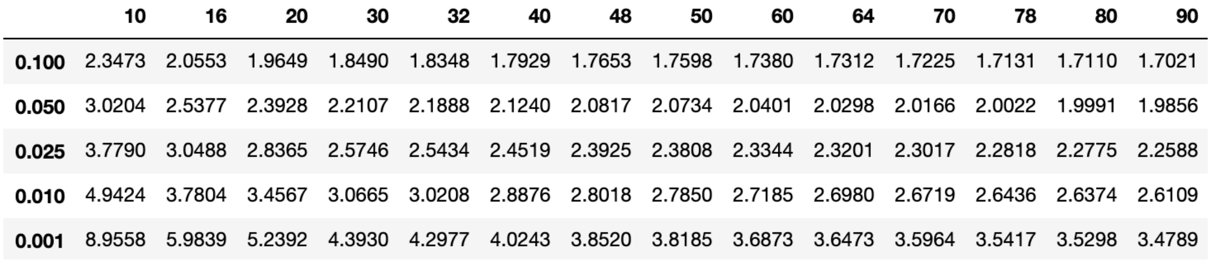
	8	16	24	32	40	48	56	64	72	80	88	96	104	112
0.100	2.6241	2.1280	1.9826	1.9132	1.8725	1.8458	1.8269	1.8128	1.8020	1.7933	1.7862	1.7803	1.7754	1.7711
0.050	3.5005	2.6572	2.4226	2.3127	2.2490	2.2074	2.1782	2.1564	2.1397	2.1263	2.1155	2.1065	2.0989	2.0924
0.025	4.5286	3.2194	2.8738	2.7150	2.6238	2.5646	2.5232	2.4925	2.4689	2.4502	2.4350	2.4223	2.4117	2.4026
0.010	6.1776	4.0259	3.4959	3.2583	3.1238	3.0372	2.9768	2.9324	2.8983	2.8713	2.8494	2.8312	2.8160	2.8030
0.001	12.3980	6.4604	5.2349	4.7186	4.4355	4.2571	4.1344	4.0449	3.9768	3.9232	3.8799	3.8442	3.8143	3.7889

$F_{8,k}$-distribution table



	9	15	18	27	30	36	45	54	60	63	72	75	81	90
0.100	2.4694	2.1185	2.0379	1.9091	1.8841	1.8471	1.8107	1.7867	1.7748	1.7697	1.7571	1.7535	1.7473	1.7395
0.050	3.2296	2.6408	2.5102	2.3053	2.2662	2.2085	2.1521	2.1152	2.0970	2.0892	2.0698	2.0644	2.0549	2.0430
0.025	4.1020	3.1987	3.0053	2.7074	2.6513	2.5691	2.4892	2.4373	2.4117	2.4008	2.3737	2.3662	2.3529	2.3363
0.010	5.4671	4.0045	3.7054	3.2558	3.1726	3.0517	2.9353	2.8602	2.8233	2.8076	2.7688	2.7580	2.7390	2.7154
0.001	10.3680	6.4707	5.7628	4.7590	4.5814	4.3281	4.0895	3.9382	3.8648	3.8338	3.7574	3.7363	3.6991	3.6531

$F_{9,k}$-distribution table



	10	16	20	30	32	40	48	50	60	64	70	78	80	90
0.100	2.3473	2.0553	1.9649	1.8490	1.8348	1.7929	1.7653	1.7598	1.7380	1.7312	1.7225	1.7131	1.7110	1.7021
0.050	3.0204	2.5377	2.3928	2.2107	2.1888	2.1240	2.0817	2.0734	2.0401	2.0298	2.0166	2.0022	1.9991	1.9856
0.025	3.7790	3.0488	2.8365	2.5746	2.5434	2.4519	2.3925	2.3808	2.3344	2.3201	2.3017	2.2818	2.2775	2.2588
0.010	4.9424	3.7804	3.4567	3.0665	3.0208	2.8876	2.8018	2.7850	2.7185	2.6980	2.6719	2.6436	2.6374	2.6109
0.001	8.9558	5.9839	5.2392	4.3930	4.2977	4.0243	3.8520	3.8185	3.6873	3.6473	3.5964	3.5417	3.5298	3.4789

Chapter 12

Solutions to exercises

Solutions to Section 1.7

Solution 1

We have

$$\mathrm{Var}(X_i) = \mathrm{E}((X_i - \mu_i)^2) = \mathrm{E}(X_i^2 - 2\mu_i X_i + \mu_i^2) = \mathrm{E}(X_i^2) - 2\mu_i \mathrm{E}(X_i) + \mu_i^2 = \mathrm{E}(X_i^2) - \mu_i^2,$$

and similarly

$$\mathrm{Cov}(X_1, X_i) = \mathrm{E}((X_1 - \mu_1)(X_2 - \mu_2)) = \mathrm{E}(X_1X_2 - \mu_1 X_2 - \mu_2 X_1 + \mu_1\mu_2) = \mathrm{E}(X_1X_2) - \mu_1\mu_2.$$

Solution 2

With

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1}, \quad Z_2 = \frac{X_2 - \mu_2}{\sigma_2},$$

we get

$$\mathrm{E}(Z_i) = \mathrm{E}(\tfrac{X_i - \mu_i}{\sigma_i}) = \tfrac{1}{\sigma_i}\mathrm{E}(X_i - \mu_i) = \tfrac{1}{\sigma_i}(\mu_i - \mu_i) = 0,$$

and

$$\mathrm{Var}(Z_i) = \mathrm{E}((\tfrac{X_i - \mu_i}{\sigma_i})^2) = \tfrac{1}{\sigma_i^2}\mathrm{E}((X_i - \mu_i)^2) = \tfrac{\sigma_i^2}{\sigma_i^2} = 1.$$

By the definition of the correlation coefficient,

$$\rho = \frac{1}{\sigma_1\sigma_2}\mathrm{E}((X_1 - \mu_1)(X_2 - \mu_2)) = \mathrm{E}\left(Z_1 Z_2\right).$$

This number is a dimensionless quantity in the interval $(-1, 1)$ by the Cauchy–Schwarz inequality, see Wikipedia. To illustrate, let X_1 and X_2 are the height and the weight of a randomly chosen person. The standardised height Z_1 and weight Z_2 would not depend on whether the numbers (X_1, X_2) are given in (centimetres, kilograms) or (inches, pounds).

Solution 3

Given $(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$, the sum $X = X_i + X_j$ can be treated as the number of outcomes among n independent trials with r possible outcomes such that either i or j are observed. Since X is the number of successes among n trials with the probability $p_i + p_j$ of a success, we conclude that

$$X_i + X_j \sim \mathrm{Bin}(n, p_i + p_j).$$

Solution 4

If $X \sim \mathrm{Gam}(\alpha, \lambda)$, then the probability density function of X is

$$f(x) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

The distribution function of the scaled random variable $Y = \lambda X$ is

$$\mathrm{P}(Y \le x) = \mathrm{P}(X \le \lambda^{-1}x) = F(\lambda^{-1}x).$$

Taking the derivatives over x on both sides, we find the probability density function of Y is

$$\lambda^{-1} f(\lambda^{-1}x) = \lambda^{-1}\frac{1}{\Gamma(\alpha)}\lambda^\alpha(\lambda^{-1}x)^{\alpha-1}e^{-\lambda\lambda^{-1}x} = \frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x},$$

indeed the density of the gamma distribution $\mathrm{Gam}(\alpha, 1)$.

Solution 5

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y)^2) - (\mathrm{E}(X+Y))^2 = \mathrm{E}((X+Y)^2) - (\mu_x + \mu_y)^2\\ &= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\mathrm{E}(XY) - \mu_x^2 - \mu_y^2 - 2\mu_x\mu_y\\ &= (\mathrm{E}(X^2) - \mu_x^2) + (\mathrm{E}(Y^2) - \mu_y^2) + 2(\mathrm{E}(XY) - \mu_x\mu_y)\\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).\end{aligned}$$

Solution 6

Because the average event rate is 2.5 goals per match, the Poisson formula with $\mu = 2.5$ gives

$$\mathrm{P}(x \text{ goals in a match}) = \frac{2.5^x e^{-2.5}}{x!}, \quad x = 0, 1 \ldots$$

Therefore, we get

$$\begin{aligned}\mathrm{P}(0 \text{ goals in a match}) &= \frac{2.5^0 e^{-2.5}}{0!} = \frac{e^{-2.5}}{1} \approx 0.082,\\ \mathrm{P}(1 \text{ goal in a match}) &= \frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5e^{-2.5}}{1} \approx 0.205,\\ \mathrm{P}(2 \text{ goals in a match}) &= \frac{2.5^2 e^{-2.5}}{2!} = \frac{6.25e^{-2.5}}{2} \approx 0.257.\end{aligned}$$

Solution 7

If X_N has the hypergeometric distribution $\mathrm{Hg}(N, n, p)$, then

$$\mathrm{P}(X_N = x) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}.$$

As $N \to \infty$, we have

$$\binom{N}{n} \sim \frac{N^n}{n!}, \quad \binom{Np}{x} \sim \frac{(Np)^x}{x!}, \quad \binom{N(1-p)}{n-x} \sim \frac{(N(1-p))^{n-x}}{(n-x)!},$$

and the binomial distribution approximation follows

$$\mathrm{P}(X_N = x) \to \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!} = \binom{n}{x} p^x (1-p)^{n-x}, \quad N \to \infty.$$

Solution 8

We have

$$\mathrm{P}(X = n) = 1 - p\sum_{x=0}^{n-1}(1-p)^x = (1-p)^n,$$

Solutions to Section 2.7

Solution 1

For the given population distribution

possible values x	1	2	4	8
probabilities $\mathrm{P}(X = x)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

the population mean and variance are computed in three steps

$$\begin{aligned}\mu &= 1 \cdot \tfrac{1}{5} + 2 \cdot \tfrac{2}{5} + 4 \cdot \tfrac{1}{5} + 8 \cdot \tfrac{1}{5} = 3.4,\\ \mathrm{E}(X^2) &= 1 \cdot \tfrac{1}{5} + 4 \cdot \tfrac{2}{5} + 16 \cdot \tfrac{1}{5} + 64 \cdot \tfrac{1}{5} = 17.8,\\ \sigma^2 &= 17.8 - \mu^2 = 6.24.\end{aligned}$$

The next table lists the possible values of the sample mean $\bar{x} = \frac{x_1 + x_2}{2}$ together with their probabilities given in brackets, obtained by multiplication of the marginal probabilities.

	$x_2 = 1$	$x_2 = 2$	$x_2 = 4$	$x_2 = 8$	marginal prob.
$x_1 = 1$	1.0 (1/25)	1.5 (2/25)	2.5 (1/25)	4.5 (1/25)	1/5
$x_1 = 2$	1.5 (2/25)	2.0 (4/25)	3.0 (2/25)	5.0 (2/25)	2/5
$x_1 = 4$	2.5 (1/25)	3.0 (2/25)	4.0 (1/25)	6.0 (1/25)	1/5
$x_1 = 8$	4.5 (1/25)	5.0 (2/25)	6.0 (1/25)	8.0 (1/25)	1/5
marginal prob.	1/5	2/5	1/5	1/5	total prob. = 1

This table yields the following sampling distribution of $\bar{X}$:

possible values $\bar{x}$	1	1.5	2	2.5	3	4	4.5	5	6	8	total prob.
probabilities $\mathrm{P}(\bar{X} = \bar{x})$	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	1

From this sampling distribution table, using the same three steps as for the computing of μ and σ^2, we find

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= 1 \cdot \tfrac{1}{25} + 1.5 \cdot \tfrac{4}{25} + 2 \cdot \tfrac{4}{25} + 2.5 \cdot \tfrac{2}{25} + 3 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{1}{25} + 4.5 \cdot \tfrac{2}{25} + 5 \cdot \tfrac{4}{25} + 6 \cdot \tfrac{2}{25} + 8 \cdot \tfrac{1}{25} = 3.4, \\
\mathrm{E}(\bar{X}^2) &= \tfrac{1}{25} + (1.5)^2 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{4}{25} + (2.5)^2 \cdot \tfrac{2}{25} + 9 \cdot \tfrac{4}{25} + 16 \cdot \tfrac{1}{25} + (4.5)^2 \cdot \tfrac{2}{25} + 25 \cdot \tfrac{4}{25} + 36 \cdot \tfrac{2}{25} + 64 \cdot \tfrac{1}{25} = 14.68, \\
\mathrm{Var}(\bar{X}) &= 14.68 - (3.4)^2 = 3.12.
\end{aligned}$$

As a result, we see that indeed,

$$\mathrm{E}(\bar{X}) = 3.4 = \mu, \quad \mathrm{Var}(\bar{X}) = 3.12 = \frac{\sigma^2}{n}.$$

Solution 2

The question is about a simple random sample, with unspecified population size N. We assume that N is much larger than the sample size $n = 1500$ and use the formulas for a random sample with independent observations.

For the given dichotomous data, we have

$$n = 1500, \quad \hat{p} = 0.55, \quad 1 - \hat{p} = 0.45, \quad s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\tfrac{0.55 \times 0.45}{1499}} = 0.013,$$

and want to estimate the population margin of victory

$$v = p - (1 - p) = 2p - 1.$$

Replacing here p by the sample proportion $\hat{p}$, we estimate margin of victory by

$$\hat{v} = \hat{p} - (1 - \hat{p}) = 2\hat{p} - 1 = 0.1.$$

(a) Since

$$\mathrm{Var}(\hat{V}) = \mathrm{Var}(2\hat{P} - 1) = \mathrm{Var}(2\hat{P}) = 4\mathrm{Var}(\hat{P}),$$

the standard error of $\hat{v}$ is twice the standard error of $\hat{p}$

$$\sqrt{\mathrm{Var}(\hat{V})} = 2\sqrt{\mathrm{Var}(\hat{P})}.$$

Therefore, the estimated standard error of $\hat{v}$ is

$$s_{\hat{v}} = 2 s_{\hat{p}} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.026.$$

(b) An approximate 95% confidence interval for v is found using the usual formula based on the point estimate $\hat{v}$ and its standard error $s_{\hat{v}}$:

$$I_v \approx \hat{v} \pm 1.96 \cdot s_{\hat{v}} = 0.10 \pm 0.05.$$

Solution 3

By the central limit theorem the t-score $\frac{\bar{X}-\mu}{S_{\bar{X}}}$ is asymptotically $\mathrm{N}(0,1)$-distributed. Therefore, referring to the normal distribution table we get

$$\begin{aligned}
0.90 &\approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} > -1.28) = \mathrm{P}(-\infty < \mu < \bar{X} + 1.28 S_{\bar{X}}), \\
0.95 &\approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} < 1.645) = \mathrm{P}(\bar{X} - 1.645 S_{\bar{X}} < \mu < \infty).
\end{aligned}$$

implying

$$k_1 = 1.28, \quad k_2 = 1.645.$$

Solution 4

It is enough to verify that for any random variable X and constant θ,

$$\mathrm{E}((X-\theta)^2) = \mathrm{Var}(X) + (\mu-\theta)^2.$$

where μ is the expected value of X. This is done by a simple transformation on the left hand side and opening the brackets

$$\mathrm{E}((X-\theta)^2) = \mathrm{E}((X-\mu+\mu-\theta)^2) = \mathrm{Var}(X) + 2\mathrm{E}((X-\mu)(\mu-\theta)) + (\mu-\theta)^2 = \mathrm{Var}(X) + (\mu-\theta)^2.$$

Solution 5

Based on the data summary

$$N = 2000, \quad n = 25, \quad \sum x_i = 2351, \quad \sum x_i^2 = 231305,$$

we apply the formulas from the Chapter 2.3.

(a) Unbiased estimate of μ is

$$\bar{x} = \frac{2351}{25} = 94.04.$$

(b) Sample variance

$$s^2 = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2) = \frac{25}{24}\Big(\frac{231305}{25} - (94.04)^2\Big) = 425.71.$$

Unbiased estimate of σ^2 is

$$\frac{N-1}{N}s^2 = \frac{1999}{2000}425.71 = 425.49.$$

Unbiased estimate of $\mathrm{Var}(\bar{X})$ is

$$s_{\bar{x}}^2 = \frac{s^2}{n}\Big(1 - \frac{n}{N}\Big) = 16.81.$$

(c) An approximate 95% confidence interval for μ

$$I_\mu = \bar{x} \pm 1.96 s_{\bar{x}} = 94.04 \pm 1.96\sqrt{16.81} = 94.04 \pm 8.04.$$

Solution 6

The bias size is computed as follows

$$\mathrm{E}(\bar{X}^2) - \mu^2 = \mathrm{E}(\bar{X}^2) - (\mathrm{E}\bar{X})^2 = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\Big(1 - \frac{n-1}{N-1}\Big).$$

For the large sample sizes n, the bias is small.

Solution 7

The problem deals with a stratified population of size $N = 2010$ having $k = 7$ strata.

(a) With $n = 100$, we get the following answers using the relevant formulas

Stratum number j	1	2	3	4	5	6	7	Weighted mean
Stratum proportion w_j	0.196	0.229	0.195	0.166	0.084	0.056	0.074	
Stratum mean μ_j	5.4	16.3	24.3	34.5	42.1	50.1	63.8	$\mu = 26.311$
Stratum standard deviation σ_j	8.3	13.3	15.1	19.8	24.5	26.0	35.2	$\bar{\sigma} = 17.018$
Optimal allocation $n\frac{w_j\sigma_j}{\bar{\sigma}}$	10	18	17	19	12	9	15	
Proportional allocation nw_j	20	23	19	17	8	6	7	

(b) Since $\bar{\sigma}^2 = 289.62$ and $\overline{\sigma^2} = 343.28$, we have

$$\mathrm{Var}(\bar{X}_{\mathrm{so}}) = \frac{\bar{\sigma}^2}{n} = 2.896, \ \mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \frac{\overline{\sigma^2}}{n} = 3.433, \ \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = 6.20,$$

where σ^2 is computed in the next item.

(c) We have $\mu = 26.311$, and

$$\sum_{j=1}^{k} w_j(\mu_j - \mu)^2 = 276.889.$$

Therefore

$$\sigma^2 = 343.28 + 276.89 = 620.17, \qquad \sigma = 24.90.$$

(d) If $n_1 = \ldots = n_7 = 10$ and $n = 70$, then

$$\mathrm{Var}(\bar{X}_\mathrm{s}) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k} = 4.45.$$

The requested sample size x is found from the equation

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{x} = \frac{620.17}{x} = 4.45, \qquad x = \frac{620.17}{4.45} = 139.364.$$

After rounding up, we find the answer $x = 140$, which is twice larger than the size $n = 70$ of the stratified sample.

(e) The proportional allocation of the sample size $n = 70$, gives the variance of the stratified sample mean $\mathrm{Var}(\bar{X}_\mathrm{sp}) = 4.90$. Notice that it is larger than the $\mathrm{Var}(\bar{X}_\mathrm{s}) = 4.45$ of the item (d). Solving the equation

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{x} = \frac{620.17}{x} = 4.90, \qquad x = \frac{620.17}{4.90} = 126.57,$$

we find that the corresponding random sample size is $x = 127$.

Solution 9

We are dealing with a stratified population having

$$N = 5, \quad k = 2, \quad w_1 = 0.6, \quad w_2 = 0.4, \quad \mu_1 = 1.67, \quad \mu_2 = 6, \quad \sigma_1^2 = 0.21, \quad \sigma_2^2 = 4.$$

Given $n_1 = n_2 = 1$ and $n = 2$, the stratified sample mean $\bar{x}_s = 0.6x_1 + 0.4x_2$ can take four possible values listed in the table below (with probabilities in the brackets).

	$x_1 = 1$	$x_1 = 2$	marginal prob.
$x_2 = 4$	2.2 (1/6)	2.8 (2/6)	1/2
$x_2 = 8$	3.8 (1/6)	4.4 (2/6)	1/2
marginal prob.	1/3	2/3	1

Using this table we find that

$$\begin{aligned}
\mathrm{E}(\bar{X}_s) &= 2.2 \cdot \tfrac{1}{6} + 2.8 \cdot \tfrac{2}{6} + 3.8 \cdot \tfrac{1}{6} + 4.4 \cdot \tfrac{2}{6} = 3.4,\\
(\mathrm{E}(\bar{X}_s))^2 &= 11.56,\\
\mathrm{E}(\bar{X}_s^2) &= (2.2)^2 \cdot \tfrac{1}{6} + (2.8)^2 \cdot \tfrac{2}{6} + (3.8)^2 \cdot \tfrac{1}{6} + (4.4)^2 \cdot \tfrac{2}{6} = 12.28,\\
\mathrm{Var}(\bar{X}_s) &= 12.28 - 11.56 = 0.72.
\end{aligned}$$

These results are in agreement with the formulas of Section 2.5:

$$\mathrm{E}(\bar{X}_s) = \mu, \quad \mathrm{Var}(\bar{X}_s) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k} = 0.36\sigma_1^2 + 0.16\sigma_2^2.$$

Solution 10

Data: a random sample of size $n = 16$ taken from a normal distribution.

(a) The summary statistics computed from the data

$$\bar{x} = 3.6109, \quad s^2 = 3.4181, \quad s_{\bar{x}} = 0.4622$$

suggest an estimate for μ to be 3.6109, and an estimate for σ^2 to be 3.4181.

(b), (c) The following exact confidence intervals are computed using the formulas of Section 2.6

	90%	95%	99%
I_μ	3.61 ± 0.81	3.61 ± 0.98	3.61 ± 1.36
I_{σ^2}	(2.05; 7.06)	(1.87; 8.19)	(1.56; 11.15)
I_σ	(1.43; 2.66)	(1.37; 2.86)	(1.25; 3.34)

(d) To find the sample size x that halves the confidence interval length, we set up an equation using the exact confidence interval formula for the mean

$$t_{15}(\tfrac{\alpha}{2}) \cdot \frac{s}{\sqrt{16}} = 2 \cdot t_{x-1}(\tfrac{\alpha}{2}) \cdot \frac{s'}{\sqrt{x}},$$

where s' is the sample standard deviation for the sample of size x. A simplified version of this equation $\frac{1}{4} = \frac{2}{\sqrt{x}}$ implies $x \approx (2 \cdot 4)^2 = 64$. Further adjustment for a 95% confidence interval is obtained using

$$t_{15}(\tfrac{\alpha}{2}) = 2.13, \quad t_{x-1}(\tfrac{\alpha}{2}) \approx 2,$$

yielding $x \approx (2 \cdot 4 \cdot \frac{2}{2.13})^2 = 56.4$. We conclude that going from a sample of size 16 to a sample of size 56 would halve the length of the confidence interval for μ.

Solutions to Section 3.4

Solution 1

A method of moment estimate of the parameter μ for the Poisson distribution model is given by the sample mean $\tilde{\mu} = 3.9$. Using this value we compute the expected counts, see table. Comparing the observed and expected counts by a naked eye we see that the Poisson model does not fit well. The sample variance is close to 5 which shows that there is over-dispersion (the estimated variance is larger than the estimated μ).

This extra variation in the data can be explained by the fact that the 300 intervals were distributed over various hours of the day and various days of the week.

x	observed counts	expected counts
0	14	6.1
1	30	23.8
2	36	46.3
3	68	60.1
4	43	58.5
5	43	45.6
6	30	29.6
7	14	16.4
8	10	8.0
9	6	3.5
10	4	1.3
11	1	0.5
12	1	0.2
13+	0	0.1

Solution 2

The likelihood function has the form

$$L(p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{t-n} p^n,$$

where

$$t = x_1 + \ldots + x_n$$

is a sufficient statistic. The log-likelihood function

$$l(p) = (t-n)\ln(1-p) + n \ln p$$

has the derivative

$$l'(p) = -\frac{t-n}{1-p} + \frac{n}{p}.$$

Thus the maximum likelihood estimate satisfies

$$\frac{n}{\hat{p}} = \frac{t-n}{1-\hat{p}},$$

yielding $\hat{p} = n/t = 1/\bar{x}$. This is a consistent estimate due to the law of large numbers.

Solution 3

The population distribution is discrete: the random variable X takes the values $x = 0, 1, 2, 3$ with probabilities

$$p_0 = \frac{2}{3} \cdot \theta, \quad p_1 = \frac{1}{3} \cdot \theta, \quad p_2 = \frac{2}{3} \cdot (1-\theta), \quad p_3 = \frac{1}{3} \cdot (1-\theta),$$

so that

$$p_0 + p_1 = \theta, \quad p_2 + p_3 = 1 - \theta.$$

We are given a random sample with

$$n = 10, \quad \bar{x} = 1.5, \quad s = 1.08,$$

and observed counts

x	0	1	2	3	Total
c_x	2	3	3	2	10

(a) Method of moments. The expression for the first population moment

$$\mu = \frac{1}{3}\cdot\theta + 2\cdot\frac{2}{3}\cdot(1-\theta) + 3\cdot\frac{1}{3}\cdot(1-\theta) = \frac{7}{3} - 2\theta,$$

leads to the equation equation

$$\bar{x} = \frac{7}{3} - 2\tilde{\theta}.$$

It gives an unbiased estimate

$$\tilde{\theta} = \frac{7}{6} - \frac{\bar{x}}{2} = \frac{7}{6} - \frac{3}{4} = 0.417.$$

(b) To find $s_{\tilde{\theta}}$, observe that

$$\mathrm{Var}(\tilde{\Theta}) = \frac{1}{4}\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{40}.$$

Thus we need to find $s_{\tilde{\theta}}$, which estimates $\sigma_{\tilde{\theta}} = \frac{\sigma}{6.325}$. Next we estimate σ using two methods.

Method 1. From

$$\sigma^2 = \mathrm{E}(X^2) - \mu^2 = \frac{1}{3}\cdot\theta + 4\cdot\frac{2}{3}\cdot(1-\theta) + 9\cdot\frac{1}{3}\cdot(1-\theta) = \frac{7}{3} - 2\theta - \left(\frac{7}{3} - 2\theta\right)^2 = \frac{2}{9} + 4\theta - 4\theta^2,$$

we can estimate σ after replacing θ with $\tilde{\theta}$:

$$\tilde{\sigma} = \sqrt{\frac{2}{9} + 4\tilde{\theta} - 4\tilde{\theta}^2} = 1.093.$$

This gives

$$s_{\tilde{\theta}} = \frac{1.093}{6.325} = 0.173.$$

Method 2. Using the sample standard deviation s we get

$$s_{\tilde{\theta}} = \frac{s}{6.325} = \frac{1.08}{6.325} = 0.171.$$

(c) Using the multinomial distribution $(C_0, C_1, C_2, C_3) \sim \mathrm{Mn}(n, p_0, p_1, p_2, p_3)$ we obtain the likelihood function to be

$$L(\theta) = \left(\tfrac{2}{3}\theta\right)^{c_0}\left(\tfrac{1}{3}\theta\right)^{c_1}\left(\tfrac{2}{3}(1-\theta)\right)^{c_2}\left(\tfrac{1}{3}(1-\theta)\right)^{c_3} = \mathrm{const}\cdot\theta^t(1-\theta)^{n-t},$$

where $t = c_0 + c_1$ is a sufficient statistic. Notice that the underlying random variable $T = C_0 + C_1$ has the binomial distribution $\mathrm{Bin}(n, \theta)$. The log-likelihood function and its derivative take the form

$$l(\theta) = \mathrm{const}\ + t\ln\theta + (n-t)\ln(1-\theta),$$

$$l'(\theta) = \frac{t}{\theta} - \frac{n-t}{1-\theta}.$$

Setting the last expression to zero, we find

$$\frac{t}{\hat{\theta}} = \frac{n-t}{1-\hat{\theta}} \quad\Rightarrow\quad \hat{\theta} = \frac{t}{n} = \frac{2+3}{10} = \frac{1}{2}.$$

Thus the maximum likelihood estimate is the sample proportion, which is an unbiased estimate of the population proportion θ.

(d) We find $s_{\hat{\theta}}$ using the formula for the standard error of the sample proportion

$$s_{\hat{\theta}} = \sqrt{\tfrac{\hat{\theta}(1-\hat{\theta})}{n-1}} = 0.167.$$

Solution 4

For a given n and $X = x$ the likelihood function is

$$L(p) = \binom{n}{x}p^x(1-p)^{n-x} \propto p^x(1-p)^{n-x}.$$

(a) We find $\hat{p}$ maximising $L(p)$ by maximising

$$\ln(p^x(1-p)^{n-x}) = x\ln p + (n-x)\ln(1-p).$$

Since

$$\frac{\partial}{\partial p}(x\ln p + (n-x)\ln(1-p)) = \frac{x}{p} - \frac{n-x}{1-p},$$

we have to solve the familiar equation $\frac{x}{p} = \frac{n-x}{1-p}$, which brings the maximum likelihood estimate formula $\hat{p} = \frac{x}{n}$.

(b) The graph of the likelihood function $L(p) = 252p^5(1-p)^5$ is symmetric around the middle point $\hat{p} = 0.5$.

Solution 5

The observed serial number $x = 888$ can be modelled by the discrete uniform distribution

$$\mathrm{P}(X = x) = N^{-1}, \quad x = 1, \ldots, N.$$

(a) Since for the uniform distribution

$$\mu = \tfrac{N+1}{2},$$

and the sample mean is $\bar{x} = x = 888$, the method of moments estimate of N is obtained from the equation

$$888 = \tfrac{\tilde{N}+1}{2}.$$

It gives $\tilde{N} = 2\bar{x} - 1 = 1775$. This is an unbiased estimate of $N = 2\mu - 1$ since $\bar{x}$ is an unbiased estimate of μ.

(b) With $x = 888$, the likelihood function

$$L(N) = \mathrm{P}(X = x) = \frac{1_{\{1 \le x \le N\}}}{N} = \frac{1_{\{N \ge 888\}}}{N}$$

reaches its maximum at $\hat{N} = 888$. We see that in this case the maximum likelihood estimate is severely biased.

Solution 6

We will treat x as the number of black balls obtained by sampling $k = 50$ balls with replacement from an urn with N balls of which $n = 100$ balls are black. Then $x = 20$ is a realisation of the binomial distribution $\mathrm{Bin}(50, \frac{100}{N})$. Thus the likelihod function is

$$L(N) = \mathrm{P}(X = 20) = \binom{50}{20}(\tfrac{100}{N})^{20}(\tfrac{N-100}{N})^{30} = \text{const} \cdot \frac{(N-100)^{30}}{N^{50}}.$$

The log-likelihood function takes the form

$$l(N) = \text{const} + 30\ln(N - 100) - 50\ln N.$$

Take the derivative and set it equal to 0:

$$\frac{30}{N - 100} = \frac{50}{N}.$$

Solving this equation we obtain the maximum likelihood estimate $\hat{N} = 250$.

Solution 7

(a) Given $\sum_{i=1}^n x_i = 58$ and $\sum_{i=1}^n x_i^2 = 260$, the likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{2\sigma^2}} = \frac{1}{(2\pi)^8\sigma^{16}} e^{-\frac{260 - 116\mu + 16\mu^2}{2\sigma^2}}.$$

(b) It is sufficient to know the values of $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$ to compute the likelihood function.

(c) Turning to the log-likelihood function

$$l(\mu, \sigma^2) := \ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{2\sigma^2}$$

take two derivatives and put them equal to zero:

$$\frac{-2\sum_{i=1}^n x_i + 2n\mu}{2\sigma^2} = 0,$$

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{\sigma^3} = 0.$$

The solution of this pair of equations gives us the maximum likelihood estimates

$$\hat{\mu} = \bar{x} = 3.63, \qquad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^n x_i^2 - \bar{x}^2 = 3.11.$$

Since

$$\mathrm{E}(n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,$$

$\hat{\sigma}^2$ is a biased estimate of σ^2.

Solution 8

Consider a random sample $(x_1, \ldots, x_n)$ drawn from the uniform distribution $\mathrm{U}(0, \theta)$ with density

$$f(x|\theta) = \tfrac{1}{\theta} 1_{\{0 \le x \le \theta\}}.$$

The population mean and variance are given by the formulas

$$\mu = \tfrac{\theta}{2}, \quad \sigma^2 = \tfrac{(a-b)^2}{12}.$$

(a) The method of moments estimate $\tilde{\theta} = 2\bar{x}$ is unbiased, and its sampling distribution has the following mean and variance

$$\mathrm{E}(\tilde{\Theta}) = \theta, \quad \mathrm{Var}(\tilde{\Theta}) = \tfrac{4\sigma^2}{n} = \tfrac{\theta^2}{3n}.$$

(b) In terms of $x_{(n)} = \max(x_1, \ldots, x_n)$, the likelihood function takes the form

$$L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \tfrac{1}{\theta^n} 1_{\{\theta \ge x_1\}} \cdots 1_{\{\theta \ge x_n\}} = \tfrac{1}{\theta^n} 1_{\{\theta \ge x_{(n)}\}},$$

so that $x_{(n)}$ is a sufficient statistic. The maximum of this function is achieved at $\hat{\theta} = x_{(n)}$.

(c) The sampling distribution of the maximum likelihood estimate $\hat{\theta} = x_{(n)}$ is computed as

$$\mathrm{P}(X_{(n)} \le x) = \mathrm{P}(X_1 \le x, \ldots, X_n \le x) = \mathrm{P}(X_1 \le x) \cdots \mathrm{P}(X_n \le x) = \Big(\frac{x}{\theta}\Big)^n,$$

where we used independence between different X_i. Taking the derivative, we find the probability density function to be

$$f_{\widehat{\Theta}}(x) = \frac{n}{\theta^n} \cdot x^{n-1}, \quad 0 \le x \le \theta.$$

Using this density function we can compute the mean and variance of $\widehat{\Theta}$:

$$\mathrm{E}(\widehat{\Theta}) = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta, \quad \mathrm{E}(\widehat{\Theta}^2) = \frac{n}{n+2}\theta^2, \quad \mathrm{Var}(\widehat{\Theta}) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

The maximum likelihood estimate is biased, but asymptotically unbiased. Notice the unusual asymptotics

$$\mathrm{Var}(\widehat{\Theta}) = \frac{\theta^2}{n^2}, \quad n \to \infty,$$

indicating that the conditions on the parametric model implying $\hat{\Theta} \approx \mathrm{N}(\theta, \frac{\sigma_\theta}{n})$ are violated.

Comparing the two mean square errors:

$$\mathrm{MSE}(\widehat{\Theta}) = \mathrm{E}(\widehat{\Theta} - \theta)^2 = \Big(-\frac{\theta}{n+1}\Big)^2 + \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{2\theta^2}{(n+1)(n+2)},$$
$$\mathrm{MSE}(\tilde{\Theta}) = \frac{\theta^2}{3n},$$

we conclude that for sufficiently large n, the maximum likelihood estimate $\widehat{\theta}$ is more efficient than the method of moments estimate $\tilde{\theta}$.

(d) The corrected maximum likelihood estimate

$$\hat{\theta}_c = \frac{n+1}{n} \cdot x_{(n)}$$

becomes unbiased $\mathrm{E}(\widehat{\Theta}_c) = \theta$ with $\mathrm{Var}(\widehat{\Theta}_c) = \frac{\theta^2}{n(n+2)}$.

Solution 9

The data in hand is summarised by four observed counts

$$c_1 = 1997, \quad c_2 = 906, \quad c_3 = 904, \quad c_4 = 32$$

obtained from a random sample of size $n = 3839$ drawn from the discrete distribution

$$p_1 = \mathrm{P}(X = 1) = \frac{2+\theta}{4}, \quad p_2 = \mathrm{P}(X = 2) = \frac{1-\theta}{4}, \quad p_3 = \mathrm{P}(X = 3) = \frac{1-\theta}{4}, \quad p_4 = \mathrm{P}(X = 4) = \frac{\theta}{4}.$$

Using the multinomial distribution $(C_1, C_2, C_3, C_4) \sim \mathrm{Mn}(n, p_1, p_2, p_3, p_4)$ for the given realisation

$$(c_1, c_2, c_3, c_4) \text{ with } c_1 + c_2 + c_3 + c_4 = n,$$

with $n = 3839$, we compute the likelihood function as

$$L(\theta) = \binom{n}{c_1, c_2, c_3, c_4} p_1^{c_1} p_2^{c_2} p_3^{c_3} p_4^{c_4} \propto (2+\theta)^{c_1}(1-\theta)^{c_2+c_3}\theta^{c_4}4^{-n} \propto (2+\theta)^{c_1}\theta^{c_4}(1-\theta)^{n-c_1-c_4},$$

where $\propto$ means that we drop the factors depending only on (n, c_1, c_2, c_3, c_4) and not involving θ. The last expression reveals that we have a case of two sufficient statistics (c_1, c_4). Putting

$$\frac{d}{d\theta}\ln L(\theta) = \frac{c_1}{2+\theta} + \frac{c_4}{\theta} - \frac{n - c_1 - c_4}{1-\theta}$$

equal to zero, we arrive at the equation

$$\frac{c_1}{2+\theta} + \frac{c_4}{\theta} = \frac{n - c_1 - c_4}{1-\theta}$$

or equivalently

$$\theta^2 n + \theta u - 2c_4 = 0,$$

where

$$u = 2c_2 + 2c_3 + c_4 - c_1 = 2n - c_4 - 3c_1.$$

We find the maximum likelihood estimate to be

$$\hat{\theta} = \frac{-u + \sqrt{u^2 + 8nc_4}}{2n} = 0.0357.$$

Solution 10

By the law of total probability the probability of the "yes" answer to the randomly generated question equals

$$p = \mathrm{P}(\text{a ``yes'' answer}) = \frac{1}{6} + \frac{4}{6}\cdot q = \frac{1+4q}{6},$$

where $\frac{1}{6}$ is the probability of the number 1 on the top face of the die, implying the answer "yes", $\frac{4}{6}$ is the probability of the numbers 2, 3, 4, or 5, and q is the probability of the honest answer "yes". For a random sample of size n, put

$$X = \text{ the number of "yes" responses.}$$

Under the independence assumptions, we have $X \sim \mathrm{Bin}(n, p)$. Given $X = x$, we estimate the unknown parameter p by the sample proportion $\hat{p} = \frac{x}{n}$. The underlying random variable $\hat{P} = \frac{X}{n}$ has the mean value p and the variance

$$\mathrm{Var}(\hat{P}) = \frac{\mathrm{Var}(X)}{n^2} = \frac{p(1-p)}{n}.$$

After computing the sample proportion $\hat{p}$, we may apply the method of moments by turning to the linear equation

$$\hat{p} = \frac{1+4\tilde{q}}{6}.$$

Its solution gives us the following estimate of the population proportion q:

$$\tilde{q} = \frac{6\hat{p} - 1}{4}.$$

This estimate is unbiased because $\hat{p}$ is an unbiased estimate of p:

$$\mathrm{E}(\tilde{Q}) = \frac{6p - 1}{4} = q.$$

The variance of the sampling distribution equals

$$\mathrm{Var}(\tilde{Q}) = \frac{9}{4}\cdot \mathrm{Var}(\hat{P}) = \frac{9}{4}\cdot\frac{p(1-p)}{n} = \frac{(1+4q)(5-4q)}{16n}.$$

To illustrate, take $n = 100$ and $x = 20$, so that the sample proportion of "yes" answers is $\hat{p} = 0.2$. In this case,

$$\tilde{q} = \frac{6\hat{p} - 1}{4} = 0.05,$$

implying that the proportion of interest q is around 5%. The corresponding estimated standard error is 6%

$$s_{\tilde{q}} = \sqrt{\frac{(1+4\tilde{q})(5-4\tilde{q})}{16n}} = 0.06,$$

indicating that it is desirable to increase the sample size if we want a more precise estimate of q.

Solutions to Section 4.6

Solution 1

The z-score

$$Z = \frac{X - 100p}{10\sqrt{p(1-p)}}$$

has a distribution that is approximated by $\mathrm{N}(0,1)$.

(a) Under H_0 we get $Z = Z_0$, where

$$Z_0 = \frac{X - 100p_0}{10\sqrt{p_0(1-p_0)}} = \frac{X-50}{5},$$

implying that the significance level in question is (using a continuity correction)

$$\begin{aligned} \alpha &= \mathrm{P}(|X-50| > 10|H_0) = \mathrm{P}(|X-50| \geq 11|H_0) \\ &\approx \mathrm{P}(|Z_0| > \tfrac{10.5}{5}|H_0) \approx 2(1-\Phi(2.1)) = 2 \cdot 0.018 = 0.036. \end{aligned}$$

(b) The power of the test is a function of the population proportion p (here for simplicity computed without continuity correction)

$$\begin{aligned} \mathrm{Pw}(p) &= \mathrm{P}(|X-50| > 10) = \mathrm{P}(X < 40) + \mathrm{P}(X > 60) \\ &= \mathrm{P}\left(Z < \frac{40-100p}{10\sqrt{p(1-p)}}\right) + \mathrm{P}\left(Z > \frac{60-100p}{10\sqrt{p(1-p)}}\right) \\ &\approx \Phi\left(\frac{4-10p}{\sqrt{p(1-p)}}\right) + \Phi\left(\frac{10p-6}{\sqrt{p(1-p)}}\right). \end{aligned}$$

Putting $\delta = 1/2 - p$, we see that the power function

$$\mathrm{Pw}(p) = \Phi\left(\frac{10\delta - 1}{\sqrt{1/4-\delta^2}}\right) + \Phi\left(-\frac{10\delta+1}{\sqrt{1/4-\delta^2}}\right) = \Phi\left(\frac{10|\delta| - 1}{\sqrt{1/4-\delta^2}}\right) + \Phi\left(-\frac{10|\delta|+1}{\sqrt{1/4-\delta^2}}\right)$$

is symmetric around $p = 1/2$

p	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
$\mathrm{Pw}(p)$	0.986	0.853	0.500	0.159	0.046	0.159	0.500	0.853	0.986

Solution 2

(a) The two composite nested hypotheses have the form

$$H_0 : \mu \leq \mu_0, \quad H : -\infty < \mu < \infty.$$

(b) The likelihood function takes the form

$$L(\mu) = \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \propto \exp\{-\sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{2}\} \propto \exp\{\tfrac{2n\bar{x}\mu - n\mu^2}{2}\} \propto \exp\{-\tfrac{n}{2}(\mu-\bar{x})^2\}.$$

Observe that in the last step, the term $\exp\{-n\overline{x^2}\}$ is treated as a constant of proportionality, as it does not involve the parameter of interest μ and will not influence the maximum likelihood estimates.

(c) The maximum likelihood estimate under H is $\hat{\mu} = \bar{x}$. The maximum likelihood estimate under H_0 is $\hat{\mu}_0 = \bar{x}$ if $\bar{x} < \mu_0$, and $\hat{\mu}_0 = \mu_0$ if $\bar{x} \geq \mu_0$. Therefore, the likelihood-ratio equals

$$w = \frac{L(\hat{\mu}_0)}{L(\hat{\mu})} = \frac{\exp\{-\frac{n}{2}(\hat{\mu}_0 - \bar{x})^2\}}{\exp\{-\frac{n}{2}(\hat{\mu} - \bar{x})^2\}} = \exp\{-\tfrac{n}{2}(\hat{\mu}_0 - \bar{x})^2\} = \begin{cases} 1 & \text{if } \bar{x} < \mu_0, \\ e^{-\frac{n}{2}(\bar{x}-\mu_0)^2} & \text{if } \bar{x} \geq \mu_0 \end{cases}$$

(d) The likelihood-ratio test rejects H_0 for small values of w or equivalently, for large positive values of $(\bar{x} - \mu_0)$. In particular, the 5% rejection region with $n = 25$ takes the form

$$\mathcal{R} = \{\bar{x} - \mu_0 \geq \tfrac{1.645}{5}\} = \{\bar{x} \geq \mu_0 + 0.33\}.$$

Here we used $\bar{X} \sim \mathrm{N}(\mu_0, \frac{1}{5})$ as the null distribution, since the other eligible choices of $\mu \in H_0$ result in a smaller area under the curve to the right of the critical value.

Solution 3

In this case tikelihood function takes the form

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{x_i!}\mu^{x_i} e^{-\mu} = e^{-\mu n}\mu^{y} \prod_{i=1}^{n} \frac{1}{x_i!}$$

where

$$y = x_1 + \ldots + x_n$$

is a sufficient statistic.

Case 1: two simple hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1.$$

Reject H_0 for small values of the likelihood-ratio

$$\frac{L(\mu_0)}{L(\mu_1)} = e^{-n(\mu_0-\mu_1)}\Big(\frac{\mu_0}{\mu_1}\Big)^{y}.$$

If $\mu_1 > \mu_0$, then we reject H_0 for large values of y. If $\mu_1 < \mu_0$, then we reject H_0 for small values of y. Test statistic Y has null distribution $\mathrm{Pois}(n\mu_0)$.

Case 2: two-sided alternative hypothesis

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Reject H_0 for small values of the generalised likelihood ratio

$$\frac{L(\mu_0)}{L(\hat{\mu})} = e^{-n(\mu_0-\hat{\mu})}\Big(\frac{\mu_0}{\hat{\mu}}\Big)^{y}, \quad \hat{\mu} = y/n.$$

Reject H_0 for the larger values of $|y|$. The test statistic Y has the null distribution $\mathrm{Pois}(n\mu_0)$.

Solution 4

We have a random sample of size $n = 25$ drawn from the population distribution $\mathrm{N}(\mu, 10)$ and would like to test two simple hypotheses

$$H_0 : \mu = 0, \quad H_1 : \mu = 1.5$$

An appropriate test statistic and its exact sampling distribution are

$$\bar{X} \sim \mathrm{N}(\mu, 2),$$

where the standard deviation 2 is obtained as $\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$. Its null and alternative distributions are

$$\bar{X} \overset{H_0}{\sim} \mathrm{N}(0, 2), \quad \bar{X} \overset{H_1}{\sim} \mathrm{N}(1.5, 2).$$

(a) The rejection region at $\alpha = 0.1$ is $\{\bar{x} > x\}$, where x is the solution of the equation

$$0.1 = \mathrm{P}(\bar{X} > x|H_0) = 1 - \mathrm{P}(\bar{X}/2 \leq x/2|H_0) = 1 - \Phi(x/2).$$

From the normal distribution table we find $x/2 = 1.28$, so that $x = 2.56$ and the rejection region is

$$\mathcal{R} = \{\bar{x} > 2.56\}.$$

The corresponding confidence interval method is based the one-sided 90% confidence interval for the mean

$$I_\mu = (\bar{x} - 2.56, \infty).$$

We reject H_0 if the interval does not cover $\mu_0 = 0$, that is when $\bar{x} - 2.56 > 0$.

(b) The power of the test (a) is

$$\mathrm{P}(\bar{X} > 2.56|H_1) = \mathrm{P}(\tfrac{\bar{X}-1.5}{2} > 0.53|H_1) = 1 - \Phi(0.53) = 1 - 0.7019 = 0.298.$$

(c) For $\alpha = 0.01$, since $1 - \Phi(2.33) = 0.01$, the rejection region is

$$\mathcal{R} = \{\bar{x} > 4.66\}.$$

The power of this test is

$$\mathrm{P}(\bar{X} > 4.66|H_1) = \mathrm{P}(\tfrac{\bar{X}-1.5}{2} > 1.58|H_1) = 1 - \Phi(1.58) = 1 - 0.9429 = 0.057.$$

Solution 5

We have a pair of alternative distribution densities (which are two beta-distribution densities)

$$f(x|H_0) = 2x, \quad f(x|H_1) = 3x^2, \quad 0 \leq x \leq 1.$$

(a) The likelihood-ratio as a function of the data value x is

$$\frac{f(x|H_0)}{f(x|H_1)} = \frac{2}{3x}, \quad 0 \leq x \leq 1.$$

The corresponding likelihood-ratio test of H_0 versus H_1 rejects H_0 for large values of x.

(b) The rejection region of a level α test is computed from the equation

$$\mathrm{P}(X > x_{\text{crit}}|H_0) = \alpha,$$

that is

$$1 - x_{\text{crit}}^2 = \alpha.$$

We conclude that

$$\mathcal{R} = \{x : x > \sqrt{1-\alpha}\}.$$

(c) The power of the test is given by

$$\mathrm{P}(X > \sqrt{1-\alpha}|H_1) = 1 - (1-\alpha)^{3/2}.$$

Solution 6

Using the confidence interval method of hypotheses testing we reject H_0 in favour of the two-sided alternative, since the value $\mu = -3$ is not covered by the two-sided confidence interval $(-2, 3)$.

Solution 7

Under the normality assumption the random variable $\frac{14 \cdot S^2}{\sigma^2}$ has the χ^2_{14}-distribution, implying

$$\mathrm{P}(\tfrac{14 \cdot S^2}{\sigma^2} \leq 6.571) = 0.05.$$

Under $H_0 : \sigma = 1$, this entails the following one-sided rejection region

$$\mathcal{R} = \{s^2 \leq \tfrac{6.571}{14}\} = \{s \leq 0.685\}.$$

Given $s = 0.7$, we do not reject $H_0 : \sigma = 1$ in favor of $H_1 : \sigma < 1$ at $\alpha = 0.05$.

Solution 8

The following analysis is the basis of the sign test.

(a) The likelihood-ratio is computed as

$$w = \frac{L(p_0)}{L(\hat{p})} = \frac{\binom{n}{x} p_0^x (1-p_0)^{n-x}}{\binom{n}{x} \hat{p}^x (1-\hat{p})^{n-x}} = \frac{(\frac{1}{2})^n}{(\frac{x}{n})^x (\frac{n-x}{n})^{n-x}} = \frac{(\frac{n}{2})^n}{x^x (n-x)^{n-x}}.$$

(b) The likelihood-ratio test rejects H_0 for small values of

$$\ln w = n \ln(n/2) - x \ln x - (n-x) \ln(n-x),$$

or equivalently, for large values of

$$x \ln x + (n-x) \ln(n-x),$$

or equivalently, for large values of

$$a(y) = (n/2 + y) \ln(n/2 + y) + (n/2 - y) \ln(n/2 - y),$$

where

$$y = |x - n/2|.$$

It remains to observe that the function $a(y)$ is monotone over $y \in [0, n/2]$, since

$$a'(y) = \ln \frac{\frac{n}{2} + y}{\frac{n}{2} - y} > 0,$$

and therefore the test rejects the null hypothesis for the large values of y.

(c) The significance level for the rejection region $|x - \frac{n}{2}| > k$ is computed by

$$\alpha = \mathrm{P}(|X - \tfrac{n}{2}| > k|H_0) = 2\sum_{i<\frac{n}{2}-k}\binom{n}{i}2^{-n}.$$

(d) In particular, for $n = 10$ and $k = 2$ we get

$$\alpha = 2^{-9}\sum_{i=0}^{2}\binom{10}{i} = \frac{1+10+45}{512} = 0.11.$$

(e) Using the normal approximation for $n = 100$ and $k = 10$, we find

$$\alpha = \mathrm{P}(|X - \tfrac{n}{2}| > k|H_0) \approx 2(1 - \Phi(\tfrac{k}{\sqrt{n/4}})) = 2(1 - \Phi(2)) = 0.046.$$

Solution 9

(a) The two-sided p-value = 0.134.

(b) The one-sided p-value = 0.067.

Solution 10

We are supposed to test

H_0 : death cannot be postponed,
H_1 : death can be postponed until after an important date.

(a) Jewish data: $n = 1919$ death dates

$x = 922$ deaths during the week before Passover,
$n - x = 997$ deaths during the week after Passover.

Under the binomial model $X \sim \mathrm{Bin}(n, p)$, we would like to test

$$H_0 : p = 0.5 \quad \text{against} \quad H_1 : p < 0.5.$$

We apply the large-sample test for proportion. Since the observed test statistic value is

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{922 - 1919 \cdot 0.5}{\sqrt{1919} \cdot 0.5} = -1.712,$$

the corresponding one-sided p-value of the test equals

$$\Phi(-1.712) = 1 - \Phi(1.712) = 1 - 0.9564 = 0.044.$$

Conclusion: we reject H_0 in favour of the one-sided H_1 at the significance level 5%.

(b) To control for the seasonal effect the Chinese and Japanese data were studied

$$n = 852, \quad x = 418, \quad n - x = 434, \quad z = -0.548.$$

The one-sided p-value for this data is 29%, showing no significant effect.

(c) Overeating during the important occasion might be a contributing factor.

Solution 11

We apply the multinomial model

$$(C_1, C_2, C_3) \sim \mathrm{Mn}(190, p_1, p_2, p_3)$$

for testing the composite null hypothesis of the Hardy-Weinberg equilibrium:

$$H_0 : p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2.$$

It yields the following likelihood function and the maximum likelihood estimate:

$$L(\theta) = \binom{190}{10, 68, 112}2^{68}\theta^{292}(1-\theta)^{88}, \quad \hat\theta = \frac{292}{380} = 0.768.$$

The Pearson chi-squared test based on the table

x	1	2	3	Total
observed counts c_x	10	68	112	190
expected counts e_x	10.23	67.71	112.07	190

results in the chi-squared test statistic $\mathrm{x}^2 = 0.0065$. With df = 1, we find the p-value to be $2(1 - \Phi(\sqrt{0.0065})) = 0.94$.

Conclusion: the Hardy-Weinberg equilibrium model fits well to the haptoglobin data.

Solution 12

month	c_j	number of days	e_j	$c_j - e_j$
Jan	1867	31	1994	−127
Feb	1789	28	1801	−12
Mar	1944	31	1994	−50
Apr	2094	30	1930	164
May	2097	31	1994	103
Jun	1981	30	1930	51
Jul	1887	31	1994	-107
Aug	2024	31	1994	30
Sep	1928	30	1930	-2
Oct	2032	31	1994	38
Nov	1978	30	1930	48
Dec	1859	31	1994	-135

Here we deal with the simple null hypothesis

$$H_0: \ p_1 = p_3 = p_5 = p_7 = p_8 = p_{10} = p_{12} = \frac{31}{365}, \ p_2 = \frac{28}{365}, \ p_4 = p_6 = p_9 = p_{11} = \frac{30}{365}.$$

The total number suicides $n = 23480$, so that the expected counts presented in the table are computed as

$$e_j = np_j^{(0)}, \quad j = 1, \ldots, 12.$$

Using the table values we find the chi-squared test statistic to be

$$\mathrm{x}^2 = \sum_j \frac{(c_j - e_j)^2}{e_j} = 47.4.$$

Since df = 12 − 1 = 11, and $\chi^2_{11}(0.005) = 26.8$, we reject H_0 of no seasonal variation. Merry Christmas!

Solution 13

We model the number of heads by

$$Y \sim \mathrm{Bin}(n, p), \quad n = 17950.$$

(a) For $H_0 : p = 0.5$ the observed z-score is

$$z_0 = \tfrac{y - np_0}{\sqrt{np_0(1-p_0)}} = 3.46.$$

According to the three-sigma rule this is a significant result and we reject H_0.

(b) Pearson's chi-squared test for the simple null hypothesis

$$H_0 : p_0 = (0.5)^5 = 0.031, \ p_1 = 5 \cdot (0.5)^5 = 0.156, \ p_2 = 10 \cdot (0.5)^5 = 0.313,$$
$$p_3 = 10 \cdot (0.5)^5 = 0.313, \ p_4 = 5 \cdot (0.5)^5 = 0.156, \ p_5 = (0.5)^5 = 0.031,$$

is applied using the table

number of heads	0	1	2	3	4	5	Total
observed counts	100	524	1080	1126	655	105	3590
expected counts	112.2	560.9	1121.9	1121.9	560.9	112.2	3590

The chi-squared test statistic takes the xalue $\mathrm{x}^2 = 21.58$. With df = 5, it gives the p-value = 0.001. We reject this H_0.

(c) For the composite null hypothesis

$$H_0 : p_x = \binom{5}{i} p^x (1-p)^{5-x}, \quad x = 0, 1, 2, 3, 4, 5,$$

we apply Pearson's chi-squared test based on the maximum likelihood estimate $\hat{p} = 0.5129$. From the table

number of heads	0	1	2	3	4	5	Total
observed counts	100	524	1080	1126	655	105	3590
expected counts	98.4	518.3	1091.5	1149.3	605.1	127.4	3590

we get $\mathrm{x}^2 = 8.74$. Here df $= 6 - 1 - 1 = 4$, yielding p-value $= 0.07$. We do not reject this H_0 at the 5% significance level.

Solutions to Section 5.5

Solution 1

Since

$$f(x|\theta) \propto \theta^5(1-\theta)^5,$$

and the prior is flat, we get

$$h(\theta|x) \propto f(x|\theta) \propto \theta^5(1-\theta)^5.$$

We conclude that the posterior distribution is Beta $(6, 6)$. This yields

$$\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{pme}} = 0.5.$$

Solution 2

The data in the table summarises a random sample

$$\boldsymbol{x} = (x_1, \ldots, x_n), \quad n = 130, \quad n\bar{x} = 363.$$

Under the geometric model $X \sim \text{Geom}(p)$ the likelihood is

$$f(\boldsymbol{x}|p) = (1-p)^{n\bar{x}-n}p^n = p^{130}(1-p)^{233}.$$

(d) Using the uniform prior $\text{U}(0, 1)$, we find the posterior to be proportional to the likelihood

$$h(p|x_1, \ldots, x_n) \propto p^{130}(1-p)^{233}.$$

It is a beta distribution $\text{Beta}(131, 234)$ with the posterior mean

$$\mu = \frac{a}{a+b} = \frac{131}{131+234} = 0.36,$$

and the standard deviation of the posterior distribution

$$\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}} = \sqrt{\frac{0.36 \cdot 0.64}{366}} = 0.025.$$

Observe that in this setting, the posterior mean estimate

$$p_{\text{pme}} = \frac{1+\frac{1}{n}}{\bar{x}+\frac{2}{n}}$$

for large n, is close to the maximum likelihood estimate $\hat{p} = 1/\bar{x}$.

Solution 3

Let us use the binomial model $X \sim \text{Bin}(n, p)$ for the number x of successes in n independent trials. Given $x = n$, we would like to estimate p using the Bayesian approach. Applying the uniform prior $\text{Beta}(1, 1)$, we find that the posterior distribution for the parameter p is $\text{Beta}(n + 1, 1)$. Since the posterior mean is $\frac{n+1}{n+2}$, we get

$$\hat{p}_{\text{pme}} = \frac{n+1}{n+2}.$$

Solution 4

Data: one observation of $X = x$. Likelihood ratio test: reject for small values of the likelihood-ratio $\frac{\text{P}(x|H_0)}{\text{P}(x|H_1)}$.

(a) The likelihood-ratio is computed on the bottom line of the table:

x	x_4	x_2	x_1	x_3
$\text{P}(x\|H_0)$	0.2	0.3	0.2	0.3
$\text{P}(x\|H_1)$	0.4	0.4	0.1	0.1
$\frac{\text{P}(x\|H_0)}{\text{P}(x\|H_1)}$	0.5	0.75	2	3

(b) The null distribution of likelihood-ratio is discrete. Under H_0 the test statistic $w = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)}$ takes values $0.5, 0.75, 2, 3$ with probabilities $0.2, 0.3, 0.2, 0.3$. Therefore

$$\mathrm{P}(W \le 0.5|H_0) = 0.2, \quad \mathrm{P}(W \le 0.75|H_0) = 0.5, \quad \mathrm{P}(W \le 2|H_0) = 0.7, \quad \mathrm{P}(W \le 3|H_0) = 1.$$

Since H_0 is rejected for the small values of w, at $\alpha = 0.2$ we reject H_0 only if $w = 0.5$, that is when $x = x_4$. Similarly, at $\alpha = 0.5$ we reject H_0 for $w \le 0.75$, that is when x is either x_4 or x_2.

(c) By the Bayes formula,

$$\mathrm{P}(H_0|x) = \frac{\mathrm{P}(x|H_0)\mathrm{P}(H_0)}{\mathrm{P}(x|H_0)\mathrm{P}(H_0) + \mathrm{P}(x|H_1)\mathrm{P}(H_1)} = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_0) + \mathrm{P}(x|H_1)}.$$

Thus the posterior odds equals the likelihood-ratio

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)},$$

and we conclude that the outcomes x_1 and x_3 favour H_0 since with these outcomes the likelihood-ratio is larger than 1.

(d) For the general prior

$$\mathrm{P}(H_0) = \pi_0, \quad \mathrm{P}(H_1) = \pi_1 = 1 - \pi_0,$$

we get

$$\mathrm{P}(H_i|x) = \frac{\mathrm{P}(x|H_i)\pi_i}{\mathrm{P}(x|H_0)\pi_0 + \mathrm{P}(x|H_1)\pi_1},$$

yielding the following relation for the posterior odds

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = \frac{\mathrm{P}(x|H_0)\pi_0}{\mathrm{P}(x|H_1)\pi_1} = w \cdot \frac{\pi_0}{\pi_1}.$$

Assuming equal costs $\mathrm{cost}_0 = \mathrm{cost}_1$, the rejection rule is

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} \le \frac{\mathrm{cost}_1}{\mathrm{cost}_0} = 1,$$

so that in terms of the likelihood-ratio, we reject H_0 when

$$w \cdot \frac{\pi_0}{\pi_1} \le 1, \quad \text{that is} \quad w \le \frac{\pi_1}{\pi_0} = \frac{1}{\pi_0} - 1, \quad \text{or equivalently} \quad \pi_0 \le \frac{1}{1+w}.$$

Recall that at $\alpha = 0.2$ the rejection region was

$$\mathcal{R} = \{x = x_4\} = \{w = 0.5\},$$

which in terms of the prior probabilities imposes the restriction

$$\pi_0 \le \frac{1}{1+w} = \frac{1}{1+0.5} = \frac{2}{3}.$$

On the other hand, at $\alpha = 0.5$ the rejection region was

$$\mathcal{R} = \{x = x_4\} \cup \{x = x_2\} = \{w = 0.5\} \cup \{w = 0.75\}.$$

To be able to reject for the value $w = 0.75$, we have to put the restriction

$$\pi_0 \le \frac{1}{1+0.75} = \frac{4}{7}.$$

Solution 5

For a single observation $X \sim \mathrm{N}(\mu, \sigma)$, where σ is known, we would like to test $H_0 : \mu = 0$ against $H_1 : \mu = 1$ using the prior probabilities

$$\mathrm{P}(H_0) = \frac{2}{3}, \quad \mathrm{P}(H_1) = \frac{1}{3}.$$

(a) Since the likelihood-ratio equals

$$\frac{f(x|0)}{f(x|1)} = \frac{e^{-\frac{x^2}{2\sigma^2}}}{e^{-\frac{(x-1)^2}{2\sigma^2}}} = e^{\frac{1-2x}{2\sigma^2}},$$

and the prior odds is $\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)} = 2$, the posterior odds takes the form

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = 2e^{\frac{1-2x}{2\sigma^2}}.$$

Choosing H_0 for x such that

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} > 1,$$

or in other words for

$$x < 0.5 + \sigma^2 \ln 2,$$

we arrive at the next answer

σ^2	0.1	0.5	1	5
choose H_0 for	$x < 0.57$	$x < 0.85$	$x < 1.19$	$x < 3.97$

(b) In the long run, the proportion of the time H_0 will be chosen is

$$\mathrm{P}(X < \tfrac{1}{2} + \sigma^2 \ln 2) = \tfrac{2}{3}\mathrm{P}(X - \mu < \tfrac{1}{2} + \sigma^2 \ln 2) + \tfrac{1}{3}\mathrm{P}(X - \mu < \sigma^2 \ln 2 - \tfrac{1}{2}) = \tfrac{2}{3}\Phi(\sigma \ln 2 + \tfrac{1}{2\sigma}) + \tfrac{1}{3}\Phi(\sigma \ln 2 - \tfrac{1}{2\sigma}).$$

Using this formula we conclude that

σ^2	0.1	0.5	1	5
proportion of the time H_0 will be chosen	0.67	0.73	0.78	0.94

Solution 6

We are given a pair of beta-densities

$$f(x|H_0) = 2x, \quad f(x|H_1) = 3x^2, \quad 0 \le x \le 1,$$

and equal prior probabilities

$$\mathrm{P}(H_0) = \mathrm{P}(H_1) = 0.5.$$

By the Bayes formula, the posterior probabilities are

$$h(H_0|x) = \frac{0.5f(x|H_0)}{0.5f(x|H_0) + 0.5f(x|H_1)} = \frac{2}{2+3x}, \quad h(H_1|x) = \frac{3x}{2+3x}.$$

Therefore, the posterior probability of H_0 is greater than that of H_1 for x satisfying $2 > 3x$, that is when $x < \frac{2}{3}$.

Solution 7

The parameters (a, b) of the prior beta distribution are found from the equations

$$\frac{a}{a+b} = \frac{1}{3}, \quad \frac{\frac{1}{3}(1-\frac{1}{3})}{a+b+1} = \frac{1}{32}.$$

The prior pseudo-counts are well approximated by $a = 2$ and $b = 4$. Thus the posterior beta distribution has parameters $(10, 16)$ giving the posterior mean estimate $\hat{p}_{\mathrm{pme}} = 0.38$.

Solution 8

(a) We are given two independent samples:

1. sample one has the sample size $n = 56$ and the sample proportion $\hat{p}_1 = \frac{8}{56} = 0.143$,
2. sample two has the sample size $m = 74$ and the sample proportion $\hat{p}_2 = \frac{12}{74} = 0.162$.

An asymptotic 95% confidence interval for the population difference is given by

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \frac{\hat{p}_2(1-\hat{p}_2)}{m-1}} = -0.019 \pm 0.125 = [-0.144, 0.106].$$

(b) To find a credibility interval for the parameter p, we can use the uniform prior Beta$(1, 1)$. Adding the pseudo-counts $(1, 1)$ to the total counts $(8+12, 48+62)$ we get Beta$(21, 111)$ as the posterior distribution. Using the posterior mean $\mu = \frac{21}{21+111} = 0.16$ and standard deviation $\sigma = \sqrt{\frac{0.16(1-0.16)}{132}} = 0.03$ we arrive at the normal approximation of the posterior distribution with mean 0.16 and standard deviation 0.03. This yields an approximate 95% credibility interval

$$J_p \approx 0.16 \pm 1.96 \cdot 0.03 = [0.10, 0.22].$$

Solution 9

(a) The exponential prior with parameter 1 is Gam$(1, 1)$. Applying the suggested updating rule consecutively four times:

$$(1, 1) \to (3, 2) \to (3, 3) \to (5, 4) \to (10, 5),$$

we find the posterior distribution to be Gam$(10, 5)$. Therefore, $\hat{\theta}_{\text{PME}} = 10/5 = 2$.

(b) The general updating rule for the random sample $(x_1, \ldots, x_n)$ taken from a Poisson distribution, becomes

- the shape parameter $\alpha_1 = \alpha + n\bar{x}$,
- the inverse scale parameter $\lambda_1 = \lambda + n$.

By the formula for the mean of the Gam(α_1, λ_1) distribution,

$$\hat{\theta}_{\text{PME}} = \frac{\alpha + n\bar{x}}{\lambda + n}.$$

Comparing this to the maximum likelihood estimator $\hat{\theta}_{\text{MLE}} = \bar{x}$, we see that

$$\hat{\theta}_{\text{PME}} - \hat{\theta}_{\text{MLE}} = \frac{\alpha + n\bar{x}}{\lambda + n} - \bar{x} = \frac{\alpha - \lambda\bar{x}}{\lambda + n} \to 0,$$

as $n \to \infty$. This demonstrates that the role of the prior is less important for large sample sizes.

Solution 10

The Beta(a, b) probability density function $g(p)$ is proportional to $p^{a-1}(1-p)^{b-1}$. Taking the derivative of this product and setting it equal to zero we arrive at the equation

$$(a-1)p^{a-2}(1-p)^{b-1} = (b-1)p^{a-1}(1-p)^{b-2}.$$

Solving this equation we obtain

$$p^* = \frac{a-1}{a+b-2}.$$

Solution 11

(a) Take the 95% equal-tailed credibility interval. By definition it is formed by the lower and upper quartiles of the posterior distribution. It implies that the interval must include the posterior median.

(b) All three options produce the same interval since the beta-posterior Beta$(5, 5)$ is symmetric about its mean.

(c) Due to symmetry of the N$(3, 0.2)$ distribution, the 95% HPDI is computed as

$$3 \pm 1.95 \cdot 0.2 = 3 \pm 0.39.$$

Solutions to Section 6.7

Solution 1

For a fixed x, the empirical distribution function $\hat{F}(x) = \hat{p}$ is the sample proportion giving an unbiased estimate of $p = F(x) = x$. The variance of $\hat{F}(x)$ is given by

$$y^2 = \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{x(1-x)}{n}, \quad x \in [0, 1].$$

Setting $n = 16$, we get the requested formula

$$(x - 0.5)^2 + 16y^2 = 0.25.$$

Solution 2

Let $0 \le x \le y \le 1$. We have

$$\hat{F}(x) = \frac{1_{\{X_1 \le x\}} + \ldots + 1_{\{X_n \le x\}}}{n}, \quad \mathrm{E}(\hat{F}(x)) = F(x) = x, \quad \mathrm{Var}(\hat{F}(x)) = \frac{x(1-x)}{n},$$
$$\hat{F}(y) = \frac{1_{\{X_1 \le y\}} + \ldots + 1_{\{X_n \le y\}}}{n}, \quad \mathrm{E}(\hat{F}(y)) = F(y) = y, \quad \mathrm{Var}(\hat{F}(y)) = \frac{y(1-y)}{n}.$$

Using

$$\hat{F}(x)\cdot\hat{F}(y)=\frac{1}{n^2}\Big[\sum_{i=1}^{n}1_{\{X_i\le x\}}1_{\{X_i\le y\}}+\sum_{i=1}^{n}\sum_{j\ne i}1_{\{X_i\le x\}}1_{\{X_j\le y\}}\Big],$$

and that

$$1_{\{X_i\le x\}}1_{\{X_i\le y\}}=1_{\{X_i\le x\}},$$

we obtain

$$\mathrm{E}(\hat{F}(x)\cdot\hat{F}(y))=\frac{1}{n^2}\Big[\sum_{i=1}^{n}x+\sum_{i=1}^{n}\sum_{j\ne i}xy\Big]=\frac{x+(n-1)xy}{n},$$

implying

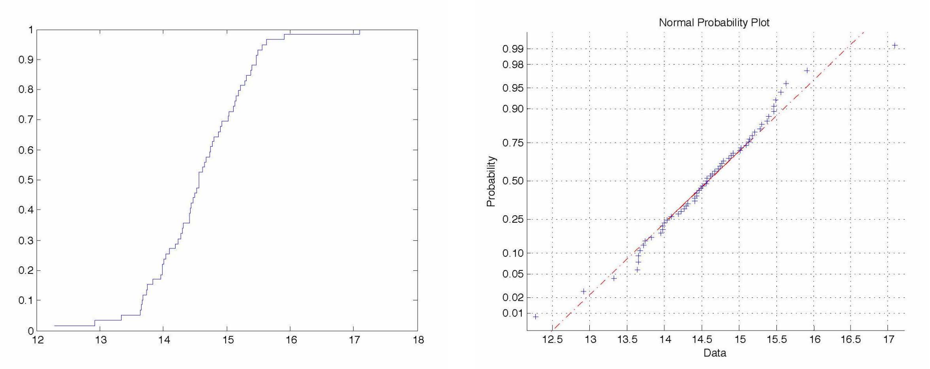
$$\mathrm{Cov}(\hat{F}(u),\hat{F}(v))=\mathrm{E}(\hat{F}(u)\cdot\hat{F}(v))-\mathrm{E}(\hat{F}(u))\cdot\mathrm{E}(\hat{F}(v))=\frac{x+(n-1)xy}{n}-xy=\frac{x(1-y)}{n}.$$

Solution 3

Take a look at the ordered sample $x_{(1)},\dots,x_{(n)}$

12.28 12.92 13.33 13.64 13.65 13.66 13.68
13.73 13.75 13.83 13.96 13.98 13.98 14.01
14.04 25% quantile
14.10 14.19 14.23 14.27 14.30 14.32 14.41
14.41 14.43 14.44 14.47 14.49 14.52 14.56
14.57 50% quantile
14.57 14.62 14.65 14.68 14.73 14.75 14.77
14.80 14.87 14.90 14.92 15.02 15.03 15.10
15.13 75% quantile
15.15 15.18 15.21 15.28 15.31 15.38 15.40
15.47 15.47 15.49 15.56 15.63 15.91 17.09

(a) The figure below shows the empirical distribution function and the normal QQ-plot.



The distribution appears to be close to normal. The 10% sample quantile equals

$$\frac{x_{(6)}+x_{(7)}}{2}=\frac{13.66+13.68}{2}=13.67.$$

(b) The expected percentages under different dilution levels are computed as follows

$$\begin{array}{lll}
1\%\text{ dilution} & \mu_1=14.58\cdot 0.99+85\cdot 0.01=15.28 & \text{can not be detected,}\\
3\%\text{ dilution} & \mu_3=14.58\cdot 0.97+85\cdot 0.03=16.69 & \text{can be detected,}\\
5\%\text{ dilution} & \mu_5=14.58\cdot 0.95+85\cdot 0.05=18.10 & \text{can be detected.}
\end{array}$$

We see that the value 15.28 can not be detected as an outlier, since it coincides with the 82% sample quantile. There is only one sample value larger than 16.69, therefore 3% dilution would be easier to detect. Obviously, 5% dilution resulting in 18.10 is very easy to detect.

Solution 4

Taking the negative derivative of the survival function

$$1 - F(x) = e^{-\alpha x^{\beta}},$$

we obtain the density function

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}},$$

and dividing the latter by the former we obtain the hazard function of the Weibull distribution

$$h(x) = \alpha\beta x^{\beta-1}.$$

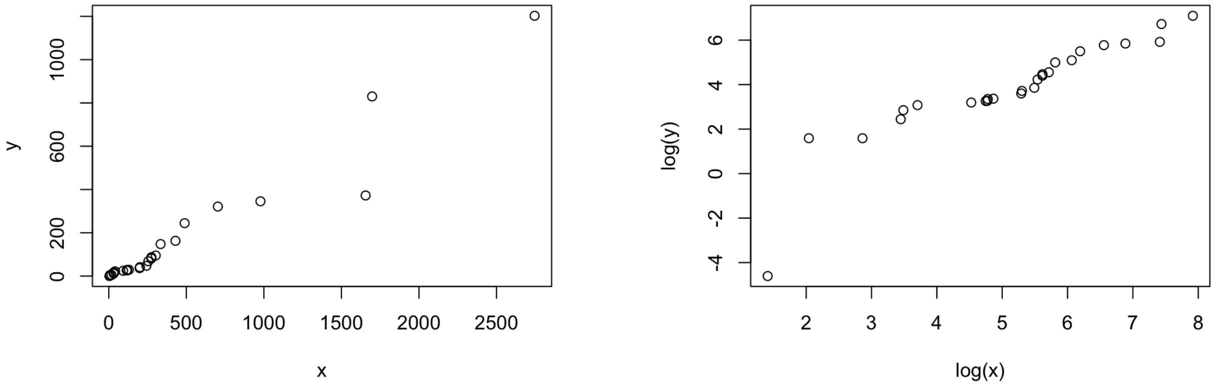
Solution 5

Take the Weibull distribution with parameters α and β.

- If $\beta = 1$, then $h(x) = \alpha$ is constant and the distribution is memoryless of the current age x.
- If $\beta > 1$, say $\beta = 2$, then $h(x) = 2\alpha x$ increases with x, implying that the older individuals die more often than the younger.
- If $0 < \beta < 1$, say $\beta = 0.5$, then $h(x) = 0.5\alpha x^{-0.5}$ decreases with x, predicting that the longer you live the healthier you become.

Solution 6

The left panel shows a QQ-plot with x for the seeded clouds and y for the control clouds. The QQ-plot ia approximated by the line $x = 2.5y$ claiming 2.5 times more rainfall from the seeded clouds.



On the right panel, the QQ-plot (constructed after the log-transformation of the data) is approximated by the line

$$\ln x = 2 + 0.8 \ln y$$

revealing a decreasing slope in the non-linear relationship $x = 7.4y^{0.8}$.

Solution 7

(a) The Laplace curve is symmetric. Its shape is formed by two exponentially declining curves: one for positive x and the other for the negative x.

(b) For $\lambda = \sqrt{2}$ the mean is 0, the standard deviation is 1, the skewness is 0, and the kurtosis is 6. Compared to the normal curve with the same mean and standard deviation but smaller kurtosis (=3), the Laplace distribution has heavier tails. Moreover, since the variances are equal, the two curves should cross 4 times as they cover the same area. This implies that the Laplace curve must also have higher peakedness.

Solution 8

(a) The null distribution of Y is $\mathrm{Bin}(25, \frac{1}{2})$ as each observation is smaller than the true median (assuming that the distribution is continuous) with probability 0.5.

(b) A non-parametric CI for the midean M is given by $(x_{(k)}, x_{(n-k+1)})$ where k is such that

$$\mathrm{P}(Y > n - k|H_0) \approx 0.025.$$

With $n = 25$ we find k using the normal approximation with continuity correction:

$$0.025 \approx \mathrm{P}(Y > 25 - k|H_0) = \mathrm{P}\Big(\frac{Y - 12.5}{2.5} > \frac{13 - k}{2.5}|H_0\Big) \approx 1 - \Phi\Big(\frac{13 - k}{2.5}\Big).$$

Thus $\frac{13-k}{2.5} \approx 1.96$ and we get $k = 8$. The approximate 95% CI for m is given by $(x_{(8)}, x_{(18)})$.

Solution 9

(a) When the population under investigation has a clear structure it is more effective to use stratified sampling for estimating the overall population mean as it has a smaller standard error. In accordance with the optimal allocation formula:

$$n_i = n\frac{w_i\sigma_i}{\bar{\sigma}},$$

the allocation of observations should follow the next two key rules: put more observations in the larger strata, and put more observations in the strata with higher variation.

(b) The sample kurtosis is computed from a sample $(x_1, \ldots, x_n)$ as

$$m_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^4,$$

where $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ is the sample mean and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance. If the corresponding coefficient of skewness is close to zero and $m_4 \approx 3$, then we get an indication that the shape of the population distribution curve is close to the normal distribution.

(c) The standard error for the sample mean is $s_{\bar{x}} = \frac{s}{\sqrt{200}}$. Roughly: the range of heights $160 - 200$ in centimeters covers 95% of the population distribution. Treating this interval as the mean plus-minus two standard deviations, we find $s \approx 10$ cm and $s_{\bar{x}}$ is something like 0.7 cm. We conclude that a random sample of size 200 may give a decent estimate of the population mean height.

Solution 10

(a) The sign test statistic $y =$ number of positive x_i has the null distribution

$$Y \overset{H_0}{\sim} \mathrm{Bin}(25, \tfrac{1}{2}) \approx \mathrm{N}(\tfrac{25}{2}, \tfrac{5}{2}).$$

At 5% significance level we would reject H_0 in favour of the one-sided alternative for $\{y \geq k\}$, where k is found from

$$0.05 = \mathrm{P}(Y \geq k|H_0) = \mathrm{P}(Y > k - 1|H_0) \approx 1 - \Phi\left(\frac{k - 0.5 - 12.5}{5/2}\right) = 1 - \Phi\left(\frac{k - 13}{2.5}\right).$$

This equation gives

$$\frac{k - 13}{2.5} = \Phi_{-1}(0.95) = 1.645,$$

which yields $k = 17$. Thus the rejection region for the sign test is

$$\mathcal{R} = \{y \geq 17\}.$$

We know that the true population distribution is $X \sim \mathrm{N}(0.3, 1)$, therefore the true probability of a positive observation is

$$\mathrm{P}(X > 0|X \sim \mathrm{N}(0.3, 1)) = 1 - \Phi(-0.3) = \Phi(0.3) = 0.62.$$

It follows that for computing the power of the sign test we should use the alternative distribution

$$Y \sim \mathrm{Bin}(25, 0.62) \approx \mathrm{N}(15.5, 2.43).$$

The power of the sign test is

$$\mathrm{P}(Y \geq 17 | Y \sim \mathrm{Bin}(25, 0.62)) \approx 1 - \Phi\left(\frac{17 - 0.5 - 15.5}{2.43}\right) = 1 - \Phi(0.41) = 0.34.$$

(b) Under the normal distribution model $X \sim \mathrm{N}(\mu, 1)$ with $n = 5$, we have $\bar{X} \sim \mathrm{N}(\mu, 1/5)$, and we reject $H_0 : \mu = 0$ for

$$5\bar{x} > 1.645, \quad \text{that is for } \ \bar{x} > 0.33.$$

The power of this test,

$$\mathrm{P}(\bar{X} > 0.33 | \bar{X} \sim \mathrm{N}(0.3, 1/5)) = 1 - \Phi\left(\frac{0.33 - 0.3}{1/5}\right) = 1 - \Phi(0.15) = 0.44,$$

is higher than the power of the sign test.

Solutions to Section 7.6

Solution 1

(a) The sample means $\bar{x} = 0.555$ and $\bar{y} = 1.624$ give unbiased estimates of μ_1 and μ_2. The difference $\mu_1 - \mu_2$ is estimated by $\bar{y} - \bar{x} = 1.069$.

(b) Using two sample variances $s_x^2 = 0.216$, $s_y^2 = 1.180$, we compute the pooled sample variance $s_{\mathrm{p}}^2 = 0.767$, which gives us an unbiased estimate of σ^2.

(c) The standard error of $\bar{y} - \bar{x} = 1.069$ is quite large $s_{\bar{y}-\bar{x}} = 0.587$ due to the small sample sizes.

(d) Based on the t_7-distribution, an exact 90% confidence interval for the mean difference is

$$I_{\mu_2 - \mu_1} = 1.069 \pm 1.113.$$

(e) It is more appropriate to use the two-sided alternative since we do not have access to any prior information about the relationship between μ_1 and μ_2.

(f) From the observed test statistic value $t_0 = 1.821$, we find the two-sided p-value to be larger than 10% using the t-distribution table for df $= 7$.

(g) No, because the obtained p-value is larger than 0.1.

(h) Given $\sigma^2 = 1$, we answer differently to some of the the above questions:

(b*) $\sigma^2 = 1$,
(c*) $s_{\bar{y}-\bar{x}} = 0.6708$,
(d*) $I_{\mu_y - \mu_x} = 1.069 \pm 1.104$,
(f*) using the observed z-score $z_0 = 1.594$, we find the two-sided p-value = 0.11 with the help of the normal distribution table.

Solution 2

If $m = n$, then by the definition of the pooled sample variance,

$$s_{\mathrm{p}}^2\left(\frac{1}{n} + \frac{1}{m}\right) = \frac{2}{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{2n - 2} = \frac{s_1^2 + s_2^2}{n} = \frac{s_x^2}{n} + \frac{s_y^2}{m} = s_{\bar{x}}^2 + s_{\bar{y}}^2.$$

Solution 3

We would like to test the null hypothesis of no drug effect

$H_0 : \mu_1 = \mu_2$, the drug is not effective for reducing high blood pressure.

The suggested measurement design: during the same $n = 10$ days take blood pressure measurements on 4 people, two on the treatment

$$x_{1,1}, \ldots, x_{1,n},$$
$$x_{2,1}, \ldots, x_{2,n},$$

and two controls

$$y_{1,1},\dots,y_{1,n},$$
$$y_{2,1},\dots,y_{2,n}.$$

Dependencies across the subjects make inappropriate both the two-sample t-test and the rank sum test, since we can not treat neither

$$(x_1,\dots,x_{2n})=(x_{1,1},\dots,x_{1,n},x_{2,1},\dots,x_{2,n}),$$

nor

$$(y_1,\dots,y_{2n})=(y_{1,1},\dots,y_{1,n},y_{2,1},\dots,y_{2,n}),$$

as random samples taken from two population distributions.

Solution 4

The paradox arises from the way how the data was collected. This data is an example of an observational study, which was properly planned, and the two samples are not random samples. The large proportion of patients in a bad condition were allocated in the hospital A, which resulted in a smaller overall survival rate for the hospital A.

Solution 5

Two independent samples

$$x_1,\dots,x_n,\quad y_1,\dots,y_n,$$

are taken from two population distributions with equal standard deviation $\sigma=10$. Approximate 95% confidence interval

$$I_{\mu_1-\mu_2}\approx\bar x-\bar y\pm 1.96\cdot 10\cdot\sqrt{\frac{2}{n}}=\bar x-\bar y\pm\frac{27.72}{\sqrt n}.$$

For the confidence interval to have width 2, requires the equality

$$\frac{27.72}{\sqrt n}=1,$$

implying $n\approx 768$.

Solution 6

We are dealing with two independent samples:

	Rank	Type I	Type II	Rank
	1	3.03	3.19	2
	8	5.53	4.26	3
	9	5.60	4.47	4
	11	9.30	4.53	5
	13	9.92	4.67	6
	14	12.51	4.69	7
	17	12.95	6.79	10
	18	15.21	9.37	12
	19	16.04	12.75	15
	20	16.84	12.78	16
Rank sum	130			80

(a) Assume equal variances and apply the two-sample t-test. Using the summary statistics

$$\bar x=10.693,\quad \bar y=6.750,\quad s_x^2=23.226,\quad s_y^2=12.978,\quad s_{\bar x-\bar y}=\sqrt{s_{\bar x}^2+s_{\bar y}^2}=1.903,$$

compute the observed test statistic

$$t_0=\frac{10.693-6.750}{1.903}=2.072.$$

For df = 18, the t-distribution table says that the two-sided p-value of test is larger than 5%. We do not reject the null hypothesis of equality at the 5% significance level.

(b) The rank sum test statistics are $r_1 = 130$, $r_2 = 80$. Using the normal approximation $R_1 \overset{H_0}{\approx} \mathrm{N}(\mu, \sigma)$ with

$$\mu = \frac{n(2n+1)}{2} = 105, \quad \sigma = \sqrt{\frac{n^2(2n+1)}{12}} = 13.23,$$

we find the two-sided p-value to be

$$2\mathrm{P}(R_1 \geq 130) \approx 2(1 - \Phi(\tfrac{129.5-105}{13.23})) = 0.064$$

greater than 5%. Again, we can not reject the null hypothesis of equality.

(c) The non-parametric test in (b) is more relevant, since the QQ-plot for the pooled deviations show non-normality.

(d) To estimate the probability π, that a type I bearing will outlast a type II bearing, we turn to the ordered pooled sample

X-YYYYYY-XX-Y-X-Y-XX-YY-XXXX.

Pick a pair (x_i, y_j) at random from the table, to estimate the probability that type II bearing will outlast the type I by

$$\mathrm{P}(X \leq Y) = \frac{\text{number of } (x_i \leq y_j)}{\text{total number of pairs } (x_i, y_j)} = \frac{10+4+4+3+2+2}{100} = 0.25.$$

This implies a point estimate $\hat{\pi} = 0.75$.

Solution 7

Consider a random sample of the differences $d_1, \ldots, d_n$ taken from a continuous population distribution which is symmetric around the unknown median m. The signed rank test statistic

$$w = \sum_{i=1}^{n} r_i \cdot 1_{\{d_i > 0\}}$$

is computed using the following steps

step 1: remove signs $|d_1|, \ldots, |d_n|$,
step 2: assign ranks $r_1, \ldots, r_n$ to $|d_1|, \ldots, |d_n|$,
step 3: attach the original signs of d_i to the ranks $r_1, \ldots, r_n$,
step 4: compute w as the sum of the positive ranks.

Under the null hypothesis of no difference $H_0 : m = 0$, on the step 3 with $n = 4$, the signs ± are assigned symmetrically at random at the ranks $(1, 2, 3, 4)$. Due to the model assumption that the population distribution is symmetric around the median. As a result there are 16 equally likely outcomes

1	2	3	4	w
−	−	−	−	0
+	−	−	−	1
−	+	−	−	2
+	+	−	−	3
−	−	+	−	3
+	−	+	−	4
−	+	+	−	5
+	+	+	−	6
−	−	−	+	4
+	−	−	+	5
−	+	−	+	6
+	+	−	+	7
−	−	+	+	7
+	−	+	+	8
−	+	+	+	9
+	+	+	+	10

Thus the null distribution of W is given by the table

k	0	1	2	3	4	5	6	7	8	9	10
p_k	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

The smallest one-sided p-value is $\frac{1}{16} = 0.06$ which is higher than 5%. We conclude that $n = 4$ is a too small sample size for the signed rank test.

Solution 8

The critical values w_α° based on the normal approximation

$$W \approx \mathrm{N}\Big(\tfrac{n(n+1)}{4}, \sqrt{\tfrac{n(n+1)(2n+1)}{24}}\Big)$$

satisfy (after the continuity correction)

$$w_{0.05}^\circ = 0.5 + \frac{n(n+1)}{4} - 1.96 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}},$$
$$w_{0.01}^\circ = 0.5 + \frac{n(n+1)}{4} - 2.58 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

This gives without rounding

	$n = 10$	$n = 20$	$n = 25$
$\frac{n(n+1)}{4}$	27.5	105	162.5
$\sqrt{\frac{n(n+1)(2n+1)}{24}}$	9.81	26.79	37.17
$\alpha = 0.05$	8.8	53.0	90.1
$\alpha = 0.01$	2.7	36.5	68.1

The last two rows should be compared with the values obtained using the exact null ditsribution

	$n = 10$	$n = 20$	$n = 25$
$\alpha = 0.05$	8	52	89
$\alpha = 0.01$	3	38	68

Solution 9

(a) Since the variance of the difference $D = X - Y$ equals

$$\mathrm{Var}(D) = \sigma_1^2 + \sigma_2^2 - 2\mathrm{Cov}(X,Y) = 100 + 100 - 100 = 100,$$

we have

$$D \overset{H_0}{\approx} \mathrm{N}(0, 10).$$

It follows that the sample mean of the differences with $n = 25$

$$\bar{D} = \bar{X} - \bar{Y} \overset{H_0}{\approx} \mathrm{N}(0, \tfrac{10}{\sqrt{25}}) = \mathrm{N}(0, 2).$$

The rejection region becomes

$$\mathcal{R} = \{\tfrac{\bar{d}}{2} > 1.645\} = \{\bar{d} > 3.29\}.$$

The power function (under the one-sided alternative $\mu_1 - \mu_2 = \delta > 0$) is computed using the alternative distribution

$$\bar{D} \approx \mathrm{N}(\delta, 2),$$

as follows

$$\mathrm{Pw}(\delta) \approx \mathrm{P}(\bar{D} > 3.29|\mathrm{N}(\delta, 2)) = 1 - \Phi(\tfrac{3.29-\delta}{2}) = \Phi(\tfrac{\delta-3.29}{2}).$$

(b) With two independent samples

$$\bar{D} \overset{H_0}{\approx} \mathrm{N}(0, \sqrt{\tfrac{100}{25} + \tfrac{100}{25}}) = \mathrm{N}(0, 2.83).$$

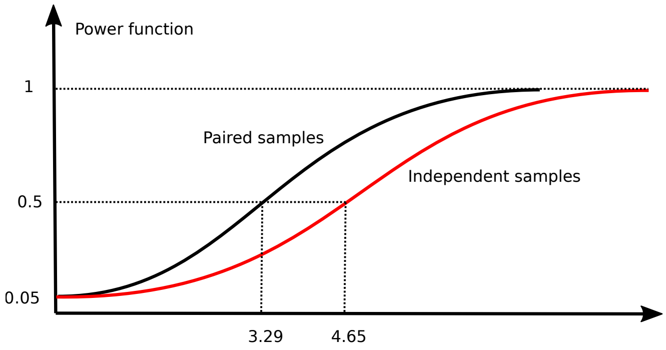
The 5% rejection region is

$$\mathcal{R} = \{\tfrac{\bar{d}}{\sqrt{8}} > 1.645\} = \{\bar{d} > 4.65\}.$$

The corresponding power function equals

$$\mathrm{Pw}(\delta) \approx \mathrm{P}(\bar{D} > 4.65|\mathrm{N}(\delta, 2.83)) = 1 - \Phi(\tfrac{4.65-\delta}{2.83}) = \Phi(\tfrac{\delta-4.65}{2.83}).$$

Conclusion. The two power functions are compared graphically on the next figure, where the x-axis represents the effect size $\delta = \mu_1 - \mu_2$. Clearly, the power of the test for the paired samples is higher.



Solution 10

Summary statistics for $n = 15$ pairs (x_i, y_i):

$$\begin{aligned}
\bar{x} &= 85.26, \quad s_1 = 21.20, \quad s_{\bar{x}} = 5.47, \\
\bar{y} &= 84.82, \quad s_2 = 21.55, \quad s_{\bar{y}} = 5.57, \\
\bar{d} &= \bar{x} - \bar{y} = 0.44, \\
s_d &= 4.63, \quad s_{\bar{x}-\bar{y}} = 1.20.
\end{aligned}$$

If the pairing had been erroneously ignored, then the two independent samples formula would give 6 times larger standard error

$$s_{\bar{x}-\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2} = 7.81.$$

To test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ assume $D \sim \mathrm{N}(\mu_1 - \mu_2, \sigma)$ and apply the one-sample t-test based on the test statistic

$$t_0 = \frac{\bar{d}}{s_{\bar{d}}} = 0.368.$$

With df $= 14$, the two-sided p-value is much larger than 20%, so that we can not reject H_0.

Solution 11

Possible explanations are given below. These are observational studies, based on the data which was not properly randomised.

(a) Rich patients get rooms with a window and higher quality of the health care.

(b) The smoker is a bad husband making miserable wife's life, increasing the risk of cancer.

(c) More stressed condition of a person may result in skipping breakfast as well as accidents.

(d) It the condition of schizophrenia that causes the use of marijuana.

(e) Being part of a community can have a positive effect on mental health and emotional well being.

Solution 12

(a) This is another example of Simpson's paradox. It is resolved by referring to the key factor of the difficulty to enter different programmes. Men tend to apply for easy programs, while women more often apply for programs with low admission rates.

(b) A simple hypothetical experimental study could be based on two independent random samples. Focus on one major program, say major F. Take n randomly chosen female candidates and n randomly chosen male candidates. Ask all of them to apply for major F. Compare two sample proportions of the admitted applicants. (Of course this experiment is impossible to perform in practice.)

Solution 13

(a) We use the binomial model for the number of females $Y \sim \text{Bin}(36, p)$. Given the observed count $y = 13$, we have to test $H_0 : p = 0.5$ against the two-sided alternative $H_1 : p \neq 0.5$. The approximate null distribution is $Y \sim \text{N}(18, 3)$, therefore, an approximate two-sided p-value becomes

$$2(1 - \Phi(\tfrac{18-13}{3}) = 2(1 - \Phi(1.67)) = 2 \times 0.048 = 9.6\%.$$

With such a high p-value we can not reject the null hypothesis of equal sex ratio.

(b) The random sample mean is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n} = \frac{13 \times 62.8 + 23 \times 69.7}{36} = 67.2,$$

and the stratified sample mean

$$\bar{x}_\text{s} = \frac{1}{2}\bar{x}_1 + \frac{1}{2}\bar{x}_2 = \frac{62.8 + 69.7}{2} = 66.3.$$

(c) The standard error of the stratified sample mean is

$$s_{\bar{x}_\text{s}} = \frac{1}{2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{1}{2}\sqrt{\frac{(6.8)^2}{13} + \frac{(11.7)^2}{23}} = 1.54.$$

Solution 14

(a) This is an example of a paired sample, therefore the signed rank test is more appropriate for testing the null hypothesis of no difference.

(b) We use the signed rank test. The observed test statistics are $w = 39$ as the sum of the positive ranks and $w = 6$ as the sum of the negative ranks.

Animal	Site I	Site II	Difference	Signed rank
1	50.6	38.0	12.6	8
2	39.2	18.6	20.6	9
3	35.2	23.2	12.0	7
4	17.0	19.0	-2.0	-2
5	11.2	6.6	4.6	4
6	14.2	16.4	-2.2	-3
7	24.2	14.4	9.8	5
8	37.4	37.6	-0.2	-1
9	35.2	24.4	10.8	6

Using the normal approximation

$$W \overset{H_0}{\approx} \text{N}\Big(\tfrac{n(n+1)}{4}, \sqrt{\tfrac{n(n+1)(2n+1)}{24}}\Big) = \text{N}(22.5, 8.44)$$

we find that the two-sided p-value

$$2\text{P}(W \leq 6|H_0) \approx 2\Phi(\tfrac{6.5-22.5}{8.44}) = 2\Phi(-1.90) = 2(1 - \Phi(1.90)) = 2 \cdot (1 - 0.9713) = 0.0574$$

is larger than 5%. Therefore, we do not reject the null hypothesis of equality in favour of the two-sided alternative.

(c) The extract from the course text book reminds that the null hypothesis for the signed rank test, beside equality of two population distributions, assumes a symmetric distribution for the differences. It also explains why such an assumption is reasonable.

Solution 15

The data is collected using the matched pairs design for 50 independent trials with four possible outcomes (slow correct, fast correct), (slow correct, fast wrong), (slow wrong, fast correct), (slow wrong, fast wrong). We are given

	fast correct	fast wrong	fotal
slow correct			42
slow wrong	0		8
total	35	15	50

which allows us to fill in the table as follows.

	fast correct	fast wrong	total
slow correct	35	7	42
slow wrong	0	8	8
total	35	15	50

We apply the McNemar test having the test statistic

$$\mathrm{x}^2 = \frac{(7-0)^2}{7+0} = 7.$$

With only 7 informative counts, the null distribution is roughly approximated by the χ^2_1−distribution. Since the square root of 7 is 2.65, the standard normal distribution gives the (two-sided) p-value 0.8%. We conclude that the observed difference is statistically significant.

Solutions to Section 8.6

Solution 1

Recall the formula for the one-way ANOVA F-test statistic

$$f = \frac{\mathrm{ms}_A}{\mathrm{ms}_E} = \frac{In(n-1)}{I-1} \cdot \frac{\sum_{i=1}^{I}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2}.$$

For $I = 2$, put

$$\bar{y}_{1\cdot} = \bar{x}, \quad \bar{y}_{2\cdot} = \bar{y}, \quad \bar{y}_{\cdot\cdot} = \frac{\bar{x}+\bar{y}}{2}.$$

In this two-sample setting, the F-test statistic becomes

$$f = 2n(n-1)\frac{(\bar{x} - \frac{\bar{x}+\bar{y}}{2})^2 + (\bar{y} - \frac{\bar{x}+\bar{y}}{2})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2} = 2n\Big(\frac{\bar{x}-\bar{y}}{2s_{\mathrm{p}}}\Big)^2 = t^2,$$

where $t = \frac{\bar{x}-\bar{y}}{s_{\mathrm{p}}\sqrt{\frac{2}{n}}}$ is the two-sample t-test statistic.

Solution 2

(a) Using the table with the data and the available summary statistics, we compute three sums of squares:

$\mathrm{ss_A} = 10((20.34 - 19.40)^2 + (18.34 - 19.40)^2 + (21.57 - 19.40)^2 + (17.35 - 19.40)^2) = 109.2,$
$\mathrm{ss_E} = 9(0.88 + 0.74 + 0.88 + 0.89) = 30.5,$
$\mathrm{ss_T} = 3.58 \cdot 39 = 139.7 = 109.2 + 30.5.$

Then the ANOVA table takes the form

Source	ss	df	ms	f
Treatment	109.2	3	36.4	42.9
Error	30.5	36	0.85	
Total	139.7	39		

Comparing of the observed test statistics 42.9 with the critical value for $\mathrm{F}_{3,36}(0.001) = 6.7436$, see Section 11.4, we see that the p-value of the F-test is much smaller than 0.001, so that we can reject the null hypothesis of no difference.

(b) The normality assumption is supported by the four skewness and kurtosis values, with the former being close to zero and the latter close to 3. On the other hand, the four sample variances are close to each other making realistic the assumption of equal variances.

(c) Since $s_{\mathrm{p}} = \sqrt{\mathrm{ms_E}} = 0.92$ and the t-distribution table gives approximately

$$t_{36}(0.0042) \approx t_{40}(0.005) = 2.7,$$

we get the following Bonferroni interval

$$B_{\mu_i - \mu_j} = \bar{y}_{i.} - \bar{y}_{j.} \pm 1.11.$$

Therefore all observed pairwise differences, except $(2, 4)$, are significant:

Pairs	$(1,2)$	$(1,3)$	$(1,4)$	$(2,3)$	$(2,4)$	$(3,4)$
Differences	2.00	-1.23	2.99	-3.23	0.99	4.22

Solution 3

Assume that under the null hypothesis

$$H_0 : \mu_1 = \ldots = \mu_I = \mu$$

all the data

$$\{y_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n\}$$

come from a single normal distribution $\mathrm{N}(\mu, \sigma)$. Then the corresponding parameter space Ω_0 has the dimension dim $\Omega_0 = 2$. On the other hand, the dimension of the general parameter space Ω is

$$\text{dim } \Omega = I + 1,$$

since the setting beyond H_0 is described by the parameters $\mu_1, \ldots, \mu_I$ and σ. The likelihood-ratio

$$w = \frac{L_0(\hat{\mu}, \hat{\sigma}_0)}{L(\hat{\mu}_1, \ldots, \hat{\mu}_I, \hat{\sigma})},$$

is expressed in terms of two likelihood functions

$$L(\mu_1, \ldots, \mu_I, \sigma) = \prod_{i=1}^{I} \prod_{j=1}^{n} \tfrac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_{ij} - \mu_i)^2}{2\sigma^2}} \propto \sigma^{-N} \exp\{-\sum_i \sum_j \tfrac{(y_{ij} - \mu_i)^2}{2\sigma^2}\},$$

$$L_0(\mu, \sigma) = L(\mu, \ldots, \mu, \sigma) \propto \sigma^{-N} \exp\{-\sum_i \sum_j \tfrac{(y_{ij} - \mu)^2}{2\sigma^2}\}.$$

where $N = In$ is the total number of observations. We find the maximum likelihood estimates to be

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\sigma}_0^2 = \frac{\text{ss}_\text{T}}{N}, \quad \hat{\mu}_i = \bar{y}_{i\cdot}, \quad \hat{\sigma}^2 = \frac{\text{ss}_\text{E}}{N},$$

so that

$$w = \frac{\hat{\sigma}_0^{-N} \exp\{-\sum\sum \frac{(y_{ij} - \hat{\mu})^2}{2\hat{\sigma}_0^2}\}}{\hat{\sigma}^{-N} \exp\{-\sum\sum \frac{(y_{ij} - \hat{\mu}_i)^2}{2\hat{\sigma}^2}\}} = \Big(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\Big)^{-N/2} \cdot \frac{\exp\{-\frac{\text{ss}_\text{T}}{2\text{ss}_\text{T}/N}\}}{\exp\{-\frac{\text{ss}_\text{E}}{2\text{ss}_\text{E}/N}\}} = \Big(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\Big)^{N/2}.$$

The likelihood ratio test rejects the null hypothesis for the smaller values of w or equivalently for the larger values of the ratio

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \frac{\text{ss}_\text{T}}{\text{ss}_\text{E}} = 1 + \frac{\text{ss}_\text{A}}{\text{ss}_\text{E}} = 1 + \frac{(I-1)\text{ms}_\text{A}}{I(n-1)\text{ms}_\text{E}} = 1 + \frac{(I-1)}{I(n-1)} \cdot f$$

that is for the larger values of F-test statistic f.

Recall that

$$-2 \ln w \overset{H_0}{\approx} \chi_k^2, \quad \text{where } k = \dim(\Omega) - \dim(\Omega_0) = I - 1.$$

To see that the exact null distribution for the F-test statistic agrees with the asymptotic null distribution of the likelihood-ratio test, observe that for large n,

$$-2 \ln w = In \cdot \ln(1 + \tfrac{(I-1)}{I(n-1)} \cdot f) \approx In \cdot \tfrac{(I-1)}{I(n-1)} \cdot f \approx (I-1)f = \frac{\text{ss}_\text{A}}{\text{ms}_\text{E}}.$$

Noticing that ms_E converges to σ^2 as $n \to \infty$, we find that

$$-2 \ln w \approx \frac{\text{ss}_\text{A}}{\sigma^2}.$$

It remains to recognise that the null distribution of the ratio $\frac{\text{ss}_\text{A}}{\sigma^2}$ is approximated by the chi-squared distribution with $I - 1$ degrees of freedom.

Solution 4

Consider the one-way layout setting with $I = 10$, $n = 7$, and a dataset generated by independent random variables

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma).$$

Then the corresponding pooled sample variance

$$s_\text{p}^2 = \text{ms}_\text{E} = \frac{1}{I(n-1)} \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

uses $I(n-1) = 60$ degrees of freedom.

(a) In this case, the 95% confidence interval for a single difference $\mu_i - \mu_j$ takes the form

$$I_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{60}(0.025) s_{\rm p}\sqrt{\tfrac{2}{n}} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm 2.82 \cdot \tfrac{s_{\rm p}}{\sqrt{n}}.$$

(b) Bonferroni simultaneous 95% confidence interval for $\binom{10}{2} = 45$ differences $(\mu_i - \mu_j)$

$$B_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{60}(\tfrac{0.025}{45}) s_{\rm p}\sqrt{\tfrac{2}{n}} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm 4.79 \cdot \tfrac{s_{\rm p}}{\sqrt{n}},$$

giving the ratio

$$\frac{\text{Bonferroni}}{\text{Single pair}} = \frac{4.79}{2.82} = 1.7.$$

(c) The Tukey simultaneous 95% confidence interval for differences $(\mu_i - \mu_j)$ is

$$T_{\mu_i-\mu_j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm q_{10,60}(0.05) \tfrac{s_{\rm p}}{\sqrt{n}} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm 4.65 \cdot \tfrac{s_{\rm p}}{\sqrt{n}},$$

giving the ratio

$$\frac{\text{Bonferroni}}{\text{Tukey}} = \frac{4.79}{4.65} = 1.03.$$

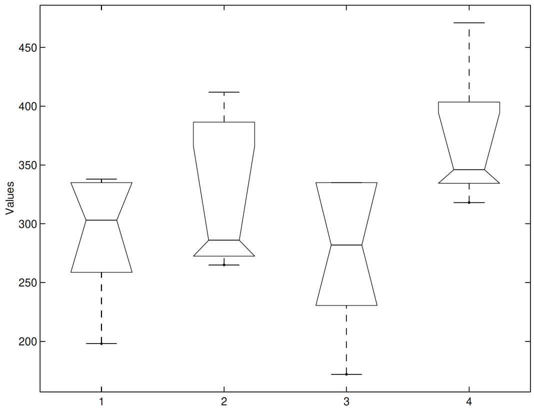
Solution 5

For $I = 4$ control groups of $n = 5$ mice each, we would like to test H_0: no systematic differences between groups. Applying the one-way F-test we get the following ANOVA table.

Source	ss	df	ms	f	p
Columns	27230	3	9078	2.271	0.12
Error	63950	16	3997		
Total	91190	19			

The p-value of 12% is obtained using the $F_{3,16}$-distribution. We do not reject H_0 at 10% significance level.

After inspecting the boxplots of the four samples we see that the assumption of normality and equal variances is clearly violated. The largest difference is between the third and the fourth boxplots. A control question: explain why the third boxplot has no upper whisker?



Turning to the non-parametric Kruskal-Wallis test, consider the pooled sample ranks and their group means:

Group I	2	6	9	11	14	$\bar{r}_{1.} = 8.4$
Group II	4	5	8	17	19	$\bar{r}_{2.} = 10.6$
Group III	1	3	7	12.5	12.5	$\bar{r}_{3.} = 7.2$
Group IV	10	15	16	18	20	$\bar{r}_{4.} = 15.8$

The observed Kruskal-Wallis test statistic takes value

$$w = \frac{12 \cdot 5}{20 \cdot 21}\left((8.4 - 10.5)^2 + (10.6 - 10.5)^2 + (7.2 - 10.5)^2 + (15.8 - 10.5)^2\right) = 6.20.$$

This is close to the critical value $x_3(0.1) = 6.25$ from the table of Section 11.3, implying that the p-value of the Kruskal-Wallis test is close to 10%. We do not reject the null hypothesis of no difference between the four control groups.

Solution 6

The two-way ANOVA table presents the results of two F-tests in the case of $I = 3$ treatments applied to $J = 10$ subjects with $n = 1$ replications.

Source	SS	df	MS	F	P
Columns (blocks)	0.517	9	0.0574	0.4683	0.8772
Rows (treatments)	1.081	2	0.5404	4.406	0.0277
Error	2.208	18	0.1227		
Total	3.806	29			

We reject

$$H_0\text{: no treatment effects}$$

at 5% significance level. Notably, the differences among the subjects are not significant.

The non-parametric Friedman test is based on the ranking within the blocks:

	Dog 1	Dog 2	Dog 3	Dog 4	Dog 5	Dog 6	Dog 7	Dog 8	Dog 9	Dog 10	$\bar{r}_{i.}$
Isof	1	2	3	2	1	2	1	3	1	3	1.9
Halo	2	1	1	3	2	1	3	1	2	2	1.8
Cycl	3	3	2	1	3	3	2	2	3	1	2.3

The observed value of the Friedman test statistic

$$q = \frac{12 \cdot 10}{3 \cdot 4}\big((1.8-2)^2 + (1.9-2)^2 + (2.3-2)^2\big) = 1.4$$

is smaller than the critical value $x_2(0.1) = 4.61$ for the chi-squared distribution with 2 degrees of freedom. Thus the p-value of the Friedman test is larger than 10% and we do not reject the null hypothesis of no difference between the three treatments.

Solution 7

The data consists of 48 survival times obtained for $I = 3$ poisons, $J = 4$ treatments, and $n = 4$ observations per cell. The cell means for the survival times are listed in the next table.

	A	B	C	D
I	4.125	8.800	5.675	6.100
II	3.200	8.150	3.750	6.625
III	2.100	3.350	2.350	3.250

If you draw by hand the three profiles connecting the means for three poisons, you will find that the profiles I and II cross each other, and the profile III is more flat than the other two. The three profiles being non-parallel indicates that there might be an interaction between the main two factors: the poisons and treatments. We have three null hypotheses of interest

H_A: no poison effect,
H_B: no treatment effect,
H_{AB}: no interaction.

(a) Referring to the two-way ANOVA table

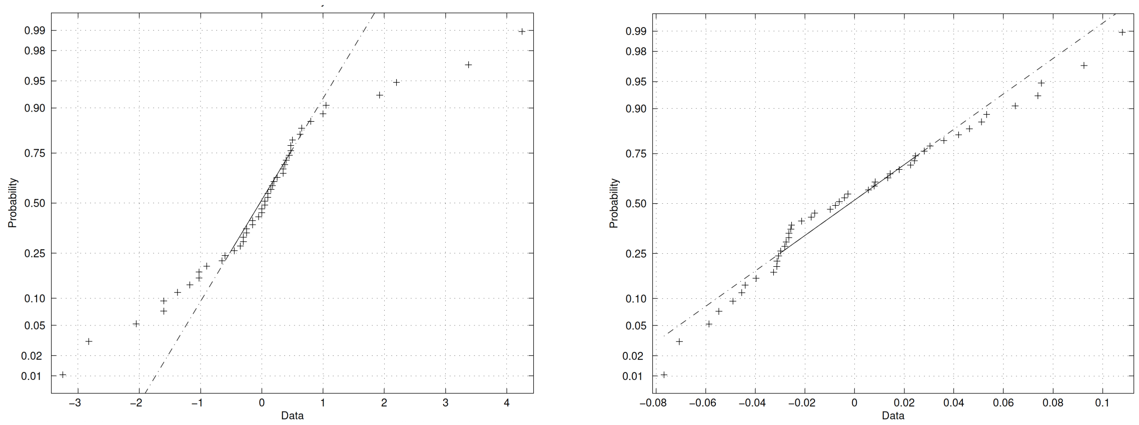
Source	SS	df	MS	F	P
Columns (treatments)	91.9	3	30.63	14.01	0.0000
Rows (poisons)	103	2	51.52	23.57	0.0000
Intercation	24.75	6	4.124	1.887	0.1100
Error	78.69	36	2.186		
Total	298.4	47			

we reject H_A and H_B at 1% significance level. We can not reject H_{AB} even at 10% significance level. Conclusion:

3 poisons act differently,
4 treatments act differently,
some indication of interaction.

The analysis of the residuals shows the following results:

the normal QQ-plot reveals non-normality (see the left panel of the figure below),
the coefficient of skewness = 0.59 is positive,
the coefficient of kurtosis = 4.1 is larger than 3.



(b) After transforming the data by death rate = 1/survival time, we obtain new cell means for the death rates given by the next table.

	A	B	C	D
I	0.249	0.116	0.186	0.169
II	0.327	0.139	0.271	0.171
III	0.480	0.303	0.427	0.309

If you now draw three profiles, you will see that they look more parallel. Using the two-way ANOVA results for the transformed data

Source	SS	df	MS	F	P
Columns (treatments)	0.204	3	0.068	28.41	0.0000
Rows (poisons)	0.349	2	0.174	72.84	0.0000
Intercation	0.01157	6	0.0026	1.091	0.3864
Error	0.086	36	0.0024		
Total	0.6544	47			

we reject H_A and H_B at 1% significance level, and do not reject H_{AB}. Conclusions:

3 poisons act differently,
4 treatments act differently,
no interaction.

The normal QQ-plot of the residuals for the transformed data approves the normality assumption of the performed three F-tests, see the right panel on the figure above.

Solution 8

(a) The assumptions of the normal theory for the two-way ANOVA require that the 20 measurements of this example (y_{ij}) are generated by the normal distributions

$$Y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma),$$

where

$$\mu_{ij} = \mu + \alpha_j + \beta_i, \quad i = 1, 2, 3, 4, 5, \ j = 1, 2, 3, 4,$$

satisfying

$$\sum_j \alpha_j = \sum_i \beta_i = 0.$$

Under this model the most interesting is the null hypothesis of no difference among different types of tires

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.$$

(b) The Friedman test based on the ranked measurements

Bus	Tire 1	Tire 2	Tire 3	Tire 4
1	1	3	4	2
2	1	3	4	2
3	1	4	3	2
4	1	3	4	2
5	1	3	4	2
Mean rank	1.0	3.2	3.8	2.0

results in the test statistic value $q = 14.04$. The null distribution is approximated by a chi-square distribution with $k = 3$ degrees of freedom, whose table gives a p-value less than 0.5%. We reject H_0 using the Friedman test.

Solution 9

(a) In terms of the two-way ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \sigma Z_{ijk}, \quad Z_{ijk} \sim \mathrm{N}(0,1)$$

(grand mean + main effects + interaction+noise), we estimate the main effects as

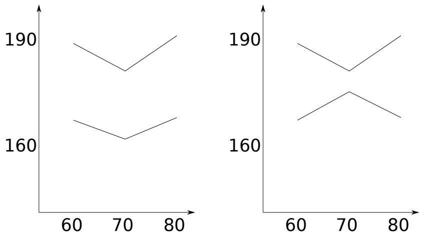
$$\hat\alpha_1 = 11.9,\ \hat\alpha_2 = -11.8, \qquad \hat\beta_1 = 1.99,\ \hat\beta_2 = -5.02,\ \hat\beta_3 = 3.04.$$

(Notice the effect of rounding errors.)

(b) Compute the cell means

	Speed		
	60	70	80
1	189.7	185.1	189.0
1	188.6	179.4	193.0
1	190.1	177.3	191.1
Cell means	189.5	180.6	191.0
2	165.1	161.7	163.3
2	165.9	159.8	166.6
2	167.6	161.6	170.3
Cell means	166.2	161.0	166.7

and draw two lines for the speed depending on two different formulations, see the left panel on the figure below. These two lines are almost parallel indicating to the absence of interaction between two main factors. This is confirmed by the ANOVA table below showing that the interaction is not significant.



One possible interaction effect could have the form on the right panel. In this case the formulation 2 interacts with the speed factor in such a way that the yield becomes largest at the speed 70.

(c) The two-way ANOVA table

Source	df	ss	ms	f	Critical values	Significance
Formulation	1	2253.44	2253.44	376.2	$F_{1,12}(0.001) = 18.6433$	Highly significant
Speed	2	230.81	115.41	19.3	$F_{2,12}(0.001) = 12.9737$	Highly significant
Interaction	2	18.58	9.29	1.55	$F_{2,12}(0.100) = 2.9245$	Not significant
Error	12	71.87	5.99			
Total	17					

(d) To check the normality assumption of the three F-tests.

Solution 10

(a)

Source of variation	ss	df	ms	f
Varieties	328.24	2	164.12	103.55
Density	86.68	3	28.89	18.23
Interaction	8.03	6	1.34	0.84
Errors	38.04	24	1.59	

(b) Using the critical values

$$F_{2,24}(0.001) = 9.3394, \qquad F_{3,24}(0.001) = 7.5545, \qquad F_{6,24}(0.100) = 2.0351,$$

we reject both null hypotheses on the main factors and do not reject the null hypothesis on interaction.

(c) $s = \sqrt{1.59} = 1.26$.

Solution 11

(a) The stratified sample mean is

$$\bar{x}_{\rm s} = 0.3 \cdot 6.3 + 0.2 \cdot 5.6 + 0.5 \cdot 6.0 = 6.01.$$

(b) We are in the one-way Anova setting with $I = 3$ and $J = 13$. The 95% Bonferroni simultaneous confidence interval for the differences $(\mu_i - \mu_j)$ has the form

$$B_{\mu_i - \mu_j} = \bar{x}_i - \bar{x}_j \pm t_{36}(^{0.05}/_{6}) s_p \sqrt{^2/_{13}},$$

involving the pooled sample variance

$$s_p^2 = \frac{12 \cdot s_1^2 + 12 \cdot s_2^2 + 12 \cdot s_3^2}{36} = \frac{2.14^2 + 2.47^2 + 3.27^2}{3} = 2.67^2.$$

This yields

$$B_{\mu_i - \mu_j} = \bar{x}_i - \bar{x}_j \pm 2.5 \cdot 2.67 \cdot 0.39 = \bar{x}_i - \bar{x}_j \pm 2.62,$$

and therefore,

$$B_{\mu_1 - \mu_2} = 0.7 \pm 2.62, \quad B_{\mu_1 - \mu_3} = 0.3 \pm 2.62, \quad B_{\mu_3 - \mu_2} = 0.4 \pm 2.62.$$

(c) We would not reject the null hypothesis of equality $\mu_1 = \mu_2 = \mu_3$, since all three simultaneous confidence intervals contain zero:

$$0.7 \pm 2.62, \quad 0.3 \pm 2.62, \quad 0.4 \pm 2.62.$$

Solution 12

(a) This is an example of a randomised block for $I = 3$ treatments, $J = 2$ blocks, and $n = 1$ observation for each of 6 combinations of levels. The subjects (blocks) were matched by gender and mail to reduce the influence of external factors.

(b) Consider the Friedman test for $I = 3$ treatments and $J = 2$ blocks. The test statistic

$$q = 2 \sum_{i=1}^{3} (\bar{r}_{i.} - 2)^2$$

is obtained from the ranks given by two subjects (r_{ij}) to the three treatments. There are $6^2 = 36$ possible rank combinations:

$$(r_{ij}) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \end{pmatrix}, \ldots, \begin{pmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 2 & 3 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 1 \end{pmatrix}$$

Under the null hypothesis, these 36 combinations are equally likely. The corresponding vector of rank averages $(\bar{r}_{1.}, \bar{r}_{2.}, \bar{r}_{3.})$ may take 5 different values (up to permutations)

$$A_1 = (1, 2, 3), A_2 = (1, 2.5, 2.5), A_3 = (1.5, 1.5, 3), A_4 = (1.5, 2, 2.5), A_5 = (2, 2, 2)$$

according to the following table

	1, 2, 3	1, 3, 2	2, 1, 3	2, 3, 1	3, 1, 2	3, 2, 1
1, 2, 3	A_1	A_2	A_3	A_4	A_4	A_5
1, 3, 2	A_2	A_1	A_4	A_3	A_5	A_4
2, 1, 3	A_3	A_4	A_1	A_5	A_2	A_4
2, 3, 1	A_4	A_3	A_5	A_1	A_4	A_2
3, 1, 2	A_4	A_5	A_2	A_4	A_1	A_3
3, 2, 1	A_5	A_4	A_4	A_2	A_3	A_1

From this table, we find the possible outcomes of interest and their probabilities due to the random sampling assuming that the null hypothesis is true:

$(\bar{r}_{1.}, \bar{r}_{2.}, \bar{r}_{3.}) =$	A_1	A_2	A_3	A_4	A_5
$q =$	4	3	3	1	0
probability =	1/6	1/6	1/6	1/3	1/6

Thus the null distribution of Q is the following one

$$\mathrm{P}(Q = 0) = 1/6, \quad \mathrm{P}(Q = 1) = 1/3, \quad \mathrm{P}(Q = 3) = 1/3, \quad \mathrm{P}(Q = 4) = 1/6.$$

Solutions to Section 9.5

Solution 1

We will test the null hypothesis of homogeneity

$$H_0\text{: equal genotype frequencies for diabetics and normal groups}$$

using the chi-squared test. The table below gives the expected counts along with the observed counts:

	diabetic	normal	total
Bb or *bb*	12 (7.85)	4 (8.15)	16
BB	39 (43.15)	49 (44.85)	88
total	51	53	104

From here the chi-squared test statistic is computed to be x^2=5.10. With df=1, the p-value of the test = 0.024 is found using the normal distribution table. We reject H_0 and conclude that the diabetics have the genotype BB less often than the normals.

Solution 2

Here we twice apply the chi-squared test of independence.

(a) We test H_0: no association of the disease and the ABO blood group, using the observed and expected counts:

	O	A	AB	B	Total
Moderate	7 (10.4)	5 (9.8)	3 (2.0)	13 (6.2)	28
Minimal	27 (30.4)	32 (29.7)	8 (6.1)	18 (18.8)	85
Not present	55 (48.6)	50 (47.5)	7 (9.8)	24 (30.0)	136
Total	89	87	18	55	249

With x^2 = 15.37 and df = 6, the chi-squared distribution table says that the p-value of the test is less than 2.5%. We reject H_0.

(b) We test H_0: no association of the disease and the MN blood group, using the observed and expected counts:

	MM	MN	NN	Total
Moderate	21 (16.7)	6 (9.4)	1 (1.9)	28
Minimal	54 (51.3)	27 (28.9)	5 (5.8)	86
Not present	74 (81.1)	51 (45.7)	11 (9.2)	136
Total	149	84	17	250

With x^2 = 4.73, df = 4, the chi-squared distribution table says that the p-value of the test is larger than 10%. We can not reject H_0.

Solution 3

(a) Applying the chi-squared test of homogeneity:

	girl	boy	total
flying fighter	51 (45.16)	38 (43.84)	89
flying transport	14 (15.22)	16 (14.78)	30
not flying	38 (42.62)	46 (41.38)	84
total	103	100	203

we get x^2 = 2.75, df = 2, giving the p-value larger than 10%. Conclusion: the data does not confirm the difference in boy proportion fathered by different professionals.

(b) We apply the chi-squared test for proportions using this example to illustrate how it works. Form the null hypothesis

$$H_0\text{: boys proportions } \pi_{12} = \pi_{22} = \pi_{32} = 0.513,$$

based on the population proportion 0.513 obtained as

$$\frac{105.37}{105.37 + 100} = 0.513.$$

In this case the observed and expected counts are computed as

	girl	boy	total
flying fighter	51 (43.34)	38 (45.66)	89
flying transport	14 (14.61)	16 (15.39)	30
not flying	38 (40.91)	46 (43.09)	84
Total	103	100	203

and we get $\mathrm{x}^2 = 3.09$, df $= 3$, giving the p-value larger than 10%. Conclusion: we can not reject H_0.

Why do we use df $= 3$ deserves an explanation. The general model is described by three independent parameters $(\pi_{12} = \pi_{22} = \pi_{32})$, this gives us

$$\dim \Omega = 3$$

degrees of freedom to start with. Since the null hypothesis model is simple, we get

$$\dim \Omega_0 = 0,$$

and the resulting number of degrees of freedom becomes

$$\text{df } = 3 - 0 = 3.$$

Solution 4

We use the chi-squared test for homogeneity involving five samples.

	no nausea	incidence of nausea	total
placebo	70 (84)	95 (81)	165
chlorpromazine	100 (78)	52 (74)	152
dimenhydrinate	33 (43)	52 (42)	85
pentobarbital (100 mg)	32 (34)	35 (33)	67
pentobarbital (150 mg)	48 (43)	37 (42)	85
total	283	271	554

The observed test statistic $\mathrm{x}^2 = 35.8$ according to the χ^2_4-distribution table gives the p-value smaller than 0.005. Comparing the observed and expected counts we conclude that chlorpromazine is most effective in ameliorating postoperative nausea.

Solution 5

(a) We test H_0: no relation between blood group and disease in London, using the chi-squared test of homogeneity.

	control	peptic ulcer	total
group A	4219 (4103.0)	579 (695.0)	4798
group O	4578 (4694.0)	911 (795.0)	5489
total	8797	1490	10287

The test statistic $\mathrm{x}^2 = 42.40$, with df $= 1$, according to the normal distribution table gives very small p-value $= 0.000$. We reject H_0. The odds ratio

$$\hat{\Delta} = \frac{4219 \cdot 911}{4578 \cdot 579} = 1.45$$

says that blood group O increases the odds of peptic ulcer by factor 1.45 compared to group A.

(b) We test H_0: no relation between blood group and disease in Manchester, using the chi-squared test of homogeneity.

	control	peptic ulcer	total
group A	3775 (3747.2)	246 (273.8)	4021
group O	4532 (4559.8)	361 (333.2)	4893
total	8307	607	8914

The test statistic $\mathrm{x}^2 = 5.52$, with df $= 1$, according to the normal distribution table gives the p-value $= 0.019$. We reject H_0. The odds ratio fo the Manchester data is $\hat{\Delta} = 1.22$.

(c) We test H_0: London group A and Manchester group A have the same propensity to Peptic Ulcer, using the chi-squared test of homogeneity.

	control	peptic ulcer	total
London group A	4219 (4349.2)	579 (448.8)	4798
Manchester group A	3775 (3644.8)	246 (376.2)	4021
total	7994	825	8819

The test statistic $\mathrm{x}^2 = 91.3$, with df = 1, gives a very small p-value = 0.000. We reject H_0.

We test H_0: London Group O and Manchester Group O have the same propensity to Peptic Ulcer, using the chi-squared test of homogeneity.

	control	peptic ulcer	total
London group O	4578 (4816.5)	911 (672.5)	5489
Manchester group O	4532 (4293.5)	361 (599.5)	4893
total	9110	1272	10382

The test statistic $\mathrm{x}^2 = 204.5$, with df = 1, gives a very small p-value = 0.000. We reject H_0.

Solution 6

Denote the two main factors of the study by

D = endometrical carcinoma,

X = estrogen taken at least 6 months prior to the diagnosis of cancer.

(a) For this retrospective case-control study with matched controls, we have the following counts.

	$\bar{D}X$	$\bar{D}\bar{X}$	total
DX	39	113	152
$D\bar{X}$	15	150	165
total	54	263	317

Applying the McNemar test we find that the test statistic

$$\mathrm{x}^2 = \tfrac{(113-15)^2}{113+15} = 75$$

is highly significant as $\sqrt{75} = 8.7$ and the corresponding two-sided p-value obtained from N(0,1) table is very small. Conclusion: since 113 is significantly larger than 15, taking estrogen increases the risk of endometrical carcinoma.

(b) Possible weak points in a retrospective case-control design

- selection bias: some patients have died prior the study,
- information bias: have to rely on other sources of information.

Solution 7

The exact Fisher test uses Hg(30,17,$\frac{16}{30}$) as the null distribution of the test statistic whose observed value is $x = 12$. We will test

$$H_0 : \text{no difference between high-anxiety and low-anxiety groups}$$

against the two-sided alternative

$$H_1 : \text{there is a difference between high-anxiety and low-anxiety groups.}$$

Personally, I do not see a natural choice of the one-sided alternative.

Next, we compute the two-sided p-value of the Fisher test using the normal approximation of the hypergeometric distribution:

$$\mathrm{Hg}(30, 17, \tfrac{16}{30}) \approx \mathrm{N}(9.1, 1.4).$$

With help of the continuity correction, we find the one-sided p-value to be

$$\mathrm{P}(X \geq 12|H_0) = \mathrm{P}(X > 11|H_0) \approx 1 - \Phi(\tfrac{11.5-9.1}{1.4}) = 1 - \Phi(1.71) = 0.044.$$

Thus the two-sided p-value is approximately 9% and we do not reject the null hypothesis.

It is interesting to compare this result with the chi-squared test yields, which in this case, gives the test statistic $\mathrm{x}^2 = 4.69$. With df = 1, the two-sided p-value of the chi-squared test

$$2(1 - \Phi(\sqrt{4.69})) = 2(1 - \Phi(2.16)) = 0.03$$

is smaller than 5% demonstrating a bad substitute of the exact Fisher test due to a small sample size.

Solution 8

Denote

π_1 = probability that red wins in boxing,
π_2 = probability that red wins in freestyle wrestling,
π_3 = probability that red wins in Greco-Roman wrestling,
π_4 = probability that red wins in Tae Kwon Do.

Is there evidence that wearing red is more favourable in some of the sports than others? We test

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 \quad \text{vs} \quad H_1 : \pi_i \neq \pi_j \quad \text{for some } i \neq j$$

using the chi-squared test of homogeneity. From

	Red	Biue	Total
Boxing	148 (147)	120 (121)	268
Freestyle wrestling	27 (28)	24 (23)	51
Greco-Roman wrestling	25 (26)	23 (22)	48
Tae Kwon Do	45 (44)	35 (36)	80
Total	245	202	447
Marginal proportions	0.55	0.45	1.00

we find that the test statistic $\mathrm{x}^2 = 0.3$ with df = 5 is not significant. We can not reject H_0. We also see that the red outfit gives an advantage over the blue one as $\hat{\pi} = 0.55$.

Solution 9

After asking the opinion of 50 female employees and 50 male employees, the company's statistician should carry out a chi-square test of homogeneity.

Solution 10

(a) Multiple testing problem.

(b) Exact Fisher's test.

(c) Nonparametric tests do not assume a particular form of the population distribution like normal distribution.

Solution 11

The null hypothesis is that everybody votes independently. Let p be the population proportion for 'yes'. Then the number of 'yes' for three voters in a household has the binomial distribution model $X \sim \mathrm{Bin}(3, p)$ with an unspecified parameter p. Thus the null hypothesis can be expressed in the following form

$$H_0 : p_0 = (1-p)^3,\ p_1 = 3p(1-p)^2,\ p_2 = 3p^2(1-p),\ p_3 = p^3.$$

We apply the Pearson chi-squared test with the expected counts based on the maximum likelihood estimate of p, the sample proportion $\hat{p} = 0.5417$:

$$e_0 = n(1-\hat{p})^3 = 19,\ e_1 = 3n\hat{p}(1-\hat{p})^2 = 68,\ e_2 = 3n\hat{p}^2(1-\hat{p}) = 81,\ e_3 = 3n\hat{p}^3 = 32.$$

The observed chi-square test statistic is $\mathrm{x}^2 = 11.8$ which has the p-value less than 0.5% according to the approximate null distribution χ^2_{df} with df $= 4 - 1 - 1 = 2$. Conclusion: we reject the null hypothesis of independent voting.

Solution 12

The equivalence of

$$H_0 : \pi_{i|j} = \pi_{i\cdot}$$

and

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

follows from the formula for the conditional probabilities

$$\pi_{i|j} = \frac{\pi_{ij}}{\pi_{\cdot j}}.$$

Solution 13

(a) This is a single sample of size $n = 441$. Each of n observations falls in of 9 groups. The multinomial distribution model

$$(C_{11}, C_{12}, C_{13}, C_{21}, C_{22}, C_{23}, C_{31}, C_{32}, C_{33}) \sim \mathrm{Mn}(n, p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33})$$

gives the likelihood function

$$\begin{aligned}
&L(p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}) \\
&= P(C_{11} = 24, C_{12} = 15, C_{13} = 17, C_{21} = 52, C_{22} = 73, C_{23} = 80, C_{31} = 58, C_{32} = 86, C_{33} = 36) \\
&= \frac{441!}{24!15!17!52!73!80!58!86!36!} p_{11}^{24} \cdot p_{12}^{15} \cdot p_{13}^{17} \cdot p_{21}^{52} \cdot p_{22}^{73} \cdot p_{23}^{80} \cdot p_{31}^{58} \cdot p_{32}^{86} \cdot p_{33}^{36}.
\end{aligned}$$

(b) The null hypothesis of independence $H_0 : p_{ij} = p_{i.} \cdot p_{.j}$ meaning that there is no relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline.

(c) The chi-squared test statistic $\mathrm{x}^2 = 22.5$ computed from the observed and expected counts should be compared with the critical values of the χ^2_4-distribution. Since it is larger than 14.86, we conclude that the p-value of the test is smaller than 0.5%. We reject the null hypothesis of independence and infer that there is a relationship between the facility conditions at gasoline stations and aggressiveness in the pricing of gasoline.

	Pricing policy			
	Aggressive	Neutral	Nonaggressive	Total
Substandard condition	24 (17)	15 (22)	17 (17)	56
Standard condition	52 (62.3)	73 (80.9)	80 (61.8)	205
Modern condition	58 (54.7)	86 (71)	36 (54.3)	180
Total	134	174	133	441

It looks like the standard conditions are coupled with the least aggressive pricing strategy.

Solution 14

(a) The risk ratio compares the chances to suffer from myocardial infarction under the aspirin treatment versus the chances to suffer from myocardial infarction under the placebo treatment:

$$RR = \frac{\mathrm{P(MyoInf|Aspirin)}}{\mathrm{P(MyoInf|Placebo)}}.$$

(b) The null hypothesis of $RR = 1$ is equivalent to the hypothesis of homogeneity.

	MyoInf	No MyoInf	Total
Aspirin	104 (146.5)	10933 (10887.5)	11037
Placebo	189 (146.5)	10845 (10887.5)	11034
Total	293	21778	22071

The corresponding chi-squared test statistic is

$$\mathrm{x}^2 = \frac{42.5^2}{146.5} + \frac{42.5^2}{146.5} + \frac{42.5^2}{10887.5} + \frac{42.5^2}{10887.5} = 25.$$

Since df = 1, we can use the normal distribution table. The square root of 25 is 5 making the result highly significant. Aspirin works!

Solutions to Section 10.5

Solution 1

Recall that the formulas for sample and population covariances:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad \mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

We have to check that

$$\mathrm{E}(S_{XY}) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

To this end, observe that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y},$$

and

$$n^2\bar{x}\bar{y} = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i y_i + \sum_{i\neq j}\sum_{j=1}^{n} x_i y_j,$$

so that

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n(n-1)}\sum_{i\neq j}\sum_{j=1}^{n} X_i Y_j.$$

It remains to see that

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i Y_i\right) = n\mathrm{E}(XY), \qquad \mathrm{E}\left(\sum_{i\neq j}\sum_{j=1}^{n} X_i Y_j\right) = n(n-1)\mathrm{E}(X)\mathrm{E}(Y).$$

Solution 2

After ordering over the x values we get

x	-1.75	-1.18	-0.88	-0.65	-0.30	0.34	0.50	0.68	1.38	1.40
y	-1.59	-0.81	-0.98	-0.53	-0.72	0.27	0.64	0.35	1.34	1.28

From this sample we compute the following five sufficient statistics

$$\bar{x} = -0.046, \quad \bar{y} = -0.075, \quad s_x = 1.076, \quad s_y = 0.996, \quad r = 0.98.$$

(a) If x is the predictor and y is the response, then the regression line takes the form

$$y - \bar{y} = r \cdot \tfrac{s_y}{s_x}(x - \bar{x}),$$

yielding

$$y = -0.033 + 0.904 \cdot x.$$

The estimated noise size is

$$s = \sqrt{\tfrac{n-1}{n-2}s_y^2(1-r^2)} = 0.22.$$

(b) If y is the predictor and x is the response, then the regression line takes the form

$$x - \bar{x} = r \cdot \tfrac{s_x}{s_y}(y - \bar{y}),$$

yielding

$$x = 0.033 + 1.055 \cdot y.$$

The estimated noise size is

$$s = \sqrt{\tfrac{n-1}{n-2}s_x^2(1-r^2)} = 0.24.$$

(c) The first fitted line

$$y = -0.033 + 0.904 \cdot x$$

is different from the second

$$y = -0.031 + 0.948 \cdot x,$$

because in (a) we minimise the vertical residuals while in (b) we minimise the horizontal residuals.

Solution 3

Using an extra explanatory variable f which equal 1 for females and 0 for males, we combine two simple linear regression models into the multiple regression

$$y = \beta_0 + \beta_1 x + \beta_2 f + \sigma Z,$$

so that

$$\beta_0 = \beta_0'', \quad \beta_2 = \beta_0' - \beta_0''.$$

Here $p = 3$ and the design matrix is

$$\mathbb{X} = \begin{pmatrix} 1 & x_1 & f_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & f_n \end{pmatrix}.$$

After $\beta_0, \beta_1, \beta_2$ are estimated, we can compare the original intercepts β_0' and β_0'' using the relations

$$\beta_0'' = \beta_0, \quad \beta_0' = \beta_0 + \beta_2.$$

We may also test the key null hypothesis $\beta_2 = 0$ saying that the grades of the female and male students satisfy the same simple linear regression model.

Solution 4

Using $\mathbb{P} = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal$, we get

$$\mathbb{P}^2 = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal\mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal = \mathbb{X}(\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal = \mathbb{P}.$$

Solution 5

(a) Given $x = 95$, we predict the final score by

$$\hat{y} = 75 + 0.5(95 - 75) = 85,$$

based on the formula for the predicted response

$$\frac{\hat{y} - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}.$$

Regression to mediocrity in action: since the predictor $x = 95$ is larger than the average 75, the predicted response value 85 is smaller than 95.

(b) Given $y = 85$, we predict the midterm score by

$$\hat{x} = 75 + 0.5(85 - 75) = 80,$$

based on the formula for the predicted response

$$\frac{\hat{x} - \bar{x}}{s_x} = r \cdot \frac{y - \bar{y}}{s_y}.$$

Again, regression to mediocrity.

Solution 6

(a) First, we find the correlation coefficient ρ between X and Y. Since $\mathrm{E}X = 0$, we have

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) = \mathrm{E}(X^2 + \beta XZ) = 1, \quad \mathrm{Var}Y = \mathrm{Var}X + \beta^2\mathrm{Var}Z = 1 + \beta^2,$$

implying that the correlation coefficient is always positive

$$\rho = \tfrac{1}{\sqrt{1+\beta^2}}.$$

(b) To generate pairs (x, y) with a given positive correlation coefficient ρ, one can proceed as follows. From $\rho = \frac{1}{\sqrt{1+\beta^2}}$, find

$$\beta = \sqrt{\rho^{-2} - 1}.$$

After generating a pair (x, z) of $\mathrm{N}(0, 1)$ numbers, the pair (x, y) with

$$y = x + z\sqrt{\rho^{-2} - 1}$$

will produce the required result.

Solution 7

Two regression models

$$y = -62.05 + 3.49 \cdot x, \quad r^2 = 0.984,$$
$$\sqrt{y} = -0.88 + 0.2 \cdot x, \quad r^2 = 0.993,$$

produce a higher coefficient of determination. The second model has a slightly better fit to the data. The kinetic energy formula explains why the second model might be better.

Solution 8

Coefficient of determination is the squared sample correlation coefficient

$$r^2 = (0.2)^2 = 0.04.$$

Solution 9

Using the formulas for b_1 and s_{b_1} we obtain

$$t = \frac{b_1}{s_{b_1}} = \frac{\frac{rs_y}{s_x}}{\sqrt{\frac{s^2}{(n-1)s_x^2}}} = \frac{rs_y}{\sqrt{\frac{s_y^2(1-r^2)}{(n-2)}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Solution 10

(a) The corresponding multiple regression model takes the form $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$, where e_i, $i = 1, \ldots, 5$ are independent realisations of a normal random variable with distribution $\mathrm{N}(0, \sigma)$. The corresponding design matrix is

$$\mathbb{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \end{pmatrix}$$

(b) Using the formula $\hat{y}_i = 111.8857 + 8.0643x_i - 1.8393x_i^2$ we get

x_i	0	2	4	6	8
y_i	110	123	119	86	62
$\hat{y}_i$	111.8857	120.6571	114.7143	94.0571	58.6857

implying

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p} = \frac{103.3}{2} = 51.65.$$

(c) The coefficient of multiple determination equals

$$R^2 = 1 - \frac{\mathrm{ss_E}}{\mathrm{ss_T}} = 1 - \frac{103.3}{2630} = 0.961.$$

Solution 11

(a) The underlying parabola makes unrealistic prediction that $\hat{y}_{40} = 139$ sec compared to $\hat{y}_{10} = 63$ sec and $\hat{y}_{20} = 34$ sec. One should be careful to extend the range of explanatory variable from that used in the data.

Empirical relationship developed in a region might break down,
if extrapolated to a wider region in which no data been observed

(b) Using $t_{17}(0.005) = 2.898$ we get the exact confidence interval (under the assumption of normality and homoscedasticity for the noise component)

$$I_\mu = 0.2730 \pm 2.898 \cdot 0.1157 = (-0.0623, 0.6083).$$

(c) Since the confidence interval from 2b covers zero, we do not reject the null hypothesis $H_0 : \beta_2 = 0$ at the 1% significance level. The observed t-test statistic $\frac{0.2730}{0.1157} = 2.36$, and according to the t_{17}-distribution table the two-sided p-value lies between 2% and 5%.

Solution 12

(b) The fitted regression line for the final score y as a function of the midterm score x is $y = 37.5 + 0.5x$. Given $x = 90$ we get a point prediction $y = 82.5$. The estimate of σ^2 is

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2) = 84.4.$$

Thus the 95% prediction interval for Carl's final score is

$$I = 82.5 \pm t_8(0.025) s \sqrt{1 + \tfrac{1}{9} + \tfrac{1}{8}(\tfrac{15}{10})^2} = 82.5 \pm 24.6.$$

(c) The fitted regression line for the midterm score x as a function of the final score y is $x = 37.5 + 0.5y$. Given $y = 80$ we get a point prediction $x = 77.5$.