Introduction to Statistical Inference

Serik Sagitov, Chalmers University of Technology and Gothenburg University

Abstract

This text is a compendium for the undergraduate course "Statistical Inference" worth of 7.5 hp, which is a second course in mathematical statistics suitable for students with different backgrounds. A main prerequisite is an introductory course in probability and statistics. The course gives 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, multiple regression.

The compendium includes a collection of solved exercises many of which are the end-of-chapter exercises from the book by John Rice, Mathematical statistics and data analysis, 3rd edition. Do not read a solution before you tried to solve an exercise on your own. Please send your corrections to serik@chalmers.se.

Last updated: March 11, 2022

Contents

Abstract **1**

1 Normal theory parametric models **4**
1.1 Normal distribution . . . 4
1.2 Mixture of normal distributions . . . 4
1.3 One-way layout model . . . 5
1.4 Two-way layout model . . . 5
1.5 Multiple regression model . . . 6

2 Discrete parametric models **6**
2.1 Binomial distribution . . . 6
2.2 Poisson distribution . . . 6
2.3 Hypergeometric distribution . . . 7
2.4 Multinomial distribution . . . 7

3 Random sampling **7**
3.1 Point estimation . . . 8
3.2 Sample mean and sample variance . . . 8
3.3 Finite population correction . . . 9
3.4 Approximate confidence intervals . . . 9
3.5 Dichotomous data . . . 10
3.6 Stratified random sampling . . . 10
3.7 Exercises . . . 11

4 Parameter estimation **13**
4.1 Method of moments . . . 13
4.2 Geometric and exponential distributions . . . 13
4.3 Maximum likelihood estimation . . . 14
4.4 Sufficiency . . . 15
4.5 Large sample properties of the maximum likelihood estimates . . . 16
4.6 Gamma distribution . . . 17
4.7 Exact confidence intervals . . . 18
4.8 Exercises . . . 19

5 Hypothesis testing **22**
5.1 Statistical significance . . . 22
5.2 Large-sample test for the proportion . . . 23
5.3 Small-sample test for the proportion . . . 24
5.4 Two tests for the mean . . . 24
5.5 Likelihood ratio test . . . 24
5.6 Chi-squared test of goodness of fit . . . 25
5.7 Case study: sex ratio . . . 26
5.8 Exercises . . . 27

6 Bayesian inference **29**
6.1 Conjugate priors . . . 30
6.2 Bayesian estimation . . . 32
6.3 Credibility interval . . . 33
6.4 Bayesian hypotheses testing . . . 33
6.5 Exercises . . . 34

7 Summarising data **35**
7.1 Empirical probability distribution . . . 35
7.2 Density estimation . . . 37
7.3 Quantiles and QQ-plots . . . 38
7.4 Testing normality . . . 39
7.5 Measures of location . . . 40
7.6 Measures of dispersion . . . 41
7.7 Exercises . . . 42

8 Comparing two samples **43**
8.1 Two independent samples: comparing population means . . . 43
8.2 Two independent samples: comparing population proportions . . . 46
8.3 Paired samples . . . 47
8.4 Paired samples: comparing population proportions . . . 49
8.5 External and confounding factors . . . 50
8.6 Exercises . . . 51

9 Analysis of variance **53**
9.1 One-way layout . . . 54
9.2 Simultaneous confidence interval . . . 56
9.3 Kruskal-Wallis test . . . 57
9.4 Two-way layout . . . 57
9.5 Case study: iron retention . . . 58
9.6 Randomised block design . . . 60
9.7 Friedman test . . . 61
9.8 Exercises . . . 62

10 Categorical data analysis **63**
10.1 Chi-squared test of homogeneity . . . 64
10.2 Chi-squared test of independence . . . 65
10.3 Matched-pairs designs . . . 65
10.4 Odds ratios . . . 67
10.5 Exercises . . . 68

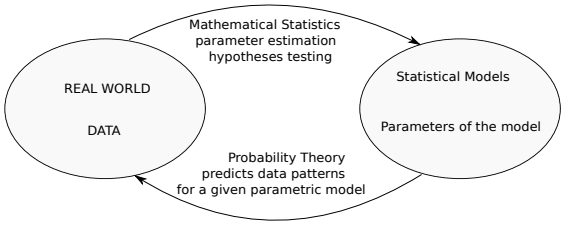
11 Multiple regression **70**
11.1 Simple linear regression model . . . 70
11.2 Residuals . . . 72
11.3 Confidence intervals and hypothesis testing . . . 73
11.4 Intervals for individual observations . . . 74
11.5 Multiple linear regression . . . 74
11.6 Exercises . . . 77

12 Course topics and distribution tables **79**
12.1 List of course topics . . . 79
12.2 Normal distribution table . . . 80
12.3 Critical values of the t-distribution . . . 82
12.4 Critical values of the chi square distribution . . . 83
12.5 Critical values of the F-distribution . . . 84

13 Solutions to exercises **86**
13.1 Solutions to Section 3 (random sampling) . . . 86
13.2 Solutions to Section 4 (parameter estimation) . . . 89
13.3 Solutions to Section 5 (hypothesis testing) . . . 95
13.4 Solutions to Section 6 (Bayesian inference) . . . 100
13.5 Solutions to Section 7 (summarising data) . . . 103
13.6 Solutions to Section 8 (two samples) . . . 105
13.7 Solutions to Section 9 (analysis of variance) . . . 110
13.8 Solutions to Section 10 (categorical data analysis) . . . 114
13.9 Solutions to Section 11 (multiple regression) . . . 119

14 Miscellaneous exercises **123**
14.1 Problems . . . 123
14.2 Numerical answers to miscellaneous exercises . . . 131

Statistical analysis consists of three parts: collection of data, summarising data, and making inferences. The graph below presents



the relationship between two sister branches of mathematics: probability theory and mathematical statistics. Example of probability versus statistics:

> PROBABILITY. Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with 99.99% chances.
>
> STATISTICS. 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.

The main focus of this course called "Statistical Inference" is on the issues of parameter estimation and hypothesis testing using properly collected, relatively small data sets. Special attention, therefore, is paid to the basic principles of experimental design: randomisation, blocking, and replication.

1 Normal theory parametric models

A statistical model represents a complicated process by a simple mathematical relationship governed by few parameters, plus random noise. A classical example is the simple linear regression model

$$Y(x) = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \text{N}(0, \sigma).$$

Here the response variable Y is a linear function of the main explanatory factor x plus noise ϵ.

1.1 Normal distribution

A key parametric statistical model is the standard normal distribution $\text{N}(0, 1)$ whose cumulative distribution function

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-y^2/2} dy$$

is computed using either the table in Section 12.2 or the R-command

```
> pnorm(1:5)
```

which gives the values of $\Phi(z)$ for $z = 1, 2, 3, 4, 5$:

```
[1] 0.8413447 0.9772499 0.9986501 0.9999683 0.9999997
```

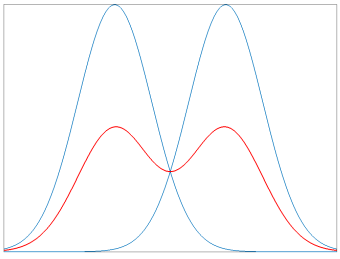
The importance of the normal model is due to the Central Limit Theorem (CLT) as well as the remarkable analytical properties of the normal density function. The CLT states that the sum of many independent and relatively small random contributions is approximately normally distributed.
We write $\epsilon \sim \text{N}(0, \sigma)$ if the random variable ϵ has a normal distribution with mean zero and standard deviation σ. In view of the CLT, it is natural to take ϵ as a model of the random noise since the noise is an accumulation of all minor independent factors neither of which having a dominating effect of the response variable. Notice that the standard deviation σ plays the role of a scale parameter of the normal distribution in that $\frac{\epsilon}{\sigma} \sim \text{N}(0, 1)$. This explains why we refer to σ as the size of the noise.
A random response variable Y having a normal distribution $\text{N}(\mu, \sigma)$ with mean μ and the standard deviation σ can be viewed as the sum

$$Y = \mu + \epsilon, \quad \epsilon \sim \text{N}(0, \sigma)$$

of a constant signal μ and a noise ϵ.

1.2 Mixture of normal distributions

A motivating example for the mixture model is the hight of people in a mixed population with two strata, women and men. Let $\text{N}(\mu_1, \sigma_1)$ be the distribution of women's heights and $\text{N}(\mu_2, \sigma_2)$ be the distribution of men's heights. Then the mixed population distribution describes the two step random experiment: first toss a coin for choosing index i to be either 1 or 2, then generate a value using $\text{N}(\mu_i, \sigma_i)$. A mixture of two bell curves may result in a "camel curve" as illustrated below (red line).



Suppose we are given k normally distributed random variables

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

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

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

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

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

The expression for σ^2 is due to the total variance formula which recognises two sources of variation for Y:

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

1.3 One-way layout model

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

$$Y(i) = \mu_i + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma), \quad i = 1, \ldots, I.$$

It is useful to write

$$\mu_i = \mu + \alpha_i, \quad \alpha_i = \mu_i - \mu, \quad \sum_{i=1}^{I} \alpha_i = 0,$$

where μ stands for overall mean and α_i represents the effect of the main factor at the level i, so that the total effect is zero by definition.

1.4 Two-way layout model

In the case of two main (categorical) factors A having I different levels, and B having J different levels,

$$Y(i,j) = \mu_{ij} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

we write

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

where μ is the overall mean, α_i is the main effect of the factor A at the level i, and β_j is the main effect of the factor B at the level j, so that

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

The term δ_{ij} defined by

$$\delta_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j$$

to describe the interaction between the two main factors, is such that

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

In general, at different combinations of levels the two factors may interact either negatively or positively. In the simple case with all $\delta_{ij} = 0$, there is no interaction and the main factors contribute additively:

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

Example: pay gap

Response variable is the salary of a person. Factor A is person's sex having $I = 2$ levels: $i = 1$ for a female and $i = 2$ for a male. Factor B is a 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 effect size α_1 represents the pay gap for women.

1.5 Multiple regression model

A model based on a linear relationship between $(p-1)$ predictors and the response variable

$$Y(x_1,\ldots,x_{p-1}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1} + \epsilon, \quad \epsilon \sim \mathrm{N}(0,\sigma).$$

This is a flexible setting, it for example covers the one-way layout model with $I = p$ as a special case. Indeed, in the framework of the one-way layout model we can define parameters $(\beta_0,\ldots,\beta_{p-1})$ by

$$\begin{aligned}
\mu_1 &= \beta_0,\\
\mu_2 &= \beta_0 + \beta_1,\\
\mu_3 &= \beta_0 + \beta_2,\\
&\ldots\\
\mu_p &= \beta_0 + \beta_{p-1},
\end{aligned}$$

and introduce dummy variables $(x_1,\ldots,x_{p-1})$ taking values 0 and 1. It is easy to see that setting the levels of the main factor of the one-way layout model as

$$\begin{aligned}
&\text{level } 1: \quad (x_1,\ldots,x_{p-1}) = (0,0,0,\ldots,0,0)\\
&\text{level } 2: \quad (x_1,\ldots,x_{p-1}) = (1,0,0,\ldots,0,0)\\
&\text{level } 3: \quad (x_1,\ldots,x_{p-1}) = (0,1,0,\ldots,0,0)\\
&\qquad\ldots\\
&\text{level } p: \quad (x_1,\ldots,x_{p-1}) = (0,0,0,\ldots,0,1)
\end{aligned}$$

we arrive at a particular example of the multiple regression model.

Example: person's height

Response variable Y is the height of a person, x_1 is the height of the person's mother, x_2 is the height of the person's father, $x_3 = 1$ if the person is female and $x_3 = 0$ if the person is male.

2 Discrete parametric models

Discrete parametric models of this section are often used in connection with categorical data analysis.

2.1 Binomial distribution

Binomial distribution $X \sim \mathrm{Bin}(n,p)$:

$$\mathrm{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \le k \le n, \quad \mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p).$$

A binomially distributed variable X is the sum of n independent random variables each having a Bernoulli distribution $\mathrm{Bin}(1,p)$. This leads to the important example of the CLT giving the normal approximation for the binomial distribution. With a *continuity correction* the claim is that for $X \sim \mathrm{Bin}(n,p)$,

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

The rule of thumb says

$\mathrm{Bin}(n,p) \approx \mathrm{N}(np, \sqrt{np(1-p)})$, if both $np \ge 5$ and $n(1-p) \ge 5$.

2.2 Poisson distribution

Poisson distribution $X \sim \mathrm{Pois}(\mu)$:

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

is an approximation of the $\mathrm{Bin}(n,p)$ distribution as

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

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

2.3 Hypergeometric distribution

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 $N = B + W$ balls, of which $B = Np$ balls are black and $W = N(1-p)$ balls are white:

$$\begin{aligned} \mathrm{P}(X = k) &= \frac{\binom{B}{k}\binom{W}{n-k}}{\binom{N}{n}}, \quad \max(0, n-W) \le k \le \min(n, B), \\ \mathrm{E}(X) &= np, \\ \mathrm{Var}(X) &= np(1-p)(1-\tfrac{n-1}{N-1}). \end{aligned}$$

Compared to the variance of the binomial distribution the last formula contains the factor $1 - \frac{n-1}{N-1}$ which is called the finite population correction. Despite the dependence between the drawings without replacement, there is a normal approximation for the hypergeometric distribution:

$$\mathrm{Hg}(N, n, p) \approx \mathrm{N}\Big(np, \sqrt{np(1-p)\tfrac{N-n}{N-1}}\Big), \text{ if both } np \ge 5 \text{ and } n(1-p) \ge 5.$$

2.4 Multinomial distribution

Multinomial distribution $(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$ is an extension of the binomial distribution

$$\mathrm{P}(X_1 = k_1, \ldots, X_r = k_r) = \binom{n}{k_1, \ldots, k_r} p_1^{k_1} \ldots p_r^{k_r}.$$

It corresponds to n independent trials with r possible outcomes labeled by $i = 1, \ldots, r$. If each trial has distribution $(p_1, \ldots, p_r)$, then X_i gives the number of trials with the outcome labeled by i. So that

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

The marginal distributions are binomial $X_i \sim \mathrm{Bin}(n; p_i)$, and the different counts are negatively correlated:

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

3 Random sampling

Statistical inference is the process of using data analysis for inferring the properties of a population distribution with help of a random sample drawn from the population in question. A finite population of size N can be viewed as a set of N elements characterised by values $\{x_1, x_2, \ldots, x_N\}$.

> If we pick at random one element from the population, then its value x is a realisation of a random variable X whose distribution is the population distribution.

In many situations finding a population distribution by enumeration is either very expensive or even impossible. However, a good guess is available using a random sample of n observations $(x_1, \ldots, x_n)$. Such a sample will be treated as a single realisation of a random vector $(X_1, \ldots, X_n)$. If the sampling experiment is repeated, the new values $(x_1', \ldots, x_n')$ usually will be different from $(x_1, \ldots, x_n)$.

> *Randomisation* is a guard against investigator's biases even unconscious.

Two important characteristics of the population distribution are the population mean and population standard deviation

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

This applies to the quantitative (continuous or discrete) data. If the data is categorical, then one may translate it to numbers. An important special case of categorical data is dichotomous data. Consider the example of $x_i \in \{\text{male, female}\}$. After converting categorical values to a quantitative form with $x_i \in \{0, 1\}$, the population distribution becomes a Bernoulli distribution $\mathrm{Bin}(1, p)$ with the parameter

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

called a population proportion.
There are two basic ways of random sampling:

1. sampling without replacement produces a so called simple random sample (finite population),
2. sampling with replacement produces what we call, a random sample (infinite population case).

Notice that the second approach produces the random variables $(X_1, \ldots, X_n)$ which are independent and identically distributed. Therefore, a random sample is easier to analyse then the simple random sample resulting in dependent observations. Importantly, if n/N is small, then the two approaches are almost indistinguishable.

Example: in class experiment

Suppose we collect data on students heights and gender using a two colour histogram. The collected hight values form a random sample taken from the population of Gothenburg students. Motivating questions:

- can this group be viewed as a simple random sample?
- what is the shape of the population distribution of heights?
- what is an estimate of population proportion of women?

3.1 Point estimation

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. Sampling distribution of $\widehat{\Theta}$ around unknown θ: different values of $\hat{\theta}$ will be observed for different samples. The sampling distribution has mean and variance

$$\mu_{\widehat{\Theta}} = \mathrm{E}(\widehat{\Theta}), \quad \sigma^2_{\widehat{\Theta}} = \mathrm{E}(\widehat{\Theta} - \mu_{\widehat{\Theta}})^2.$$

The quality of the the point estimator $\widehat{\Theta}$ is measured by the mean square error

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

The mean square error has two components involving

$\mu_{\widehat{\Theta}} - \theta$ is the size of systematic error, bias, lack of accuracy,
$\sigma_{\widehat{\Theta}}$ is the size of the random error, lack of precision.

Desired properties of point estimates:

$\hat{\theta}$ is an unbiased estimate of θ, that is $\mu_{\widehat{\Theta}} = \theta$,
$\hat{\theta}$ is a consistent estimate, in that the mean square error

$$\mathrm{E}(\widehat{\Theta} - \theta)^2 \to 0 \text{ as the sample size } n \to \infty.$$

The standard error for an estimator $\widehat{\Theta}$ is its standard deviation $\sigma_{\widehat{\Theta}} = \sqrt{\mathrm{Var}(\widehat{\Theta})}$.

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

3.2 Sample mean and sample variance

For a given random sample $(x_1, \ldots, x_n)$, the most basic summary statistics are the sample mean and sample variance

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

An alternative formula for the sample variance

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

In the same way as $(x_1, \ldots, x_n)$ is a realisation of a random vector $(X_1, \ldots, X_n)$, the summary statistics $\bar{x}$ and s^2 are realisation of random variables

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

Consider a random sample. The sample mean $\bar{x}$ and sample variance s^2 are unbiased and consistent estimators for the population mean μ and variance σ^2 respectively

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

Notice that the sample standard deviation s systematically underestimates the population standard deviation σ since

$$\mathrm{E}(S^2) = \sigma^2 \quad \text{and} \quad (\mathrm{E}(S))^2 < \mathrm{E}(S^2), \quad \text{so that} \quad \mathrm{E}(S) < \sigma.$$

Estimated standard error for the sample mean $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

3.3 Finite population correction

Now consider *simple random sampling*, when there is dependence between observations. In this case the sample mean $\bar{x}$ is again an unbiased and consistent estimate for the population mean:

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

Here N is the finite population size and the extra factor

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

can be called a finite population correction. It reflects the negative dependence due to sampling without replacement. However, the sample variance s^2 is a biased estimate of σ^2, since

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

This is verified by

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

Since

$$\sigma_{\bar{X}}^2 = \tfrac{\sigma^2}{n}(1 - \tfrac{n-1}{N-1}),$$

in the light of the previous equality we get an unbiased estimate of $\mathrm{Var}(\bar{X})$ to be

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

Thus, for the sampling without replacement, the formula for the estimated standard error of $\bar{X}$ for the simple random sample takes the new form

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

3.4 Approximate confidence intervals

By the Central Limit Theorem, the sample mean distribution is approximately normal

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

in that for large sample sizes n, we have

$$\mathrm{P}(|\tfrac{\bar{X}-\mu}{\sigma_{\bar{X}}}| > z) \approx 2(1 - \Phi(z)).$$

Since $S_{\bar{X}} \approx \sigma_{\bar{X}}$, we derive that

$$\mathrm{P}(\bar{X} - zS_{\bar{X}} < \mu < \bar{X} + zS_{\bar{X}}) = \mathrm{P}(|\tfrac{\bar{X}-\mu}{S_{\bar{X}}}| > z) \approx \mathrm{P}(|\tfrac{\bar{X}-\mu}{\sigma_{\bar{X}}}| > z) \approx 2(1 - \Phi(z)).$$

This yields the following formula of an approximate 100(1–α)% two-sided confidence interval for μ:

$$I_\mu \approx \bar{x} \pm z_{\alpha/2} \cdot s_{\bar{x}},$$

where z_α stands for the normal quantile. These two formulas are valid even for the sampling without replacement due to a more advanced version of the central limit theorem. According to the normal distribution table we have

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

The higher is confidence level the wider is the confidence interval. On the other hand, the larger is sample the narrower is the confidence interval.
The exact meaning of the confidence level is a bit tricky. For example, a 95% confidence interval is a random interval, such that out of 100 intervals I_μ computed for 100 samples, only 95 are expected cover the true value of μ. Notice that the random number of successful realisations of the confidence interval has distribution Bin(100,0.95) which is approximately normal with mean $\mu = 95$ and standard deviation $\sigma = 2.2$.

3.5 Dichotomous data

We will pay a special attention to the dichotomous case, when the population distribution is a Bernoulli distribution $\text{Bin}(1, p)$, so that

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

In this case, the sample values x_i are either 0 or 1, and the sample mean turns into a sample proportion $\hat{p} = \bar{x}$ giving an unbiased and consistent estimate of p. In the dichotomous case, $x_i^2 = x_i$ and therefore

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

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

For the sampling without replacement, the formula for the estimated standard errors of $\hat{p}$ for the simple random sample take the new form

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

An approximate 100(1–α)% two-sided confidence interval for p:

$$I_p \approx \hat{p} \pm z_{\alpha/2} \cdot s_{\hat{p}}.$$

3.6 Stratified random sampling

Given additional information on population structure, one can reduce the sampling error using the method of stratified sampling. Assume that a population consists of k strata with known strata fractions $(w_1, \ldots, w_k)$ such that

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

(Example: Swedish population consists of two strata, females and males with $k = 2$ and $w_1 = w_2 = 0.5$.) In terms of unknown strata means and standard deviations

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

we have the following expressions for the population mean and variance

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

where the last equality is due to the total variance formula, with

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

being the average variance. Stratified random sampling consists of taking k independent iid-samples from each stratum with sample sizes $(n_1, \ldots, n_k)$ and sample means $\bar{x}_1, \ldots, \bar{x}_k$.

Stratified sample mean: $\bar{x}_\text{s} = w_1\bar{x}_1 + \ldots + w_k\bar{x}_k$

Observe that for any allocation $(n_1, \ldots, n_k)$, the stratified sample mean is an unbiased estimate of μ

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

The variance of $\bar{X}_\text{s}$

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

is estimated by

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

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

Approximate confidence interval $I_\mu \approx \bar{x}_\text{s} \pm z_{\alpha/2} \cdot s_{\bar{x}_\text{s}}$

Optimisation problem: allocate $n = n_1 + \ldots + n_k$ observations among different strata to minimise the sampling error of $\bar{x}_{\mathrm{s}}$.

$$\boxed{\text{Optimal allocation: } n_j = n\frac{w_j\sigma_j}{\bar{\sigma}}, \text{ gives the minimum variance } \mathrm{Var}(\bar{X}_{\mathrm{so}}) = \frac{1}{n}\cdot\bar{\sigma}^2}$$

Here $\bar{\sigma}^2$ is the squared average standard deviation

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

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 knowledge of the standard deviations σ_j. If σ_j are unknown, then a sensible choice is to allocate observations proportionally to the strata sizes.

$$\boxed{\text{Proportional allocation: } n_j = nw_j,\ \mathrm{Var}(\bar{X}_{\mathrm{sp}}) = \frac{1}{n}\cdot\overline{\sigma^2}}$$

Comparing three unbiased estimates of the population mean we see that their variances are ordered in the following way

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

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

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

Variability in μ_j across strata makes proportional allocation more effective than iid-sample

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

Difference between the proportional and random allocation

Observe that with $n_i = nw_i$, we get

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

However, this is not the mean of a truly random sample, since the n observations are forcefully allocated among the 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 $\mathrm{Mn}(n, w_1, \ldots, w_k)$.

3.7 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 sample of size 2 by generating all possible such samples. From them, find the mean and variance of the sampling distribution, and compare the results to those obtained by the formulas from this section.

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?

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 a 90% confidence interval for μ? How should k_2 be chosen so that

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

is a 95% confidence interval for μ?

Problem 4

Warner (1965) introduced the method of randomised response to deal with surveys asking sensitive questions. Suppose we want to estimate the proportion q of illegal drug users among prison inmates. We are interested in the population as a whole - not in punishing particular individuals. Randomly chosen n inmates have responded yes/no to a randomised statement (after rolling a die):

"I use heroin" (with probability 5/6)
"I do not use heroin" (with probability 1/6).

Suggest a probability model for this experiment, find a method of moments estimate for q and its standard error.

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.) Compute the bias of this point estimate.

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

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

(c) What are the population mean and variance?

(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 the 70 samples.

Problem 8

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

(a) A survey of household expenditures 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.

(d) A survey of programs on a television station, taken to estimate the proportion of time taken up by advertising on Monday through Friday from 6 pm until 10 pm. Assume that 52 weeks of recorded broadcasts are available for the analysis.

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 1.4.

4 Parameter estimation

Given a parametric model with unknown parameter(s) θ, we wish to estimate θ from a random sample. 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 a simple intuitive method, that also can be used for finding an initial value for the maximum likelihood method, which is preferable for large samples.

4.1 Method of moments

Suppose we are given a random sample $(x_1, \ldots, x_n)$ from a population distribution characterised by a pair of parameters (θ_1, θ_2). Suppose we have the following formulas for the first and second population moments:

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

Method of moments estimates $(\tilde{\theta}_1, \tilde{\theta}_2)$ are found after replacing the population moments with the corresponding sample moments, as a solution of the obtained equations

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

This approach is justified by the Law of Large Numbers telling that

$$\tfrac{X_1+\ldots+X_n}{n} \to \mu, \quad \tfrac{X_1^2+\ldots+X_n^2}{n} \to \mathrm{E}(X^2), \quad n \to \infty.$$

4.2 Geometric and exponential distributions

We say $X \sim \mathrm{Geom}(p)$ if

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

so that

$$\mathrm{E}(X) = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

The geometric random variable X counts the number of Bernoulli trials until the first success is observed. From

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

its easy to see that as $p \to 0$, the distribution of pX converges to the exponential distribution Exp(1).

We say $T \sim \mathrm{Exp}(\lambda)$ if the random variable T has density

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0,$$

which implies

$$\mathrm{E}(T) = \frac{1}{\lambda}, \quad \mathrm{Var}(T) = \frac{1}{\lambda^2}.$$

and

$$\mathrm{P}(T > t) = e^{-\lambda t}, \quad t \geq 0.$$

Example: geometric model

Consider the dataset summarising a random sample $(x_1, \ldots, x_n)$ of the hop counts for $n = 130$ birds, where

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

Number of hops j	1	2	3	4	5	6	7	8	9	10	11	12	Tot
Observed frequency O_j	48	31	20	9	6	5	4	2	1	1	2	1	130

The observed frequency O_j is the number of birds, who hopped j times before flying up

$$O_j = 1_{\{x_1=j\}} + \ldots + 1_{\{x_n=j\}}.$$

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

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

By inspecting the histogram of the data values, we observe geometrically descending frequencies suggesting a geometric model for the number of jumps for a random bird. Geometric model $X \sim \text{Geom}(p)$ assumes 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. 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

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

We can compute an approximate 95% confidence interval for p using the above mentioned I_μ:

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

To answer the question of how well does the geometric distribution fit the data, let us compare the observed frequencies to the frequencies expected from the geometric distribution with parameter $\tilde{p}$:

j	1	2	3	4	5	6	7+
O_j	48	31	20	9	6	5	11
E_j	46.5	29.9	19.2	12.3	7.9	5.1	9.1

Expected frequencies e_j are computed in terms of independent geometric random variables $(X_1, \ldots, X_n)$

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

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

$$\text{X}^2 = \sum_{j=1}^{7} \tfrac{(O_j - E_j)^2}{E_j} = 1.86.$$

As it will be explained later on, the observed small value test statistic allows us to conclude that the geometric model fits the data well.

4.3 Maximum likelihood estimation

In a parametric setting, given a parameter value θ, the observed sample $(x_1, \ldots, x_n)$ is a realisation of the random vector $(X_1, \ldots, X_n)$ which has a certain joint distribution

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

as a function of possible values $(y_1, \ldots, y_n)$. Fixing the variables $(y_1, \ldots, y_n) = (x_1, \ldots, x_n)$ and allowing the parameter value θ to vary, we obtain the so-called likelihood function

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

Notice, that the likelihood function usually is not a density function over θ. To illustrate this construction, draw three density curves for three parameter values $\theta_1 < \theta_2 < \theta_3$, then show how for a given observed value x, the likelihood curve connects the three points on the plane

$$(\theta_1, f(x|\theta_1)), \quad (\theta_2, f(x|\theta_2)), \quad (\theta_3, f(x|\theta_3)).$$

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

Example: binomial model

Consider the binomial distribution model $X \sim \text{Bin}(n, p)$, with a single observation corresponding to n observations in the Ber(p) model. From $\mu = np$, we see that the method of moment estimate

$$\tilde{p} = \tfrac{x}{n}$$

is the sample proportion. To maximise the likelihood function

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

we can maximise the log-likelihood function

$$l(p) = \ln L(p) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p).$$

Take its derivative and solving the equation $l'(p) = 0$

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p} = 0,$$

we find that the MLE of population proportion is the sample proportion $\hat{p} = \frac{x}{n}$.

4.4 Sufficiency

Suppose there is a 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 $\propto$ means proportional. 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: Bernoulli distribution model

For a single Bernoulli trial with probability of success p, we have

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

and for n independent Bernoulli trials,

$$f(x_1, \ldots, x_n | p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{n\bar{x}} (1-p)^{n-n\bar{x}},$$

where $\prod$ stands for the product. This implies that for the Bernoulli model, the number of successes

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

is a sufficient statistic whose distribution is $T \sim$Bin(n, p).

Example: normal distribution model

The two-parameter normal distribution model N(μ, σ) has a two-dimensional sufficient statistic (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

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

4.5 Large sample properties of the maximum likelihood estimates

For a random sample $(x_1, \ldots, x_n)$ taken from a parametric population distribution $f(x|\theta)$, the likelihood function is given by the product

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

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

$$\boxed{\text{Normal approximation } \hat{\Theta} \approx \mathrm{N}(\theta, \tfrac{1}{\sqrt{n\mathbb{I}(\theta)}}), \text{ as } n \gg 1}$$

where $\mathbb{I}(\theta)$ is the Fisher information in a single observation defined as follows. The larger is

$$g(x, \theta) = -\tfrac{\partial^2}{\partial \theta^2} \ln f(x|\theta)$$

at the top of the log-likelihood curve , the more information on the parameter θ is contained at a single observation x. The Fisher information in a single observation is the average curvature

$$\mathbb{I}(\theta) = \mathrm{E}[g(X, \theta)] = \int g(x, \theta) f(x|\theta) dx.$$

The larger information $n\mathbb{I}(\theta)$ is in n observations, the smaller is the asymptotic variance of $\hat{\Theta}$.

$$\boxed{\text{Approximate } 100(1-\alpha)\% \text{ confidence interval } I_\theta \approx \hat{\theta} \pm \tfrac{z_{\alpha/2}}{\sqrt{n\mathbb{I}(\hat{\theta})}}}$$

It turns out that the maximum likelihood estimators are asymptotically unbiased, consistent, and asymptotically efficient (have minimal variance) in the following sense.

$$\boxed{\text{Cramer-Rao inequality: if } \theta^* \text{ is an unbiased estimator of } \theta, \text{ then } \mathrm{Var}(\Theta^*) \geq \tfrac{1}{n\mathbb{I}(\theta)}.}$$

Example: exponential model

For lifetimes of five batteries measured in hours

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

we propose an exponential model $X \sim \mathrm{Exp}(\theta)$, where θ is the battery death rate per hour. Method of moment estimate: from $\mu = 1/\theta$, we find

$$\tilde{\theta} = \tfrac{1}{\bar{x}} = \tfrac{5}{28.5} = 0.175.$$

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

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

is biased since

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

but asymptotically unbiased since

$$\mathrm{E}(\widehat{\Theta}) \approx \theta$$

for large samples. The latter holds due to the Law of Large Numbers $\bar{X} \approx \mu$.
Fisher information for the exponential model is easy to compute:

$$g(x, \theta) = -\tfrac{\partial^2}{\partial \theta^2} \ln f(x|\theta) = \tfrac{1}{\theta^2}, \qquad \mathbb{I}(\theta) = \mathrm{E}[g(X, \theta)] = \tfrac{1}{\theta^2}.$$

This yields

$$\mathrm{Var}(\widehat{\Theta}) \approx \tfrac{\theta^2}{n}$$

and we get an approximate 95% confidence interval

$$I_\theta \approx 0.175 \pm 1.96 \cdot \tfrac{0.175}{\sqrt{5}} = 0.175 \pm 0.153.$$

4.6 Gamma distribution

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

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

involves the gamma function

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

which is an extension of the function $(\alpha - 1)!$ to non-integer α, in that

$$\Gamma(k) = (k-1)! \quad \text{for } k = 1, 2, \ldots$$

It brings the mean and variance values

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

The gamma distribution model is more flexible than the normal distribution model as for different values of the shape parameter α the density curves have different shapes. If $\alpha = 1$, then

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

If $\alpha = k$ is integer, and $X_i \sim \mathrm{Exp}(\lambda)$ are independent, then

$$X_1 + \ldots + X_k \sim \mathrm{Gam}(k, \lambda).$$

Parameter λ does not influence the shape of the density influencing only the scaling of the random variable. It is easy to see that

$$X \sim \mathrm{Gam}(\alpha, \lambda) \Rightarrow \lambda X \sim \mathrm{Gam}(\alpha, 1).$$

Normal approximation for the gamma distribution:

$$\mathrm{Gam}(\alpha, \lambda) \approx \mathrm{N}(\tfrac{\alpha}{\lambda}, \tfrac{\alpha}{\lambda^2}), \qquad \alpha \gg 1.$$

Turning to the likelihood function

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

where

$$(t_1, t_2) = (x_1 + \ldots + x_n, x_1 \cdots x_n)$$

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. Then we solve numerically the system of two equations

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

using the method of moment estimates $(\tilde{\alpha}, \tilde{\lambda})$ as the initial values.

Example: male heights

We illustrate the gamma model by applying it to a male height sample of size $n = 24$ given below in an 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 would like to estimate the parameters (λ, α) using the data. First, compute two sample moments

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

To apply the method of moments we use formulas

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

Replacing here $\mathrm{E}X$ by $\bar{x}$ and $\mathrm{E}(X^2)$ by $\overline{x^2}$, we get

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

or

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

which give the method of moments estimates

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

Mathematica command

$$\text{FindRoot[Log[a] == 0.00055+Gamma}'[a]\text{/Gamma[a], \{a, 907.3\}]}$$

gives the maximum likelihood estimates

$$\hat{\alpha} = 908.76, \quad \hat{\lambda} = 5.01.$$

4.7 Exact confidence intervals

If we put a restrictive assumption on the population distribution and assume that an iid-sample $(x_1, \ldots, x_n)$ is taken from a normal distribution $\mathrm{N}(\mu, \sigma)$ with unspecified parameters μ and σ, then instead of the approximate confidence interval formula for the mean we may apply an exact confidence interval formula based on the following probability theory fact. If random variables $X_1, \ldots, X_n$ are independent and have the same distribution $\mathrm{N}(\mu, \sigma)$, then

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} \sim t_{n-1}$$

has the so-called t-distribution with $n - 1$ degrees of freedom. Here

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

stands for the random variable whose realisation is $s_{\bar{x}}$.

Exact $100(1-\alpha)\%$ confidence interval for the mean $I_\mu = \bar{x} \pm t_{n-1}(\frac{\alpha}{2}) \cdot s_{\bar{x}}$

Here if $n = 10$ and $\alpha = 0.05$, then $t_{n-1}(\frac{\alpha}{2}) = t_9(0.025)$ is found as a quantile of the t-distribution from the table in Section 12.3: row df $= 9$ and column 0.025 give $t_9(0.025) = 2.262$.
Using R-command

```
> qt(0.995, c(10,20,30))
```

we confirm the values in the table for $t_k(0.005)$ with $k = 10, 20, 30$:

```
[1] 3.169273 2.845340 2.749996
```

A t_k-distribution curve looks similar to N(0,1)-curve. Its density function is symmetric around zero:

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

It has a larger spread than the standard normal distribution. If the number of degrees of freedom $k \geq 3$, then the variance is $\frac{k}{k-2}$. Connection to the standard normal distribution: if $Z, Z_1, \ldots, Z_k$ are N(0,1) and independent, then

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

Let $\alpha = 0.05$. The exact confidence interval for μ is wider than the approximate confidence interval $\bar{x} \pm 1.96 \cdot s_{\bar{x}}$ valid for the very large n. For example

$$I_\mu = \bar{x} \pm 2.26 \cdot s_{\bar{x}} \text{ for } n = 10, \qquad I_\mu = \bar{x} \pm 2.13 \cdot s_{\bar{x}} \text{ for } n = 16,$$
$$I_\mu = \bar{x} \pm 2.06 \cdot s_{\bar{x}} \text{ for } n = 25, \qquad I_\mu = \bar{x} \pm 2.00 \cdot s_{\bar{x}} \text{ for } n = 60.$$

Moreover, in the N(μ, σ) case we get access to an exact confidence interval formula for the variance thanks to the following result.

$$\boxed{\text{Exact distribution } \tfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}}$$

The chi-squared distribution with k degrees of freedom is the gamma distribution with $\alpha = \frac{k}{2}, \lambda = \frac{1}{2}$. It is connected to the standard normal distribution as follows: if $Z_1, \ldots, Z_k$ are N(0,1) and independent, then

$$Z_1^2 + \ldots + Z_k^2 \sim \chi^2_k.$$

$$\boxed{\text{Exact } 100(1-\alpha)\% \text{ confidence interval } I_{\sigma^2} = \Big(\tfrac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}; \tfrac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\Big)}$$

Here the function $\chi^2_k(\alpha)$ stands for the quantiles of the chi-square distribution summarised by the table in Section 12.4. The exact confidence interval for σ^2 is non-symmetric. Examples of 95% confidence intervals for σ^2:

$$I_{\sigma^2} = (0.47s^2, 3.33s^2) \text{ for } n = 10,$$
$$I_{\sigma^2} = (0.55s^2, 2.40s^2) \text{ for } n = 16,$$
$$I_{\sigma^2} = (0.61s^2, 1.94s^2) \text{ for } n = 25,$$
$$I_{\sigma^2} = (0.72s^2, 1.49s^2) \text{ for } n = 60.$$

$$\boxed{\text{Under the normality assumption } \mathrm{Var}(S^2) = \tfrac{2\sigma^4}{n-1}, \text{ estimated standard error for } s^2 \text{ is } \sqrt{\tfrac{2}{n-1}}s^2.}$$

4.8 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. Fit a Poisson distribution. Comment on 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.

n	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

Problem 2

One of the earliest applications of the Poisson distribution was made by Student (1907) in studying errors made in counting yeast cells.In this study, yeast cells were killed and mixed with water and gelatin; the mixture was then spread on a glass and allowed to cool. Four different concentrations were used. Counts were made on 400 squares, and the data are summarised in the following table:

Number of cells	Concent. 1	Concent. 2	Concent. 3	Concent. 4
0	213	103	75	0
1	128	143	103	20
2	37	98	121	43
3	18	42	54	53
4	3	8	30	86
5	1	4	13	70
6	0	2	2	54
7	0	0	1	37
8	0	0	0	18
9	0	0	1	10
10	0	0	0	5
11	0	0	0	2
12	0	0	0	2

(a) Estimate the parameter λ for each of the four sets of data.

(b) Find an approximate 95% confidence interval for each estimate.

(c) Compare observed and expected counts.

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 $\theta \in [0,1]$ is parameter. The following 10 independent observations were taken from such a distribution:

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

(a) Find the method of moments estimate of θ.

(b) Find an approximate standard error for your estimate.

(c) What is the maximum likelihood estimate of θ?

(d) What is an approximate standard error of the maximum likelihood estimate?

Problem 4

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

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

(b) Show that $\hat{p} = \frac{x}{n}$ attains the Cramer-Rao lower bound.

(c) If $n = 10$ and $X = 5$, plot the log-likelihood function.

Problem 5

A company has manufactured certain objects and has printed a serial number on each manufactured 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

Capture-recapture method for estimating the number N of fish living in a lake follows the two-step procedure:

1. capture and tag say $n = 100$ fish, then release them in the lake,
2. recapture say $k = 50$ fish and count the number of tagged fish.

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

Problem 7

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?

Problem 8

Let $X_1, \ldots, X_n$ be independent random variables uniformly distributed on $[0, \theta]$.

(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 estimate 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	$x_1 = 1997$
Starchy white	$x_2 = 906$
Sugary green	$x_3 = 904$
Sugary white	$x_4 = 32$

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 maximum likelihood estimate of θ and its asymptotic variance.

(b) For an approximate 95% confidence interval for θ based on part (a).

5 Hypothesis testing

5.1 Statistical significance

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

null hypothesis H_0: the effect of interest is zero,

alternative H_1: the effect of interest is not zero.

H_0 represents an established theory that must be discredited in order to demonstrate some effect H_1.
A decision rule for hypotheses 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 $\mathcal{R}$ and

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

Making a decision we can commit type I or type II error, see the table:

	Negative decision: do not reject H_0	Positive decision: reject H_0 in favor 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

Four important conditional probabilities:

$\alpha = \mathrm{P}(T \in \mathcal{R}|H_0)$ significance level of the test, conditional probability of type I error,
$1 - \alpha = \mathrm{P}(T \notin \mathcal{R}|H_0)$ specificity of the test,
$\beta = \mathrm{P}(T \notin \mathcal{R}|H_1)$ conditional probability of type II error,
$1 - \beta = \mathrm{P}(T \in \mathcal{R}|H_1)$ sensitivity of the test or power.

If test statistic and sample size are fixed, one can not make smaller both α and β by changing $\mathcal{R}$.

A significance test tries to control the type I error:

1. fix an appropriate significance level α, commonly used significance levels are 5%, 1%, 0.1%,
2. find $\mathcal{R}$ from $\alpha = \mathrm{P}(T \in \mathcal{R}|H_0)$ using the null distribution of the test statistic T.

P-value of the test

A p-value is the probability of obtaining a test statistic value as extreme or more extreme than the observed one, given that H_0 is true. For a given significance level α,

$$\text{reject } H_0, \text{ if p-value} \leq \alpha, \text{ and do not reject } H_0, \text{ if p-value} > \alpha.$$

Observe that the p-value depends on the data 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). Suppose the null hypothesis is true. Given $z_{\text{obs}} = z$, the p-value is

$$\mathrm{p} = \mathrm{P}(Z > z) = 1 - \Phi(z),$$

and in terms of the random variables

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

Now, under H_0 the observed value for different samples has distribution N(0,1) so that

$$\mathrm{P}(\mathrm{P} > \mathrm{p}) = \mathrm{P}(1 - \Phi(Z_{\text{obs}}) > 1 - \Phi(z)) = \mathrm{P}(\Phi(Z_{\text{obs}}) < \Phi(z)) = \mathrm{P}(Z_{\text{obs}} < z) = \Phi(z) = 1 - \mathrm{p},$$

and we conclude that the P-value has a uniform null distribution.

5.2 Large-sample test for the proportion

Binomial model $X \sim \text{Bin}(n, p)$. The corresponding sample proportion $\hat{p} = \frac{x}{n}$.

For H_0: $p = p_0$ use the test statistic $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

Three different composite alternative hypotheses:

$$\text{one-sided } H_1\text{: } p > p_0, \quad \text{one-sided } H_1\text{: } p < p_0, \quad \text{two-sided } H_1\text{: } p \neq p_0.$$

By the central limit theorem, the null distribution of the Z-score is approximately normal: $Z \overset{a}{\sim} \text{N}(0,1)$
The corresponding rejection region depends on the alternative hypothesis

Alternative H_1	Rejection rule	P-value
$p > p_0$	$z \geq z_\alpha$	$\text{P}(Z \geq z_{\text{obs}})$
$p < p_0$	$z \leq -z_\alpha$	$\text{P}(Z \leq z_{\text{obs}})$
$p \neq p_0$	$z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	$2 \cdot \text{P}(Z \geq \lvert z_{\text{obs}} \rvert)$

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

Power function

Consider two simple hypotheses

$$H_0\text{: } p = p_0 \text{ and } H_1\text{: } p = p_1, \text{ assuming } p_1 > p_0.$$

The power function of the one-sided test can be computed using the normal approximation for $\frac{X - np_1}{\sqrt{np_1(1-p_1)}}$ under H_1:

$$\begin{aligned}
\text{Pw}(p_1) &= \text{P}\Big(\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \geq z_\alpha | H_1\Big) \\
&= \text{P}\Big(\frac{X - np_1}{\sqrt{np_1(1-p_1)}} \geq \frac{z_\alpha\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}} | H_1\Big) \\
&\approx 1 - \Phi\big(\frac{z_\alpha\sqrt{p_0(1-p_0)} + \sqrt{n}(p_0 - p_1)}{\sqrt{p_1(1-p_1)}}\big).
\end{aligned}$$

Planning of sample size: given α and β, choose sample size n such that

$$\sqrt{n} = \frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}}{|p_1 - p_0|}.$$

Example: extrasensory perception

An experiment: guess the suits of $n = 100$ cards chosen at random with replacement from a deck of cards with four suits. Binomial model: the number of cards guessed correctly $X \sim \text{Bin}(100, p)$. Hypotheses of interest

$$H_0 : p = 0.25 \text{ (pure guessing)}, \quad H_1 : p > 0.25 \text{ (ESP ability)}.$$

Rejection rule at 5% significance level

$$\mathcal{R} = \{\tfrac{\hat{p} - 0.25}{0.0433} \geq 1.645\} = \{\hat{p} \geq 0.32\} = \{x \geq 32\}.$$

With a simple alternative $H_1 : p = 0.30$ the power of the test is

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

The sample size required for the 90% power is

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

Suppose the observed sample count is $x_{\text{obs}} = 30$, then

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

and the one-sided p-value becomes

$$\text{P}(Z \geq 1.15) = 12.5\%.$$

The result is not significant, we do not reject H_0.

5.3 Small-sample test for the proportion

Binomial model $X \sim \text{Bin}(n, p)$ with H_0: $p = p_0$. For small n, use exact null distribution $X \sim \text{Bin}(n, p_0)$.

Example: extrasensory perception

ESP test: guess the suits of $n = 20$ cards. Model: the number of cards guessed correctly is

$$X \sim \text{Bin}(20, p).$$

For $H_0 : p = 0.25$, the null distribution of the test statistic x is

Bin(20,0.25) table

x	8	9	10	11
$\text{P}(X \geq x)$	.101	.041	.014	0.004

For the one-sided alternative $H_1 : p > 0.25$ and $\alpha = 5\%$, the rejection rule is $\mathcal{R} = \{x \geq 9\}$. Notice that the exact significance level = 4.1%.

Power function

p	0.27	0.30	0.40	0.5	0.60	0.70
$\text{P}(X \geq 9)$	0.064	0.113	0.404	0.748	0.934	0.995

5.4 Two tests for the mean

We wish to test H_0: $\mu = \mu_0$ against either the two-sided or a one-sided alternative for continuous or discrete data. As the test statistic we use the t-score

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

Large-sample test for mean is used when the population distribution is not necessarily normal but the sample size n is sufficiently large. Compute the rejection region using an approximate null distribution

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

One-sample t-test is used for small n, under the assumption that the population distribution is normal. Compute the rejection region using an exact null distribution

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

Confidence interval method of hypotheses testing

Observe that at significance level α the rejection rule can be expressed

$$\mathcal{R} = \{\mu_0 \notin I_\mu\}$$

in terms of a 100(1-α)% confidence interval for the mean. Having such confidence interval, reject H_0: $\mu = \mu_0$ if the interval does not cover the value μ_0.

5.5 Likelihood ratio test

A general method of finding asymptotically optimal tests (having the largest power for a given α) takes 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 $\Lambda = \frac{L(\theta_0)}{L(\theta_1)}$ as a test statistic. Large values of Λ suggest that H_0 explains the data set better than H_1, while a small Λ indicates that H_1 explains the data set better. Likelihood ratio test rejects H_0 for small values of Λ.

Neyman-Pearson lemma: the likelihood ratio test is optimal in the case of two simple hypothesis.

Nested hypotheses

With a pair of nested parameter sets $\Omega_0 \subset \Omega$ we get two composite alternatives

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

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 $\theta \in \Omega$.

Generalised likelihood ratio test rejects H_0 for small values of

$$\Lambda = \tfrac{L(\hat{\theta}_0)}{L(\hat{\theta})},$$

or equivalently for large values of

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

It turns out that the test statistic $-2\ln\Lambda$ has a nice approximate null distribution

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

5.6 Chi-squared test of goodness of fit

Suppose that each of n independent observations belongs 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

$$(O_1, \ldots, O_J) \sim \mathrm{Mn}(n; p_1, \ldots, p_J), \qquad \mathrm{P}(O_1 = k_1, \ldots, O_J = k_J) = \frac{n!}{k_1! \cdots k_J!} p_1^{k_1} \cdots p_J^{k_J}.$$

Consider a parametric model for the data

$$H_0 : (p_1, \ldots, p_J) = (v_1(\lambda), \ldots, v_J(\lambda)) \text{ with unknown parameters } \lambda = (\lambda_1, \ldots, \lambda_r).$$

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

$$E_j = n \cdot v_j(\hat{\lambda}),$$

where "expected" means expected under the null hypothesis model. In the current setting, the likelihood ratio test statistic

$$-2\ln\Lambda \approx \mathrm{X}^2$$

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

$$\mathrm{X}^2 = \sum_{j=1}^{J} \frac{(O_j - E_j)^2}{E_j}.$$

The approximate null distribution of the chi-squared test statistic is χ^2_{J-1-r}, since

$$\dim(\Omega_0) = r \quad \text{and} \quad \dim(\Omega) = J - 1,$$

where dim stands for dimension or the number of independent parameters. A mnemonic rule for the number of degrees of freedom:

$$\text{df} = \text{(number of cells)} - 1 - \text{(number of independent parameters estimated from the data)}.$$

Since the chi-squared test is approximate, 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

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

Using $\hat{p} = 0.358$ and $J = 7$ we obtain $\mathrm{X}^2 = 1.86$. With df $= 5$ we find p-value $= 0.87$, therefore we conclude a good fit of the geometric distribution model to the data.

5.7 Case study: sex ratio

A 1889 study made in Germany recorded the numbers of boys $(y_1, \ldots, y_n)$ for $n = 6115$ families with 12 children each. Consider three nested models for the distribution of the number of boys Y in a family with 12 children

$$\begin{aligned} \text{Model 1: } & Y \sim \text{Bin}(12,\, 0.5) \\ & \cap \\ \text{Model 2: } & Y \sim \text{Bin}(12,\, \text{p}) \\ & \cap \\ \text{General model: } & p_j = \text{P}(Y = j), \quad j = 0, 1, \ldots, 12. \end{aligned}$$

Model 1 leads to a simple null hypothesis

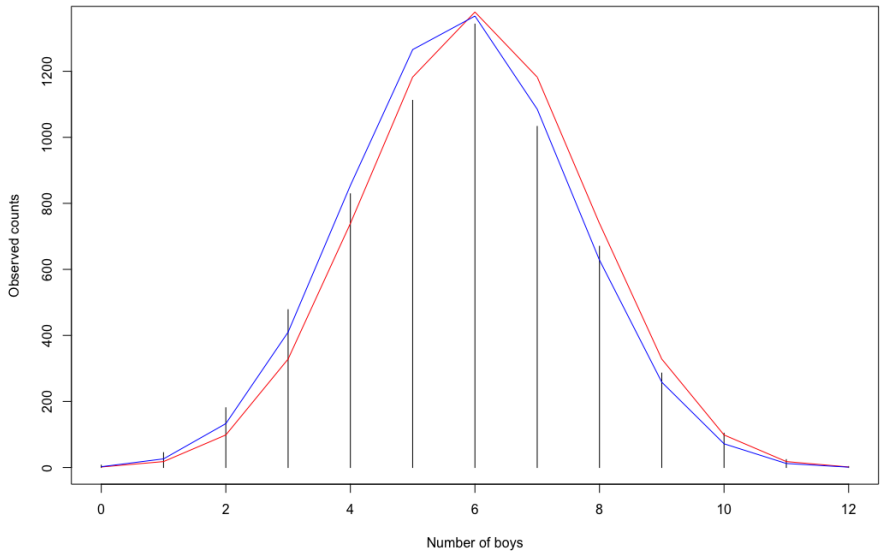
$$H_0 : p_j = \binom{12}{j} \cdot 2^{-12}, \quad j = 0, 1, \ldots, 12.$$

The expected cell counts

$$E_j = 6115 \cdot \binom{12}{j} \cdot 2^{-12}, \quad j = 0, 1, \ldots, 12,$$

are summarised in the table below. Observed chi-squared test statistic $\text{X}^2 = 249.2$, $\text{df} = 12$. Since $\chi^2_{12}(0.005) = 28.3$, we reject H_0 at 0.5% level.

cell j	O_j	Model 1: E_j	and $\frac{(O_j - E_j)^2}{E_j}$	Model 2: E_j	and $\frac{(O_j - E_j)^2}{E_j}$
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	$\text{X}^2 = 249.2$	6115	$\text{X}^2 = 110.5$



Model 2 is more flexible and leads to a composite null hypothesis

$$H_0 : p_j = \binom{12}{j} \cdot p^j (1-p)^{12-j}, \quad j = 0, \ldots, 12, \quad 0 \le p \le 1.$$

The expected cell counts

$$E_j = 6115 \cdot \binom{12}{j} \cdot \hat{p}^j \cdot (1-\hat{p})^{12-j}$$

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

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

The observed chi-squared test statistic is $\mathrm{X}^2 = 110.5$. Since $r = 1$, $\mathrm{df} = 11$, and the table value is $\chi^2_{11}(0.005) = 26.76$, we reject even model 2 at 0.5% level.
We see that what is needed is an even more flexible model addressing large variation in the observed cell counts. Suggestion: allow the probability of a male child p to differ from family to family. Namely, assume that for each family the value p is generated by a Beta-distribution, see Section 6.1.

5.8 Exercises

Problem 1

Suppose that $X \sim \mathrm{Bin}(100, p)$. 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 answer the following:

(a) What is α?

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

Problem 2

Let X have 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, Λ, for each x_i and order the x_i according to Λ.

(b) What is the likelihood ratio test of H_0 versus H_1 at level $\alpha = 0.2$? What is the test at level $\alpha = 0.5$?

Problem 3

Let $(x_1, \ldots, x_n)$ be a sample from a Poisson distribution. Find the likelihood ratio for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$. Use the fact that the sum of independent Poisson random variables follows a Poisson distribution to explain how to determine a rejection region for a test at 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 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 a level α test?

(c) What is the power of the test?

Problem 6

An iid-sample from $\mathrm{N}(\mu, \sigma)$ gives a 99% confidence interval for μ to be $(-2, 3)$. Test

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

at $\alpha = 0.01$.

Problem 7

Let $(x_1, \ldots, x_{15})$ 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

Binomial model for the data value x:

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

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

(b) Show that the test rejects for large values of $|x - \frac{n}{2}|$.

(c) How the significance level corresponding to the rejection region

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

can be determined?

(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 if $n = 100$ and $k = 10$.

Problem 9

Suppose that a test statistic Z has a standard normal null-distribution.

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

(b) Answer the same question if the test rejects for large 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

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

Plato et al. (1964) published the following data on haptoglobin type in a sample of 190 people

Genotype	Hp 1-1	Hp 1-2	Hp 2-2
Observed count x_i	10	68	112

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

Problem 12

US suicides in 1970. Check for the seasonal variation

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 $H_0 : p = \frac{1}{2}$?

(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 $\frac{1}{2}$?

6 Bayesian inference

The statistical tools introduced in this course so far are based on the so called *frequentist approach*. In the parametric case the data x is assumed to be randomly generated by a distribution $f(x|\theta)$ and the unknown population parameter θ is estimated using the maximum likelihood estimate. This section presents basic concepts of the *Bayesian approach* when it is assumed that the parameter of interest θ is itself randomly generated using a given prior distribution $g(\theta)$. The prior distribution brings into the model our knowledge (or lack of knowledge) on θ before data x is generated using $f(x|\theta)$, 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

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

where

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

gives the marginal distribution of the random data X. For a fixed x, the denominator $\phi(x)$ is treated as a constant and the Bayes formula can be summarised as

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

where $\propto$ means proportional, as the coefficient of proportionality $\phi(x)$ does not explicitly depend on θ.
When we have no prior knowledge of θ, the prior distribution is often modelled by the uniform distribution. In this case of uninformative prior, given $g(\theta)$ is a constant, we have $h(\theta|x) \propto f(x|\theta)$ so that all the posterior knowledge comes from the likelihood function.

Example: IQ measurement

A randomly chosen individual has an unknown true intelligence quotient value θ. Suppose an IQ test is calibrated in such a way that the prior distribution of θ is normal N(100, 15). This normal distribution describes the population distribution of people's IQ with population mean of $\mu_0 = 100$ 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 N(θ, 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 get

$$g(\theta) f(x|\theta) = \frac{1}{2\pi\sigma_0\sigma} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2} - \frac{(x-\theta)^2}{2\sigma^2}} \propto e^{-\frac{\theta^2 - 2\mu_1\theta}{2\sigma_1^2}},$$

where

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

Thus the posterior becomes proportional to

$$h(\theta|x) \propto e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}},$$

implying that the posterior distribution is also normal N(μ_1, σ_1). The parameter $\gamma \in (0, 1)$ is called a shrinkage factor, as it gives the ratio between the posterior variance to the prior variance: $\sigma_1^2 = \gamma\sigma_0^2$.
In particular, if the observed IQ result is $x = 130$, then the posterior distribution becomes N(120.7, 8.3). We see that the prior expectation $\mu_0 = 100$ has corrected the observed result $x = 130$ down to 120.7. 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 is less uncertain than the prior knowledge.

6.1 Conjugate priors

Suppose we have two parametric families of probability distributions $\mathcal{G}$ and $\mathcal{H}$.

$\mathcal{G}$ is called a family of conjugate priors to $\mathcal{H}$, if a $\mathcal{G}$-prior and a $\mathcal{H}$-likelihood give a $\mathcal{G}$-posterior.

Below we present five models involving conjugate priors. For this we need to introduce another two parametric distributions: Beta and Dirichlet.

Beta distribution

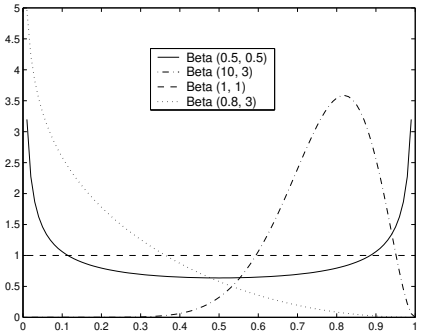
Beta distribution Beta(a, b) is determined by two parameters $a > 0$, $b > 0$ which are called pseudo-counts. It has density,

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

with mean and variance having the form

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

Beta distribution is a convenient prior distribution for a population frequency $p \in (0, 1)$. The uniform distribution U(0, 1) is obtained with $a = b = 1$.



Exercise: verify that for given $a > 1$ and $b > 1$, the maximum of density function $f(p)$ is attained at

$$\hat{p} = \frac{a-1}{a+b-2}.$$

Dirichlet distribution

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

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

Positive parameters $\alpha_1, \ldots, \alpha_r$ are also called pseudo-counts. It has density

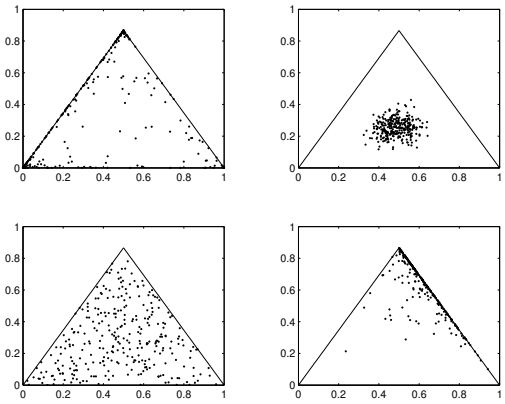
$$f(p_1, \ldots, p_r) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_r)} p_1^{\alpha_1 - 1} \ldots p_r^{\alpha_r - 1}, \quad \alpha_0 = \alpha_1 + \ldots + \alpha_r.$$

The marginal distributions are

$$P_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j), \quad j = 1, \ldots, r,$$

Beta distributions, and we have negative covariances

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



The figure illustrates four examples of $\text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ 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 (x_1, x_2, x_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).

List of conjugate priors

Data distribution	Prior	Posterior distribution
$X_1, \ldots, X_n \sim \text{N}(\mu, \sigma^2)$	$\mu \sim \text{N}(\mu_0, \sigma_0)$	$\text{N}(\gamma_n \mu_0 + (1 - \gamma_n)\bar{x}; \sigma_0 \sqrt{\gamma_n})$
$X \sim \text{Bin}(n, p)$	$p \sim \text{Beta}(a, b)$	$\text{Beta}(a + x, b + n - x)$
$(X_1, \ldots, X_r) \sim \text{Mn}(n; p_1, \ldots, p_r)$	$(p_1, \ldots, p_r) \sim \text{Dir}(\alpha_1, \ldots, \alpha_r)$	$\text{Dir}(\alpha_1 + x_1, \ldots, \alpha_r + x_r)$
$X_1, \ldots, X_n \sim \text{Geom}(p)$	$p \sim \text{Beta}(a, b)$	$\text{Beta}(a + n, b + n\bar{x} - n)$
$X_1, \ldots, X_n \sim \text{Pois}(\mu)$	$\mu \sim \text{Gam}(\alpha_0, \lambda_0)$	$\text{Gam}(\alpha_0 + n\bar{x}, \lambda_0 + n)$
$X_1, \ldots, X_n \sim \text{Gam}(\alpha, \lambda)$	$\lambda \sim \text{Gam}(\alpha_0, \lambda_0)$	$\text{Gam}(\alpha_0 + \alpha n, \lambda_0 + n\bar{x})$

For the Normal-Normal model, the shrinkage factor

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

becomes smaller for larger n. Notice that the posterior variance is always smaller than the prior variance. This list of conjugate prior models illustrates that the contribution of the prior distribution becomes smaller for larger samples. For the Binomial-Beta and Multinomial-Dirichlet models the update rule has the form

posterior pseudo-counts = prior pseudo-counts plus sample counts

Example: Binomial-Beta model

First we show a simple demonstration that beta distribution gives a conjugate prior to the binomial likelihood. Indeed, if

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

and

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

then obviously posterior is also a beta distribution:

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

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: $n_1 = 10$ tosses, counts $x_1 = 2$, $n_1 - x_1 = 8$. We apply the uninformative prior distribution Beta$(1, 1)$ with mean $\mu_0 = 0,5$ and standard deviation $\sigma_0 = 0.29$. It gives a posterior distribution Beta(3, 9) with mean $\hat{p} = \frac{3}{12} = 0.25$ and standard deviation $\sigma_1 = 0.12$.

Experiment 2: $n_2 = 40$ tosses, counts $x_2 = 9$, $n_2 - x_2 = 31$. 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 mean $\hat{p} = \frac{12}{52} = 0.23$ and standard deviation $\sigma_2 = 0.06$.

6.2 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 depends on the so-called loss function $l(\theta, a)$. The so-called Bayes action minimises the posterior risk

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

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 Bayesian estimators.

Zero-one loss function and maximum a posteriori probability

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

Using the zero-one loss function we find that the posterior risk is 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. We define $\hat{\theta}_{\text{map}}$ as the value of θ that maximises $h(\theta|x)$. Observe that with the non-informative prior, $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$.

Squared error loss function and posterior mean estimate

$$\boxed{\text{Squared error loss: } l(\theta, a) = (\theta - a)^2}$$

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

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

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

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

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 $\theta = (p_1, \ldots, p_6)$. The maximum likelihood estimate based on the sample counts is

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

The maximum likelihood estimate assigns value zero to p_6, thereby excluding sixes in future observations. Now take the uninformative prior distribution $\text{Dir}(1, 1, 1, 1, 1, 1)$ and compare two Bayesian estimates

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

The latter has an advantage of assigning a positive value to p_6.

6.3 Credibility interval

Let x stand for the data in hand. For a confidence interval formula

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

the parameter θ is an unknown constant and the confidence interval is random (due to the random sampling procedure)

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

A credibility interval (or credible interval)

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

is treated as a nonrandom interval, while θ is understood to be generated by the posterior distribution of a random variable Θ. A credibility interval is computed from the posterior distribution

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

Example: IQ measurement

Given $n = 1$, we have $\bar{X} \sim \text{N}(\mu; 10)$ and an exact 95% confidence interval for μ takes the form

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

Posterior distribution of the mean is $\text{N}(120.7; 8.3)$ and therefore a 95% credibility interval for μ is

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

6.4 Bayesian hypotheses testing

We consider the case of two simple hypotheses. Choose between H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ using not only the two likelihood functions $f(x|\theta_0)$, $f(x|\theta_1)$ but also the prior probabilities of the two alternative hypotheses

$$\text{P}(H_0) = \pi_0, \quad \text{P}(H_1) = \pi_1 = 1 - \pi_0.$$

In terms of the rejection region $\mathcal{R}$, the 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	c_1
$x \in \mathcal{R}$	Reject H_0	c_0	0

where c_0 is the error type I cost and c_1 is the error type II cost. For a given set $\mathcal{R}$, the average cost is the weighted mean of two values c_0 and c_1

$$c_0\pi_0\text{P}(X \in \mathcal{R}|H_0) + c_1\pi_1\text{P}(X \notin \mathcal{R}|H_1) = c_1\pi_1 + \int_{\mathcal{R}} \Big(c_0\pi_0 f(x|\theta_0) - c_1\pi_1 f(x|\theta_1)\Big)dx.$$

Now observe that

$$\int_{\mathcal{R}} \Big(c_0\pi_0 f(x|\theta_0) - c_1\pi_1 f(x|\theta_1)\Big)dx \geq \int_{\mathcal{R}^*} \Big(c_0\pi_0 f(x|\theta_0) - c_1\pi_1 f(x|\theta_1)\Big)dx,$$

where

$$\mathcal{R}^* = \{x : c_0\pi_0 f(x|\theta_0) < c_1\pi_1 f(x|\theta_1)\}.$$

It follows that the rejection region minimising the average cost is $\mathcal{R} = \mathcal{R}^*$. With the rejection region $\mathcal{R}^*$, we should reject H_0 for small values of the likelihood ratio:

$$\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{c_1\pi_1}{c_0\pi_0},$$

or in other terms, we reject H_0 for small values of the posterior odds

$$\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{c_1}{c_0}.$$

Example of Bayesian hypothesis testing

The defendant A 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: A is innocent, H_1: A is guilty.
A non-informative prior probability

$$\pi_1 = \tfrac{1}{200000}, \text{ so that } \tfrac{\pi_0}{\pi_1} = 200000,$$

takes into account the number of males who theoretically could have committed the crime without any evidence taken into account. There were three conditionally independent pieces of evidence

E_1: a DNA match,

E_2: defendant A is not recognised by the victim,

E_3: an alibi supported by the girlfriend.

The reliability of these pieces of evidence was quantified as

$\mathrm{P}(E_1\|H_0) = \frac{1}{200{,}000{,}000}$, $\mathrm{P}(E_1\|H_1)$=1, so that $\frac{\mathrm{P}(E_1\|H_0)}{\mathrm{P}(E_1\|H_1)} = \frac{1}{200{,}000{,}000}$	very strong evidence in favour of H_1
$\mathrm{P}(E_2\|H_1) = 0.1$, $\mathrm{P}(E_2\|H_0) = 0.9$, so that $\frac{\mathrm{P}(E_2\|H_0)}{\mathrm{P}(E_2\|H_1)} = 9$	evidence in favour of H_0
$\mathrm{P}(E_3\|H_1) = 0.25$, $\mathrm{P}(E_3\|H_0) = 0.5$, so that $\frac{\mathrm{P}(E_3\|H_0)}{\mathrm{P}(E_3\|H_1)} = 2$	evidence in favour of H_0

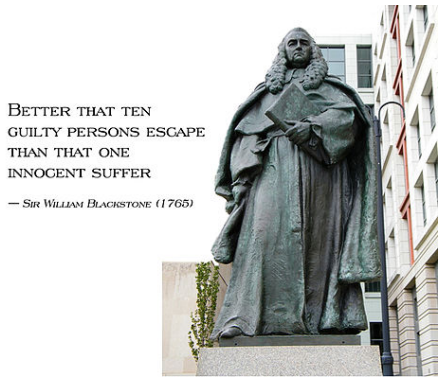
The posterior odds was computed as

$$\frac{\mathrm{P}(H_0|E)}{\mathrm{P}(H_1|E)} = \frac{\pi_0\mathrm{P}(E|H_0)}{\pi_1\mathrm{P}(E|H_1)} = \frac{\pi_0\mathrm{P}(E_1|H_0)\mathrm{P}(E_2|H_0)\mathrm{P}(E_3|H_0)}{\pi_1\mathrm{P}(E_1|H_1)\mathrm{P}(E_2|H_1)\mathrm{P}(E_3|H_1)} = 0.018.$$

Conclusion: we would reject H_0 if the cost values are assigned by the jury are such that

$$\frac{c_1}{c_0} = \frac{\text{cost for unpunished crime}}{\text{cost for punishing an innocent}} > 0.018.$$

Prosecutor's fallacy: $\mathrm{P}(H_0|E) = \mathrm{P}(E|H_0)$,
which is only true if $\mathrm{P}(E) = \pi_0$.
Example: $\pi_0 = \pi_1 = 1/2$, $\mathrm{P}(E|H_0) \approx 0$, $\mathrm{P}(E|H_1) \approx 1$.



6.5 Exercises

Problem 1

This is a continuation of the Problem 3 from Section 4.8.

(e) Assume uniform prior $\Theta \sim \mathrm{U}(0, 1)$ and find the posterior density. Plot it. What is the mode of the posterior?

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 frequency O_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 a uniform prior for p. What is then the posterior distribution and what are the posterior mean and standard deviation?

Problem 3

Laplace's 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?

Problem 4

This is a continuation of the Problem 2 from Section 5.8.
Let X have 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

(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 $\mathrm{P}(H_0|x) > \mathrm{P}(H_1|x)$. For $\sigma^2 = 0.1, 0.5, 1.0, 5.0$:

(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 has a cumulative distribution function $F(x) = x^2, \quad 0 \le x \le 1$, and under H_1, it has a cumulative distribution function $F(x) = x^3, \quad 0 \le x \le 1$.
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?

7 Summarising data

7.1 Empirical probability distribution

Consider an iid-sample $(x_1, \ldots, x_n)$ from the population distribution $F(x) = \mathrm{P}(X \le x)$.

$$\boxed{\text{Empirical distribution function } \hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}}.}$$

For a fixed x,

$$\hat{F}(x) = \hat{p}$$

is the sample proportion estimating the population proportion $p = F(x)$.
On the other hand for variable x, the function $\hat{F}(x)$ is a cumulative distribution function for a random variable Y with the discrete 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 $\hat{F}(\cdot)$ is a cumulative distribution function with mean $\bar{x}$ and variance $\frac{n-1}{n}s^2$ even if some of x_i coincide. We 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.

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 T has distribution function $F(t) = \mathrm{P}(T \le t)$, then its survival function is

$$S(t) = \mathrm{P}(T > t) = 1 - F(t).$$

Hazard function $h(t) = \frac{f(t)}{S(t)}$, where $f(t) = F'(t)$ is the probability density function.

The hazard function is the mortality rate at age t:

$$\mathrm{P}(t < T \le t + \delta | T \ge t) = \frac{\mathrm{P}(t < T \le t + \delta)}{\mathrm{P}(T \ge t)} = \frac{F(t+\delta) - F(t)}{S(t)} \sim \delta \cdot h(t), \quad \delta \to 0.$$

The hazard function can be viewed as the negative of the slope of the log survival function:

$$h(t) = -\tfrac{d}{dt} \ln S(t) = -\tfrac{d}{dt} \ln(1 - F(t)).$$

A constant hazard rate $h(t) = \lambda$ corresponds to the exponential distribution $\mathrm{Exp}(\lambda)$.

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 (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

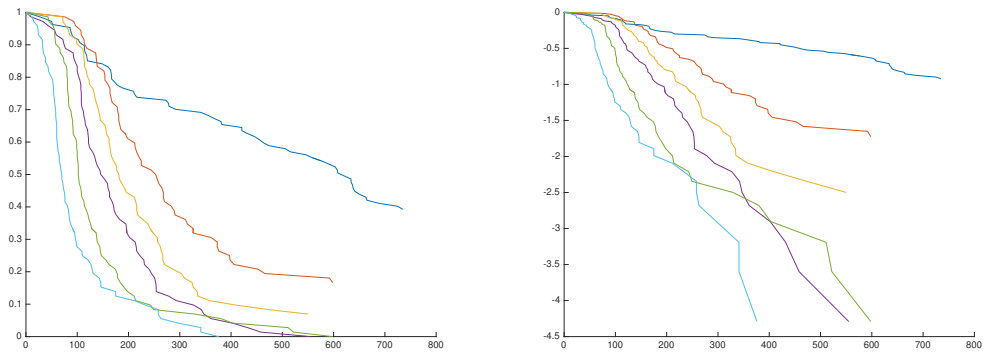
Dose IV lifetimes

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 groups just looking at numbers. The data is illuminated by two graphs: one for the survival functions and the other for the log-survival functions.



The negative slopes of the curves to the right illustrate the hazard rates for different groups.

7.2 Density estimation

A histogram displays the observed counts

$$O_j = \sum_{i=1}^{n} 1_{\{x_i \in \text{cell}_j\}}$$

over the adjacent cells of width h. The choice of a balanced width h is important: smaller h give ragged profiles, larger h give obscured profiles. Put

$$f_h(x) = \frac{O_j}{nh}, \quad \text{for } x \in \text{cell}_j,$$

and notice that

$$\int f_h(x)dx = \frac{h}{nh} \sum_j O_j = 1.$$

The scaled histogram given by the graph of $f_h(x)$ is a density estimate. Kernel density estimate with bandwidth h produces a smooth curve

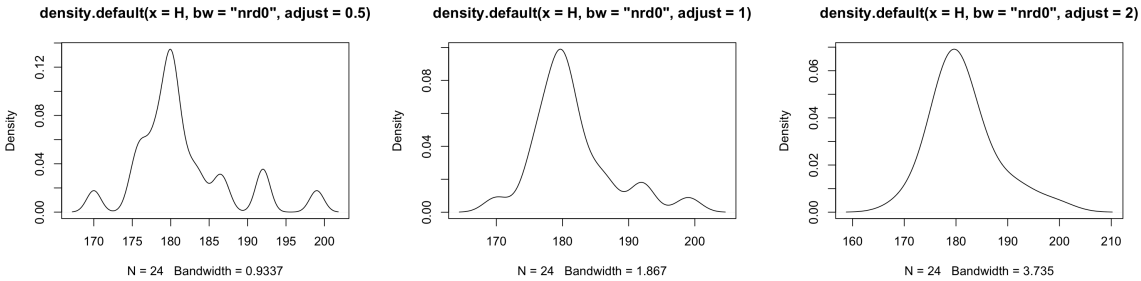
$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi(\tfrac{x-x_i}{h}), \text{ where } \phi(x) = \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Example: male heights

Let H stand for the column of 24 male heights. Applying the R code

```
> H = c(170,175,176,176,177,178,178,179,179,180,180,180,180,180,
        181,181,182,183,184,186,187,192,192,199)
> plot(density(H, bw = "nrd0", adjust = 0.5))
> plot(density(H, bw = "nrd0", adjust = 1))
> plot(density(H, bw = "nrd0", adjust = 2))
```

we get the following plots for three different bandwidths, where bw = "nrd0" stands for a default bandwidth.



7.3 Quantiles and QQ-plots

The inverse of the cumulative distribution function $F(x)$ is called the quantile function

$$Q(p) = F^{-1}(p), \quad 0 < p < 1.$$

The quantile function Φ^{-1} for the standard normal distribution Φ is called the *probit* function (from *prob*ability un*it*).

For a given distribution F and $0 \le p \le 1$, the p-quantile is $x_p = Q(p)$.

Special quantiles:

median $m = x_{0.5} = Q(0.5)$,
lower quartile $x_{0.25} = Q(0.25)$,
upper quartile $x_{0.75} = Q(0.75)$.

Quantile x_p cuts off proportion p of smallest values of a random variable X with $\mathrm{P}(X \le x) = F(x)$:

$$\mathrm{P}(X \le x_p) = F(x_p) = F(Q(p)) = p.$$

By sorting a random sample $(x_1, \ldots, x_n)$ we arrive at the ordered sample values

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)},$$

where in particular,

$$x_{(1)} = \min\{x_1, \ldots, x_n\}, \quad x_{(n)} = \max\{x_1, \ldots, x_n\}.$$

In the continuous case, we obtain

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)},$$

strictly ordered n jump points for the empirical distribution function, so that

$$F_n(x_{(k)}) = \tfrac{k}{n}, \qquad F_n(x_{(k)} - \epsilon) = \tfrac{k-1}{n}.$$

This observation leads to the following definition of empirical quantiles

$x_{(k)}$ is called the empirical $(\frac{k-0.5}{n})$-quantile

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 F_X and F_Y. A relevant null hypothesis H_0: $F_X \equiv F_Y$ is equivalent to H_0: $Q_X \equiv Q_Y$. It can be tested graphically using a QQ-plot.

QQ-plot is a scatter plot of n 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 can claim that H_0: $F_X \equiv F_Y$ is true.
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,

$$F_X(x) = F_Y(a + bx),$$

so that putting $Q_X(p) = x$, we get $Q_Y(p) = a + bx$, which yields

$$Q_Y(p) = a + bQ_X(p), \quad 0 < p < 1.$$

7.4 Testing normality

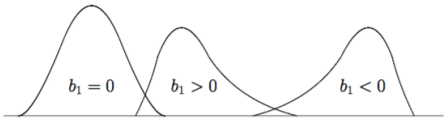
The normality hypothesis H_0 states that the population distribution for an iid-sample $(x_1, \ldots, x_n)$ is normal $\mathrm{N}(\mu, \sigma)$ with unspecified parameter values. A QQ-plot used for testing this hypothesis is called a normal probability plot. The normal probability plot is the scatter plot for

$$(x_{(1)}, y_1), \ldots, (x_{(n)}, y_n), \quad \text{where } y_k = \Phi^{-1}(\tfrac{k-0.5}{n}).$$

If the normal probability plot is close to a straight line $y = a + bx$, then we accept H_0 and use the point estimates $\hat{\mu} = -\frac{a}{b}$, $\hat{\sigma} = \frac{1}{b}$. If normality does not hold, draw a straight line via empirical lower and upper quartiles to detect a light tails profile or heavy tails profile.
Another simple way of testing normality relies on two summary statistics: skewness and kurtosis.

Coefficient of skewness: $\beta_1 = \frac{\mathrm{E}[(X-\mu)^3]}{\sigma^3}$, sample skewness: $b_1 = \frac{1}{s^3 n}\sum_{i=1}^{n}(x_i - \bar{x})^3$

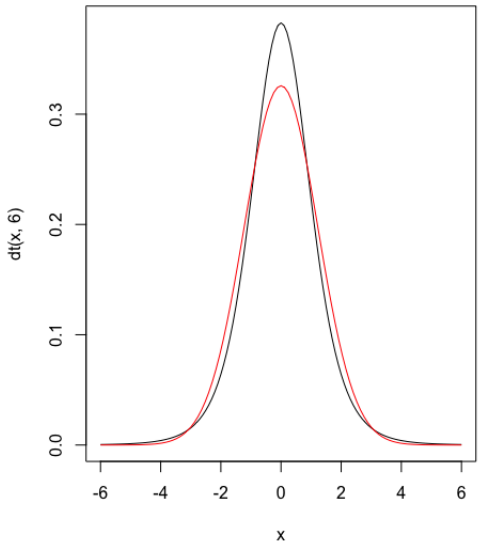
Depending on the sign of the coefficient of skewness, we distinguish between symmetric $\beta_1 = 0$, skewed to the right $\beta_1 > 0$, and skewed to the left $\beta_1 < 0$ distributions.

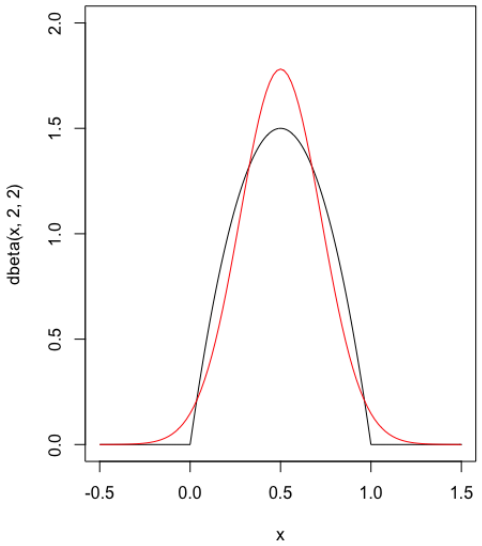


Given that b_1 is close to zero, kurtosis can be used as an indication of the curve profile to be close to that of the normal distribution.

Kurtosis $\beta_2 = \frac{\mathrm{E}[(X-\mu)^4]}{\sigma^4}$, sample kurtosis: $b_2 = \frac{1}{s^4 n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

For the normal distribution, kurtosis coefficient takes value $\beta_2 = 3$. Leptokurtic distribution: $\beta_2 > 3$ (heavy tails). Platykurtic distribution: $\beta_2 < 3$ (light tails).
On the left panel the t-distribution with df = 6 is compared to the normal distribution with a matching variance (here $\beta_2 = 6$). On the right panel the beta-distribution with parameters (2,2) is compared to the normal distribution with a matching variance (here $\beta_2 = 2.143$).





Example: male heights

Summary statistics:

$$\bar{x} = 181.46, \quad \hat{m} = 180, \quad b_1 = 1.05, \quad b_2 = 4.31.$$

Good to know: the distribution of the heights of adult males is positively skewed, so that $m < \mu$, or in other terms,

$$\mathrm{P}(X < \mu) > \mathrm{P}(X \leq m) \geq 0.50,$$

implying that more than half of heights are below the average.

The gamma distribution $\mathrm{Gam}(\alpha, \lambda)$ is positively skewed $\beta_1 = \frac{2}{\sqrt{\alpha}}$, and leptokurtic $\beta_2 = 3 + \frac{6}{\alpha}$.

7.5 Measures of location

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 median m is estimated by the sample median.

Sample median: $\hat{m} = x_{(k)}$, if $n = 2k - 1$, and $\hat{m} = \frac{x_{(k)} + x_{(k+1)}}{2}$, if $n = 2k$.

The sample mean $\bar{x}$ is sensitive to outliers, while the sample median $\hat{m}$ is not. Therefore, we say that $\hat{m}$ is a robust estimator (robust to outliers).

Confidence interval for the median

Consider an iid-sample $(x_1, \ldots, x_n)$ without assuming any parametric model for the unknown population distribution. Let

$$y = \sum_{i=1}^{n} 1_{\{x_i \leq m\}}$$

be the number of observations below the true median, then

$$p_k = \mathrm{P}(X_{(k)} < m < X_{(n-k+1)}) = \mathrm{P}(k \leq Y \leq n - k) = \sum_{i=k}^{n-k} \binom{n}{i} 2^{-n}$$

can be computed from the symmetric binomial distribution $Y \sim \mathrm{Bin}(n, 0.5)$.
This yields the following non-parametric formula for an exact confidence interval for the median.

$I_m = (x_{(k)}, x_{(n-k+1)})$ is a $100 \cdot p_k\%$ 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
$100 \cdot p_k$	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: $m \neq m_0$.
The sign test statistic

$$y_0 = \sum_{i=1}^{n} 1_{\{x_i \leq m_0\}}$$

counts the number of observations below the null hypothesis value. It has a simple null distribution $Y_0 \overset{H_0}{\sim} \mathrm{Bin}(n, 0.5)$. Connection to the above confidence interval formula: reject H_0 if m_0 falls outside the corresponding confidence interval

$$I_m = (x_{(k)}, x_{(n-k+1)}).$$

Trimmed means

A trimmed mean is a robust measure of location computed from a central portion of the data.

α-trimmed mean $\bar{x}_\alpha$ = sample mean without $\frac{n\alpha}{2}$ smallest and $\frac{n\alpha}{2}$ largest observations

Example: male heights

Ignoring 20% of largest and 20% of smallest observations we compute $\bar{x}_{0.4}$=180.36. The trimmed mean is between $\bar{x} = 181.46$ and $\hat{m} = 180$.

When summarizing data compute several measures of location and compare the results.

7.6 Measures of dispersion

Sample variance s^2 and sample range

$$R = x_{(n)} - x_{(1)}$$

are sensitive to outliers. Two robust measures of dispersion:

interquartile range IQR $= x_{0.75} - x_{0.25}$ is the difference between the upper and lower quartiles,

MAD = Median of Absolute values of Deviations from sample median $|x_i - \hat{m}|$, $i = 1, \ldots, n$.

Three estimates of σ for the normal distribution $\mathrm{N}(\mu, \sigma)$ model: s, $\frac{\mathrm{IQR}}{1.35}$, $\frac{\mathrm{MAD}}{0.675}$

Indeed, for the standard normal distribution, we have

$$\Phi(0.675) = 0.75, \quad \Phi^{-1}(0.75) = 0.675,$$

so that for the general normal distribution $\mathrm{N}(\mu, \sigma)$ the theoretical quartiles are

$$\mathrm{LQ} = \mu - 0.675 \cdot \sigma \text{ and } \mathrm{UQ} = \mu + 0.675 \cdot \sigma.$$

Therefore,

$$\mathrm{IQR} = (\mu + 0.675 \cdot \sigma) - (\mu - 0.675 \cdot \sigma) = 1.35 \cdot \sigma.$$

On the other hand, since

$$\mathrm{P}(|X - \mu| \leq 0.675 \cdot \sigma) = (\Phi(0.675) - 0.5) \cdot 2 = 0.5,$$

we obtain $\mathrm{MAD} = 0.675 \cdot \sigma$.

Boxplots

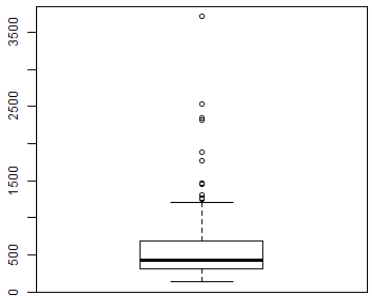
Boxplots are convenient to use for comparing different samples. A boxplot is built of the following components: box, whiskers and outliers.

Box

- upper edge of the box = upper quartile (UQ)
- box center = median
- lower edge of the box = lower quartile (LQ)

Wiskers

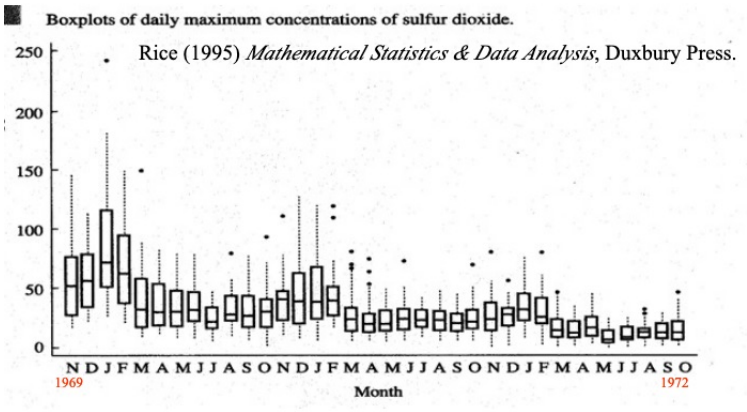
- upper whisker end = {maximal data point not exceeding UQ + 1.5 × IQR}
- lower whisker end = {min data point ≥ LQ – 1.5 × IQ



Outliers

- upper dots = {data points ≥ UQ + 1.5 × IQR}
- lower dots = {data points ≤ LQ – 1.5 × IQR}

Example of parallel boxplots



7.7 Exercises

Problem 1

Suppose that $(X_1, \ldots, X_n)$ are independent uniform U$(0, 1)$ random variables. The corresponding empirical distribution function $\hat{F}(x)$ is a random variable with mean $F(x) = x$ and standard deviation $\sqrt{\frac{x(1-x)}{n}}$.

(a) Sketch the population distribution function $F(x)$ and the standard deviation of the empirical distribution function $\hat{F}(x)$.

(b) Generate many samples of size 16. For each sample, plot the difference $F(x) - \hat{F}(x)$ and relate what you see to your answer to (a).

Problem 2

Let $(X_1, \ldots, X_n)$ be independent random variables with the same distribution F, and let $\hat{F}$ denote the empirical distribution function. Show that for $u < v$,

$$\text{Cov}(\hat{F}(u), \hat{F}(v)) = \tfrac{1}{n} F(u)(1 - F(v)).$$

It follows that $\hat{F}(u)$ and $\hat{F}(v)$ are positively correlated: if $\hat{F}(u)$ overshoots $F(u)$, then $\hat{F}(v)$ will tend to overshoot $F(v)$.

Problem 3

A random sample $x_1, \ldots, x_n$, $n = 59$:

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

are the percentages of hydrocarbons in each sample of beeswax.

(a) Plot the empirical distribution function, a histogram, and a normal probability 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

$$F(t) = 1 - e^{-\alpha t^\beta}, \quad t \geq 0,$$

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. 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, .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 rainfall versus rainfall and log rainfall versus log rainfall. What do these plots suggest about the effect, if any, of seeding?

Problem 7

Express the survival function in terms of the cumulative hazard function

$$H(t) = \int_0^t h(x)dx.$$

8 Comparing two samples

Suppose we wish to compare two population distributions with means and standard deviations (μ_1, σ_1) and (μ_2, σ_2) based on two iid-samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_m)$ from these two populations. We start by computing two sample means $\bar{x}, \bar{y}$, and their standard errors

$$s_{\bar{x}} = \tfrac{s_1}{\sqrt{n}}, \quad s_1^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$
$$s_{\bar{y}} = \tfrac{s_2}{\sqrt{m}}, \quad s_2^2 = \tfrac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2.$$

The difference $(\bar{x} - \bar{y})$ is an unbiased estimate of $\mu_1 - \mu_2$. We are interested in finding the standard error of $\bar{x} - \bar{y}$ and an interval estimate 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: two independent samples and paired samples (sampling the differences).

8.1 Two independent samples: comparing population means

If $(X_1, \ldots, X_n)$ is independent from $(Y_1, \ldots, Y_m)$, then

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

and

$$s_{\bar{x}-\bar{y}}^2 = s_{\bar{x}}^2 + s_{\bar{y}}^2 = \frac{s_1^2}{n} + \frac{s_2^2}{m}$$

gives an unbiased estimate of $\mathrm{Var}(\bar{X} - \bar{Y})$. Therefore, $s_{\bar{x}-\bar{y}}$ will be called the (estimated) standard error of the point estimate $\bar{x} - \bar{y}$.

Large sample test for the difference between two means

If n and m are large, we can use a normal approximation

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_{\bar{X}-\bar{Y}}} \approx \mathrm{N}(0,1).$$

The hypothesis

$$H_0 : \mu_1 = \mu_2$$

is tested using the test statistic

$$z = \frac{\bar{x}-\bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}}$$

whose null distribution is approximated by the standard normal N(0,1).

Approximate confidence interval formula $I_{\mu_1-\mu_2} \approx \bar{x}-\bar{y} \pm z_{\alpha/2} \cdot \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}$.

Two-sample t-test

The key assumption of the two-sample t-test:

two normal population distributions $X \sim \mathrm{N}(\mu_1, \sigma)$, $Y \sim \mathrm{N}(\mu_2, \sigma)$ have equal variances.

Given $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the pooled sample variance

$$s_{\mathrm{p}}^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2 + \sum_{i=1}^m (y_i-\bar{y})^2}{n+m-2} = \frac{n-1}{n+m-2} \cdot s_1^2 + \frac{m-1}{n+m-2} \cdot s_2^2$$

is an unbiased estimate of the variance with

$$\mathrm{E}(S_{\mathrm{p}}^2) = \frac{n-1}{n+m-2}\mathrm{E}(S_1^2) + \frac{m-1}{n+m-2}\mathrm{E}(S_2^2) = \sigma^2.$$

In the equal variance two sample setting, the variance

$$\mathrm{Var}(\bar{X}-\bar{Y}) = \sigma^2 \cdot \frac{n+m}{nm},$$

has the following unbiased estimate

$$s_{\bar{x}-\bar{y}}^2 = s_p^2 \cdot \frac{n+m}{nm}.$$

Exact distribution $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}} \sim t_{n+m-2}$

Exact confidence interval formula

$$I_{\mu_1-\mu_2} = \bar{x}-\bar{y} \pm t_{n+m-2}(\alpha/2) \cdot s_{\mathrm{p}} \cdot \sqrt{\tfrac{n+m}{nm}}.$$

Two sample t-test uses the test statistic $t = \frac{\bar{x}-\bar{y}}{s_{\mathrm{p}}} \cdot \sqrt{\frac{nm}{n+m}}$ for testing H_0: $\mu_1 = \mu_2$. The null distribution of the test statistic is

$$T \sim t_{n+m-2}.$$

Example: iron retention

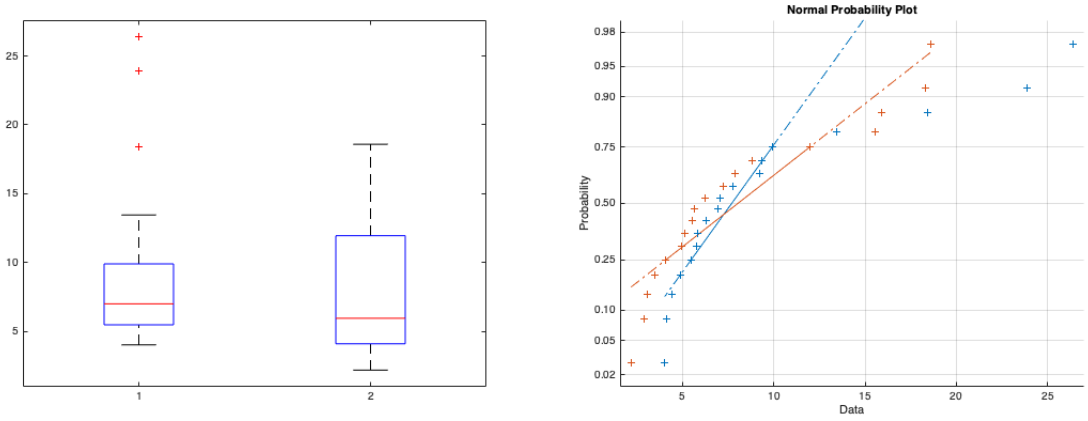
Percentage of Fe^{2+} and Fe^{3+} retained by mice data at concentration 1.2 millimolar. (The two samples are taken out from a larger dataset given in Section 9.5.) Summary of the data:

Fe^{2+}: $n = 18$, $\bar{x} = 9.63$, $s_1 = 6.69$, $s_{\bar{x}} = 1.58$
Fe^{3+}: $m = 18$, $\bar{y} = 8.20$, $s_2 = 5.45$, $s_{\bar{y}} = 1.28$

The graphs below show that the population distributions are not normal. Therefore, to test H_0: $\mu_1 = \mu_2$ we use the large sample test. Using the observed value

$$z_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2}} = 0.7,$$

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 probability plots.

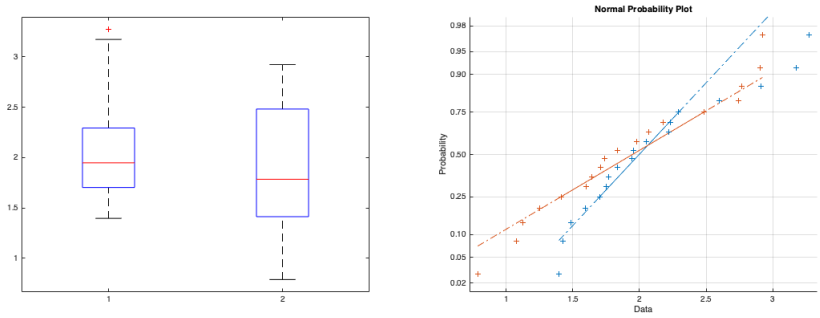
After the log transformation the data look more like normally distributed, as seen from the graphs below. For the transformed data, we get

$n = 18$, $\bar{x}' = 2.09$, $s_1' = 0.659$, $s_{\bar{x}'} = 0.155$,
$m = 18$, $\bar{y}' = 1.90$, $s_2' = 0.574$, $s_{\bar{y}'} = 0.135$.

Two sample t-test for the transformed data with

$$t_{\text{obs}} = 0.917, \quad \text{df} = 34, \quad \text{p-value} = 0.366,$$

also results in non-significant difference. Boxplots and normal probability plots for logs of percentages:



We return to this example in Section 9.5.

Rank sum test

The rank sum test is a nonparametric test for two independent samples, which does not assume normality of population distributions.
Assume continuous population distributions F_1 and F_2, and consider

$$H_0\colon\ F_1 = F_2 \text{ against } H_1\colon\ F_1 \neq F_2.$$

The rank sum test procedure:

> 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,

$$r_1 + r_2 = 1 + 2 + \ldots + (n+m) = \tfrac{(n+m)(n+m+1)}{2}.$$

The null distributions for R_1 and R_2 depend only on the sample sizes n and m.
For $n \geq 10$, $m \geq 10$ apply the normal approximation for the null distributions of R_1 and R_2 with

$$\mathrm{E}(R_1) = \frac{n(n+m+1)}{2},\ \mathrm{E}(R_2) = \frac{m(n+m+1)}{2},\ \mathrm{Var}(R_1) = \mathrm{Var}(R_2) = \frac{mn(n+m+1)}{12}.$$

Example: in class experiment

Height distributions for females F_1, and males F_2. For $n = m = 3$, compute r_1, r_2 and one-sided p-value.

8.2 Two independent samples: comparing population proportions

For the binomial model $X \sim \mathrm{Bin}(n, p_1)$, $Y \sim \mathrm{Bin}(m, p_2)$, two independently generated values (x, y) give sample proportions

$$\hat{p}_1 = \frac{x}{n}, \qquad \hat{p}_2 = \frac{y}{m},$$

which are unbiased estimates of p_1, p_2 and have 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 95 % confidence interval for the difference $p_1 - p_2$ is given by

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n-1} + \tfrac{\hat{p}_2(1-\hat{p}_2)}{m-1}}.$$

With help of this formula we can test the null hypothesis of equality

$$H_0 : p_1 = p_2.$$

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

$$|\hat{p}_1 - \hat{p}_2| > 1.96 \cdot \sqrt{2 \cdot \tfrac{0.4 \cdot 0.6}{5000}} \approx 1.9\%.$$

This should 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

$$I_p \approx \hat{p} \pm 1.96 \cdot \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}},$$

and if $\hat{p} \approx 0.4$, then the difference is significant if

$$|\hat{p} - 0.4| > 1.96 \cdot \sqrt{\tfrac{0.4 \cdot 0.6}{5000}} \approx 1.3\%.$$

Fisher's exact test

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 binomial distributions. We summarise the data of two independent samples as a 2×2 table of sample counts

	Sample 1	Sample 2	Total
Number of successes	x	y	$Np = x + y$
Number of failures	$n - x$	$m - y$	$Nq = n + m - x - y$
Sample sizes	n	m	$N = n + m$

Fisher's idea for this case, was to use X as a test statistic conditionally on the total number of successes $x+y$. Under the null hypothesis, the conditional distribution of X is hypergeometric

$$X \sim \mathrm{Hg}(N, n, p)$$

with parameters (N, n, p) defined by

$$N = n + m, \quad p = \tfrac{x+y}{N}.$$

This is a discrete distribution with probability mass function

$$\mathrm{P}(X = x) = \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}, \quad \max(0, n - Nq) \le x \le \min(n, Np).$$

This null distribution should be used for determining the rejection rule of the Fisher test.

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.

	Male	Female	Total
Promote	21	14	35
Hold file	3	10	13
Total	24	24	48

Each expert decision had two possible outcomes: promote or hold file. We wish to test

$$H_0\text{: } p_1 = p_2 \text{ no gender bias,}$$

against

$$H_1\text{: } p_1 > p_2 \text{ males are favoured.}$$

Fisher's test would reject H_0 in favour of the one-sided alternative H_1 for large values of x under the null distribution

$$\mathrm{P}(X = x) = \frac{\binom{35}{x}\binom{13}{24-x}}{\binom{48}{24}} = \frac{\binom{35}{35-x}\binom{13}{x-11}}{\binom{48}{24}}, \quad 11 \le x \le 24.$$

This is a symmetric distribution with

$$\mathrm{P}(X \le 14) = \mathrm{P}(X \ge 21) = 0.025.$$

so that a one-sided p-value $= 0.025$, and a two-sided p-value $= 0.05$. We conclude that there is a significant evidence of sex bias, and reject the null hypothesis.

8.3 Paired samples

Examples of paired observations:

- different drugs for two patients matched by age, sex,
- a fruit weighed before and after shipment,
- two types of tires tested on the same car.

Two paired samples forms a vector of iid-pairs

$$(x_1, y_1), \ldots, (x_n, y_n).$$

As before, our main question is whether the difference $\mu_1 - \mu_2$ is statistically significant. To this end, we turn a one-dimensional iid-sample of differences

$$(d_1, \ldots, d_n), \quad d_i = x_i - y_i.$$

The population mean difference $\mu_1 - \mu_2$ is estimated by $\bar{d} = \bar{x} - \bar{y}$. This is an unbiased estimate whose variance value takes into account dependence between X and Y. Observe that

$$\begin{aligned}\mathrm{Var}(\bar{X} - \bar{Y}) &= \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) - 2\mathrm{Cov}(\bar{X}, \bar{Y}) \\ &= \tfrac{\sigma_1^2}{n} + \tfrac{\sigma_2^2}{n} - \tfrac{2}{n^2}\mathrm{Cov}(X_1 + \ldots + X_n, Y_1 + \ldots + Y_n).\end{aligned}$$

Since X_i and Y_j are independent for $i \neq j$, we get

$$\mathrm{Cov}(X_1 + \ldots + X_n, Y_1 + \ldots + Y_n) = \mathrm{Cov}(X_1, Y_1) + \ldots + \mathrm{Cov}(X_n, Y_n) = n\mathrm{Cov}(X, Y) = n\sigma_1\sigma_2\rho,$$

where

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_1\sigma_2}$$

is the correlation coefficient for the joint population distribution of (X, Y). Thus

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho).$$

If samples are independent and have equal sizes, then $\rho = 0$ and

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \tfrac{1}{n}(\sigma_1^2 + \sigma_2^2).$$

Importantly, if $\rho > 0$, then

$$\mathrm{Var}(\bar{X} - \bar{Y}) < \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}),$$

which demonstrates that the the pared sampling with $\rho > 0$ ensures a smaller standard error for the estimate $\bar{x} - \bar{y}$ as compared with the two independent samples case.

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 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	$d_i = x_i - y_i$	Rank of $\lvert d_i \rvert$	Signed rank
25	27	2	2	+2
25	29	4	3.5	+3.5
27	37	10	6	+6
28	43	15	8.5	+8.5
30	46	16	10	+10
44	56	12	7	+7
52	61	9	5	+5
53	57	4	3.5	+3.5
53	80	27	11	+11
60	59	–1	1	–1
67	82	15	8.5	+8.5

We have $n = 11$ pairs of measurements of individuals. Using the data we estimate correlation as $\rho \approx 0.90$.
Assuming that the population distribution for differences $D \sim \mathrm{N}(\mu, \sigma)$ is normal with $\mu = \mu_1 - \mu_2$, we apply the one-sample t-test for

$$H_0\colon\ \mu_1 - \mu_2 = 0 \text{ against } H_1\colon\ \mu_1 - \mu_2 \neq 0.$$

The observed test statistic value

$$t_{\mathrm{obs}} = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10.27}{2.40} = 4.28$$

reveals a small two-sided P-value, showing that smoking has a significant health effect

```
> 2*(1-pt(4.28,10))
[1] 0.001611392
```

Without assumption of normality, having an iid-sample of difference, we can apply the non-parametric sign test for a pair of hypothesis for the median $m = m_D$ of difference D

$$H_0\colon\ m = 0 \text{ against } H_1\colon\ m \neq 0.$$

Test statistics: either

$$y_+ = \sum 1_{\{d_i > 0\}} \quad \text{or} \quad y_- = \sum 1_{\{d_i < 0\}}.$$

Assuming the distribution of D being continuous, we find that both Y_+ and Y_- have null distribution $\mathrm{Bin}(n, 0.5)$.
To take care of ties $d_i = 0$:

- you either discard the tied observations and reduce n respectively,
- or dissolve the ties by randomisation.

For the data on platelet aggregation, the observed test statistic is $y_- = 1$. Thus a two-sided p-value of the sign test is

$$\text{p-value} = 2[(0.5)^{11} + 11(0.5)^{11}] = 0.012.$$

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 sizes of positive and negative differences. This is a non-parametric test for the null hypothesis of no difference

$$H_0: \text{ distribution of } D \text{ is symmetric about its median } m = 0.$$

The null hypothesis consists of two parts: symmetry of the distribution and $m = 0$. Test statistics: either

$$w_+ = \sum_{i=1}^{n} \text{rank}(|d_i|) \cdot 1_{\{d_i > 0\}}$$

or

$$w_- = \sum_{i=1}^{n} \text{rank}(|d_i|) \cdot 1_{\{d_i < 0\}}.$$

Assuming no ties, that is $d_i \neq 0$, we always get

$$w_+ + w_- = \frac{n(n+1)}{2}.$$

The null distributions of W_+ and W_- are the same and tabulated for smaller values of n (not shown). For $n \geq 20$, one can use the normal approximation of the null distribution with mean and variance

$$\mu_W = \frac{n(n+1)}{4}, \qquad \sigma_W^2 = \frac{n(n+1)(2n+1)}{24}.$$

The signed rank test uses more data information than the sign test but requires symmetric distribution of differences.

Example: platelet aggregation

Observed value of the test statistic $w_- = 1$. It gives a two-sided p-value p = 0.002. The null hypothesis can be rejected for two reasons, therefore it is important to check the symmetry property of the distribution of differences (d_i) before we conclude that there is a significant treatment effect.

8.4 Paired samples: comparing population proportions

Suppose we have two Bernoulli random variables $X \sim \text{Bin}(1, p_1)$, $Y \sim \text{Bin}(1, p_2)$ which depend on each other. The vector (X, Y) has four possible values $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ with probabilities $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$. With n independent paired observations, we count $(W_{00}, W_{01}, W_{10}, W_{11})$ the numbers of different outcomes. The corresponding joint distribution is multinomial $\text{Mn}(n, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ with

$$\text{Var}(W_{10} - W_{01}) = n\pi_{10}(1 - \pi_{10}) + n\pi_{01}(1 - \pi_{01}) + 2n\pi_{10}\pi_{01} = n(\pi_{10} + \pi_{01} - (\pi_{10} - \pi_{01})^2).$$

In this section we produce an approximate confidence interval formula for the difference

$$p_1 - p_2 = \pi_{10} - \pi_{01}.$$

An unbiased point estimate of this difference is given by

$$\hat{p}_1 - \hat{p}_2 = \hat{\pi}_{10} - \hat{\pi}_{01}, \quad \hat{\pi}_{10} = \frac{w_{10}}{n}, \quad \hat{\pi}_{01} = \frac{w_{01}}{n}.$$

The standard error of this point estimate is given by

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{\pi}_{10} + \hat{\pi}_{01} - (\hat{\pi}_{10} - \hat{\pi}_{01})^2}{n-1}}.$$

Again, referring to the central limit theorem we arrive at the following approximate $100(1-\alpha)\%$ confidence interval formula

$$I_{p_1-p_2} \approx \hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} s_{\hat{p}_1-\hat{p}_2}.$$

A significant difference between p_1 and p_2 corresponds to the case when the above confidence interval does not cover zero, that is when

$$|\hat{p}_1 - \hat{p}_2| > z_{\frac{\alpha}{2}} s_{\hat{p}_1-\hat{p}_2},$$

or in other words, the rejection region for $H_0 : \pi_{10} = \pi_{01}$ against $H_1 : \pi_{10} \neq \pi_{01}$ has the form

$$\mathcal{R} = \left\{ \frac{|\hat{\pi}_{10} - \hat{\pi}_{01}|}{\sqrt{\frac{\hat{\pi}_{10}+\hat{\pi}_{01}-(\hat{\pi}_{10}-\hat{\pi}_{01})^2}{n-1}}} > z_{\frac{\alpha}{2}} \right\}.$$

Now notice that the squared left hand side equals

$$\frac{(\hat{\pi}_{10} - \hat{\pi}_{01})^2}{\frac{\hat{\pi}_{10}+\hat{\pi}_{01}-(\hat{\pi}_{10}-\hat{\pi}_{01})^2}{n-1}} \approx \frac{1}{\frac{\hat{\pi}_{10}+\hat{\pi}_{01}}{n(\hat{\pi}_{10}-\hat{\pi}_{01})^2} - \frac{1}{n}} = \frac{1}{\frac{w_{10}+w_{01}}{(w_{10}-w_{01})^2} - \frac{1}{n}} \approx \frac{(w_{10} - w_{01})^2}{w_{10} + w_{01}}.$$

This observation leads to the McNemar test introduced in Section 10.3.

8.5 External and confounding factors

Double-blind, randomised controlled experiments are used to balance out such external factors as

placebo effect,
time factor,
background variables like temperature,
location factor.

Example: portocaval shunt

Portocaval shunt is an operation used to lower blood pressure in the liver. People believed in its high efficiency until the controlled experiments were performed.

Enthusiasm level	Marked	Moderate	None
No controls	24	7	1
Nonrandomized controls	10	3	2
Randomized controls	0	1	3

Example: platelet aggregation

Further parts of the experimental design: control group 1 smoked lettuce cigarettes, control group 2 "smoked" unlit cigarettes.

Simpson's paradox

Hospital A has higher overall death rate than hospital B. However, if we split the data in two parts, patients in good (+) and bad (−) conditions, for both parts hospital A performs better.

Hospital:	A	B	A+	B+	A–	B–
Died	63	16	6	8	57	8
Survived	2037	784	594	592	1443	192
Total	2100	800	600	600	1500	200
Death Rate	.030	.020	.010	.013	.038	.040

Here, the external factor, patient condition, is an example of a confounding factor:

$$\text{Hospital performance} \leftarrow \text{Patient condition} \rightarrow \text{Death rate}$$

Always remember that

CORRELATION DOES NOT IMPLY CAUSATION

8.6 Exercises

Problem 1

Four random numbers generated from a normal distribution

$$x_1 = 1.1650, \quad x_2 = 0.6268, \quad x_3 = 0.0751, \quad x_4 = 0.3516,$$

along with five random numbers with the same variance σ^2 but perhaps a different mean

$$y_1 = 0.3035, \quad y_2 = 2.6961, \quad y_3 = 1.0591, \quad y_4 = 2.7971, \quad y_5 = 1.2641.$$

(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) What do you think the variance of the random number generator was?

(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 $\sigma_x = \sigma_y$,

(b) $\frac{s_x^2}{n} + \frac{s_y^2}{m}$ 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

Let $x_1, \ldots, x_{25}$ be an iid-sample drawn from $\mathrm{N}(0.3, 1)$. Consider testing at $\alpha = 0.05$

$$H_0 : \mu = 0, \qquad H_1 : \mu > 0.$$

Compare

(a) the power of the sign test , and

(b) the power of the test based on the normal theory assuming that σ is known.

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

Data: millions of cycles until failure for two types of engine bearings

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 normal theory to test the null hypothesis of no difference against the two-sided alternative

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x \neq \mu_y.$$

(b) Test the hypothesis that there is no difference between the two types of bearing using a nonparametric method.

(c) Which of the methods (a) or (b) do you think is better in this case?

(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 two-sided 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 $\mathrm{P}(W \leq w_\alpha)$ 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 those obtained using the normal approximation of the null distribution.

Problem 9

Two population distributions with $\sigma_x = \sigma_y = 10$. Two samples of sizes $n = 25$ can be taken in two ways

(a) paired with $\mathrm{Cov}(X_i, Y_i) = 50$, $i = 1, \ldots, 25$,
(b) unpaired $x_1, \ldots, x_{25}$ and $y_1, \ldots, y_{25}$.

Compare the power curves for testing

$$H_0 : \mu_x = \mu_y, \qquad H_1 : \mu_x > \mu_y, \qquad \alpha = 0.05.$$

Problem 10

Lin, Sutton, and Qurashi (1979) compared microbiological and hydroxylamine methods for the analysis of ampicillin dosages. In one series of experiments, pairs of tablets were analysed by the two methods. The data in the table give the percentages of claimed amount of ampicillin found by the two methods in several pairs of tablets.

Microbiological method	Hydroxylamine method
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

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 the two methods.

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 school integration program involved busing children from minority schools to majority (primarily white) schools. Participation in the program was voluntary. It was found that the students who were bused scored lower on standardised tests than did their peers who chose not to be bused.

(e) When a group of students were asked to match pictures of newborns and with pictures of their mothers, they were correct 36% of the time.

(f) A survey found that that those who drank a moderate amount of beer were healthier than those who totally abstained from alcohol.

(g) 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.

(h) A University of Wisconsin study showed that within 10 years of wedding, 38% of those who had lived together before marriage had split up, compared to 27% of those who had married without a "trial period".

(i) 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.

9 Analysis of variance

The two sample setting from the previous section is the case of a single main factor having two levels. In this section, we extend the setting first to a single main factor with arbitrary many levels (one-way layout) and then to two main factors (two-way layout). Afterwards we introduce the two-way layout which generalises the paired sample setting of the previous section.

9.1 One-way layout

Consider the one-way layout model from Section 1.3. For each of the I levels of the main factor A, we independently collect an iid-sample $(y_{i1}, \ldots, y_{in})$ of the same size n. Having such I independent samples we want to develop a utility test of

$$H_0 : \mu_1 = \ldots = \mu_I, \text{ against } H_1 : \mu_u \neq \mu_v \text{ for some } (u, v).$$

Suppose 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 (so that the factor A has no influence of the measured response and the suggested one-way layout model is not useful).

Example: seven labs

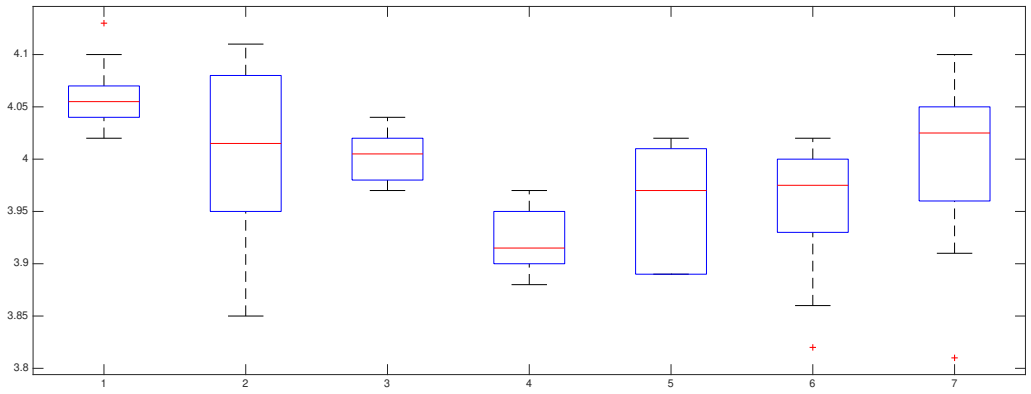
Data: 70 measurements of chlorpheniramine maleate in tablets with a nominal dosage of 4 mg. Seven labs made ten measurements each: $I = 7$, $n = 10$.

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 summarised below in the form of seven boxplots. Ordered 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

Here the null hypothesis of interest states that there is no significant difference between the output of the seven laboratories.



Normal theory model

Given $I \cdot n$ independent random variables

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \quad \epsilon_{ik} \sim \mathrm{N}(0, \sigma),$$

where $\alpha_i = \mu_i - \mu$ are the main effects such that

$$\alpha_1 + \ldots + \alpha_I = 0,$$

one can find the following maximum likelihood estimates

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\mu}_i = \bar{y}_{i.}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..},$$

expressed in terms of the sample means

$$\bar{y}_{i\cdot} = \tfrac{1}{n}\sum_k y_{ik}, \quad \bar{y}_{\cdot\cdot} = \tfrac{1}{I}\sum_i \bar{y}_{i\cdot} = \tfrac{1}{I\cdot n}\sum_i\sum_k y_{ik}.$$

The observed response values can be represented as

$$y_{ik} = \hat{\mu} + \hat{\alpha}_i + \hat{\epsilon}_{ik}, \quad \hat{\epsilon}_{ik} = y_{ik} - \bar{y}_{i\cdot}, \quad \sum_{i=1}^{I}\hat{\alpha}_i = 0,$$

where $\hat{\epsilon}_{ik}$ are the so-called residuals.
The ANOVA tests are built around the following observation:

$$\boxed{\text{Decomposition of the total sum of squares: } SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{E}}}$$

where
$SS_{\mathrm{T}} = \sum_i\sum_k (y_{ik} - \bar{y}_{\cdot\cdot})^2$ is the total sum of squares for the pooled sample with $\mathrm{df}_{\mathrm{T}} = I\cdot n - 1$,
$SS_{\mathrm{A}} = n\sum_i \hat{\alpha}_i^2$ is the factor A sum of squares with $\mathrm{df}_{\mathrm{A}} = I - 1$,
$SS_{\mathrm{E}} = \sum_i\sum_k \hat{\epsilon}_{ik}^2$ is the error sum of squares with $\mathrm{df}_{\mathrm{E}} = I\cdot(n-1)$.

This decomposition says that the total variation in response values is the sum of the between-group variation and the within-group variation. After normalising by the numbers of degrees of freedom, we obtain so-called the mean squares

$$MS_{\mathrm{A}} = \frac{SS_{\mathrm{A}}}{\mathrm{df}_{\mathrm{A}}}, \quad MS_{\mathrm{E}} = \frac{SS_{\mathrm{E}}}{\mathrm{df}_{\mathrm{E}}}.$$

If treated as random variables, they lead the following formulas for the expected values

$$\mathrm{E}(MS_{\mathrm{A}}) = \sigma^2 + \tfrac{n}{I-1}\sum_i \alpha_i^2, \qquad \mathrm{E}(MS_{\mathrm{E}}) = \sigma^2,$$

which suggest looking for the ratio between the two mean squares $\frac{MS_{\mathrm{A}}}{MS_{\mathrm{E}}}$ to find an evidence against the null hypothesis

$$H_0 : \alpha_1 = \ldots = \alpha_I = 0.$$

One-way F-test

The pooled sample variance

$$s_{\mathrm{p}}^2 = MS_{\mathrm{E}} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{n}(y_{ik} - \bar{y}_{i\cdot})^2}{I(n-1)}$$

is an unbiased estimate of σ^2. F-test rejection rule: use $F = \frac{MS_{\mathrm{A}}}{MS_{\mathrm{E}}}$ as a test statistic for and reject H_0 for large values of F based on the null distribution

$$F \overset{H_0}{\sim} F_{\mathrm{df}_1,\mathrm{df}_2}, \text{ where } \mathrm{df}_1 = I - 1,\ \mathrm{df}_2 = I(n-1).$$

F-distribution F_{n_1,n_2} with degrees of freedom (n_1, n_2) is the distribution for the ratio

$$\tfrac{X_1/n_1}{X_2/n_2} \sim F_{\mathrm{df}_1,\mathrm{df}_2}, \text{ where } X_1 \sim \chi^2_{\mathrm{df}_1} \text{ and } X_2 \sim \chi^2_{\mathrm{df}_2} \text{ are two independent random variables.}$$

The critical values of the F-distribution are given in Section 12.5.

Example: seven labs

The normal probability plot of residuals $\hat{\epsilon}_{ik}$ supports the normality assumption. Noise size σ is estimated by $s_{\mathrm{p}} = \sqrt{0.0037} = 0.061$. One-way Anova table

Source	df	SS	MS	F	P
Labs	6	.125	.0210	5.66	.0001
Error	63	.231	.0037		
Total	69	.356			

Conclusion: at least one of the $c = \binom{7}{2} = 21$ pairwise differences is significant.

9.2 Simultaneous confidence interval

Using the 95% confidence interval for a single pair of independent samples $(\mu_u - \mu_v)$ we get

$$I_{\mu_u-\mu_v} = (\bar{y}_{u.} - \bar{y}_{v.}) \pm t_{63}(0.025) \cdot \tfrac{s_\mathrm{p}}{\sqrt{5}} = (\bar{y}_{u.} - \bar{y}_{v.}) \pm 0.055,$$

where $t_{63}(0.025) = 2.00$. Notice that here we use the t-distribution with df $= 63$, since the corresponding pooled sample variance $s_\mathrm{p}^2 = 0.0037$ is based on $I = 7$ samples each of size $n = 10$. This formula yields 9 significant differences:

Labs	1–4	1–6	1–5	3–4	7–4	2–4	1–2	1–7	1–3	5–4
Diff	0.142	0.107	0.105	0.083	0.078	0.077	0.065	0.064	0.059	0.047

The multiple comparison problem: the above confidence interval formula is aimed at a single difference, and may produce false discoveries. We need a simultaneous confidence interval formula for all $c = 21$ pairwise comparisons.

Bonferroni method

Think of a statistical test repeatedly applied to c independent samples of size n. The overall result is positive if we get at least one positive result among these k tests. Observe that the overall significance level α is obtained, if each single test is performed at significance level $\alpha_c = \alpha/c$. Indeed, assuming the null hypothesis is true, the number of positive results is $X \sim \mathrm{Bin}(c, \alpha_c)$. Thus for small values of α_c,

$$\mathrm{P}(X \geq 1 | H_0) = 1 - (1 - \alpha_c)^c \approx c\alpha_c = \alpha.$$

This yields Bonferroni's formula of a $100(1-\alpha)\%$ simultaneous confidence interval which can be used as an first approximation for $c = \binom{I}{2}$ pairwise differences $(\mu_u - \mu_v)$:

$$B_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm t_\mathrm{df}(\tfrac{\alpha_c}{2}) \cdot s_\mathrm{p}\sqrt{\tfrac{2}{n}}, \quad 1 \leq u < v \leq I.$$

where df $= I(n-1)$ and $\alpha_c = \frac{2\alpha}{I(I-1)}$. Warning: pairwise differences $\mu_u - \mu_v$ are not independent as required by Bonferroni method, for example

$$\mu_1 - \mu_2 + \mu_2 - \mu_3 = \mu_1 - \mu_3,$$

Bonferroni method gives slightly wider intervals compared to the Tukey method introduced below.

Example: seven labs

Bonferroni 95% takes the form

$$B_{\mu_u-\mu_v} = (\bar{y}_{u.} - \bar{y}_{v.}) \pm t_{63}(\tfrac{.025}{21}) \cdot \tfrac{s_\mathrm{p}}{\sqrt{5}} = (\bar{y}_{u.} - \bar{y}_{v.}) \pm 0.086,$$

where $t_{63}(0.0012) = 3.17$, detects 3 significant differences between labs (1,4), (1,5), (1,6).

Tukey method

Under current assumptions, we have i.i.d random variables

$$Z_i = \bar{Y}_{i.} - \mu_i \sim \mathrm{N}(0, \tfrac{\sigma}{\sqrt{n}}), \quad i = 1, \ldots, I.$$

Consider the range of differences

$$R = \max\{Z_1, \ldots, Z_I\} - \min\{Z_1, \ldots, Z_I\}$$

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

$$\frac{R}{S_\mathrm{p}/\sqrt{n}} \sim \mathrm{SR}(I, \mathrm{df}), \quad \mathrm{df} = I(n-1).$$

The so-called studentised range distribution SR has two parameters: the number of samples and the number of degrees of freedom used in the variance estimate s_p^2. Tukey's simultaneous confidence interval is built using an appropriate quantile $q_{I,\mathrm{df}}(\alpha)$ of the studentised range distribution. In contrast to Bonferroni, Tukey takes into account the dependences between the differences $\mu_u - \mu_v$.

Tukey's $100(1-\alpha)\%$ simultaneous confidence interval $T_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm q_{I,\mathrm{df}}(\alpha) \cdot \frac{s_\mathrm{p}}{\sqrt{n}}$

Example: seven labs

Using R

```
> qtukey(0.95,7,63)
[1] 4.307136
```

we find $q_{7,63}(0.05) = 4.307$ so that

$$T_{\mu_u-\mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm q_{7,63}(0.05) \cdot \frac{0.061}{\sqrt{10}} = \bar{y}_{u.} - \bar{y}_{v.} \pm 0.083,$$

which puts forward four significant pairwise differences: (1,4), (1,5), (1,6), (3,4).

9.3 Kruskal-Wallis test

A nonparametric test, without assuming normality, for no treatment effect

$$H_0: \text{ all observations are equal in distribution.}$$

Extending the idea of the rank-sum test, consider the pooled sample of size $N = I \cdot n$. Let r_{ik} be the pooled ranks of the sample values y_{ik}, so that

$$\sum_i \sum_k r_{ik} = 1 + 2 + \ldots + N = \tfrac{N(N+1)}{2},$$

implying that the mean rank is $\bar{r}_{..} = \frac{N+1}{2}$.

$$\boxed{\text{Kruskal-Wallis test statistic } W = \tfrac{12n}{N(N+1)} \textstyle\sum_{i=1}^{I} (\bar{r}_{i.} - \frac{N+1}{2})^2}$$

Reject H_0 for large W using the null distribution table. For $I = 3$, $n \geq 5$ or $I \geq 4$, $n \geq 4$, 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 $1 \div 70$. (There is a tie 4.06 between laboratories 1 and 7, however, this is due to rounding.) With the observed Kruskal-Wallis test statistic $W = 28.17$ and df $= 6$, using χ^2_6-distribution table we get a p-value of approximately 0.0001.

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

9.4 Two-way layout

We assume that the data is generated in the following way

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n,$$

where $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma)$ are independent and have the same variance. Here we assume that for each combination of levels (i, j) of two main factors, $n \geq 2$ replications are performed. The maximum likelihood estimates

$$\hat{\mu} = \bar{y}_{...} = \tfrac{1}{IJn} \sum_i \sum_j \sum_k y_{ijk},$$

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \quad \bar{y}_{i..} = \tfrac{1}{Jn} \sum_j \sum_k y_{ijk},$$

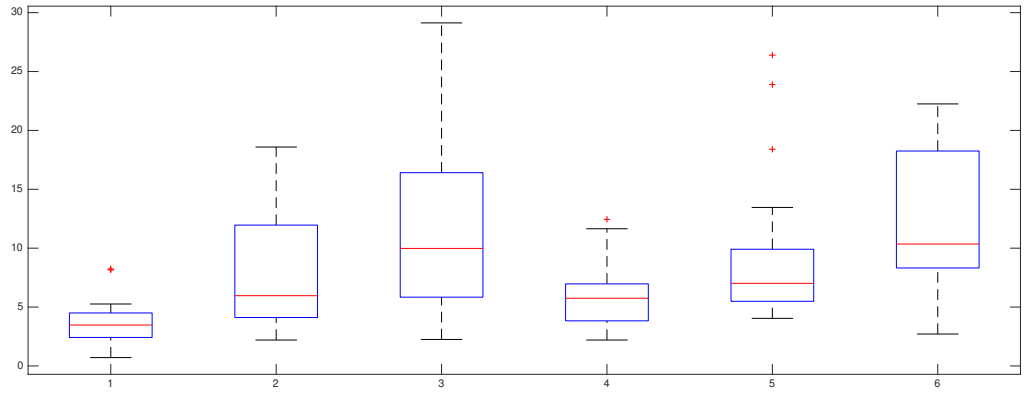
$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}, \quad \bar{y}_{.j.} = \tfrac{1}{In} \sum_i \sum_k y_{ijk},$$

$$\hat{\delta}_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}, \quad \bar{y}_{ij.} = \tfrac{1}{n} \sum_k y_{ijk},$$

bring a decomposition involving residuals

$$y_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\delta}_{ij} + \hat{\epsilon}_{ijk}, \quad \hat{\epsilon}_{ijk} = y_{ijk} - \bar{y}_{ij.}$$

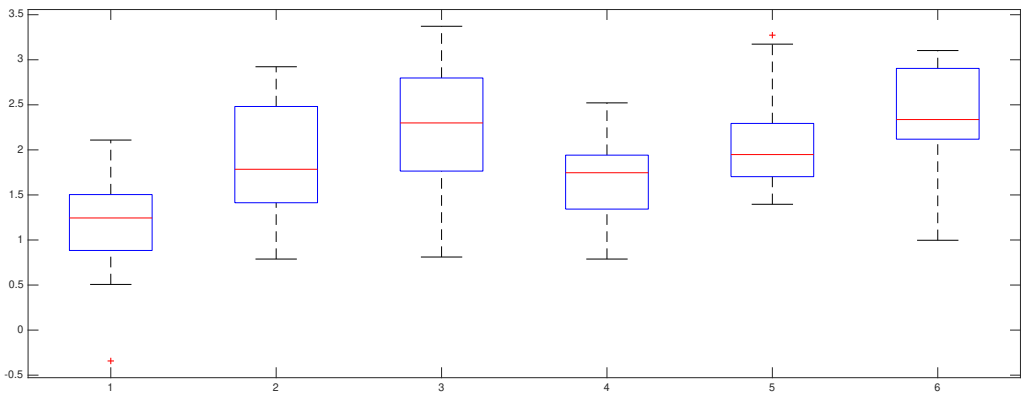
9.5 Case study: iron retention



The raw data z_{ijk} is the percentage of iron retained in mice. Factor A has two levels $I = 2$ representing two iron forms, while factor B has three levels $J = 3$ representing dosage concentrations. Six samples are collected with $n = 18$ replications for each (iron form, dosage) combination. Six boxplots for these six samples (see above) show that the raw data is not normally distributed having different variances across six samples.

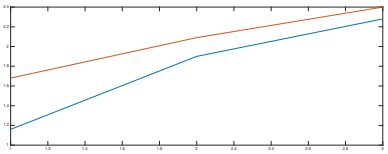
Fe^{3+} (10.2)	Fe^{3+} (1.2)	Fe^{3+} (0.3)	Fe^{2+} (10.2)	Fe^{2+} (1.2)	Fe^{2+} (0.3)
0.71	2.20	2.25	2.20	4.04	2.71
1.66	2.93	3.93	2.69	4.16	5.43
2.01	3.08	5.08	3.54	4.42	6.38
2.16	3.49	5.82	3.75	4.93	6.38
2.42	4.11	5.84	3.83	5.49	8.32
2.42	4.95	6.89	4.08	5.77	9.04
2.56	5.16	8.50	4.27	5.86	9.56
2.60	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

However, the transformed data $y_{ijk} = \ln(z_{ijk})$ produce more satisfactory boxplots.



The six sample means for the transformed data ($\bar{y}_{ij.}$)

	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



when depicted as two profiles for the two rows are not parallel which indicates possible interaction. The maximum likelihood estimates are

$$\bar{y}_{\cdots} = 1.92, \qquad \hat{\alpha}_1 = -0.14, \qquad \hat{\alpha}_2 = 0.14, \qquad \hat{\beta}_1 = -0.50, \qquad \hat{\beta}_2 = 0.08, \qquad \hat{\beta}_3 = 0.42,$$

and

$$(\hat{\delta}_{ij}) = \begin{pmatrix} -0.12 & 0.04 & 0.08 \\ 0.12 & -0.04 & -0.08 \end{pmatrix}$$

A two-way ANOVA table for the transformed iron retention data:

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			

According to the rightmost column

- the dosage effect is undoubtably significant, however, this is something expected,
- interaction is not statistically significant,
- there is significant effect due to iron form (compare to the previous analysis of two samples).

The estimated log scale difference $\hat{\alpha}_2 - \hat{\alpha}_1 = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = 0.28$ yields the multiplicative effect of $e^{0.28} = 1.32$ on the original scale, implying that the retention percentage of Fe^{2+} is 1.32 higher than that of Fe^{3+}.

Three F-tests

We explain the ANOVA table above by starting with the sums of squares decomposition

$$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{AB}} + SS_{\mathrm{E}},$$

where

$$SS_{\mathrm{T}} = \textstyle\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{\cdots})^2, \qquad \mathrm{df}_{\mathrm{T}} = IJn - 1$$

$$SS_{\mathrm{A}} = Jn \textstyle\sum_i \hat{\alpha}_i^2, \qquad \mathrm{df}_{\mathrm{A}} = I - 1$$

$$SS_{\mathrm{B}} = In \textstyle\sum_j \hat{\beta}_j^2, \qquad \mathrm{df}_{\mathrm{B}} = J - 1$$

$$SS_{\mathrm{AB}} = n \textstyle\sum_i \sum_j \hat{\delta}_{ij}^2, \qquad \mathrm{df}_{\mathrm{AB}} = (I-1)(J-1)$$

$$SS_{\mathrm{E}} = \textstyle\sum_i \sum_j \sum_k \hat{\epsilon}_{ijk}^2, \qquad \mathrm{df}_{\mathrm{E}} = IJ(n-1)$$

The mean sums of squares and their expected values

$$MS_{\mathrm{A}} = \tfrac{SS_A}{\mathrm{df}_{\mathrm{A}}}, \qquad \mathrm{E}(MS_{\mathrm{A}}) = \sigma^2 + \tfrac{Jn}{I-1} \textstyle\sum_i \alpha_i^2$$

$$MS_{\mathrm{B}} = \tfrac{SS_{\mathrm{B}}}{\mathrm{df}_{\mathrm{B}}}, \qquad \mathrm{E}(MS_{\mathrm{B}}) = \sigma^2 + \tfrac{In}{J-1} \textstyle\sum_j \beta_j^2$$

$$MS_{\mathrm{AB}} = \tfrac{SS_{\mathrm{AB}}}{\mathrm{df}_{\mathrm{AB}}}, \qquad \mathrm{E}(MS_{\mathrm{AB}}) = \sigma^2 + \tfrac{n}{(I-1)(J-1)} \textstyle\sum_i \sum_j \delta_{ij}^2$$

$$MS_{\mathrm{E}} = \tfrac{SS_{\mathrm{E}}}{\mathrm{df}_{\mathrm{E}}}, \qquad \mathrm{E}(MS_{\mathrm{E}}) = \sigma^2$$

Pooled sample variance $s_{\mathrm{p}}^2 = MS_{\mathrm{E}}$ is an unbiased estimate of σ^2.

Null hypothesis	No-effect property	Test statistics and null distribution
H_{A}: $\alpha_1 = \ldots = \alpha_I = 0$	$\mathrm{E}(MS_{\mathrm{A}}) = \sigma^2$	$F_{\mathrm{A}} = \frac{MS_{\mathrm{A}}}{MS_{\mathrm{E}}} \sim F_{\mathrm{df_A},\mathrm{df_E}}$
H_{B}: $\beta_1 = \ldots = \beta_J = 0$	$\mathrm{E}(MS_{\mathrm{B}}) = \sigma^2$	$F_{\mathrm{B}} = \frac{MS_{\mathrm{B}}}{MS_{\mathrm{E}}} \sim F_{\mathrm{df_B},\mathrm{df_E}}$
H_{AB}: all $\delta_{ij} = 0$	$\mathrm{E}(MS_{\mathrm{AB}}) = \sigma^2$	$F_{\mathrm{AB}} = \frac{MS_{\mathrm{AB}}}{MS_{\mathrm{E}}} \sim F_{\mathrm{df_{AB}},\mathrm{df_E}}$

Reject null hypothesis for large values of the respective test statistic.
Inspect normal probability plot for the residuals $\hat{\epsilon}_{ijk}$.

9.6 Randomised block design

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.

> Block what you can, randomise what you cannot.

Experimental design: randomly assign I treatments within each of J blocks.
Test the null hypothesis of no treatment effect using the two-way layout Anova.
The block effect is anticipated and is not of major interest. Examples:

Block	Treatments	Observation
A homogeneous plot of land divided into I subplots	I fertilizers 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 sinlings	the weight gain

Additive model

For the rest of this section suppose that $n = 1$. With only one replication per cell, then we cannot estimate interaction. This restricts us to the additive model without interaction

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma).$$

For the given data (y_{ij}), find the maximum likelihood estimates and residuals

$$\hat{\mu} = \bar{y}_{..}, \qquad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \qquad \hat{\beta}_i = \bar{y}_{.j} - \bar{y}_{..},$$
$$\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{..} - \hat{\alpha}_i - \hat{\beta}_i = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..},$$

yields a representation

$$y_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\epsilon}_{ij}.$$

Sums of squares decomposition takes a reduced form

$$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{E}},$$

with

$$SS_{\mathrm{T}} = \textstyle\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{..})^2, \qquad \mathrm{df_T} = IJ - 1$$

$$SS_{\mathrm{A}} = J \textstyle\sum_i \hat{\alpha}_i^2, \qquad \mathrm{df_A} = I - 1$$

$$SS_{\mathrm{B}} = I \textstyle\sum_j \hat{\beta}_j^2, \qquad \mathrm{df_B} = J - 1$$

$$SS_{\mathrm{E}} = \textstyle\sum_i \sum_j \hat{\epsilon}_{ij}^2, \qquad \mathrm{df_E} = (I-1)(J-1)$$

and

$$MS_{\mathrm{A}} = \tfrac{SS_{\mathrm{A}}}{\mathrm{df_A}}, \qquad \mathrm{E}(MS_{\mathrm{A}}) = \sigma^2 + \tfrac{J}{I-1} \textstyle\sum_i \alpha_i^2$$

$$MS_{\mathrm{B}} = \tfrac{SS_{\mathrm{B}}}{\mathrm{df_B}}, \qquad \mathrm{E}(MS_{\mathrm{B}}) = \sigma^2 + \tfrac{I}{J-1} \textstyle\sum_j \beta_j^2$$

$$MS_{\mathrm{E}} = \tfrac{SS_{\mathrm{E}}}{\mathrm{df_E}}, \qquad \mathrm{E}(\mathrm{MS_E}) = \sigma^2$$

We can apply two F-tests for two different null hypotheses

$$H_{\mathrm{A}} : \alpha_1 = \ldots = \alpha_I = 0, \qquad F_{\mathrm{A}} = \frac{MS_{\mathrm{A}}}{MS_{\mathrm{E}}} \overset{H_{\mathrm{A}}}{\sim} F_{\mathrm{df_A},\mathrm{df_E}},$$
$$H_{\mathrm{B}} : \beta_1 = \ldots = \beta_J = 0, \qquad F_{\mathrm{B}} = \frac{MS_{\mathrm{B}}}{MS_{\mathrm{E}}} \overset{H_{\mathrm{B}}}{\sim} F_{\mathrm{df_B},\mathrm{df_E}}.$$

Example: itching

Data: the duration of the itching in seconds y_{ij}, with $n = 1$ observation per cell, $I = 7$ treatments to relieve itching applied to $J = 10$ male volunteers aged 20-30.

Subject	No Drug	Placebo	Papaverine	Morphine	Aminophylline	Pentabarbital	Tripelennamine
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

Boxplots indicate violations of the assumptions of normality and equal variance. Notice much bigger variance for the placebo group. 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			

9.7 Friedman test

Here we introduce another nonparametric test, which does not require that ϵ_{ij} are normally distributed, for testing H_0: no treatment effect. The Friedman test is based on within block ranking. Let ranks within j-th block be:

$$(r_{1j}, \ldots, r_{Ij}) = \text{ranks of } (y_{1j}, \ldots, y_{Ij}),$$

so that

$$r_{1j} + \ldots + r_{Ij} = 1 + 2 + \ldots + I = \frac{I(I+1)}{2}.$$

For these ranks, we have $\frac{1}{I}(r_{1j} + \ldots + r_{Ij}) = \frac{I+1}{2}$ and therefore $\bar{r}_{\cdot\cdot} = \frac{I+1}{2}$.

Friedman test statistic $Q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} (\bar{r}_{i\cdot} - \frac{I+1}{2})^2$ has an approximate null distribution $Q \overset{H_0}{\approx} \chi^2_{I-1}$.

Test statistic Q is a measure of agreement between J rankings, so we reject H_0 for large values of Q.

Example: itching

From the rank values r_{ij} and $\bar{r}_{i\cdot}$ given in the next table and $\frac{I+1}{2} = 4$, we find the Friedman test statistic value to be $Q = 14.28$. Using the chi-squared distribution table with df $= 6$ we obtain the p-value is approximately 2.67%. We reject the null hypothesis of no effect even in the non-parametric setting.

Subject	No Drug	Placebo	Papaverine	Morphine	Aminophylline	Pentabarbital	Tripelennamine
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
$\bar{r}_{i\cdot}$	5.10	4.90	2.30	3.50	3.05	4.90	4.25

9.8 Exercises

Problem 1

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 data result:

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

Consider the null hypothesis of equality between the four treatment means of tensile strength.

(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 suggests the following formula for computing simultaneous 95% confidence intervals for six pairwise differences between four treatment means

$$B_{\mu_u - \mu_v} = (\bar{y}_{u.} - \bar{y}_{v.}) \pm t_{36}(\tfrac{0.025}{6}) \cdot 0.4472 \cdot s_{\mathrm{p}}.$$

Explain this formula and using it check which of the pairs of treatments have significantly different means.

Problem 2

For a one-way analysis of variance with two treatment groups, show that the F statistic is t^2, where t is the test statistic for a two-sample t-test.

Problem 3

Derive the likelihood ratio test for the null hypothesis of the one-way layout, 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.

10 Categorical data analysis

Categorical data appear in the form of a contingency table containing the sample counts for k various categories. The categorical population distribution is an extension of the Bernoulli distribution with several possible outcomes. Assuming n independent trials with the same probabilities of k outcomes $(\pi_1, \ldots, \pi_k)$ we arrive at a multinomial statistical model $\mathrm{Mn}(n, \pi_1, \ldots, \pi_k)$.

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 the 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	...	b_J
a_1	$\pi_{1\|1}$	$\pi_{1\|2}$	...	$\pi_{1\|J}$
a_2	$\pi_{2\|1}$	$\pi_{2\|2}$	...	$\pi_{2\|J}$
...	...	...	...	...
a_I	$\pi_{I\|1}$	$\pi_{I\|2}$	...	$\pi_{I\|J}$
Total	1	1	...	1

The null hypothesis of homogeneity states the equality of J population distributions

$$H_0 : \pi_{i|j} = \pi_i \quad \text{for all pairs } (i, j).$$

In this sense, the hypothesis of homogeneity is equivalent to the hypothesis of independence.

10.1 Chi-squared test of homogeneity

Consider a table of $I \times J$ observed counts obtained from J independent samples taken from J population distributions:

	Pop. 1	Pop. 2	...	Pop. J	Total
Category 1	n_{11}	n_{12}	...	n_{1J}	$n_{1\cdot}$
Category 2	n_{21}	n_{22}	...	n_{2J}	$n_{2\cdot}$
...	...	...	...	...	...
Category I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\cdot}$
Sample sizes	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	$n_{\cdot\cdot}$

This model is described by J multinomial distributions

$$(N_{1j}, \ldots, N_{Ij}) \sim \mathrm{Mn}(n_{\cdot j}; \pi_{1|j}, \ldots, \pi_{I|j}), \quad j = 1, \ldots, J.$$

The total number of degrees of freedom for J independent samples of size I 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

$$\hat{\pi}_i = n_{i\cdot}/n_{\cdot\cdot}, \quad i = 1, \ldots, I.$$

These estimates consumes $(I-1)$ degrees of freedom, since their sum is 1. Using these maximum likelihood estimates we compute the expected cell counts

$$E_{ij} = n_{\cdot j} \cdot \hat{\pi}_i = n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}$$

and the chi-squared test statistic becomes

$$\mathrm{X}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}}$$

We reject H_0 for large values of χ^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 (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

The chi-squared test statistic is

$$\mathrm{X}^2 = 27.24 \text{ with df} = (3-1)\cdot(3-1) = 4.$$

After comparing X^2 with the table value $\chi^2_4(0.005) = 14.86$, we reject the hypothesis of homogeneity at 0.5% significance level. Persons who saw themselves as cautious conservatives are more likely to express a favourable opinion of small cars.

10.2 Chi-squared test of independence

Data: a single cross-classifying sample is summarised in terms of the observed counts,

	b_1	b_2	...	b_J	Total
a_1	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
a_2	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...
a_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	$n_{..}$

whose joint distribution is multinomial

$$(N_{11}, \ldots, N_{IJ}) \sim \text{ Mn}(n_{..}; \pi_{11}, \ldots, \pi_{IJ})$$

The maximum likelihood estimates of $\pi_{i.}$ and $\pi_{.j}$ are $\hat{\pi}_{i.} = \frac{n_{i.}}{n_{..}}$ and $\hat{\pi}_{.j} = \frac{n_{.j}}{n_{..}}$. Therefore, under the hypothesis of independence $\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}^2}$ implying the same expected cell counts as before

$$E_{ij} = n_{..}\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

with the same

$$\text{df} = (IJ-1) - (I-1+J-1) = (I-1)(J-1).$$

The same chi-squared test rejection rule for the homogeneity test and independence test.

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

The observed chi-squared test statistic is $\mathrm{X}^2 = 16.01$. With df = 1 we can use the normal distribution table, since $Z \sim \mathrm{N}(0,1)$ is equivalent to $Z^2 \sim \chi^2_1$. Thus

$$\mathrm{P}(\mathrm{X}^2 > 16.01) \approx \mathrm{P}(|Z| > 4.001) = 2(1 - \Phi(4.001)).$$

We see that the p-value of the test is less that 0.1%, and we reject the null hypothesis of independence. College-educated women, once they marry, are less likely to divorce.

10.3 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

	X	X^c
D	n_{11}	n_{12}
D^c	n_{21}	n_{22}

where the four counts distinguish among sampled individual who are

either D = affected (have the **D**isease) or D^c = unaffected,
and either X = e**X**posed (had tonsillectomy) or X^c = non-exposed.

Three possible sampling designs:

- simple random sampling would give counts like $\begin{pmatrix} 0 & 0 \\ 0 & n \end{pmatrix}$,
- prospective study (take an X-sample and a control X^c-sample, then watch who gets affected) would give $\begin{pmatrix} 0 & 0 \\ n_1 & n_2 \end{pmatrix}$,
- retrospective study (take a D-sample and a control D^c-sample, then find who had been exposed) would give informative counts.

Two retrospective case-control studies produced opposite results. Study A (Vianna, Greenwald, Davis, 1971) gave a cross classification table

Study A	X	X^c
D	67	34
D^c	43	64

The chi-squared test of homogeneity was applied. With $\mathrm{X}^2_{\mathrm{A}} = 14.29$ and df = 1, the p-value was found to be very small

$$\mathrm{P}(\mathrm{X}^2_{\mathrm{A}} \geq 14.29) \approx 2(1 - \Phi(\sqrt{14.29})) = 0.0002.$$

Study B (Johnson and Johnson, 1972) was summarised by a table

Study B	X	X^c
D	41	44
D^c	33	52

and the chi-squared tests of homogeneity was applied. With $\mathrm{X}^2_{\mathrm{B}} = 1.53$ and df = 1, the p-value was strikingly different

$$\mathrm{P}(\mathrm{X}^2_{\mathrm{B}} \geq 1.53) \approx 2(1 - \Phi(\sqrt{1.53})) = 0.215.$$

It turned out that the study B was based on a matched-pairs design violating the assumption of the chi-squared test of homogeneity. The sample consisted of $m = 85$ 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: (X, X), (X, X^c), (X^c, X), (X^c, X^c)

	unaffected X	unaffected X^c	Total
affected X	$m_{11} = 26$	$m_{12} = 15$	41
affected X^c	$m_{21} = 7$	$m_{22} = 37$	44
Total	33	52	85

Notice that this contingency table contains more information than the previous one.
An appropriate test in this setting is McNemar's test (see below). For the data of study B, the McNemar's test statistic is

$$\mathrm{X}^2 = \frac{(m_{12} - m_{21})^2}{m_{12} + m_{21}} = 2.91,$$

giving the p-value of

$$\mathrm{P}(\mathrm{X}^2 \geq 2.91) \approx 2(1 - \Phi(\sqrt{2.91})) = 0.09.$$

The correct p-value is much smaller than that of 0.215 computed using the test of homogeneity. Since there are very few informative, only $m_{12} + m_{21} = 22$, observations, more data is required.

McNemar's test

Consider data obtained by matched-pairs design for the population distribution

	unaffected X	unaffected X^c	Total
affected X	p_{11}	p_{12}	$p_{1.}$
affected X^c	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

The relevant null hypothesis is not the hypothesis of independence but rather

$$H_0\text{: } p_{1.} = p_{.1}, \text{ or equivalently, } H_0\text{: } p_{12} = p_{21} = p \text{ for an unspecified } p.$$

The maximum likelihood estimates for the population frequencies under the null hypothesis are

$$\hat{p}_{11} = \frac{m_{11}}{m}, \quad \hat{p}_{22} = \frac{m_{22}}{m}, \quad \hat{p}_{12} = \hat{p}_{21} = \hat{p} = \frac{m_{12} + m_{21}}{2m}.$$

These yield the McNemar test statistic of the form

$$\mathrm{X}^2 = \sum_i \sum_j \frac{(m_{ij} - m\hat{p}_{ij})^2}{m\hat{p}_{ij}} = \frac{(m_{12} - m_{21})^2}{m_{12} + m_{21}},$$

whose approximate null distribution is χ^2_1, where

$$\mathrm{df} = 4 - 1 - 2$$

because 2 independent parameters are estimated from the data. We reject the H_0 for large values of the test statistic.

10.4 Odds ratios

Odds and probability of a random event A:

$$\mathrm{odds}(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(\bar{A})} \quad \text{and} \quad \mathrm{P}(A) = \frac{\mathrm{odds}(A)}{1 + \mathrm{odds}(A)}.$$

Notice that for small $\mathrm{P}(A)$,

$$\mathrm{odds}(A) \approx \mathrm{P}(A).$$

Conditional odds for A given B are defined as

$$\mathrm{odds}(A|B) = \frac{\mathrm{P}(A|B)}{\mathrm{P}(A^c|B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A^cB)}.$$

Odds ratio for a pair of events defined by

$$\Delta_{AB} = \frac{\mathrm{odds}(A|B)}{\mathrm{odds}(A|B^c)} = \frac{\mathrm{P}(AB)\mathrm{P}(A^cB^c)}{\mathrm{P}(A^cB)\mathrm{P}(AB^c)},$$

has the properties

$$\Delta_{AB} = \Delta_{BA}, \quad \Delta_{AB^c} = \frac{1}{\Delta_{AB}}.$$

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 $\mathrm{P}(A|B) > \mathrm{P}(A|B^c)$ so that B increases probability of A,
if $\Delta_{AB} < 1$, then $\mathrm{P}(A|B) < \mathrm{P}(A|B^c)$ so that B decreases probability of A.

Odds ratios for case-control studies

Return to conditional probabilities and observed counts

	X	X^c	Total
D	$\mathrm{P}(X\|D)$	$\mathrm{P}(X^c\|D)$	1
D^c	$\mathrm{P}(X\|D^c)$	$\mathrm{P}(X^c\|D^c)$	1

	X	X^c	Total
D	n_{11}	n_{12}	$n_{1.}$
D^c	n_{21}	n_{22}	$n_{2.}$

The corresponding odds ratio

$$\Delta_{DX} = \frac{\mathrm{P}(X|D)\mathrm{P}(X^c|D^c)}{\mathrm{P}(X^c|D)\mathrm{P}(X|D^c)} = \frac{\mathrm{odds}(D|X)}{\mathrm{odds}(D|X^c)}$$

quantifies the influence of eXposition to a certain factor on the onset of the Disease in question. This odds ratio can be estimated using the observed counts as

$$\widehat{\Delta}_{DX} = \frac{(n_{11}/n_{1\cdot})(n_{22}/n_{2\cdot})}{(n_{12}/n_{1\cdot})(n_{21}/n_{2\cdot})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Example: Hodgkin's disease

Study A gives the odds ratio

$$\widehat{\Delta}_{DX} = \frac{67 \cdot 64}{43 \cdot 34} = 2.93.$$

Conclusion: tonsillectomy increases the odds for Hodgkin's onset by factor 2.93.

10.5 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 nondiabetics. The data is shown in the following table:

	Diabetic	Normal	Total
Bb or *bb*	12	4	16
BB	39	49	88
Total	51	53	104

Are the relative frequencies of the alleles significantly different in the two groups?

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?

Problem 8

Hill and Barton (2005): 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

(a) Let π_R denote the probability that the contestant wearing red wins. Test the null hypothesis that $\pi_R = 0.5$ versus the alternative hypothesis that π_R is the same in each sport, but $\pi_R \neq 0.5$.

(b) Test the null hypothesis that $\pi_R = 0.5$ versus the alternative hypothesis that allows π_R to be different in different sports, but not equal to 0.5.

(c) Are these hypothesis tests equivalent to that which would test the null hypothesis $\pi_R = 0.5$ versus the alternative hypothesis $\pi_R \neq 0.5$, using as data the total numbers of wins summed over all the sports?

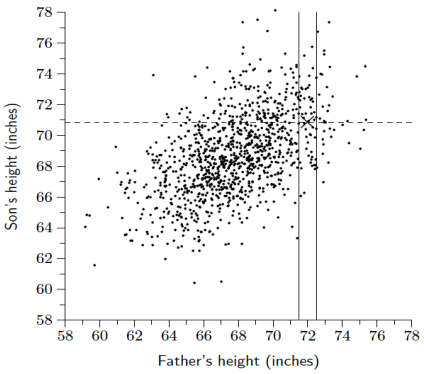
(d) Is there any evidence that wearing red is more favourable in some of the sports than others?

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?

11 Multiple regression

Pearson's father-son data. The following scatter diagram shows the heights of 1,078 fathers and their full-grown sons, in England, circa 1900. There is one dot for each father-son pair.



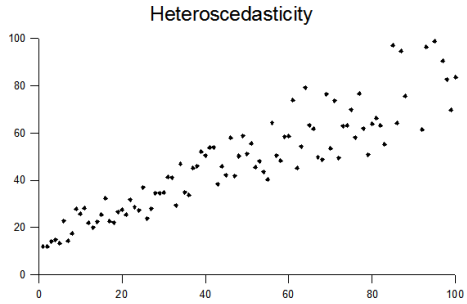
Focussing on 6 feet tall fathers (above average height), we see that their sons on average are shorter than their fathers. Francis Galton called this phenomenon *regression to mediocrity*.

11.1 Simple linear regression model

A simple linear regression model is based on the linear relation

$$Y(x) = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \sigma),$$

where ϵ is the noisy part of the response, that is not explained by the value x of the main explanatory variable. The key assumption of *homoscedasticity* requires that the standard deviation of ϵ is independent of the x-value. Whenever this assumption is violated, the situation is described by the term *heteroscedasticity*.



For a given collection of x-values $(x_1, \ldots, x_n)$, and a vector $(e_1, \ldots, e_n)$ of independent realisations of the noise component, we get a sample of response values

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n.$$

The corresponding likelihood is a function of the three-dimensional parameter $\theta = (\beta_0, \beta_1, \sigma^2)$

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \Big\} = C\sigma^{-n} e^{-\frac{S(\beta_0,\beta_1)}{2\sigma^2}},$$

where

$$C = (2\pi)^{-n/2}, \qquad S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This implies the following expression for the log-likelihood function $l(\theta) = \ln L(\theta)$

$$l(\theta) = \ln C - n \ln \sigma - \tfrac{S(\beta_0,\beta_1)}{2\sigma^2}.$$

Observe that

$$n^{-1} S(\beta_0, \beta_1) = n^{-1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \beta_0^2 + 2\beta_0\beta_1 \bar{x} - 2\beta_0 \bar{y} - 2\beta_1 \overline{xy} + \beta_1^2 \overline{x^2} + \overline{y^2},$$

where

$$\bar{x} = \tfrac{x_1 + \ldots + x_n}{n}, \quad \bar{y} = \tfrac{y_1 + \ldots + y_n}{n}, \quad \overline{x^2} = \tfrac{x_1^2 + \ldots + x_n^2}{n}, \quad \overline{y^2} = \tfrac{y_1^2 + \ldots + y_n^2}{n}, \quad \overline{xy} = \tfrac{x_1 y_1 + \ldots + x_n y_n}{n}$$

delineate a set of sufficient statistics.
To obtain the maximum likelihood estimates of $\theta = (\beta_0, \beta_1, \sigma^2)$ compute the derivatives

$$\begin{aligned} \tfrac{\partial l}{\partial \beta_0} &= -\tfrac{1}{2\sigma^2} \tfrac{\partial S}{\partial \beta_0}, \\ \tfrac{\partial l}{\partial \beta_1} &= -\tfrac{1}{2\sigma^2} \tfrac{\partial S}{\partial \beta_1}, \\ \tfrac{\partial l}{\partial \sigma^2} &= -\tfrac{n}{2\sigma^2} + \tfrac{S(\beta_0,\beta_1)}{2\sigma^4}, \end{aligned}$$

and set them equal to zeros. Putting $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$, we get the so-called normal equations:

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y}, \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy}, \end{cases} \qquad \text{implying} \qquad \begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{r s_y}{s_x}, \\ b_0 = \bar{y} - b_1 \bar{x}. \end{cases}$$

where

$$s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2.$$

As a result, the fitted regression line $y = b_0 + b_1 x$ takes the form

$$y = \bar{y} + r \tfrac{s_y}{s_x} (x - \bar{x}),$$

involving the sample correlation coefficient defined using the sample covariance:

$$r = \frac{s_{xy}}{s_x s_y}, \qquad s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

Notice that the maximum likelihood estimates (b_0, b_1) of parameters (β_0, β_1) are obtained by minimising the sum of squares $S(\beta_0, \beta_1)$. Therefore, they are called the least squares estimates. Warning: the least squares estimates (b_0, b_1) are not robust against outliers exerting leverage on the fitted line.
Putting $\frac{\partial l}{\partial \sigma^2} = 0$, and replacing (β_0, β_1) with (b_0, b_1), we find the maximum likelihood estimate of σ^2 to be

$$\hat{\sigma}^2 = \frac{S(b_0, b_1)}{n},$$

where

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

and

$$\hat{y}_i = b_0 + b_1 x_i$$

are the so-called predicted responses. The maximum likelihood estimate of $\hat{\sigma}^2$ is a biased but asymptotically unbiased estimate of σ^2. An unbiased estimate of σ^2 is given by

$$s^2 = \frac{S(b_0, b_1)}{n-2}.$$

11.2 Residuals

The fitted regression line

$$y = b_0 + b_1 x = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x}),$$

is used for prediction of the response to a given predictor value x. For the given sample $(x_1, y_1), \ldots, (x_n, y_n)$, we now compare the actual responses y_i with the predicted responses $\hat{y}_i$. Introducing residuals by

$$\hat{e}_i = y_i - \hat{y}_i,$$

we can write

$$y_i = \hat{y}_i + \hat{e}_i, \quad i = 1, \ldots, n,$$

The residuals $(\hat{e}_1, \ldots, \hat{e}_n)$ are linearly connected via

$$\hat{e}_1 + \ldots + \hat{e}_n = 0, \qquad x_1\hat{e}_1 + \ldots + x_n\hat{e}_n = 0, \qquad \hat{y}_1\hat{e}_1 + \ldots + \hat{y}_n\hat{e}_n = 0,$$

so we can say that $\hat{e}_i$ are uncorrelated with x_i and $\hat{e}_i$ are uncorrelated with $\hat{y}_i$. The residuals $\hat{e}_i$ are realisations of random variables having normal distributions with zero means and

$$\mathrm{Var}(\hat{e}_i) = \sigma^2\Big(1 - \frac{\sum_k (x_k - x_i)^2}{n(n-1)s_x^2}\Big), \qquad \mathrm{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 \cdot \frac{\sum_k (x_k - x_i)(x_k - x_j)}{n(n-1)s_x^2}.$$

The error sum of squares

$$SS_{\mathrm{E}} = S(b_0, b_1) = \sum (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2$$

can be expressed as

$$SS_{\mathrm{E}} = \sum_i (y_i - \bar{y})^2 - 2r\frac{s_y}{s_x} n(\overline{xy} - \bar{y}\bar{x}) + r^2\frac{s_y^2}{s_x^2}\sum_i (x_i - \bar{x})^2 = (n-1)s_y^2(1 - r^2).$$

This leads to the following useful expression

$$s^2 = \frac{n-1}{n-2}\, s_y^2(1 - r^2).$$

Using

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{e}_i$$

we obtain a decomposition

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{E}},$$

where

$$SS_{\mathrm{T}} = \sum_i (y_i - \bar{y})^2 = (n-1)s_y^2$$

is the total sum of squares, and

$$SS_{\mathrm{R}} = \sum_i (\hat{y}_i - \bar{y})^2 = (n-1) b_1^2 s_x^2$$

is the regression sum of squares. Combining these relations, we find that

$$r^2 = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{T}}} = 1 - \frac{SS_{\mathrm{E}}}{SS_{\mathrm{T}}}.$$

This relation explains 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. Therefore, r^2 has a more intuitive meaning than the sample correlation coefficient r.
To test the normality assumption, use the normal distribution plot for the standardised residuals

$$\tilde{e}_i = \frac{\hat{e}_i}{s_i}, \quad i = 1, \ldots, n,$$

where

$$s_i = s\sqrt{1 - \frac{\sum_k (x_k - x_i)^2}{n(n-1)s_x^2}}$$

are the estimated standard deviations of $\hat{E}_i$. Within the simple linear regression model, the scatter plot of the standardised residuals versus x_i should look as a horizontal blur. If the linear model is not valid, it will show up in a somewhat bowed shape of the residual scatter plot. In some cases, the non-linearity problem can fixed by a log-log transformation of the data.

11.3 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

$$B_0 \sim \mathrm{N}(\beta_0, \sigma_0), \qquad \sigma_0^2 = \frac{\sigma^2 \sum x_i^2}{n(n-1)s_x^2}, \qquad s_{b_0}^2 = \frac{s^2 \sum x_i^2}{n(n-1)s_x^2}, \qquad \frac{B_0 - \beta_0}{S_{B_0}} \sim t_{n-2},$$

$$B_1 \sim \mathrm{N}(\beta_1, \sigma_1), \qquad \sigma_1^2 = \frac{\sigma^2}{(n-1)s_x^2}, \qquad s_{b_1}^2 = \frac{s^2}{(n-1)s_x^2}, \qquad \frac{B_1 - \beta_1}{S_{B_1}} \sim t_{n-2}.$$

There is a weak correlation between the two estimators:

$$\mathrm{Cov}(B_0, B_1) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2}$$

which is negative, if $\bar{x} > 0$, and positive, if $\bar{x} < 0$.

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 would like to the 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.

1. 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,

   $$t = \frac{b_1}{s_{b_1}} = \frac{\frac{rs_y}{s_x}}{\sqrt{\frac{s^2}{(n-1)s_x^2}}} = \frac{rs_y}{\sqrt{\frac{s_y^2(1-r^2)}{(n-2)}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

   The corresponding null distribution is t_{n-2}.

2. Zero-intercept test aims at

$$H_0 : \beta_0 = 0.$$

Compute its t-value

$$t = b_0/s_{b_0},$$

and find whether this value is significant using the t-distribution with df $= n - 2$.

11.4 Intervals for individual observations

Given the earlier sample of size n consider a new value $x = x_0$ of the predictor variable. We wish to say something on the response value

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon.$$

Its expected value

$$\mu_0 = \beta_0 + \beta_1 x_0$$

is estimated by

$$\hat{\mu}_0 = b_0 + b_1 x_0.$$

The standard error of $\hat{\mu}_0$ is computed as the square root of

$$\mathrm{Var}(\hat{\mu}_0) = \tfrac{\sigma^2}{n} + \tfrac{\sigma^2}{n-1} \cdot (\tfrac{x_0 - \bar{x}}{s_x})^2.$$

Exact $100(1-\alpha)\%$ confidence interval $I_{\mu_0} = b_0 + b_1 x_0 \pm t_{n-2}(\frac{\alpha}{2}) \cdot s\sqrt{\frac{1}{n} + \frac{1}{n-1}(\frac{x_0 - \bar{x}}{s_x})^2}$
Exact $100(1-\alpha)\%$ prediction interval $I_{Y_0} = b_0 + b_1 x_0 \pm t_{n-2}(\frac{\alpha}{2}) \cdot s\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}(\frac{x_0 - \bar{x}}{s_x})^2}$

Prediction interval has wider limits since it contains the uncertainty due the noise factors:

$$\mathrm{Var}(Y_0 - \hat{\mu}_0) = \mathrm{Var}(\mu_0 + \epsilon - \hat{\mu}_0) = \sigma^2 + \mathrm{Var}(\hat{\mu}_0) = \sigma^2(1 + \tfrac{1}{n} + \tfrac{1}{n-1} \cdot (\tfrac{x_0 - \bar{x}}{s_x})^2).$$

Graphically compare the formulas for I_{μ_0} and prediction interval I_{Y_0} by drawing the confidence bands around the regression line both for the individual observation Y_0 and the mean μ_0.

11.5 Multiple linear regression

With $p-1$ predictors, the corresponding data set consists of n vectors $(x_{i,1}, \ldots, x_{i,p-1}, y_i)$ with $n > p$ and

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \ldots + \beta_{p-1} x_{1,p-1} + e_1, \\ &\ldots \\ y_n &= \beta_0 + \beta_1 x_{n,1} + \ldots + \beta_{p-1} x_{n,p-1} + e_n, \end{aligned}$$

where $e_1, \ldots, e_n$ are independent realisations of a homoscedastic random noise

$$\epsilon \sim \ \mathrm{N}(0, \sigma).$$

It is very convenient to use the matrix notation

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where

$$\mathbf{y} = (y_1, \ldots, y_n)^\intercal, \quad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\intercal, \quad \boldsymbol{e} = (e_1, \ldots, e_n)^\intercal,$$

are column vectors, and $\mathbb{X}$ is the so called design matrix

$$\mathbb{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,p-1} \end{pmatrix}$$

assumed to have rank p.
The machinery developed for the simple linear regression model works well for the multiple regression. The least squares estimates $\mathbf{b} = (b_0, \ldots, b_{p-1})^\intercal$ minimise

$$S(\mathbf{b}) = \|\mathbf{y} - \mathbb{X}\mathbf{b}\|^2,$$

where

$$\|\mathbf{a}\|^2 = a_1^2 + \ldots + a_k^2, \qquad \mathbf{a} = (a_1, \ldots, a_k).$$

Solving the normal equations $\mathbb{X}^\intercal\mathbb{X}\mathbf{b} = \mathbb{X}^\intercal\mathbf{y}$ we find the least squares estimates:

$$\mathbf{b} = \mathbb{M}\mathbb{X}^\intercal\mathbf{y}, \quad \mathbb{M} = (\mathbb{X}^\intercal\mathbb{X})^{-1}.$$

In particular, in the simple linear regression case with $p = 2$, we have

$$\mathbb{X}^\intercal = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix}, \qquad \mathbb{X}^\intercal\mathbb{X} = \begin{pmatrix} n & x_1 + \ldots + x_n \\ x_1 + \ldots + x_n & x_1^2 + \ldots + x_n^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}$$

so that

$$\mathbb{M} = (\mathbb{X}^\intercal\mathbb{X})^{-1} = \frac{1}{n(\overline{x^2} - (\bar{x})^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}, \qquad \mathbf{b} = \mathbb{M}\mathbb{X}^\intercal\mathbf{y} = \frac{1}{\overline{x^2} - (\bar{x})^2} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

Least squares multiple regression: predicted responses $\hat{\mathbf{y}} = \mathbb{X}\mathbf{b} = \mathbb{P}\mathbf{y}$, where $\mathbb{P} = \mathbb{X}\mathbb{M}\mathbb{X}^\intercal$.

Check that $\mathbb{P}$ is a projection matrix such that $\mathbb{P}^2 = \mathbb{P}$.
Turning to the random vector $\mathbf{B}$ behind the the least squares estimates $\mathbf{b}$, we find that

$$\mathrm{E}(\mathbf{B}) = \boldsymbol{\beta}.$$

Furthermore, the covariance matrix, the $p \times p$ matrix with elements $\mathrm{Cov}(B_i, B_j)$, is given by

$$\mathrm{E}\{(\mathbf{B} - \boldsymbol{\beta})(\mathbf{B} - \boldsymbol{\beta})^\intercal\} = \sigma^2\mathbb{M}.$$

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 $\sigma^2(\mathbb{I} - \mathbb{P})$.

An unbiased estimate of σ^2 is given by $s^2 = \frac{SS_\mathrm{E}}{n-p}$, where $SS_\mathrm{E} = \|\hat{\boldsymbol{e}}\|^2$.

Denote by

$$m_{11}, m_{22}, \ldots, m_{p-1,p-1}, m_{pp}$$

the diagonal elements of matrix $\mathbb{M}$. Then the standard error of b_j is computed as

$$s_{b_j} = s\sqrt{m_{j+1,j+1}}.$$

Exact sampling distributions $\frac{B_j - \beta_j}{S_{B_j}} \sim t_{n-p}, \quad j = 0, 1, \ldots, p-1.$

To check the underlying normality assumption inspect the normal probability plot for the standardised residuals $\frac{\hat{e}_i}{s\sqrt{1-p_{ii}}}$, where p_{ii} are the diagonal elements of $\mathbb{P}$.

Coefficient of multiple determination can be computed similarly to the simple linear regression model as

$$R^2 = 1 - \frac{SS_\mathrm{E}}{SS_\mathrm{T}},$$

where $SS_\mathrm{T} = (n-1)s_y^2$. The problem with R^2 is that it increases even if irrelevant variables are added to the model. To punish for irrelevant variables it is better to use the adjusted coefficient of multiple determination

$$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SS_\mathrm{E}}{SS_\mathrm{T}} = 1 - \frac{s^2}{s_y^2}.$$

The adjustment factor $\frac{n-1}{n-p}$ gets larger for the larger values of predictors p.

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	.34	.29	.28	.42	.29	.41	.76	.73	.46	.40
Flow rate y	.64	.32	.73	1.33	.49	.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 $x_2 = x^2$.

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

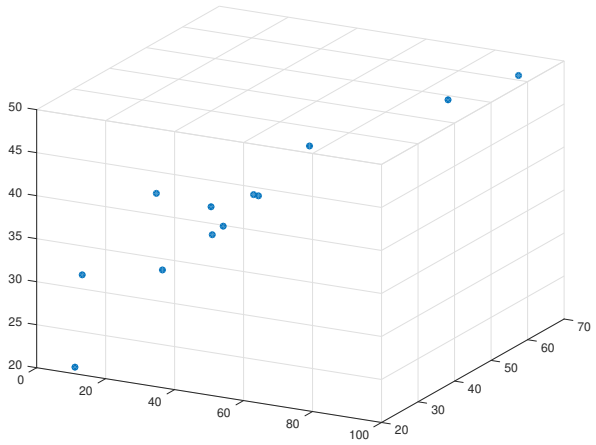
The residuals show no sign of systematic misfit. The test statistic $t = 5.51$ of the utility test of $H_0 : \beta_2 = 0$ shows that the quadratic term in the model is statistically significant.

> Empirical relationship developed in a region might break down,
> if extrapolated to a wider region in which no data been observed

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



We start with two simple linear regressions

$$\text{H-model: } L = \beta_0 + \beta_1 H + \epsilon, \qquad \text{W-model: } L = \beta_0 + \beta_1 W + \epsilon.$$

The analysis of these two models is summarised as follows

Estimate	H-model	t value	W-model	t value
$b_0(s_{b_0})$	12.1(4.3)	2.8	25.6(2.0)	12.8
$b_1(s_{b_1})$	0.60(0.10)	6.0	0.28(0.04)	7.0
s	4.0		3.8	
r^2	0.78		0.80	
R_a^2	0.76		0.78	

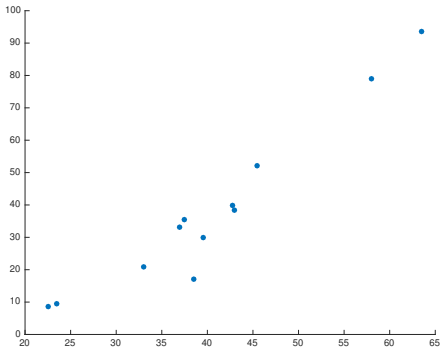
The plots of standardised residuals do not contradict the normality assumptions.
These two simple regression models should be compared to the multiple regression model

$$L = \beta_0 + \beta_1 H + \beta_2 W + \epsilon,$$

which gives

$$
\begin{array}{lll}
b_0 = 21, & s_{b_0} = 8.8, & b_0/s_{b_0} = 2.39, \\
b_1 = 0.20, & s_{b_1} = 0.36, & b_1/s_{b_1} = 0.56, \\
b_2 = 0.19, & s_{b_2} = 0.17, & b_2/s_{b_2} = 1.12, \\
s = 3.9, & R^2 = 0.81, & R_a^2 = 0.77.
\end{array}
$$

In contrast to the simple models, we can not reject neither $H_1 : \beta_1 = 0$ nor $H_2 : \beta_2 = 0$. This paradox is explained by 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.
Collinearity problem: height and weight have a strong linear relationship. The fitted plane has a well resolved slope along the line about which the (H, W) points fall and poorly resolved slopes along the H and W axes.



Conclusion: since the simple W-model

$$L = \beta_0 + \beta_1 W + \epsilon$$

gives the highest adjusted coefficient of determination, there is little or no gain from adding H to the regression model model with a single explanatory variable W.

11.6 Exercises

Problem 1

Suppose we are given a two-dimensional iid-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 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) Are the lines on (a) and (b) the same? If not, why not?

Problem 3

Two consecutive grades

X = the high school GPA (grade point average),
Y = the freshman GPA.

Allow two different intercepts for females

$$Y = \beta_F + \beta_1 X + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma)$$

and for males

$$Y = \beta_M + \beta_1 X + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma).$$

Give the form of the design matrix for such a model.

Problem 4

Simple linear regression model

$$Y(x) = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma).$$

Using n pairs of (x_i, y_i) we fit a regression line by

$$y = b_0 + b_1 x, \quad \mathrm{Var}(B_0) = \frac{\sigma^2 \overline{x^2}}{(n-1)s_x^2}, \quad \mathrm{Var}(B_1) = \frac{\sigma^2}{(n-1)s_x^2}, \quad \mathrm{Cov}(B_0, B_1) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2}.$$

For a given $x = x_0$, we wish to predict the value of a new observation

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$$

by

$$\hat{y}_0 = b_0 + b_1 x_0.$$

(a) Find an expression for the variance of $\hat{Y}_0 - Y_0$, and compare it to the variance of $\hat{Y}_0$. Find a_n, the standard deviation of $\frac{\hat{Y}_0 - Y_0}{\sigma}$.

(b) Derive a formula for 95% prediction interval I such that

$$\mathrm{P}(Y_0 \in I) = 0.95$$

using

$$\frac{Y_0 - \hat{Y}_0}{S a_n} \sim t_{n-2}.$$

Problem 5

Data collected for

x = midterm grade,
y = final grade,

gave

$$r = 0.5, \quad \bar{x} = \bar{y} = 75, \quad s_x = s_y = 10.$$

(a) Given $x = 95$, predict the final score.

(b) Given $y = 85$ and not knowing the midterm score, predict the 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

$$\rho = \frac{1}{\sqrt{1+\beta^2}}.$$

(b) Use the result of part (a) to generate bivariate samples (x_i, y_i) of size 20 with population correlation coefficients -0.9, -0.5, 0, 0.5, and 0.9. Compute the sample correlation coefficients.

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 velocity, and examine the residuals in each case. Which fit is better? Can you suggest any physical reason that explains why?

12 Course topics and distribution tables

12.1 List of course topics

Statistical inference vs probability theory. Statistical models.
Population distribution. Population mean and standard deviation, population proportion.
Randomisation.
Sampling with replacement, random (iid) sample.
Sampling without replacement, simple random sample.

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 interval for the mean.
Stratified random sampling. Optimal allocation of observations, proportional allocation.

Parametric models, population parameters.
Binomial, geometric, Poisson, discrete uniform models.
Continuous uniform, exponential, gamma models.
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 and t-distributions.

Statistical hypotheses, simple and composite, null and alternative.
Rejection region. Two types of error.
Significance level, test power.
P-value of the test, one-sided and two-sided p-values.
Large-sample test for the proportion. Small-sample test for the proportion.
Large-sample test for the mean. One-sample t-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.
Loss function, posterior risk, 0-1 loss function and squared error loss.
Conjugate priors. Normal-normal model.
Beta and Dirichlet distributions. Beta-binomial model and Dirichlet-multinomial model.
Bayesian estimation, MAP and PME. Credibility interval.
Posterior odds. Bayesian hypotheses testing.

Empirical cumulative distribution function. Empirical variance.
Survival function and hazard function. Weibull distribution. Empirical survival function.
Kernel density estimate.
Population quantiles. Ordered sample and empirical quantiles.
QQ-plots, normal probability 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.
Trimmed means.
Sample range, quartiles, IQR and MAD. Boxplots.

Two independent 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 vs exact measurements. Rank sum test. Signed rank test.
Approximate confidence interval for the difference $p_1 - p_2$.
Large sample test for two proportions.
Fisher's exact test.
Double-blind randomised controlled experiments.
Confounding factors, Simpson's paradox.

One-way ANOVA, sums of squares and mean squares.
Normal theory model, F-test, F-distribution.
Normal probability plots for the residuals.
Multiple comparison or multiple testing problem.
Simultaneous CI, Bonferroni's method and Tukey's method.
Two-way ANOVA, main effects and interaction. Three F-tests.
Additive model. Randomised block design.
Kruskal-Wallis test. Friedman's test.

Categorical data.
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 correlation coefficient, sample covariance.
Corrected MLE of the noise variance. Coefficient of determination.
confidence interval and hypotheses testing for the intercept and slope. Model utility test.
Prediction interval for a new observation.
Standardised residuals.
Linear regression and ANOVA.
Multiple regression. Design matrix.
Coefficient of multiple determination. Adjusted coefficient of multiple determination. Collinearity problem.

12.2 Normal distribution table

If Z has a standard normal distribution N(0,1), then the following table gives

$$\Phi(z) = \mathrm{P}(Z \leq z).$$

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

12.3 Critical values of the t-distribution

If X has a t-distribution with df degrees of freedom, then for a given α the following table gives the value of x such that $\mathrm{P}(X > x) = \alpha$. For example, if $\alpha = 0.05$ and $df = 5$, then $x = 2.015$.

	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

12.4 Critical values of the chi square distribution

If X has a χ^2-distribution with df degrees of freedom, then for a given α the following table gives the value of x such that $\mathrm{P}(X > x) = \alpha$. For example, if $\alpha = 0.05$ and $df = 5$, then $x = 11.070$.

df	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

12.5 Critical values of the F-distribution

If X has a $F_{k,df}$-distribution, then the tables below give the critical value x such that $\mathrm{P}(X > x) = \alpha$. Different df are shown as columns and different α are shown as rows. For example, if $\alpha = 0.025$, $k = 2$, and $df = 6$, then $x = 7.2599$.

$F_{1,df}$-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

$F_{2,df}$-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

$F_{3,df}$-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

$F_{4,df}$-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

$F_{5,df}$-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

$F_{6,df}$-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

$F_{7,df}$-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

$F_{8,df}$-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

$F_{9,df}$-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

13 Solutions to exercises

13.1 Solutions to Section 3 (random sampling)

Solution 1

Here we consider sampling with replacement. For the given population distribution

Values	1	2	4	8
Probab.	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

the population mean and variance are computed in three steps

$$\mu = 1 \cdot \tfrac{1}{5} + 2 \cdot \tfrac{2}{5} + 4 \cdot \tfrac{1}{5} + 8 \cdot \tfrac{1}{5} = 3.4$$
$$\mathrm{E}(X^2) = 1 \cdot \tfrac{1}{5} + 4 \cdot \tfrac{2}{5} + 16 \cdot \tfrac{1}{5} + 64 \cdot \tfrac{1}{5} = 17.8$$
$$\sigma^2 = 17.8 - \mu^2 = 6.24.$$

The list of $\bar{X}$ values (and their probabilities in brackets) for $n = 2$ observations taken with replacement:

	1	2	4	8	Total prob.
1	1.0 (1/25)	1.5 (2/25)	2.5 (1/25)	4.5 (1/25)	1/5
2	1.5 (2/25)	2.0 (4/25)	3.0 (2/25)	5.0 (2/25)	2/5
4	2.5 (1/25)	3.0 (2/25)	4.0 (1/25)	6.0 (1/25)	1/5
8	4.5 (1/25)	5.0 (2/25)	6.0 (1/25)	8.0 (1/25)	1/5
Tot. prob.	1/5	2/5	1/5	1/5	1

This yields the followig sampling distribution of $\bar{X}$:

Values	1	1.5	2	2.5	3	4	4.5	5	6	8
Probab.	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

Using the same three steps we find

$$\mathrm{E}(\bar{X}) = 1 \cdot \tfrac{1}{25} + 1.5 \cdot \tfrac{4}{25} + 2 \cdot \tfrac{4}{25} + 2.5 \cdot \tfrac{2}{25} + 3 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{1}{25} + 4.5 \cdot \tfrac{2}{25} + 5 \cdot \tfrac{4}{25} + 6 \cdot \tfrac{2}{25} + 8 \cdot \tfrac{1}{25} = 3.4$$
$$\mathrm{E}(\bar{X}^2) = \tfrac{1}{25} + (1.5)^2 \cdot \tfrac{4}{25} + 4 \cdot \tfrac{4}{25} + (2.5)^2 \cdot \tfrac{2}{25} + 9 \cdot \tfrac{4}{25} + 16 \cdot \tfrac{1}{25} + (4.5)^2 \cdot \tfrac{2}{25} + 25 \cdot \tfrac{4}{25} + 36 \cdot \tfrac{2}{25} + 64 \cdot \tfrac{1}{25} = 14.68$$
$$\mathrm{Var}(\bar{X}) = 14.68 - (3.4)^2 = 3.12.$$

We see that indeed,

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = 3.12 = \frac{\sigma^2}{n}.$$

Solution 2

The question is about a simple random sample but N is not specifies. We assume N is very large and use the formulas for a random sample (independent observations).
Dichotomous data

$$n = 1500, \quad \hat{p} = 0.55, \quad 1 - \hat{p} = 0.45, \quad s_{\hat{p}} = \sqrt{\tfrac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\tfrac{0.55 \times 0.45}{1499}} = 0.013.$$

Population margin of victory

$$v = p - (1 - p) = 2p - 1.$$

Estimated margin of victory

$$\hat{v} = \hat{p} - (1 - \hat{p}) = 2\hat{p} - 1 = 0.1.$$

(a) Since

$$\mathrm{Var}(\hat{V}) = \mathrm{Var}(2\hat{P}) = 4\mathrm{Var}(\hat{P}),$$

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, in terms of the estimated standard errors, we obtain

$$s_{\hat{v}} = 2s_{\hat{p}} = 0.026.$$

(b) Approximate 95% confidence interval for v is

$$I_v \approx \hat{v} \pm 1.96 s_{\hat{v}} = 0.10 \pm 0.05.$$

Solution 3

Normal approximation: $\frac{\bar{X}-\mu}{S_{\bar{X}}}$ is asymptotically N(0,1)-distributed. From

$$0.90 \approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} > -1.28) = \mathrm{P}(-\infty < \mu < \bar{X} + 1.28S_{\bar{X}}),$$
$$0.95 \approx \mathrm{P}(\tfrac{\bar{X}-\mu}{S_{\bar{X}}} < 1.645) = \mathrm{P}(\bar{X} - 1.645S_{\bar{X}} < \mu < \infty).$$

we find

$$k_1 = 1.28, \quad k_2 = 1.645.$$

Solution 4

Randomised response method. Using the law of total probability we find the probability of "yes" answer to the randomly generated question

$$p = \mathrm{P}(\text{a "yes" answer}) = \frac{5}{6} \cdot q + \frac{1}{6} \cdot (1-q) = \frac{1+4q}{6},$$

where q is the population proportion of illegal drug users among prison inmates. For a random sample of size n, put

$$x = \text{ number of "yes" responses for } n \text{ inmates.}$$

Under the independence assumptions, we can use a Bin (n,p) distribution model. Given x, we estimate p by the sample proportion $\hat{p} = \frac{x}{n}$. Treated as random variable, the point estimate $\hat{P} = \frac{X}{n}$ is a scaled binomial random variable $X \sim$ Bin (n,p) implying

$$\mathrm{Var}(\hat{P}) = \frac{\mathrm{Var}(X)}{n^2} = \frac{p(1-p)}{n}.$$

Using $\hat{p}$ as a given number, we get an equation

$$\hat{p} = \frac{1+4\tilde{q}}{6},$$

whose solution gives a method of moments estimate $\tilde{q}$ 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:

$$\mathrm{E}(\tilde{Q}) = \frac{6p-1}{4} = q.$$

Its variance 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 for example $n = 40$, $x = 8$. Then $\hat{p} = 0.2$ and

$$\tilde{q} = \frac{6\hat{p}-1}{4} = 0.05.$$

The estimated standard error

$$s_{\tilde{q}} = \sqrt{\frac{(1+4\tilde{q})(5-4\tilde{q})}{16n}} = 0.095.$$

Conclusion: the estimate has way too big standard error. We have to increase the sample size.

Solution 5

Data summary

$$N = 2000, \quad n = 25, \quad \sum x_i = 2351, \quad \sum x_i^2 = 231305.$$

(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

$$s^2_{\bar{x}} = \frac{s^2}{n}\Big(1 - \frac{n}{N}\Big) = 16.81.$$

(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 is

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

For large n, the bias is small.

Solution 7

Stratified population of size $N = 2010$ with $k = 7$ 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

$$\mathrm{Var}(\bar{X}_{\mathrm{s}}) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k} = 4.450.$$

The requested sample size 139 is found from the equation

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{x} = \frac{620.17}{x} = 4.45, \qquad x = \frac{620.17}{4.45} = 139.364.$$

(e) If $n = 70$, then $\mathrm{Var}(\bar{X}_{\mathrm{sp}}) = 4.90$. 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 127 which is smaller than that of the item (d).

Solution 9

Stratified population with

$$N = 5, \quad k = 2, \quad w_1 = 0.6, \quad w_2 = 0.4, \quad \mu_1 = 1.67, \quad \mu_2 = 6, \quad \sigma_1^2 = 0.21, \quad \sigma_2^2 = 4.$$

Given $n_1 = n_2 = 1$ and $n = 2$, the sampling distribution of the stratified sample mean $\bar{x}_s = 0.6x_1 + 0.4x_2$ is

	$x_1 = 1$	$x_1 = 2$	Total 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
Tot. prob.	1/3	2/3	1

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

$$\mathrm{E}(\bar{X}_s) = \mu, \quad \mathrm{Var}(\bar{X}_s) = \tfrac{w_1^2\sigma_1^2}{n_1} + \ldots + \tfrac{w_k^2\sigma_k^2}{n_k} = 0.36\sigma_1^2 + 0.16\sigma_2^2.$$

13.2 Solutions to Section 4 (parameter estimation)

Solution 1

A method of moment estimate of the parameter λ for the Poisson distribution model is given by the sample mean $\tilde{\lambda} = 3.9$. Using this value we compute the expected counts, see table. Comparing 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 in that the variance is larger than λ.
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.

n	Frequency	Expected frequency
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

Number X of yeast cells on a square. Test the Poisson model $X \sim \mathrm{Pois}(\lambda)$.

Concentration 1.

$$\bar{x} = 0.6825, \quad \overline{x^2} = 1.2775, \quad s^2 = 0.8137, \quad s = 0.9021, \quad s_{\bar{x}} = 0.0451.$$

Approximate 95% confidence interval

$$I_\mu = 0.6825 \pm 0.0884.$$

Pearson's chi-squared test based on $\hat{\lambda} = 0.6825$:

x	0	1	2	3	4+	Total
Observed	213	128	37	18	4	400
Expected	202.14	137.96	47.08	10.71	2.12	400

Observed test statistic $\mathrm{X}^2 = 10.12$, df $= 5 - 1 - 1 = 3$, p-value < 0.025. Reject the model.

Concentration 2.

$$\bar{x} = 1.3225, \quad \overline{x^2} = 3.0325, \quad s = 1.1345, \quad s_{\bar{x}} = 0.0567.$$

Approximate 95% confidence interval

$$I_\mu = 1.3225 \pm 0.1112.$$

Pearson's chi-squared test: observed test statistic $\mathrm{X}^2 = 3.16$, df $= 4$, p-value > 0.10. Do not reject the model.

Concentration 3.

$$\bar{x} = 1.8000, \quad s = 1.1408, \quad s_{\bar{x}} = 0.0701.$$

Approximate 95% confidence interval for

$$I_\mu = 1.8000 \pm 0.1374.$$

Pearson's chi-squared test: observed test statistic $\mathrm{X}^2 = 7.79$, df $= 5$, p-value > 0.10. Do not reject the model.

Concentration 4.

$$n = 410, \quad \bar{x} = 4.5659, \quad s^2 = 4.8820, \quad s_{\bar{x}} = 0.1091.$$

Approximate 95% confidence interval

$$I_\mu = 4.566 \pm 0.214.$$

Pearson's chi-squared test: observed test statistic $\mathrm{X}^2 = 13.17$, df $= 10$, p-value > 0.10. Do not reject the model.

Solution 3

Population distribution: X takes values $0, 1, 2, 3$ 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 an iid-sample with

$$n = 10, \quad \bar{x} = 1.5, \quad s = 1.08,$$

and observed counts

x	0	1	2	3	Total
O_x	2	3	3	2	10

(a) Method of moments. Using

$$\mu = \frac{1}{3} \cdot \theta + 2 \cdot \frac{2}{3} \cdot (1-\theta) + 3 \cdot \frac{1}{3} \cdot (1-\theta) = \frac{7}{3} - 2\theta,$$

derive an 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 estimate σ as

$$\sqrt{\frac{2}{9} + 4\tilde{\theta} - 4\tilde{\theta}^2} = 1.093.$$

This gives

$$s_{\tilde{\theta}} = \frac{1.093}{6.325} = 0.173.$$

Method 2:

$$s_{\tilde{\theta}} = \frac{s}{6.325} = \frac{1.08}{6.325} = 0.171.$$

(c) Likelihood function is obtained using $(O_0, O_1, O_2, O_3) \sim \text{Mn}(n, p_0, p_1, p_2, p_3)$

$$L(\theta) = \left(\tfrac{2}{3}\theta\right)^{O_0} \left(\tfrac{1}{3}\theta\right)^{O_1} \left(\tfrac{2}{3}(1-\theta)\right)^{O_2} \left(\tfrac{1}{3}(1-\theta)\right)^{O_3} = \text{const } \theta^t (1-\theta)^{n-t},$$

where $t = O_0 + O_1$ is a sufficient statistic. Notice that $T = O_0 + O_1$ has $\text{Bin}(n, \theta)$ distribution.
Log-likelihood and its derivative

$$l(\theta) = \text{const } + t \ln\theta + (n-t)\ln(1-\theta),$$
$$l'(\theta) = \frac{t}{\theta} - \frac{n-t}{1-\theta}.$$

Setting the last expression to zero, we find

$$\frac{t}{\hat{\theta}} = \frac{n-t}{1-\hat{\theta}}, \quad \hat{\theta} = \frac{t}{n} = \frac{2+3}{10} = \frac{1}{2}.$$

The maximum likelihood estimate is the sample proportion, an unbiased estimate of the population proportion θ.

(d) We find $s_{\hat{\theta}}$ using the formula for the standard error of sample proportion

$$s_{\hat{\theta}} = \sqrt{\tfrac{\hat{\theta}(1-\hat{\theta})}{n-1}} = 0.167.$$

A similar answer is obtained using the formula

$$s_{\hat{\theta}} = \sqrt{\tfrac{1}{n\mathbb{I}(\hat{\theta})}}, \quad \mathbb{I}(\theta) = \text{E}(g(Y,\theta)), \quad g(y,\theta) = -\tfrac{\partial^2}{\partial\theta^2} \ln f(y|\theta),$$

where $Y \sim \text{Ber}(\theta)$. Since $f(1|\theta) = \theta$, $f(0|\theta) = 1-\theta$, we have

$$g(1,\theta) = -\frac{\partial^2}{\partial\theta^2} \ln\theta = \frac{1}{\theta^2}, \quad g(0,\theta) = -\frac{\partial^2}{\partial\theta^2} \ln(1-\theta) = \frac{1}{(1-\theta)^2},$$

we get

$$\mathbb{I}(\theta) = \text{E}\left(g(Y,\theta)\right) = g(1,\theta)f(1|\theta) + g(0,\theta)f(0|\theta) = \frac{1}{\theta^2} \cdot \theta + \frac{1}{(1-\theta)^2} \cdot (1-\theta) = \frac{1}{\theta(1-\theta)}.$$

Solution 4

Likelihood function of $X \sim \text{Bin}(n,p)$ for a given n and $X = x$ is

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x} \propto p^x (1-p)^{n-x}.$$

(a) To maximise $L(p)$ we minimise

$$\ln(p^x(1-p)^{n-x}) = x\ln p + (n-x)\ln(1-p).$$

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 $\frac{x}{p} = \frac{n-x}{1-p}$, which brings the maximum likelihood estimate formula $\hat{p} = \frac{x}{n}$.

(b) We have $X = Y_1 + \ldots + Y_n$, where $(Y_1, \ldots, Y_n)$ are iid Bernoulli random variables with

$$f(y|p) = p^y(1-p)^{1-y}, \quad y = 0, 1.$$

By Cramer-Rao, if $\tilde{p}$ is an unbiased estimate of p, then

$$\mathrm{Var}(\tilde{P}) \geq \frac{1}{n\mathbb{I}(p)},$$

where

$$\mathbb{I}(p) = -\mathrm{E}\left(\frac{\partial^2}{\partial p^2} \ln f(Y|p)\right) = \frac{1}{p(1-p)},$$

see Solution 3 (d). We conclude that the variance sample proportion $\hat{p}$ attains the Cramer-Rao lower bound since

$$\mathrm{Var}(\hat{P}) = \tfrac{p(1-p)}{n}.$$

(c) Plot $L(p) = 252p^5(1-p)^5$. The top of the curve is in the middle $\hat{p} = 0.5$.

Solution 5

The observed serial number $x = 888$ can be modeled by the discrete uniform distribution $X \sim \mathrm{U}(N)$.

(a) The method of moments estimate of N is obtained from

$$\mu = \tfrac{N+1}{2}, \quad 888 = \tfrac{\tilde{N}+1}{2}.$$

It gives $\tilde{N} = 2x - 1 = 1775$. This is an unbiased estimate.

(b) The likelihood function

$$L(N) = \mathrm{P}(X = x) = \frac{1_{\{1 \leq x \leq N\}}}{N} = \frac{1_{\{N \geq 888\}}}{N}$$

reaches its maximum at $\hat{N} = 888$. We see that in this case the MLE is severely biased.

Solution 6

Statistical model 1: x is the number of black balls obtained by sampling k balls without replacement from an urn with N balls of which n balls are black. Hypergeometric distribution

$$\mathrm{P}(X = 20) = \frac{\binom{n}{20}\binom{N-n}{30}}{\binom{N}{50}}.$$

The likelihood function

$$L(N) = \frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}} = \mathrm{const} \cdot \frac{(N-100)(N-101)\cdots(N-129)}{N(N-1)\cdots(N-49)}.$$

Consider the ratio

$$\frac{L(N)}{L(N-1)} = \frac{(N-100)(N-50)}{N(N-130)}.$$

We find the value of $N = \hat{N}$ that maximises $L(N)$ by solving the equation

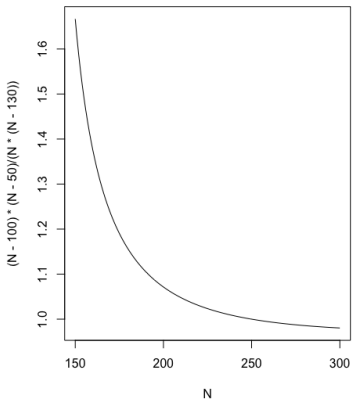
$$\frac{L(\hat{N})}{L(\hat{N}-1)} = 1.$$

It gives

$$(\hat{N}-100)(\hat{N}-50) = \hat{N}(\hat{N}-130),$$

and we arrive at the maximum likelihood estimate $\hat{N} = 250$.
The obtained solution is very intuitive, it says that that the proportion of captured fish is similar to the proportion of recaptured

$$100 : N \approx 20 : 50.$$



Equation $\frac{L(\hat{N})}{L(\hat{N}-1)} = 1$ is justified using the graph. It says, if $N < \hat{N}$, then $\frac{L(N+1)}{L(N)} > 1$, so that

$$L(N) < L(\hat{N}),$$

and on the other hand, if $N > \hat{N}$, then $\frac{L(N)}{L(N-1)} < 1$, so that again

$$L(N) < L(\hat{N}).$$

Statistical model 2: x is the number of black balls obtained by sampling k balls with replacement from an urn with N balls of which n balls are black. Binomial distribution

$$\mathrm{P}(X = 20) = \binom{k}{20} (\tfrac{100}{N})^{20} (\tfrac{N-100}{N})^{30}.$$

The likelihood function

$$L(N) = \text{const} \cdot \frac{(N-100)^{30}}{N^{50}}.$$

Log-likelihood

$$l(N) = \text{const} + 30 \ln(N-100) - 50 \ln N.$$

Take the derivative and set to be equal 0

$$\frac{30}{N-100} = \frac{50}{N},$$

giving the same maximum likelihood estimate as Model 1 $\hat{N} = 250$.

Solution 7

An iid-sample of size $n = 16$ from a normal distribution.

(a) The summary statistics

$$\bar{x} = 3.6109, \quad s^2 = 3.4181, \quad s_{\bar{x}} = 0.4622$$

suggest an estimate for μ to be 3.6109, and an estimate for σ^2 to be 3.4181.

(b), (c) Exact confidence intervals

	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 sample size x that halves the confidence interval length we set up an equation using the exact confidence interval formula for the mean

$$t_{15}(\alpha/2) \cdot \frac{s}{\sqrt{16}} = 2 \cdot t_{x-1}(\alpha/2) \cdot \frac{s'}{\sqrt{x}},$$

where s' is the sample standard deviation for the sample of size x. A simplistic 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

$$t_{15}(\alpha/2) = 2.13, \quad t_{x-1}(\alpha/2) \approx 2,$$

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.

Solution 8

An iid-sample $(x_1, \ldots, x_n)$ from the uniform distribution $\mathrm{U}(0, \theta)$ with density

$$f(x|\theta) = \tfrac{1}{\theta} 1_{\{0 \le x \le \theta\}}.$$

(a) Method of moments estimate $\tilde{\theta}$ is unbiased

$$\mu = \tfrac{\theta}{2}, \quad \tilde{\theta} = 2\bar{x}, \quad \mathrm{E}(\tilde{\Theta}) = \theta, \quad \mathrm{Var}(\tilde{\Theta}) = \tfrac{4\sigma^2}{n} = \tfrac{\theta^2}{3n}.$$

Here we used the variance formula for the continuous uniform $\mathrm{U}(a, b)$ distribution

$$\sigma^2 = \tfrac{(a-b)^2}{12}.$$

(b) Denote $x_{(n)} = \max(x_1, \ldots, x_n)$. Likelihood function takes the form

$$L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta) = \tfrac{1}{\theta^n} 1_{\{\theta \geq x_1\}} \cdots 1_{\{\theta \geq x_n\}} = \tfrac{1}{\theta^n} 1_{\{\theta \geq x_{(n)}\}},$$

so that $x_{(n)}$ is a sufficient statistic. The maximum is achieved at $\hat{\theta} = x_{(n)}$.

(c) Sampling distribution of the maximum likelihood estimate $\hat{\theta} = x_{(n)}$:

$$\mathrm{P}(X_{(n)} \leq x) = \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) = \mathrm{P}(X_1 \leq x) \cdots \mathrm{P}(X_n \leq x) = \left(\frac{x}{\theta}\right)^n$$

with pdf

$$f_{\widehat{\Theta}}(x) = \frac{n}{\theta^n} \cdot x^{n-1}, \quad 0 \leq x \leq \theta.$$

The maximum likelihood estimate is biased

$$\mathrm{E}(\widehat{\Theta}) = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta, \quad \mathrm{E}(\widehat{\Theta}^2) = \frac{n}{n+2}\theta^2, \quad \mathrm{Var}(\widehat{\Theta}) = \frac{n\theta^2}{(n+1)^2(n+2)},$$

but asymptotically unbiased. Notice the unusual asymptotics indicating that the conditions on the parametric model implying $\hat{\Theta} \approx \mathrm{N}(\theta, \frac{1}{n\mathbb{I}(\theta)})$ are violated:

$$\mathrm{Var}(\widehat{\Theta}) = \frac{\theta^2}{n^2}, \quad n \to \infty.$$

Compare two mean square errors:

$$\mathrm{MSE}(\widehat{\Theta}) = \mathrm{E}(\widehat{\Theta} - \theta)^2 = \left(-\frac{\theta}{n+1}\right)^2 + \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{2\theta^2}{(n+1)(n+2)},$$
$$\mathrm{MSE}(\tilde{\Theta}) = \frac{\theta^2}{3n}.$$

(d) 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

Data

$$x_1 = 1997, \quad x_2 = 906, \quad x_3 = 904, \quad x_4 = 32$$

and model

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

(a) Sample counts $(X_1, X_2, X_3, X_4) \sim \mathrm{Mn}(n, p_1, p_2, p_3, p_4)$ with $n = 3839$. Given a realisation

$$(x_1, x_2, x_3, x_4) \text{ with } x_1 + x_2 + x_3 + x_4 = n,$$

the likelihood function takes the form

$$L(\theta) = \binom{n}{x_1, x_2, x_3, x_4} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \propto (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} 4^{-n} \propto (2+\theta)^{x_1} \theta^{x_4} (1-\theta)^{n-x_1-x_4},$$

where $\propto$ means that we drop factors depending only on (n, x_1, x_2, x_3, x_4). The last expression reveals that we have a case of two sufficient statistics (x_1, x_4). Putting

$$\frac{d}{d\theta} \ln L(\theta) = \frac{x_1}{2+\theta} + \frac{x_4}{\theta} - \frac{n - x_1 - x_4}{1-\theta}$$

equal to zero, we arrive at the equation

$$\frac{x_1}{2+\theta} + \frac{x_4}{\theta} = \frac{n - x_1 - x_4}{1-\theta}$$

or equivalently

$$\theta^2 n + \theta u - 2x_4 = 0,$$

where

$$u = 2x_2 + 2x_3 + x_4 - x_1 = 2n - x_4 - 3x_1.$$

We find the maximum likelihood estimate to be

$$\hat{\theta} = \frac{-u + \sqrt{u^2 + 8nx_4}}{2n} = 0.0357.$$

Asymptotic variance

$$\mathrm{Var}(\widehat{\Theta}) \approx \frac{1}{n\mathbb{I}(\theta)}, \quad \mathbb{I}(\theta) = \mathrm{E}(g(Y_1, Y_2, Y_3, Y_4, \theta)).$$

where $(Y_1, Y_2, Y_3, Y_4) \sim \mathrm{Mn}(1, p_1, p_2, p_3, p_4)$ with

$$f(y_1, y_2, y_3, y_4|\theta) = p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} = (2+\theta)^{y_1}(1-\theta)^{y_2+y_3}\theta^{y_4} 4^{-n}$$

and

$$g(y_1, y_2, y_3, y_4, \theta) = -\frac{\partial^2}{\partial\theta^2} \ln f(y_1, y_2, y_3, y_4|\theta) = \frac{y_1}{(2+\theta)^2} + \frac{y_2 + y_3}{(1-\theta)^2} + \frac{y_4}{\theta^2}.$$

Since $\mathrm{E}(Y_i) = p_i$, we find

$$\mathbb{I}(\theta) = \frac{1}{4(2+\theta)} + \frac{2}{4(1-\theta)} + \frac{1}{4\theta} = \frac{1+2\theta}{2\theta(2+\theta)(1-\theta)},$$

and get

$$n\mathbb{I}(\hat{\theta}) = 29345.8, \qquad s_{\hat{\theta}} = \sqrt{\tfrac{1}{n\mathbb{I}(\hat{\theta})}} = 0.0058.$$

(b) $I_\theta = 0.0357 \pm 1.96 \cdot 0.0058 = 0.0357 \pm 0.0114$.

13.3 Solutions to Section 5 (hypothesis testing)

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 have

$$Z = \frac{X - 100p_0}{10\sqrt{p_0(1-p_0)}} = \frac{X - 50}{5},$$

and the significance level in question is (using a continuity correction)

$$\begin{aligned}\alpha &= \mathrm{P}(|X - 50| > 10|H_0) = \mathrm{P}(|X - 50| \geq 11|H_0) \\ &\approx \mathrm{P}(|Z| > \tfrac{10.5}{5}|H_0) \approx 2(1 - \Phi(2.1)) = 2 \cdot 0.018 = 0.036.\end{aligned}$$

(b) The power of the test is a function of the parameter value p (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

Data: one observation of $X = x$. Likelihood ratio test: reject for small values of $\Lambda = \frac{\mathrm{P}(x|H_0)}{\mathrm{P}(x|H_1)}$.

(a) See the bottom line of the table:

X-values	x_4	x_2	x_1	x_3
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.2	0.3
$\mathrm{P}(x\|H_1)$	0.4	0.4	0.1	0.1
Likelihood ratio $\Lambda = \frac{\mathrm{P}(x\|H_0)}{\mathrm{P}(x\|H_1)}$	0.5	0.75	2	3

(b) The null distribution of Λ is discrete. Under H_0 the test statistic Λ takes values $0.5, 0.75, 2, 3$ with probabilities $0.2, 0.3, 0.2, 0.3$, see the table below

X-values	x_4	x_2	x_1	x_3
Likelihood ratio Λ	0.5	0.75	2	3
$\mathrm{P}(x\|H_0)$	0.2	0.3	0.2	0.3
Cumulative probab.	0.2	0.5	0.7	1

Since H_0 is rejected for the small values of the likelihood ratio Λ, at $\alpha = 0.2$ we reject H_0 only if $\Lambda = 0.5$, that is when $X = x_4$.
At $\alpha = 0.5$ we reject H_0 for $\Lambda \leq 0.75$, that is when X is either x_4 or x_2.

Solution 3

Likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \frac{1}{x_i!} \lambda^{x_i} e^{-\lambda} = e^{-\lambda n} \lambda^{y} \prod_{i=1}^{n} \frac{1}{x_i!}$$

where

$$y = x_1 + \ldots + x_n$$

is a sufficient statistic.
Case 1: two simple hypotheses

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda = \lambda_1.$$

Reject H_0 for small values of the likelihood ratio

$$\frac{L(\lambda_0)}{L(\lambda_1)} = e^{-n(\lambda_0 - \lambda_1)} (\tfrac{\lambda_0}{\lambda_1})^{y}.$$

If $\lambda_1 > \lambda_0$, then we reject H_0 for large values of y. If $\lambda_1 < \lambda_0$, then we reject H_0 for small values of y. Test statistic Y has null distribution $\mathrm{Pois}(n\lambda_0)$.
Case 2: two-sided alternative hypothesis

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda \neq \lambda_0.$$

Reject H_0 for small values of the generalised likelihood ratio

$$\frac{L(\lambda_0)}{L(\hat{\lambda})} = e^{-n(\lambda_0 - \hat{\lambda})} (\tfrac{\lambda_0}{\hat{\lambda}})^{y}, \quad \hat{\lambda} = y/n.$$

Reject H_0 for the both small and large values of y. Test statistic Y has null distribution $\mathrm{Pois}(n\lambda_0)$.

Solution 4

We have an iid-sample from $\mathrm{N}(\mu, 10)$ of size $n = 25$. Two simple hypotheses

$$H_0 : \mu = 0, \quad H_1 : \mu = 1.5$$

Test statistic and its exact sampling distribution

$$\bar{X} \sim \mathrm{N}(\mu, 2),$$

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 uses 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 beta-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 data value x is

$$\Lambda = \frac{f(x|H_0)}{f(x|H_1)} = \tfrac{2}{3x}, \quad 0 \leq x \leq 1.$$

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

$$\mathrm{P}(X > x_{\mathrm{crit}}|H_0) = \alpha,$$

that is

$$1 - x_{\mathrm{crit}}^2 = \alpha.$$

We conclude that

$$\mathcal{R} = \{x : x > \sqrt{1-\alpha}\}.$$

(c) The power of the test is

$$\mathrm{P}(X > \sqrt{1-\alpha}|H_1) = 1 - (1-\alpha)^{3/2}$$

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 $\frac{14 \cdot S^2}{\sigma^2}$ has a χ^2_{14}-distribution, so that

$$\mathrm{P}(\tfrac{14 \cdot S^2}{\sigma^2} \leq 6.571) = 0.05.$$

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 analysis is the basis of the sign test.

(a) Generalised likelihood ratio

$$\Lambda = \frac{L(p_0)}{L(\hat{p})} = \frac{\binom{n}{x} p_0^x (1-p_0)^{n-x}}{\binom{n}{x} \hat{p}^x (1-\hat{p})^{n-x}} = \frac{(\frac{1}{2})^n}{(\frac{x}{n})^x (\frac{n-x}{n})^{n-x}} = \frac{(\frac{n}{2})^n}{x^x (n-x)^{n-x}}.$$

(b) The generalised likelihood ratio test rejects H_0 for small values of

$$\ln \Lambda = n \ln(n/2) - x \ln x - (n-x) \ln(n-x),$$

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

$$y = |x - n/2|.$$

The function $a(y)$ is monotonely increasing over $y \in [0, n/2]$, since

$$a'(y) = \ln \frac{\frac{n}{2} + y}{\frac{n}{2} - y} > 0.$$

We conclude that the test rejects for large values of y.

(c) Compute the significance level for the rejection region $|x - \frac{n}{2}| > k$:

$$\alpha = \mathrm{P}(|X - \tfrac{n}{2}| > k | H_0) = 2 \sum_{i < \frac{n}{2} - k} \binom{n}{i} 2^{-n}.$$

(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) Two-sided p-value = 0.134.

(b) 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 $X \sim \text{Bin}(n, p)$, 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. Observed test statistic

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{922 - 1919 \cdot 0.5}{\sqrt{1919} \cdot 0.5} = -1.712.$$

One-sided p-value of the test

$$\Phi(-1.712) = 1 - \Phi(1.712) = 1 - 0.9564 = 0.044.$$

Reject H_0 in favour of one-sided H_1 at the significance level 5%.

(b) To control for the seasonal effect the Chinese and Japanese data were studied

$$n = 852, \quad x = 418, \quad n - x = 434, \quad z = -0.548.$$

One-sided p-value is 29%, showing no significant effect.

(c) Overeating during the important occasion might be a contributing factor.

Solution 11

Multinomial model

$$(X_1, X_2, X_3) \sim \text{Mn}(190, p_1, p_2, p_3).$$

Composite null hypothesis (Hardy-Weinberg Equilibrium)

$$H_0 : p_1 = (1-\theta)^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = \theta^2.$$

Likelihood function and maximum likelihood estimate

$$L(\theta) = \binom{190}{10, 68, 112} 2^{68} \theta^{292} (1-\theta)^{88}, \quad \hat{\theta} = \frac{292}{380} = 0.768.$$

Pearson's chi-squared test:

cell	1	2	3	Total
observed	10	68	112	190
expected	10.23	67.71	112.07	190

Observed chi-squared test statistic $\text{X}^2 = 0.0065$, df = 1, p-value $= 2(1 - \Phi(\sqrt{0.0065})) = 0.94$.

Conclusion: the Hardy-Weinberg Equilibrium model fits well the haptoglobin data.

Solution 12

Month	O_j	Days	E_j	$O_j - E_j$
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

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 are

$$E_j = np_j^{(0)}, \quad j = 1, \ldots, 12.$$

The χ^2-test statistic

$$\mathrm{X}^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = 47.4.$$

Since $\mathrm{df} = 12 - 1 = 11$, and $\chi^2_{11}(0.005) = 26.8$, we reject H_0 of no seasonal variation. Merry Christmas!

Solution 13

Number of heads

$$Y \sim \mathrm{Bin}(n, p), \quad n = 17950.$$

(a) For $H_0 : p = 0.5$ the observed z-score

$$z = \frac{y - np_0}{\sqrt{np_0(1-p_0)}} = 3.46.$$

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

$$H_0 : p_0 = (0.5)^5 = 0.031, \ p_1 = 5 \cdot (0.5)^5 = 0.156, \ p_2 = 10 \cdot (0.5)^5 = 0.313,$$
$$p_3 = 10 \cdot (0.5)^5 = 0.313, \ p_4 = 5 \cdot (0.5)^5 = 0.156, \ p_5 = (0.5)^5 = 0.031.$$

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590
expected	112.2	560.9	1121.9	1121.9	560.9	112.2	3590

Observed $\mathrm{X}^2 = 21.58$, df $= 5$, p-value $= 0.001$.

(c) Composite null hypothesis

$$H_0 : p_i = \binom{5}{i} p^i (1-p)^{5-i}, \quad i = 0, 1, 2, 3, 4, 5.$$

Pearson's chi-squared test based on the maximum likelihood estimate $\hat{p} = 0.5129$

number of heads	0	1	2	3	4	5	Total
observed	100	524	1080	1126	655	105	3590
expected	98.4	518.3	1091.5	1149.3	605.1	127.4	3590

Observed $\mathrm{X}^2 = 8.74$, df $= 6 - 1 - 1 = 4$, p-value $= 0.07$. Do not reject H_0 at 5% level.

13.4 Solutions to Section 6 (Bayesian inference)

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

$$\hat{\theta}_{\mathrm{map}} = \hat{\theta}_{\mathrm{pme}} = \frac{1}{2}.$$

Solution 2

Number of bird hops $X \sim \text{Geom}(p)$

$$f(x|p) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots.$$

Data in the table summarises an iid-sample

$$(x_1, \ldots, x_n), \quad n = 130.$$

(d) Using a uniform prior $P \sim \text{U}(0,1)$, we find the posterior to be

$$h(p|x_1, \ldots, x_n) \propto f(x_1|p) \cdots f(x_n|p) = (1-p)^{n\bar{x}-n}p^n, \quad n = 130, \quad n\bar{x} = 363.$$

It is a beta distribution

$$\text{Beta}(n+1, n\bar{x} - n + 1) = \text{Beta}(131, 234).$$

Posterior mean

$$\mu = \frac{a}{a+b} = \frac{131}{131+234} = 0.36.$$

Observe that

$$\mu = \frac{1 + \frac{1}{n}}{\bar{x} + \frac{2}{n}},$$

gets closer to the method of moments estimate of p as $n \to \infty$. The standard deviation of the posterior distribution

$$\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}} = \sqrt{\frac{0.36 \cdot 0.64}{366}} = 0.025.$$

Solution 3

We use the binomial model $X \sim \text{Bin}(n, p)$, with p being the probability that the event will occur at a given trial. Use an uninformative conjugate prior $P \sim \text{Beta}(1,1)$. Given $X = n$, the posterior becomes $P \sim \text{Beta}(n+1, 1)$. Since the posterior mean is $\frac{n+1}{n+2}$, we get

$$\hat{p}_{\text{pme}} = \frac{n+1}{n+2}.$$

Solution 4

Recall solutions of parts (a) and (b).

(c) By the Bayes formula,

$$\text{P}(H_0|x) = \frac{\text{P}(x|H_0)\text{P}(H_0)}{\text{P}(x|H_0)\text{P}(H_0) + \text{P}(x|H_1)\text{P}(H_1)} = \frac{\text{P}(x|H_0)}{\text{P}(x|H_0) + \text{P}(x|H_1)}.$$

Thus the posterior odds equals the likelihood ratio

$$\frac{\text{P}(H_0|x)}{\text{P}(H_1|x)} = \Lambda,$$

and we conclude that outcomes x_1 and x_3 favour H_0 since with these outcomes $\Lambda > 1$.

(d) For the general prior

$$\text{P}(H_0) = \pi_0, \quad \text{P}(H_1) = \pi_1 = 1 - \pi_0,$$

we get

$$\text{P}(H_i|x) = \frac{\text{P}(x|H_i)\pi_i}{\text{P}(x|H_0)\pi_0 + \text{P}(x|H_1)\pi_1},$$

yielding a relation for the posterior odds

$$\frac{\text{P}(H_0|x)}{\text{P}(H_1|x)} = \frac{\text{P}(x|H_0)\pi_0}{\text{P}(x|H_1)\pi_1} = \Lambda \cdot \frac{\pi_0}{\pi_1}.$$

Assuming equal costs $c_0 = c_1$, the rejection rule is

$$\frac{\text{P}(H_0|x)}{\text{P}(H_1|x)} \leq \frac{c_1}{c_0} = 1,$$

so that in terms of the likelihood ratio, we reject H_0 if

$$\Lambda \cdot \frac{\pi_0}{\pi_1} \leq 1, \quad \text{that is} \quad \Lambda \leq \frac{\pi_1}{\pi_0} = \frac{1}{\pi_0} - 1, \quad \text{or equivalently} \quad \pi_0 \leq \frac{1}{1+\Lambda}.$$

Recall that at $\alpha = 0.2$ the rejection region was

$$\mathcal{R} = \{X = x_4\} = \{\Lambda = \tfrac{1}{2}\},$$

which in terms of the prior probabilities imposes the restriction

$$\pi_0 \leq \frac{1}{1+\Lambda} = \frac{2}{3}.$$

On the other hand, at $\alpha = 0.5$ the rejection region was

$$\mathcal{R} = \{X = x_4\} \cup \{X = x_2\} = \{\Lambda = \tfrac{1}{2}\} \cup \{\Lambda = \tfrac{3}{4}\}.$$

To be able to reject for the value $\Lambda = \frac{3}{4}$, we have to put the restriction

$$\pi_0 \leq \frac{1}{1+\frac{3}{4}} = \frac{4}{7}.$$

Solution 5

For a single observation $X \sim \mathrm{N}(\mu, \sigma)$, where σ is known, test $H_0 : \mu = 0$ vs $H_1 : \mu = 1$. Prior probabilities

$$\mathrm{P}(H_0) = \frac{2}{3}, \quad \mathrm{P}(H_1) = \frac{1}{3}.$$

(a) Likelihood ratio

$$\frac{f(x|0)}{f(x|1)} = \frac{e^{-\frac{x^2}{2\sigma^2}}}{e^{-\frac{(x-1)^2}{2\sigma^2}}} = e^{\frac{\frac{1}{2}-x}{\sigma^2}}.$$

Choose H_0 for x such that

$$\frac{\mathrm{P}(H_0|x)}{\mathrm{P}(H_1|x)} = 2e^{\frac{\frac{1}{2}-x}{\sigma^2}} > 1, \qquad x < \frac{1}{2} + \sigma^2 \ln 2.$$

We conclude that

	$\sigma^2 = 0.1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 5$
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

$$\mathrm{P}(X < \tfrac{1}{2} + \sigma^2 \ln 2) = \tfrac{2}{3}\mathrm{P}(X - \mu < \tfrac{1}{2} + \sigma^2 \ln 2) + \tfrac{1}{3}\mathrm{P}(X - \mu < \sigma^2 \ln 2 - \tfrac{1}{2}) = \tfrac{2}{3}\Phi(\sigma \ln 2 + \tfrac{1}{2\sigma}) + \tfrac{1}{3}\Phi(\sigma \ln 2 - \tfrac{1}{2\sigma}).$$

We conclude that

	$\sigma^2 = 0.1$	$\sigma^2 = 0.5$	$\sigma^2 = 1$	$\sigma^2 = 5$
Proportion of the time H_0 will be chosen	0.67	0.73	0.78	0.94

Solution 6

We have a pair of beta-densities

$$f(x|H_0) = 2x, \quad f(x|H_1) = 3x^2, \quad 0 \leq x \leq 1.$$

If the two hypotheses have equal prior probabilities, then the posterior probabilities equal

$$h(H_0|x) = \frac{\frac{1}{2}f(x|H_0)}{\frac{1}{2}f(x|H_0) + \frac{1}{2}f(x|H_1)} = \frac{x}{x + \frac{3}{2}x^2} = \frac{2}{2+3x}, \quad h(H_1|x) = \frac{3x}{2+3x}.$$

Therefore, the posterior probability of H_0 is greater than that of H_1 for x satisfying

$$2 > 3x, \text{ that is when } x < \tfrac{2}{3}.$$

13.5 Solutions to Section 7 (summarising data)

Solution 1

Recall that for a fixed x, the empirical distribution function $\hat{F}(x) = \hat{p}$ is the sample proportion estimate of $p = F(x) = x$.

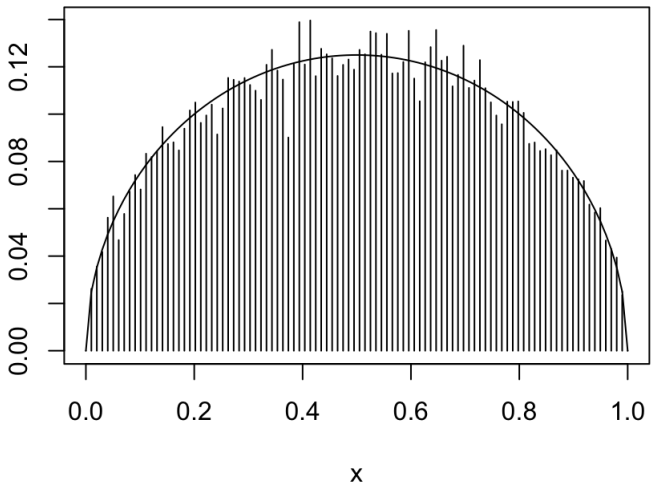
(a) The variance of $\hat{F}(x)$ is

$$\sigma^2_{\hat{F}(x)} = \sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{x(1-x)}{n},$$

so that the standard deviation is

$$\sigma_{\hat{F}(x)} = \sqrt{\frac{x(1-x)}{n}}, \quad x \in [0,1].$$

(b) 100 samples of size $n = 16$ generated from the uniform distribution have produced the standard errors of $\hat{F}(x)$ for different values of x in the following figure.



Solution 2

We have

$$\hat{F}(u) = \frac{1_{\{X_1 \le u\}} + \ldots + 1_{\{X_n \le u\}}}{n}, \quad \mathrm{E}(\hat{F}(u)) = F(u),$$
$$\hat{F}(v) = \frac{1_{\{X_1 \le v\}} + \ldots + 1_{\{X_n \le v\}}}{n}, \quad \mathrm{E}(\hat{F}(v)) = F(v).$$

Assuming $u < v$, we get

$$\begin{aligned}
\mathrm{E}(\hat{F}(u) \cdot \hat{F}(v)) &= \frac{1}{n^2}\Big[\sum_{i=1}^{n} \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_i \le v\}}) + \sum_{i=1}^{n}\sum_{j \ne i} \mathrm{E}(1_{\{X_i \le u\}} 1_{\{X_j \le v\}})\Big] \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n} F(u) + \sum_{i=1}^{n}\sum_{j \ne i} F(u)F(v)\Big] \\
&= \frac{1}{n}\Big[F(u) + (n-1)F(u)F(v)\Big].
\end{aligned}$$

Finish by using

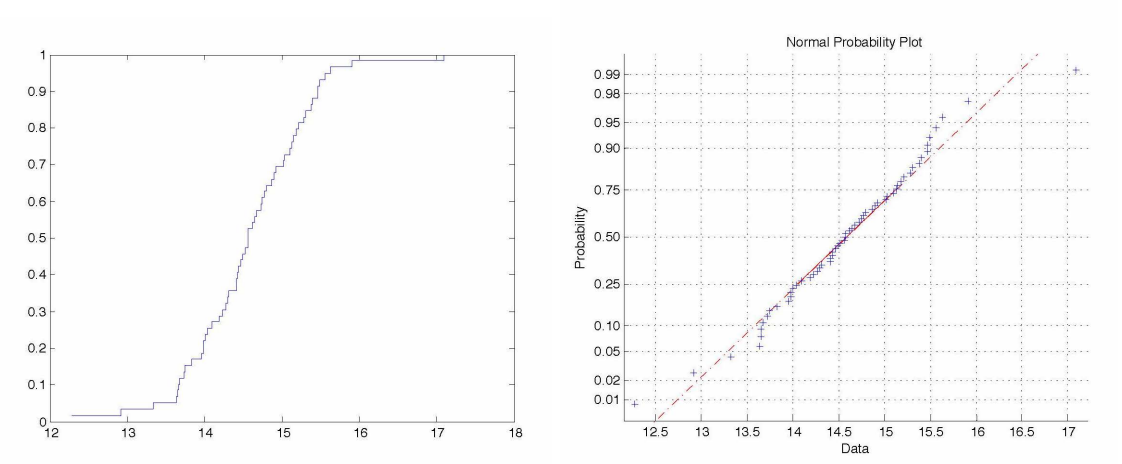
$$\begin{aligned}
\mathrm{Cov}(\hat{F}(u), \hat{F}(v)) &= \mathrm{E}(\hat{F}(u) \cdot \hat{F}(v)) - \mathrm{E}(\hat{F}(u)) \cdot \mathrm{E}(\hat{F}(v)) \\
&= \frac{1}{n}[F(u) + (n-1)F(u)F(v)] - F(u)F(v) \\
&= \tfrac{1}{n}F(u)(1 - F(v)).
\end{aligned}$$

Solution 3

Ordered sample $x_{(1)}, \ldots, x_{(n)}$

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 shows the empirical distribution function and a normal probability plot.



The distribution appears to be rather close to normal. The 10% quantile

$$\frac{x_{(6)} + x_{(7)}}{2} = \frac{13.66 + 13.68}{2} = 13.67.$$

(b) Expected percentages under different dilution levels:

1% dilution	$\mu_1 = 14.58 \cdot 0.99 + 85 \cdot 0.01 = 15.28$	can not be detected
3% dilution	$\mu_3 = 14.58 \cdot 0.97 + 85 \cdot 0.03 = 16.69$	can be detected
5% dilution	$\mu_5 = 14.58 \cdot 0.95 + 85 \cdot 0.05 = 18.10$	can be detected

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 derivative of

$$1 - F(t) = e^{-\alpha t^\beta},$$

we find the density

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta},$$

and dividing the latter by the former we obtain the hazard function

$$h(t) = \alpha\beta t^{\beta-1}.$$

Solution 5

Take the Weibull distribution with parameters α and β.

- If $\beta = 1$, then $h(t) = \alpha$ is constant and the distribution is memoryless.
- If $\beta > 1$, then $h(t)$ increases with t meaning that the older individuals die more often than the younger.
- If $0 < \beta < 1$, then $h(t)$ decreases with t meaning that the longer you live the healthier you become.

Solution 6

R-command (y = control and x = seeded data)

$$> \ \mathbf{qqplot}(\mathrm{x}, \mathrm{y})$$

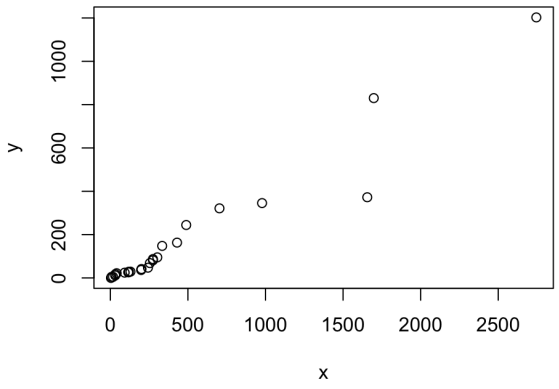
produces a QQ-plot that fits the line $x = 2.5y$ claiming 2.5 times more rainfall from seeded clouds. On the other hand, after the log-transformation

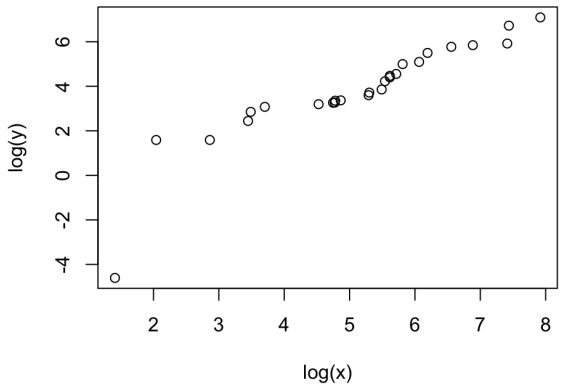
$$> \ \mathbf{qqplot}(\mathbf{log}(\mathrm{x}), \mathbf{log}(\mathrm{y}))$$

we get a QQ-plot that fits the line

$$\ln x = 2 + 0.8 \ln y$$

meaning a decreasing slope in the non-linear relationship $x = 7.4y^{0.8}$.





Solution 7

Using

$$h(t) = -\frac{d \log S(t)}{dt},$$

we get

$$\log S(t) = -H(t),$$

which implies

$$S(t) = e^{-H(t)}.$$

13.6 Solutions to Section 8 (two samples)

Solution 1

(a) $\bar{x} = 0.5546$, $\bar{y} = 1.6240$, $\bar{y} - \bar{x} = 1.0694$

(b) We have $s_x^2 = 0.2163$, $s_y^2 = 1.1795$, $s_{\mathrm{p}}^2 = 0.7667$. The latter is an unbiased estimate of σ^2.

(c) $s_{\bar{y}-\bar{x}} = 0.5874$

(d) Based on t_7-distribution, an exact 90% confidence interval for mean difference is

$$I_{\mu_y - \mu_x} = 1.0694 \pm 1.1128.$$

(e) More appropriate to use a two-sided test.

(f) From the observed test statistic value $t = 1.8206$, we find the two-sided p = 0.1115 using the R command

```
> 2*pt(-1.8206,7)
[1] 0.1114705
```

(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: $s_{\bar{y}-\bar{x}} = 0.0.6708$,
d: $I_{\mu_y-\mu_x} = 1.0694 \pm 1.1035$,
f: $z = 1.5942$ two-sided p-value = 0.11.

Solution 2

If $m = n$, then

$$s_{\mathrm{p}}^2\left(\frac{1}{n}+\frac{1}{m}\right) = \frac{2}{n}\cdot\frac{\sum_{i=1}^n(x_i-\bar{x})^2+\sum_{i=1}^n(y_i-\bar{y})^2}{2n-2} = \frac{s_x^2+s_y^2}{n} = \frac{s_x^2}{n}+\frac{s_y^2}{m} = s_{\bar{x}}^2+s_{\bar{y}}^2.$$

Solution 3

Test the null hypothesis of no drug effect

$H_0 : \mu_1 = \mu_2$, the drug is not effective for reducing high blood pressure.

Suggested measurement design: during the same $n = 10$ days take blood pressure measurements on 4 people, two on the treatment

$$x_{11}, \ldots, x_{1n},$$
$$x_{21}, \ldots, x_{2n},$$

and two controls

$$x_{31}, \ldots, x_{3n},$$
$$x_{41}, \ldots, x_{4n}.$$

Dependencies across the days and the people make inappropriate both two-sample t test and rank sum test. Proper design for 40 measurements is that of two independent samples: 20 people on the treatment and 20 controls:

$$x_1, \ldots, x_{20},$$
$$y_1, \ldots, y_{20}.$$

Solution 4

(a) The sign test statistic

$$t = \text{number of positive } x_i, \qquad T \overset{H_0}{\sim} \mathrm{Bin}(25, \tfrac{1}{2}) \approx \mathrm{N}(\tfrac{25}{2}, \tfrac{5}{2}).$$

Reject H_0 for $t \geq k$, where k is found from

$$0.05 = \mathrm{P}(T \geq k|H_0) = \mathrm{P}(T > k-1|H_0) \approx 1-\Phi\left(\frac{k-0.5-12.5}{5/2}\right) = 1-\Phi\left(\frac{k-13}{2.5}\right),$$

which gives

$$\frac{k-13}{2.5} = 1.645, \qquad k = 17.$$

We know the true population distribution is $\mathrm{N}(0.3, 1)$. Since

$$\mathrm{P}(X > 0|\mathrm{N}(0.3,1)) = 1-\Phi(-0.3) = \Phi(0.3) = 0.62,$$

we can use

$$T \sim \text{Bin}(25, 0.62) \approx \text{N}(15.5, 2.43)$$

to find the power of the sign test by

$$1 - \beta = \text{P}(T \geq 17) \approx 1 - \Phi\left(\frac{17 - 0.5 - 15.5}{2.43}\right) = 1 - \Phi(0.41) = 0.34.$$

(b) Normal distribution model $X \sim \text{N}(\mu, 1)$. Since $\frac{\bar{X}-\mu}{1/5} \sim \text{N}(0, 1)$, we reject H_0 for

$$5\bar{x} > 1.645, \quad \text{that is for} \quad \bar{x} > 0.33.$$

The power of the test

$$1 - \beta = \text{P}(\bar{X} > 0.33 | \mu = 0.3) = 1 - \Phi\left(\frac{0.33 - 0.3}{1/5}\right) = 1 - \Phi(0.15) = 0.44$$

is higher than the power of the sign test.

Solution 5

Two independent samples

$$x_1, \ldots, x_n, \quad y_1, \ldots, y_n,$$

are taken from two population distributions with equal standard deviation $\sigma = 10$. Approximate 95% confidence interval

$$I_{\mu_1 - \mu_2} \approx \bar{x} - \bar{y} \pm 1.96 \cdot 10 \cdot \sqrt{\tfrac{2}{n}} = \bar{x} - \bar{y} \pm \tfrac{27.72}{\sqrt{n}}.$$

If the confidence interval has width 2, then

$$\tfrac{27.72}{\sqrt{n}} = 1,$$

implying $n \approx 768$.

Solution 6

	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) Two-sample t-test

$$\bar{x} = 10.693, \quad \bar{y} = 6.750, \quad s_x^2 = 23.226, \quad s_y^2 = 12.978, \quad s_{\bar{x}-\bar{y}} = \sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2} = 1.903.$$

Assume equal variances. The observed test statistic

$$t = \frac{10.693 - 6.750}{1.903} = 2.072.$$

With df = 18, the two-sided p-value = 0.053 is found using the R-command 2*pt(-2.072,18).

(b) Rank sum test statistics $R_x = 130$, $R_y = 80$. From the table for rank sum test we find that the two-sided p-value is between $0.05 <$ p-value < 0.10.

(c) The non-parametric test in (b) is more relevant, since both normplot(x) and normplot(y) show non-normality of the data distribution.

(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, Y) at random, then by the division rule of probability

$$\mathrm{P}(X < Y) = \frac{\text{number of } (x_i < y_j)}{\text{total number of pairs } (x_i, y_j)} = \frac{10+4+4+3+2+2}{100} = 0.25.$$

This implies a point estimate $\hat{\pi} = 0.75$.

Solution 7

Model: iid-sample of the differences $d_1, \ldots, d_n$ whose population distribution is symmetric around the unknown median m. Test the null hypothesis of no difference $H_0 : m = 0$ using the signed ranks test statistic w_+ defined as follows:

step 1: remove signs $|d_1|, \ldots, |d_n|$,
step 2: assign ranks $1, \ldots, n$ to $|d_1|, \ldots, |d_n|$,
step 3: attach the original signs of d_i to the ranks $1, \ldots, n$,
step 4: compute w_+ as the sum of the positive ranks.

Under $H_0 : m = 0$, on the step 3, the signs $\pm$ are assigned symmetrically at random (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%. Thus $n = 4$ is too small sample size. Therefore, the table for the critical values of the signed rank test starts from $n = 5$.

Solution 8

Using

$$W_{0.05}(n) = \frac{n(n+1)}{4} - 1.96 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}},$$
$$W_{0.01}(n) = \frac{n(n+1)}{4} - 2.58 \cdot \sqrt{\frac{n(n+1)(2n+1)}{24}},$$

we find (table/normal approximation)

	$n = 10$	$n = 20$	$n = 25$
$\frac{n(n+1)}{4}$	27.5	105	162.5
$\sqrt{\frac{n(n+1)(2n+1)}{24}}$	9.81	26.79	37.17
$\alpha = 0.05$	8/8.3	52/53.5	89/89.65
$\alpha = 0.01$	3/2.2	38/36.0	68/67.6

Solution 9

(a) The variance of a difference

$$\mathrm{Var}(D) = \mathrm{Var}(X - Y) = \sigma_x^2 + \sigma_y^2 - 2\mathrm{Cov}(X, Y) = 100 + 100 - 100 = 100.$$

Using the normal approximation we get

$$\bar{D} = \bar{X} - \bar{Y} \approx \mathrm{N}(\mu_x - \mu_y, \sqrt{\tfrac{100}{25}}) = \mathrm{N}(\delta, 2).$$

The rejection region becomes

$$\mathcal{R} = \{\tfrac{\bar{d}}{2} > 1.645\} = \{\bar{d} > 3.29\}.$$

The power function (under the one-sided alternative $\delta > 0$)

$$\mathrm{Pw}(\delta) \approx \mathrm{P}(\bar{D} > 3.29 | \mathrm{N}(\delta, 2)) = 1 - \Phi(\tfrac{3.29-\delta}{2}) = \Phi(\tfrac{\delta-3.29}{2}).$$

(b) Two independent samples

$$\bar{D} \approx \mathrm{N}(\mu_x - \mu_y, \sqrt{\tfrac{100}{25} + \tfrac{100}{25}}) = \mathrm{N}(\delta, 2.83).$$

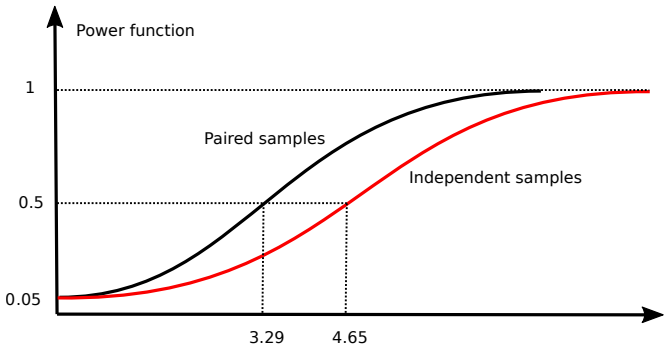
The rejection region

$$\mathcal{R} = \{\tfrac{\bar{d}}{\sqrt{8}} > 1.645\} = \{\bar{d} > 4.65\}.$$

The power function

$$\mathrm{Pw}(\delta) \approx \mathrm{P}(\bar{D} > 4.65 | \mathrm{N}(\delta, 2.83)) = 1 - \Phi(\tfrac{4.65-\delta}{2.83}) = \Phi(\tfrac{\delta-4.65}{2.83}).$$

The two power functions are compared graphically on the next figure.



Solution 10

Paired samples

$$\begin{aligned} \bar{x} &= 85.26, \quad s_x = 21.20, \quad s_{\bar{x}} = 5.47, \quad n_x = 15, \\ \bar{y} &= 84.82, \quad s_y = 21.55, \quad s_{\bar{y}} = 5.57, \quad n_y = 15, \\ \bar{d} &= \bar{x} - \bar{y} = 0.44, \\ s_d &= 4.63, \quad s_{\bar{x}-\bar{y}} = 1.20. \end{aligned}$$

If the pairing had been erroneously ignored, then the two independent samples formula would give 6 times larger standard error

$$s_{\bar{x}-\bar{y}} = 7.81.$$

To test $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x \neq \mu_y$ assume $D \sim \mathrm{N}(\mu, \sigma)$ and apply one-sample t-test

$$t = \frac{\bar{d}}{s_{\bar{d}}} = 0.368.$$

With df $= 14$, two-sided p-value $= 0.718$, we can not reject H_0.
Without normality assumption we apply the signed rank test. The R-command

```
> wilcox.test(x, y, paired = TRUE)
```

computes the two-sided p-value $= 0.604$. We can not reject H_0.

Solution 11

Possible explanations

(a) room with a window ← rich patient → recovers faster,

(b) besides passive smoking: smoker ← the man is a bad husband → wife gets cancer,

(c) no breakfast ← more stress → accident,

(d) choose to change the school and to be bused ← lower grades before → lower grades after,

(e) match two babies with two mothers (or even 3 babies with 3 mothers) then it is pure chance,

(f) abstain from alcohol ← poor health,

(g) marijuana ← schizophrenia,

(h) total time together = time before wedding + time after wedding,

(i) being part of a community can have a positive effect on mental health and emotional wellbeing.

13.7 Solutions to Section 9 (analysis of variance)

Solution 1

(a) The sums of squares: between samples, within samples and total:

$\mathrm{SS_A} = 10((20.34-19.40)^2 + (18.34-19.40)^2 + (21.57-19.40)^2 + (17.35-19.40)^2) = 109.2$
$\mathrm{SS_E} = 9(0.88 + 0.74 + 0.88 + 0.89) = 30.5$
$\mathrm{SS_T} = 3.58 \cdot 39 = 139.7 = 109.2 + 30.5$

Source	SS	df	MS	F
Treatment	109.2	3	36.4	42.9
Error	30.5	36	0.85	
Total	139.7	39		

Comparison of the observed test statistics 42.9 with the critical value for $\mathrm{F}_{3,36}(0.001) = 6.7436$, see Section 12.5, shows that the P-value is maller than 0.001, so that we can reject the null hypothesis.

(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

$$B_{\mu_u - \mu_v} = (\bar{y}_{u\cdot} - \bar{y}_{v\cdot}) \pm 1.11.$$

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 2

Consider one-way ANOVA test statistic

$$F = \frac{MS_A}{MS_E} = \frac{\frac{n}{I-1}\sum_{i=1}^{I}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\frac{1}{I(n-1)}\sum_{i=1}^{I}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2}$$

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

$$F = \frac{n[(\bar{x} - \frac{\bar{x}+\bar{y}}{2})^2 + (\bar{y} - \frac{\bar{x}+\bar{y}}{2})^2]}{\frac{1}{2(n-1)}[\sum_{j=1}^{n}(x_j - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2]} = \frac{2n(\frac{\bar{x}-\bar{y}}{2})^2}{s_\mathrm{p}^2} = \Big(\frac{\bar{x}-\bar{y}}{s_\mathrm{p}\sqrt{\frac{2}{n}}}\Big)^2.$$

This equals t^2, where $t = \frac{\bar{x}-\bar{y}}{s_\mathrm{p}\sqrt{\frac{2}{n}}}$ is the two-sample t-test statistic.

Solution 3

The null hypothesis says that the data

$$(y_{ij}), \quad i = 1, \ldots, I, \quad j = 1, \ldots, n,$$

come from a single normal distribution

$$H_0 : \mu_1 = \ldots = \mu_I = \mu$$

described by two parameters μ and σ^2, so that dim $\Omega_0 = 2$, while

$$\dim \Omega = I + 1,$$

since the general setting is described by parameters $\mu_1, \ldots, \mu_I$ and σ^2. The likelihood ratio

$$\Lambda = \frac{L_0(\hat{\mu}, \hat{\sigma}_0^2)}{L(\hat{\mu}_1, \ldots, \hat{\mu}_I, \hat{\sigma}^2)},$$

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

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\sigma}_0^2 = \frac{SS_{\mathrm{T}}}{N}, \quad \hat{\mu}_i = \bar{y}_{i\cdot}, \quad \hat{\sigma}^2 = \frac{SS_{\mathrm{E}}}{N},$$

which yields

$$\Lambda = \frac{\hat{\sigma}_0^{-N} \exp\{-\sum\sum \frac{(y_{ij}-\hat{\mu})^2}{2\hat{\sigma}_0^2}\}}{\hat{\sigma}^{-N} \exp\{-\sum\sum \frac{(y_{ij}-\hat{\mu}_i)^2}{2\hat{\sigma}^2}\}} = \Big(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\Big)^{-N/2} \cdot \frac{\exp\{-\frac{SS_{\mathrm{T}}}{2SS_{\mathrm{T}}/N}\}}{\exp\{-\frac{SS_{\mathrm{E}}}{2SS_{\mathrm{E}}/N}\}} = \Big(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\Big)^{N/2}.$$

The likelihood ratio test rejects the null hypothesis for small values of Λ or equivalently for large values of

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \tfrac{SS_{\mathrm{T}}}{SS_{\mathrm{E}}} = 1 + \tfrac{SS_{\mathrm{A}}}{SS_{\mathrm{E}}} = 1 + \tfrac{(I-1)MS_{\mathrm{A}}}{I(n-1)MS_{\mathrm{E}}} = 1 + \tfrac{(I-1)}{I(n-1)} \cdot F$$

that is for large values of F-test statistics.

Recall that

$$-2 \ln \Lambda \overset{H_0}{\approx} \chi^2_{\mathrm{df}}, \quad \text{where df} = \dim(\Omega) - \dim(\Omega_0) = I - 1.$$

To see that the exact null distribution for F agrees with the asymptotic null distribution for the LRT, observe that for large n,

$$-2 \ln \Lambda = (I \cdot n) \ln(1 + \tfrac{(I-1)}{I(n-1)} \cdot F) \approx (I-1)F = \frac{SS_{\mathrm{A}}}{MS_{\mathrm{E}}}.$$

Notice that MS_{E} converges to σ^2 as $n \to \infty$, and therefore,

$$-2 \ln \Lambda \approx \frac{SS_{\mathrm{A}}}{\sigma^2},$$

where the null distribution of the ratio $\frac{SS_{\mathrm{A}}}{\sigma^2}$ is approximated by the chi-squared distribution with $I - 1$ degrees of freedom.

Solution 4

One-way layout with $I = 10$, $n = 7$,

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma).$$

Pooled sample variance

$$s_{\mathrm{p}}^2 = MS_{\mathrm{E}} = \frac{1}{I(n-1)} \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$$

uses df $= I(n-1) = 60$.

(a) A 95% confidence interval for a single difference $\mu_u - \mu_v$

$$I_{\mu_u-\mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm t_{60}(0.025) s_{\mathrm{p}} \sqrt{\tfrac{2}{n}}$$

has the half-width of

$$2.82 \cdot \tfrac{s_{\mathrm{p}}}{\sqrt{n}}.$$

(b) Bonferroni simultaneous 95% confidence interval for $\binom{10}{2} = 45$ differences $\mu_u - \mu_v$

$$B_{\mu_u-\mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm t_{60}(\tfrac{0.025}{45}) s_{\mathrm{p}} \sqrt{\tfrac{2}{n}}$$

has the half-width of

$$4.79 \cdot \tfrac{s_{\mathrm{p}}}{\sqrt{n}},$$

giving the ratio

$$\tfrac{4.79}{2.82} = 1.7.$$

(c) Since

```
> qtukey(0.95,10,60)
[1] 4.646324
```

the Tukey simultaneous 95% confidence interval for differences $\mu_u - \mu_v$

$$T_{\mu_u-\mu_v} = \bar{y}_{u\cdot} - \bar{y}_{v\cdot} \pm q_{10,60}(0.05) \tfrac{s_{\mathrm{p}}}{\sqrt{n}}$$

has the half-width of

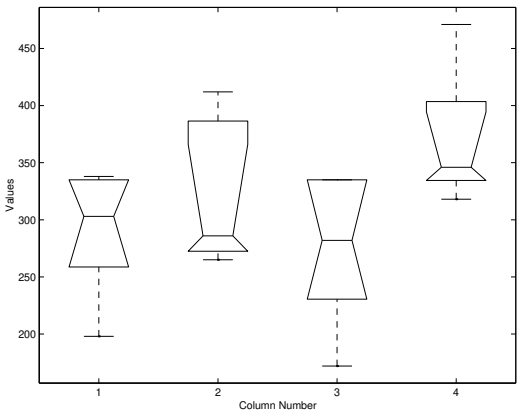
$$4.65 \cdot \tfrac{s_{\mathrm{p}}}{\sqrt{n}},$$

giving the ratio

$$\tfrac{\text{Bonferroni}}{\text{Tukey}} = \tfrac{4.79}{4.65} = 1.03.$$

Solution 5

For $I = 4$ control groups of $n = 5$ mice each, test H_0: no systematic differences between groups.



One way ANOVA table

Source	SS	df	MS	F	P
Columns	27230	3	9078	2.271	0.12
Error	63950	16	3997		
Total	91190	19			

Do not reject H_0 at 10% significance level. Boxplots show non-normality. The largest difference is between the third and the fourth boxplots. Control question: why the third boxplot has no upper whisker?
Kruskal-Wallis test. Pooled sample ranks

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	$\bar{r}_{4.} = 15.8$

Kruskal-Wallis test statistic

$$W = \frac{12 \cdot 5}{20 \cdot 21}\left((8.4 - 10.5)^2 + (10.6 - 10.5)^2 + (7.2 - 10.5)^2 + (15.8 - 10.5)^2\right) = 6.20.$$

Since $\chi^2_3(0.1) = 6.25$, we do not reject H_0 at 10% significance level.

Solution 6

Two-way layout with $I = 3$ treatments on $J = 10$ subjects with $n = 1$ observations per cell. ANOVA table

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			

Reject

H_0: no treatment effects

at 5% significance level. (Interestingly, no significant differences among the blocks.)

Friedman's test. Ranking within 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

$$Q = \frac{12 \cdot 10}{3 \cdot 4}\left((1.8 - 2)^2 + (1.9 - 2)^2 + (2.3 - 2)^2\right) = 1.4.$$

Since $\chi^2_2(0.1) = 4.61$, we can not reject H_0 even at 10% significance level.

Solution 7

Forty eight survival times: $I = 3$ poisons and $J = 4$ treatments with $n = 4$ observations per cell. Cell means for the survival times

	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

Draw three profiles: I and II cross each other, and profile III is more flat. Three null hypotheses of interest

H_A: no poison effect,
H_B: no treatment effect,
H_{AB}: no interaction.

(a) Survival in hours x data matrix. ANOVA2 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			

Reject H_A and H_B at 1% significance level, we can not reject H_{AB} even at 10% significance level:

3 poisons act differently,
4 treatments act differently,
some indication of interaction.

Analysis of the residuals

normal probability plot reveals non-normality,
skewness = 0.59,
kurtosis = 4.1.

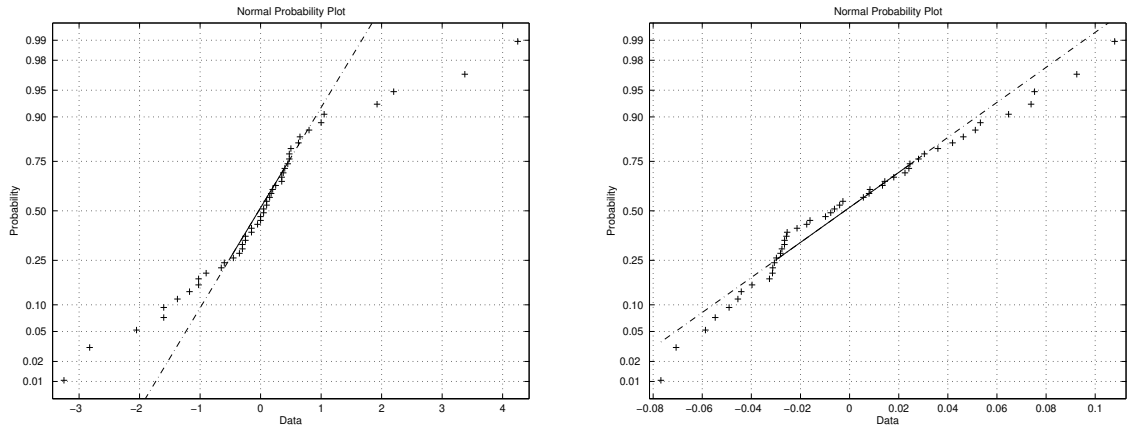


Figure 1: Left panel: survival times. Right panel: death rates.

(b) Transformed data: death rate = 1/survival time. Cell means for the death rates

	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. The two-way ANOVA results

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			

Reject H_A and H_B at 1% significance level. Do not reject H_{AB}. Conclusions

3 poisons act differently,
4 treatments act differently,
no interaction,
the normal probability plot of residuals reveals a closer fit to normality assumption.

13.8 Solutions to Section 10 (categorical data analysis)

Warning: in some of the contingency tables the expected counts are rounded. If you then will compute the chi-squared test statistic X^2 from the table, you will often get a somewhat different value.

Solution 1

Test

H_0: same genotype frequencies for diabetics and normal

using the chi-squared test of homogeneity. 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

Observed X^2=5.10, df=1, p-value = 0.024. Reject H_0. Diabetics have genotype BB less often.

The exact Fisher test uses Hg(104,51,$\frac{16}{104}$) as the null distribution of the test statistic $N_{11} = 12$

one-sided P-value:

```
> 1-phyper(11,16,88,51)
[1] 0.02245863
```

two-sided P-value $P = 0.045$.

Notice that

$$\mathrm{P}(N_{11} \geq 12) = 1 - \mathrm{P}(N_{11} \leq 11),$$

which explains why the number 11 (and not 12) appears in the command "phyper(11,16,88,51)".

Normal approximation of the null distribution

$$\mathrm{Hg}(104, 51, \tfrac{16}{104}) \approx \mathrm{N}(7.85, \sqrt{3.41}).$$

Since $z_{\mathrm{obs}} = \frac{12-7.85}{1.85}$=2.245, the approximate two-sided p-value = 0.025.

Solution 2

(a) H_0: no association of the disease and the ABO blood group:

	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

Observed X^2=15.37, df=6, p-value = 0.018. Reject H_0.

(b) H_0: no association of the disease and the MN blood group:

	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

Observed X^2=4.73, df=4, p-value = 0.42. Can not reject H_0.

Solution 3

(a) Apply 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

Observed X^2=2.75, df=2, p-value = 0.25. Can not reject H_0.

(b) Goodness of fit chi-squared test for the same sex ratio for three father's activities

$$H_0\text{: boys proportions } p_{12} = p_{22} = p_{32} = 0.513.$$

Here 0.513 is obtained as

$$\frac{105.37}{105.37 + 100} = 0.513.$$

Observed and expected counts

	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

Observed $X^2 = 3.09$, df $= 3$, p-value $= 0.38$. Can not reject H_0.
Why we use df $= 3$ is explained next. The general model is described by three independent parameters (p_{12}, p_{22}, p_{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

	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 (150 mg)	283	271	554

The observed test statistic $X^2 = 35.8$ according to the χ^2_4-distribution table gives p-value $= 3 \cdot 10^{-7}$. Comparing the observed and expected counts we conclude that Chlorpromazine is most effective in ameliorating postoperative nausea.

Solution 5

(a) H_0: no relation between blood group and disease in London:

	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

Observed X^2=42.40, df=1, p-value $= 0.000$. Reject H_0. Odds ratio $\hat{\Delta} = 1.45$.

(b) H_0: no relation between blood group and disease in Manchester:

	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

Observed X^2=5.52, df=1, p-value $= 0.019$. Reject H_0. Odds ratio $\hat{\Delta} = 1.22$.

(c) H_0: London Group A and Manchester Group A have the same propensity to Peptic Ulcer:

	C and A	PU and A	Total
London	4219 (4349.2)	579 (448.8)	4798
Manchester	3775 (3644.8)	246 (376.2)	4021
Total	7994	825	8819

Observed X^2=91.3, df=1, p-value $= 0.000$. Reject H_0.

H_0: London Group O and Manchester Group O have the same propensity to Peptic Ulcer:

	C and O	PU and O	Total
London	4578 (4816.5)	911 (672.5)	5489
Manchester	4532 (4293.5)	361 (599.5)	4893
Total	9110	1272	10382

Observed X^2=204.5, df=1, p-value = 0.000. Reject H_0.

Solution 6

D = endometrical carcinoma, X = estrogen taken at least 6 months prior to the diagnosis of cancer.

(a) Matched controls, retrospective case-control study

	$\bar{D}X$	$\bar{D}\bar{X}$	Total
DX	39	113	152
$D\bar{X}$	15	150	165
Total	54	263	317

Apply McNemar test for

$$H_0 : \pi_{1\cdot} = \pi_{\cdot 1} \quad \text{vs} \quad H_1 : \pi_{1\cdot} \neq \pi_{\cdot 1}.$$

Observed value of the test statistic

$$X^2 = \frac{(113-15)^2}{113+15} = 75$$

is highly significant as $\sqrt{75} = 8.7$ and the corresponding two-sided P-value obtained from N(0,1) table is very small.

(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

(a) The exact Fisher test uses Hg(30,17,$\frac{16}{30}$) as the null distribution of the test statistic whose observed value is $x = 12$. It gives

one-sided p-value:

```
> 1-phyper(11,16,14,17)
[1] 0.03548226
```

two-sided p-value $P = 0.071$.

(b) Using normal approximation

$$\text{Hg}(30, 17, \tfrac{16}{30}) \approx \text{N}(9.1, 1.4)$$

and continuity correction, we find the one-sided p-value to be

$$\text{P}(X \geq 12|H_0) = \text{P}(X > 11|H_0) \approx 1 - \Phi(\tfrac{11.5-9.1}{1.4}) = 1 - \Phi(1.71) = 0.044.$$

(c) Approximate chi-squared test yields: observed $X^2 = 4.69$, df = 1, two-sided p-value

$$2(1 - \Phi(\sqrt{4.69})) = 2(1 - \Phi(2.16)) = 0.03.$$

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.

(a, c) Assuming

$$H_{eq} : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi,$$

we test

$$H_0 : \pi = \tfrac{1}{2} \quad \text{vs} \quad H_1 : \pi \neq \tfrac{1}{2}.$$

We use the large sample test for proportion based on the statistic $X = 245$ whose null distribution is $\text{Bin}(n, \frac{1}{2})$, $n = 447$. The two-sided P-value is approximated by

$$2\left(1 - \Phi\left(\frac{245 - \frac{447}{2}}{\sqrt{447} \cdot \frac{1}{2}}\right)\right) = 2(1 - \Phi(2.034)) = 0.042.$$

At 5% level we reject the $H_0 : \pi = \frac{1}{2}$. The maximum likelihood estimate is $\hat{\pi} = \frac{245}{447} = 0.55$.

(d) Is there evidence that wearing red is more favourable in some of the sports than others? We test

$$H_{eq} : \pi_1 = \pi_2 = \pi_3 = \pi_4 \quad \text{vs} \quad H_{ineq} : \pi_i \neq \pi_j \quad \text{for some } i \neq j$$

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 $\text{X}^2 = 0.3$ is not significant. We can not reject H_{eq}, which according to (a) leads to $\hat{\pi} = 0.55$.

(b) Now we state the hypotheses of interest directly: consider

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \tfrac{1}{2} \quad \text{vs} \quad H_1 : (\pi_1, \pi_2, \pi_3, \pi_4) \neq (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}).$$

Here we need a new chi-squared test, a chi-squared test for k proportions with $k = 4$ (see below). Given four observed counts $x_1 = 148$, $x_2 = 27$, $x_3 = 25$, $x_4 = 45$, we obtain the following table of observed and expected counts

	Red	Biue	Total
Boxing	148 (134)	120 (134)	268
Freestyle wrestling	27 (25.5)	24 (25.5)	51
Greco-Roman wrestling	25 (24)	23 (24)	48
Tae Kwon Do	45 (40)	35 (40)	80
H_0 proportions	0.5	0.5	1.00

It gives $\text{X}^2_{\text{obs}} = 4.4$. Since $\chi^2_4(0.1) = 7.8$, we do not reject $H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{2}$.

Chi-square test for k proportions

Given k independent values $(x_1, \ldots, x_k)$ drawn from k binomial distributions with parameters $(n_1, \pi_1), \ldots, (n_k, \pi_k)$ we test (for some specified values $(p_1, \ldots, p_k)$)

$$H_0 : \pi_1 = p_1, \ldots, \pi_k = p_k \quad \text{vs} \quad H_1 : (\pi_1, \pi_2, \pi_3, \pi_4) \neq (p_1, \ldots, p_k)$$

using the test statistic

$$\text{X}^2 = \sum_{j=1}^{2k} \frac{(O_j - E_j)^2}{E_j} = \sum_{i=1}^{k} \left(\frac{(x_i - n_i p_i)^2}{n_i p_i} + \frac{(n_i - x_i - n_i(1-p_i))^2}{n_i(1-p_i)} \right),$$

whose null distribution is approximately χ^2_k. The last fact follows from

$$\sum_{i=1}^{k} \left(\frac{(x_i - n_i p_i)^2}{n_i p_i} + \frac{(n_i - x_i - n_i(1-p_i))^2}{n_i(1-p_i)} \right) = \sum_{i=1}^{k} \frac{(x_i - n_i p_i)^2}{n_i p_i (1-p_i)} = \sum_{i=1}^{k} z_i^2,$$

where z_i are realisations of independent random variables

$$Z_i = \frac{X_i - n_i p_i}{\sqrt{n_i p_i (1-p_i)}}, \quad i = 1, \ldots, k$$

which are approximately N(0,1) distributed, provided $X_i \sim \mathrm{Bin}(n_i, p_i)$.
We derive this test statistic using the likelihood ratio approach. The likelihood function based on the binomial model has the form

$$L(\pi_1, \ldots, \pi_k) = \prod_{i=1}^{k} \binom{n_i}{x_i} \pi_i^{x_i} (1-\pi_i)^{n_i - x_i}.$$

Using $\hat{\pi}_i = \frac{x_i}{n_i}$, we compute the likelihood ratio as

$$\Lambda = \frac{L(p_1, \ldots, p_k)}{L(\hat{\pi}_1, \ldots, \hat{\pi}_k)} = \frac{\prod_{i=1}^{k} p_i^{x_i} (1-p_i)^{n_i - x_i}}{\prod_{i=1}^{k} (\frac{x_i}{n_i})^{x_i} (\frac{n_i - x_i}{n_i})^{n_i - x_i}} = \prod_{i=1}^{k} (\frac{n_i p_i}{x_i})^{x_i} (\frac{n_i (1-p_i)}{n_i - x_i})^{n_i - x_i}.$$

Turn to the logarithms,

$$-\ln \Lambda = \sum_{i=1}^{k} x_i \ln \tfrac{x_i}{n_i p_i} + (n_i - x_i) \ln \tfrac{n_i - x_i}{n_i (1-p_i)},$$

observe that under H_0 we have

$$\tfrac{x_i}{n_i} \approx p_i, \qquad \tfrac{n_i - x_i}{n_i} \approx 1 - p_i.$$

Using a Taylor expansion

$$\ln \tfrac{x_i}{n_i p_i} \approx \tfrac{x_i - n_i p_i}{n_i p_i}, \qquad \ln \tfrac{n_i - x_i}{n_i (1-p_i)} \approx \tfrac{n_i p_i - x_i}{n_i (1-p_i)},$$

we find that

$$-\ln \Lambda \approx \sum_{i=1}^{k} \tfrac{(x_i - n_i p_i)^2}{n_i p_i (1-p_i)} = \mathrm{X}^2.$$

13.9 Solutions to Section 11 (multiple regression)

Solution 1

Recall that the sample covariance and the population covariance are

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \quad \mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

It is enough to check that

$$\mathrm{E}\left(\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right) = (n-1)\mathrm{E}(XY) - (n-1)\mathrm{E}(X)\mathrm{E}(Y).$$

To do this, 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

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{n-1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i \neq j} \sum_{j=1}^{n} x_i y_j.$$

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

We have after ordering

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

and

$$\bar{x} = -0.046, \quad \bar{y} = -0.075, \quad s_x = 1.076, \quad s_y = 0.996, \quad r = 0.98.$$

(a) Simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma).$$

Fitting a straight line using

$$y - \bar{y} = r \cdot \tfrac{s_y}{s_x}(x - \bar{x})$$

we get the predicted response

$$\hat{y} = -0.033 + 0.904 \cdot x.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_y^2 (1 - r^2) = 0.05.$$

(b) Simple linear regression model

$$X = \beta_0 + \beta_1 y + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma).$$

Fitting a straight line using

$$x - \bar{x} = r \cdot \tfrac{s_x}{s_y}(y - \bar{y})$$

we get the predicted response

$$\hat{x} = 0.033 + 1.055 \cdot y.$$

Estimated σ^2

$$s^2 = \tfrac{n-1}{n-2} s_x^2 (1 - r^2) = 0.06.$$

(c) First fitted line

$$y = -0.033 + 0.904 \cdot x$$

is different from the second

$$y = -0.031 + 0.948 \cdot x.$$

They are different since in (a) we minimise the vertical residuals while in (b) - horizontal.

Solution 3

Using an extra explanatory variable f which equal 1 for females and 0 for males, we rewrite this model in the form of a multiple regression

$$Y = f\beta_F + (1 - f)\beta_M + \beta_1 x + \epsilon = \beta_0 + \beta_1 x + \beta_2 f + \epsilon,$$

where

$$\beta_0 = \beta_M, \quad \beta_2 = \beta_F - \beta_M.$$

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 compute

$$\beta_M = \beta_0, \quad \beta_F = \beta_0 + \beta_2.$$

A null hypothesis of interest $\beta_2 = 0$.

Solution 4

(a) The predicted value $\hat{Y}_0$ and actual observation Y_0 are independent random variables, therefore

$$\mathrm{Var}(Y_0 - \hat{Y}_0) = \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{Y}_0) = \sigma^2 + \mathrm{Var}(b_0 + b_1 x_0) = \sigma^2 a_n^2,$$

where

$$a_n^2 = 1 + \frac{\mathrm{Var}(b_0)+\mathrm{Var}(b_1)x_0^2-2x_0\mathrm{Cov}(b_0,b_1)}{\sigma^2} = 1 + \frac{\overline{x^2}+x_0^2-2\bar{x}x_0}{(n-1)s_x^2} = 1 + \frac{\overline{x^2}-\bar{x}^2+(x_0-\bar{x})^2}{(n-1)s_x^2} = 1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}.$$

(b) A 95% prediction interval I for the new observation Y_0 is obtained from

$$\frac{Y_0-\hat{Y}_0}{Sa_n} \sim t_{n-2}.$$

Since

$$0.95 = \mathrm{P}(|Y_0 - \hat{Y}_0| \le t_{n-2}(0.025) \cdot S \cdot a_n) = \mathrm{P}(Y_0 \in \hat{Y}_0 \pm t_{n-2}(0.025) \cdot S \cdot a_n),$$

we conclude that a 95% prediction interval for the new observation Y_0 is given by

$$I = b_0 + b_1 x_0 \pm t_{n-2}(0.025) \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}.$$

The further from $\bar{x}$ lies x_0, the more uncertain becomes the prediction.

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.

(b) Given $y = 85$ and we do not know the midterm score, 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) Find the correlation coefficient ρ for (X, Y). Since $\mathrm{E}X = 0$, we have

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) = \mathrm{E}(X^2 + \beta XZ) = 1, \quad \mathrm{Var}Y = \mathrm{Var}X + \beta^2 \mathrm{Var}Z = 1 + \beta^2,$$

and we see that the correlation coefficient is always positive

$$\rho = \frac{1}{\sqrt{1+\beta^2}}.$$

(b) Use (a) to generate five samples

$$(x_1, y_1), \ldots, (x_{20}, y_{20})$$

with different

$$\rho = -0.9, \quad -0.5, \quad 0, \quad 0.5, \quad 0.9,$$

and compute the sample correlation coefficients.
From $\rho = \frac{1}{\sqrt{1+\beta^2}}$, we get $\beta = \sqrt{\rho^{-2} - 1}$ so that

$$\rho = 0.5 \Rightarrow \beta = 1.73, \qquad \rho = 0.9 \Rightarrow \beta = 0.48.$$

How to generate a sample with $\rho = -0.9$ using R:

```
> X=rnorm(20)
> Y=-X+0.48*rnorm(20)
> r=cor(X,Y)
```

Simulation results

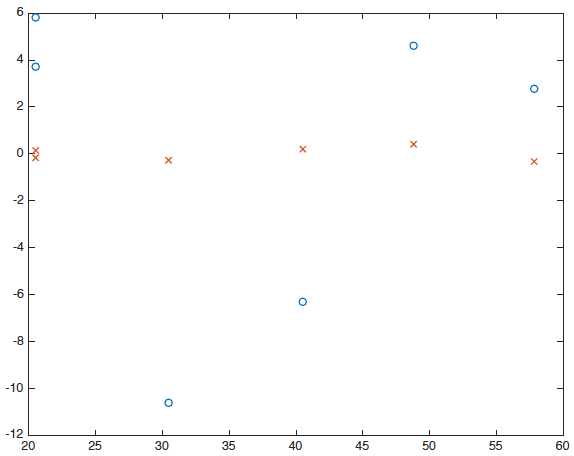
ρ	-0.9	-0.5	0	0.5	0.9
r	-0.92	-0.45	-0.20	0.32	0.92

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 two residual plots



Kinetic energy formula explains why the second model is better.

14 Miscellaneous exercises

14.1 Problems

Problem 1

From 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 the last sentence.

Problem 2

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 $x \in (-\infty, \infty)$.

(a) The variance of this distribution is $2\lambda^{-2}$ and kurtosis is 6. Prove this using the formula $\int_0^\infty x^k e^{-x} dx = k!$ valid for any natural number k.

(b) Take $\lambda = \sqrt{2}$. Plot carefully the density $f(x)$ together with the standard normal distribution density.

(c) 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 3

The following 16 numbers came from 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 for the mean and variance of the generating normal distribution. (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 MLE for the variance unbiased?

Problem 4

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 5

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 6

Suppose you have a data of size n for which the linear regression model seems to work well. The key summary statistics are represented by sample means $\bar{x}, \bar{y}$, sample standard deviations s_x, s_y, and a sample correlation coefficient r.
An important use of the linear regression model is forecasting. Assume we are interested in the response to a particular value x of the explanatory variable.

(a) The exact $100(1-\alpha)\%$ confidence interval for the mean response value is given by the formula:

$$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{1}{n-1}\left(\frac{x-\bar{x}}{s_x}\right)^2}.$$

Explain carefully the meaning and role of each of the terms.

(b) Another important formula in this context

$$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{1}{n-1}\left(\frac{x-\bar{x}}{s_x}\right)^2}$$

is called the exact $100(1-\alpha)\%$ prediction interval. Explain the difference between these two formulae. Illustrate by a simple example.

(c) Comment on the predictor properties depending on the distance from the given value x to the sample mean $\bar{x}$. Illustrate using appropriate plots.

Problem 7

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 randomized.

(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.

Problem 8

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.

(c) Compute the standard errors for these point estimates.

(d) What assumptions do you make for these calculations?

Problem 9

Identify important statistical terms hiding behind "bla-bla-bla" in the following extracts from the Internet (one term per item).

(a) Bla-bla-bla 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 bla-bla-bla.

(b) Note that bla-bla-bla is the probability of finding a difference that does exist, as opposed to the likelihood of declaring a difference that does not exist (which is known as a Type I error, or "false positive").

(c) The bla-bla-bla is a measure of the tendency to fail; the greater the value of the bla-bla-bla, the greater the probability of impending failure... The bla-bla-bla is also known as the instantaneous failure rate.

(d) Naive interpretation of statistics derived from data sets that include bla-bla-bla may be misleading. For example, if one is calculating the average temperature of 10 objects in a room, and most are between 20 and 25 degrees Celsius, but an oven is at 175 C, the median of the data may be 23 C but the mean temperature will be between 35.5 and 40 C. In this case, the median better reflects the temperature of a randomly sampled object than the mean; however, naively interpreting the mean as "a typical sample", equivalent to the median, is incorrect. As illustrated in this case, bla-bla-bla may be indicative of data points that belong to a different population than the rest of the sample set.

Problem 10

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 11

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?

Problem 12

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 13

The data of 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

Depth	Flow rate
0.34	0.64
0.29	0.32
0.28	0.73
0.42	1.33
0.29	0.49
0.41	0.92
0.76	7.35
0.73	5.89
0.46	1.98
0.40	1.12

(a) Draw the scatter plot for the given data using the x axis for depth. Fit by eye a regression line and plot the residuals against the depth. What does it say to you about the relevance of the simple linear regression model for this particular data?

(b) The least square estimates for the parameters of the simple linear regression model are $b_0 = -3.98$, $b_1 = 13.83$. Given the standard deviations are $s_x = 0.17$ and $s_y = 2.46$ estimate the noise size (σ) and find the coefficient of determination

(c) The statistics for a quadratic model are given in the following table:

Coefficient	Estimate	Standard error	t value
β_0	1.68	1.06	1.59
β_1	-10.86	4.52	-2.40
β_2	23.54	4.27	5.51

Compute a 95 procent confidence interval for β_0.

(d) Is the quadratic term statistically significant? Carefully explain.

Problem 14

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, in kr/16 hr) and dry weight of roots (y, in mg$\times 10^{-1}$):

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 15

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
Formulation	1	189.7	185.1	189.0	
	1	188.6	179.4	193.0	187.03
	1	190.1	177.3	191.1	
	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

$$SSForm = 2253.44, \quad SSSpeed = 230.81, \quad SSForm * Speed = 18.58, \quad SSE = 71.87.$$

(a) Calculate estimates of the main effects.

(b) Does there appear to be interaction between the factors? In which various ways interaction between such 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 scatter plot of residuals?

Problem 16

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 17

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 18

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 Risk Ratio

$$RR = R_A/R_P = \frac{104/11037}{189/11034} = 0.55.$$

(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?

Problem 19

Given a sample $(x_1, \ldots, x_n)$ of independent and identically distributed observations, we are interested in testing $H_0 : m = m_0$ against the two-sided alternative $H_1 : m \neq m_0$ concerning the population median m. No parametric model is assumed. As a test statistic we take $y = \sum_{i=1}^{n} 1_{\{x_i \leq m_0\}}$, the number of observations below the null hypothesis value.

(a) Find the exact null distribution of Y. What are your assumptions?

(b) Suppose $n = 25$. Suggest an approximate confidence interval formula for m.

Problem 20

Consider the problem of comparison of two simple hypotheses H_0: $p = p_0$, H_1: $p = p_1$ with $p_1 > p_0$ using the large-sample test for the proportion.

(a) Let Y have a binomial distribution with parameters (n, p). The power function of the one-sided test is given by

$$\mathrm{Pw}(p_1) = \mathrm{P}\left(\frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \geq z_\alpha \,|\, p = p_1\right).$$

Explain in detail all parts of this formula.

(b) Suppose we want to plan for the sample size n to control the sizes of two types of error at levels α and β. Derive the following formula for the optimal sample size

$$\sqrt{n} = \frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_1(1-p_1)}}{|p_1 - p_0|}.$$

Hint: under the alternative hypothesis, $\frac{Y - np_1}{\sqrt{np_1(1-p_1)}}$ is approximately normally distributed with parameters (0,1).

(c) What happens to the planned sample size if the alternatives are very close to each other? What happens if we decrease the levels α and β?

Problem 21

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 22

In the Bayesian estimation framework we search for an optimal action

$$a = \{\text{assign value } a \text{ to unknown parameter } \theta\}.$$

The optimal choice depends on the particular form of the loss function $l(\theta, a)$. Bayes action minimizes the posterior risk

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

(a) Explain the meaning of the posterior risk function. What does $h(\theta|x)$ stand for? How is $h(\theta|x)$ computed?

(b) The zero-one loss function is defined by $l(\theta, a) = 1_{\{\theta \neq a\}}$. Compute the posterior risk using the discrete distribution formula. Why is it called the probability of misclassification?

(c) What Bayesian estimator corresponds to the optimal action with the zero-one loss function? Compare this estimator to the maximum likelihood estimator.

Problem 23

Study the picture to the right.
From this observation we would like to estimate the amount of work required to clean a street from chewing gums.

(a) Describe a Poisson distribution model suitable for this particular observation. Summarise the data in a convenient way.

(b) Write down the likelihood function for this particular observation. Find the maximum likelihood estimate.

(c) Without performing the required statistical test describe how to check whether the Poisson model fits to the data.

(d) Estimate the proportion of tiles free from chewing gums using the fitted Poisson model.

Problem 24

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 25

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
H	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
P	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 26

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 27

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)?

Problem 28

The gamma distribution Gamma(α, λ) is a conjugate prior for the Poisson data distribution with a parameter θ. If x is a single observed value randomly sampled from the Poisson distribution, then the parameters (α', λ') for the posterior gamma distribution of θ are found by the following updating rule:
- the shape parameter $\alpha' = \alpha + x$,
- the inverse scale parameter $\lambda' = \lambda + 1$.

(a) Find $\hat{\theta}_{\text{PME}}$, the posteriori mean estimate for the θ, under the exponential prior with parameter 1, given the following iid sample values from the Poisson(θ) 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 $\hat{\theta}_{\text{PME}}$ with the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ as $n \to \infty$. Your conclusions?

Problem 29

Extracorporeal membrane oxygenation (ECMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment was conducted in which 29 babies were treated with ECMO and 10 babies were treated with conventional medical therapy (CMT). In the ECMO group only 1 patient died, while in the CMT group 4 patients died.

(a) Suggest a statistical model and compute the likelihood function for the data as a function of two parameters: p - the probability to die under the ECMO treatment and q - the probability to die under the CMT treatment.

(b) Write down a relevant pair of statistical hypotheses in the parametric form. Perform the exact Fisher test.

Problem 30

Suppose that we have an iid sample of size 100 from the normal distribution with mean μ and standard deviation $\sigma = 10$. For $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$ we use the absolute value of the sample mean $T = |\bar{X}|$ as the test statistic. Denote by V the P-value of the test.

(a) Show that $V = 2(1 - \Phi(T_{\text{obs}}))$, where T_{obs} is the observed value of the test statistic and $\Phi(x)$ is the standard normal distribution function. Plot the null distribution curve for $\bar{X}$ and graphically illustrate this formula.

(b) In what sense the P-value V is a random variable? Using (a) show that

$$P(V \leq 0.05) = P(\bar{X} < -1.96) + P(\bar{X} > 1.96).$$

(c) Suppose that the true value of the population mean is $\mu = 4$. Using (b) show that $P(V \leq 0.05) \approx 0.975$. Illustrate by drawing the density curve for the true distribution of $\bar{X}$.

(d) Comment on the result (c) in the light of the statement: "P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume".

Problem 31

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 iid 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 simultaneous confidence intervals for the three 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 32

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 33

Represent the large sample test for a proportion as a chi-square test.

14.2 Numerical answers to miscellaneous exercises

Answer 1

Coefficient of determination is the squared sample correlation $r^2 = (0.2)^2 = 0.04$.

Answer 2

(a) Since the mean is 0, the variance is computed as

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx = 2\lambda^{-2}.$$

The kurtosis is the scaled fourth moment

$$\beta_2 = \sigma^{-4} \int_{-\infty}^{\infty} x^4 f(x)dx = \frac{\lambda^5}{4} \int_0^{\infty} x^2 e^{-\lambda x} dx = 6.$$

(b) 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.

(c) For $\lambda = \sqrt{2}$ the mean is 0, the skewness is 0, and the kurtosis is 6. Compared to the normal curve with the same mean but smaller kurtosis (=3), the Laplace distribution has heavier tails. Moreover, since the variances are equal, the two curves should cross 4 times. This implies that the Laplace curve must also have higher peakedness.

Answer 3

(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 $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$ to compute the likelihood function.

(c) The MLE for the mean is $\bar{x} = 3.63$ and the MLE for the variance $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n x_i^2 - \bar{x}^2 = 3.11$. These are computed by taking the derivative of the log-likelihood

$$l(\mu, \sigma^2) := \ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{2\sigma^2}$$

and solving a pair of equations

$$\frac{-2\sum_{i=1}^n x_i + 2n\mu}{2\sigma^2} = 0,$$
$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{\sigma^3} = 0.$$

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.

Answer 4

(a) Multiple testing.
(b) Exact Fisher's test.
(c) Nonparametric tests do not assume a particular form of the population distribution like normal distribution.

Answer 5

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. So the null hypothesis can be expressed in the following form

$$H_0 : p_0 = (1-p)^3,\ p_1 = 3p(1-p)^2,\ p_2 = 3p^2(1-p),\ p_3 = p^3.$$

The MLE of p is the sample mean $\hat{p} = 0.5417$. We use the Pearson chi-square test with expected counts

$$E_0 = n(1-\hat{p})^3 = 19,\ E_1 = 3n\hat{p}(1-\hat{p})^2 = 68,\ E_2 = 3n\hat{p}^2(1-\hat{p}) = 81,\ E_3 = 3n\hat{p}^3 = 32.$$

The observed chi-square test statistic is $\mathrm{X}^2 = 11.8$ which has a P-value less than 0.5% according to the approximate null distribution χ^2_{df} with $\mathrm{df} = 4 - 1 - 1 = 2$.
Reject the null hypothesis of independent voting.

Answer 7

(b) Friedman's test for $I = 3$ treatments and $J = 2$ blocks. The test statistic

$$Q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} (\bar{R}_{i\cdot} - \tfrac{I+1}{2})^2$$

is obtained from the ranks given by two subjects (R_{ij}) to the three treatments. Under the null distribution all 36 possible rank combinations

$$(R_{ij}) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \end{pmatrix}, \ldots, \begin{pmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 2 & 3 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 1 \end{pmatrix}$$

are equally likely. The corresponding vector of rank averages $(\bar{R}_{1\cdot}, \bar{R}_{2\cdot}, \bar{R}_{3\cdot})$ takes 5 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

Next we have

$(\bar{R}_{1\cdot}, \bar{R}_{2\cdot}, \bar{R}_{3\cdot}) =$	A_1	A_2	A_3	A_4	A_5
$\sum_{i=1}^{3} (\bar{R}_{i\cdot} - 2)^2 =$	2	1.5	1.5	0.5	0
Probability =	1/6	1/6	1/6	1/3	1/6

Thus the null distribution of Q is the following one

$$\mathrm{P}(Q = 0) = 1/6, \quad \mathrm{P}(Q = 1) = 1/3, \quad \mathrm{P}(Q = 2) = 1/3, \quad \mathrm{P}(Q = 3) = 1/6.$$

Answer 8

(a) Binomial model for the number of females $Y \sim \mathrm{Bin}(36, p)$. Given $Y_{obs} = 13$ 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 $Y \sim \mathrm{N}(18, 3)$, therefore, an approximate two-sided p-value becomes

$$2 \times (1 - \Phi(\tfrac{18-13}{3}) = 2(1 - \Phi(1.67)) = 2 \times 0.048 = 9.6\%.$$

With such a high p-value we can not reject the null hypothesis of equal sex ratio.

(b) A simple sample mean

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n} = \frac{13 \times 62.8 + 23 \times 69.7}{36} = 67.2,$$

and a stratified sample mean

$$\bar{x}_s = \frac{1}{2}\bar{x}_1 + \frac{1}{2}\bar{x}_2 = \frac{62.8 + 69.7}{2} = 66.3.$$

(c) The standard error of the stratified sample mean is

$$s_{\bar{x}_s} = \frac{1}{2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{1}{2}\sqrt{\frac{(6.8)^2}{13} + \frac{(11.7)^2}{23}} = 1.54.$$

To compute the sample variance from the simple random sample take some effort. First we observe that

$$\sum (x_{1i} - \bar{x}_1)^2 = \sum x_{1i}^2 - n\bar{x}_1^2.$$

It follows,

$$\begin{aligned}\sum x_{1i}^2 &= \sum (x_{1i} - \bar{x}_1)^2 + n_1 \bar{x}_1^2 = (n_1 - 1)s_1^2 + n_1 \bar{x}_1^2 \\ &= 12 \times (6.8)^2 + 13 \times (62.8)^2 = 51825, \\ \sum x_{2j}^2 &= (n_2 - 1)s_2^2 + n_2 \bar{x}_2^2 \\ &= 22 \times (11.7)^2 + 23 \times (69.7)^2 = 114748, \\ s^2 &= \frac{\sum x_{1i}^2 + \sum x_{2j}^2}{35} - \frac{36}{35}\bar{x}^2 = 114.4.\end{aligned}$$

So that $s_{\bar{x}} = \sqrt{\frac{114.4}{36}} = 1.78$.

Answer 9

(a) Kurstosis.
(b) Power of the test.
(c) Hazard function.
(d) Outliers.

Answer 10

(a) Under the two-way ANOVA model the most interesting is $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ the null hypothesis of no difference among different types of tires.
(b) The Friedman test

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 a test statistics $Q = 14.04$. The null distribution is approximated by a chi-square distribution with df $= 3$, whose table gives a P-value less than 0.5%. Reject H_0.

Answer 11

Matched pairs design for 50 independent trials with four possible outcomes (correct, correct), (correct, wrong), (wrong, correct), (wrong, wrong). Assuming that in the slow regime the "talking computer" recognizes correctly all correct answers made in the fast regime we can summarize the results as follows

	Fast correct	Fast wrong	Totals
Slow correct	35	7	42
Slow wrong	0	8	8
Totals	35	15	50

McNemara's test statistics is $\frac{(7-0)^2}{7+0} = 7$. The null distribution is approximated by the χ_1^2-distribution. Since the square root of 7 is 2.65, the standard normal distribution gives a (two-sided) P-value 0.8%. We conclude that the observed difference is statistically significant.
The conclusion will be different if our assumption is wrong. In the worst case the slow regime correct answers are totally different and the table of the outcomes looks as

	Fast correct	Fast wrong	Totals
Slow correct	27	15	42
Slow wrong	8	0	8
Totals	35	15	50

McNemara's test statistics is then $\frac{(7-0)^2}{8+15} = 2.13$. Since the square root of 2.13 is 1.46, the standard normal distribution gives a two-sided p-value 14%. We can not reject the null hypothesis in this case.

Answer 12

We use a Beta prior with parameters (a, b) satisfying

$$\frac{a}{a+b} = \frac{1}{3}, \quad \frac{\frac{1}{3}(1-\frac{1}{3})}{a+b+1} = \frac{1}{32}.$$

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 $\hat{p}_{\text{pme}} = 0.38$.

Answer 13

(b) First we find the sample correlation coefficient by $r = b_1 \frac{s_x}{s_y} = 0.96$. The coefficient of determination is $r^2 = 0.91$. Using formula

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2) = 0.589$$

the noise size is estimated as $s = \sqrt{0.589} = 0.77$.

(c) An exact 95% CI for β_0 is $b_0 \pm t_{n-3}(0.025) s_{b_0} = 1.68 \pm 2.365 \times 1.06 = [-0.83, 4.19]$.

(d) The observed test statistic $t = 5.51$ for the model utility test for $H_0 : \beta_2 = 0$ has an exact null distribution t_7. After consulting the t_7-distribution we reject this null hypothesis at 0.5% significance level. The quadratic term is therefore highly statistically significant.

Answer 14

(a) Multiple regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, where the random variables ϵ_i, $i = 1, \ldots, 5$ are independent and have the same normal distribution $\mathrm{N}(0, \sigma)$. The corresponding design matrix has the form

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \end{pmatrix}$$

(b) Using the formula $\hat{y}_i = 111.8857 + 8.0643 x_i - 1.8393 x_i^2$ 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

and then $s^2 = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{n-p} = \frac{103.3}{2} = 51.65$.

(c) Coefficient of multiple determination

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{103.3}{2630} = 0.961.$$

Answer 15

(a) In terms of the two-way ANOVA model $Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijk}$ (grand mean + main effects + interaction+noise), we estimate the main effects as

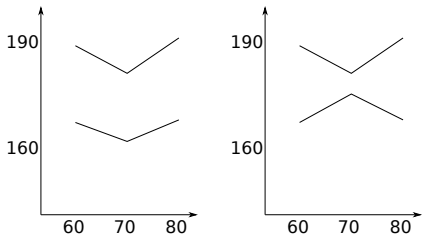
$$\hat{\alpha}_1 = 11.9, \hat{\alpha}_2 = -11.8, \qquad \hat{\beta}_1 = 1.99, \hat{\beta}_2 = -5.02, \hat{\beta}_3 = 3.04.$$

(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) Anova-2 table

Source	df	SS	MS	F	Critical values	Significance
Formulation	1	2253.44	2253.44	376.2	$F_{1,12} = 4.75$	Highly significant
Speed	2	230.81	115.41	19.3	$F_{2,12} = 3.89$	Highly significant
Interaction	2	18.58	9.29	1.55	$F_{2,12} = 3.89$	Not significant
Error	12	71.87	5.99			
Total	17					

(d) To check the normality assumption for the noise with the same variance across different values of the explanatory variable.

Answer 16

(a) This is a single sample of size $n = 441$. Each of n observations falls in of 9 groups. The multinomial distribution model

$$(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33}) \sim \text{Mn}(n, p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33})$$

gives the likelihood function

$$\begin{aligned} &L(p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}) \\ &= P(n_{11} = 24, n_{12} = 15, n_{13} = 17, n_{21} = 52, n_{22} = 73, n_{23} = 80, n_{31} = 58, n_{32} = 86, n_{33} = 36) \\ &= \frac{441!}{24!15!17!52!73!80!58!86!36!} p_{11}^{24} \cdot p_{12}^{15} \cdot p_{13}^{17} \cdot p_{21}^{52} \cdot p_{22}^{73} \cdot p_{23}^{80} \cdot p_{31}^{58} \cdot p_{32}^{86} \cdot p_{33}^{36}. \end{aligned}$$

(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-square test statistic $\text{X}^2 = 22.5$ should be compared with the critical values of χ^2_4-distribution. Even though the corresponding table is not given we may guess that the result must be significant as the square root of 22.5 is quite large. We reject the null hypothesis of independence and conclude that that there is a relationship between 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.

Answer 17

(a) Two independent dichotomous samples with $n = 56$, $\hat{p}_1 = \frac{8}{56} = 0.143$ and $m = 74$, $\hat{p}_2 = \frac{12}{74} = 0.162$. 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) For a credibility interval we can use the non-informative uniform prior $p \in \text{Beta}(1,1)$. Adding the pseudo-counts $(1,1)$ to the total counts $(8+12, 48+62)$ we get $p \in \text{Beta}(21, 111)$ as the posterior distribution. Using Matlab one can find the exact 95% credibility interval $[a, b]$ for p by finding the 2.5% and 97.5% quantiles of the posterior distribution.

Posterior mean $\mu = \frac{21}{21+111} = 0.16$ and standard deviation $\sigma = \sqrt{\frac{0.16(1-0.16)}{132}} = 0.03$ leads to the normal approximation of the posterior distribution with mean 0.16 and standard deviation 0.03. This yield an approximate 95% credibility interval

$$J_p \approx 0.16 \pm 1.96 \cdot 0.03 = [0.10, 0.22].$$

Answer 18

(a) The risk ratio compares the chances to suffer from myocardial infarction under the aspirin treatment vs the chances to suffer from myocardial infarction under the placebo treatment:

$$RR = \frac{\text{P(MyoInf|Aspirin)}}{\text{P(MyoInf|Placebo)}}.$$

(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-square test statistic is

$$\text{X}^2 = \frac{42.5^2}{146.5} + \frac{42.5^2}{146.5} + \frac{42.5^2}{10887.5} + \frac{42.5^2}{10887.5} = 25.$$

Since df=1 we can use the normal distribution table. The square root of 25 is 5 making the result highly significant. Aspirin works!

Answer 19

(a) The null distribution of Y is $\text{Bin}(n, \frac{1}{2})$ 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

$$\text{P}_{H_0}(Y > n-k) \approx 0.025.$$

With $n = 25$ we find k using the normal approximation with continuity correction:

$$0.025 \approx \text{P}_{H_0}(Y > 25-k) = \text{P}_{H_0}\Big(\frac{Y-12.5}{2.5} > \frac{13-k}{2.5}\Big) \approx \text{P}\Big(Z > \frac{13-k}{2.5}\Big).$$

Thus $\frac{13-k}{2.5} \approx 1.96$ and we get $k = 8$. The approximate 95% CI for M is given by $(X_{(8)}, X_{(18)})$.

Answer 20

(a) The null distribution of Y is approximately normally distributed with parameters (np_0, np_0q_0), where $q_0 = 1 - p_0$. At the significance level α, the rejection region for the one-sided alternative is

$$\frac{y - np_0}{\sqrt{np_0q_0}} \geq z_\alpha.$$

The power function is the probability of rejecting the null hypothesis given the alternative one is true

$$\mathrm{Pw}(p_1) = \mathrm{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} \geq z_\alpha \,|\, p = p_1).$$

(b) To compute the required sample size observe first that

$$\beta = \mathrm{P}(\tfrac{Y-np_0}{\sqrt{np_0q_0}} < z_\alpha \,|\, p = p_1) = \mathrm{P}(\tfrac{Y-np_1}{\sqrt{np_1q_1}} < \tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}} \,|\, p = p_1).$$

Now, since under the alternative hypothesis Y is approximately normally distributed with parameters (np_1, np_1q_1), we get

$$\beta \approx \Phi(\tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}}).$$

Combining this with

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

we arrive at the equation

$$\tfrac{z_\alpha\sqrt{p_0q_0}+\sqrt{n}(p_0-p_1)}{\sqrt{p_1q_1}} = -z_\beta,$$

which brings the desired formula for the optimal sample size

$$\sqrt{n} = \frac{z_\alpha\sqrt{p_0q_0} + z_\beta\sqrt{p_1q_1}}{|p_1 - p_0|}.$$

(c) If the alternatives are very close to each other, the denominator goes to zero and the sample size becomes very large. This is very intuitive as it becomes more difficult to distinguish between two close parameter values. If we decrease the levels α and β, the values z_α and z_β from the normal distribution table become larger and the sample size will be larger as well. Clearly, if you want have more control over both types of errors, you have to pay by collecting more data.

Answer 21

(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.

(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 $\frac{0.2730}{0.1157} = 2.36 \in (2.110, 2.567)$, and according to the t_{17}-distribution table says that the two-sided p-value is between 2% and 5%.

Answer 22

(a) For a given action a, the posterior risk function

$$R(a|x) = \sum_\theta l(\theta, a)h(\theta|x) = \mathrm{E}(l(\Theta, a)|x).$$

is the expected loss when the unknown parameter θ is treated as a random variable Θ with the posterior distribution:

$$\mathrm{P}(\Theta = \theta|x) = h(\theta|x).$$

(b) For the 0-1 loss function in the discrete distribution case,

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

is the probability of misclassification, that is the posterior probability that the chosen action a is different from the true value of the parameter θ.

(c) The corresponding Bayesian estimator minimizing the risk function $R(a|x) = 1 - h(a|x)$ maximizes $h(a|x)$, the posterior probability. It is denoted $\hat{\theta}_{\text{MAP}}$ and called the maximum a posteriori probability estimate. In the case the prior distribution is non-informative, so that the posterior distribution is proportional to the likelihood function, we have $\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}}$.

Answer 23

(a) The numbers of chewing gums for different tiles are summarized in the form of observed counts

Number of gums per tile	0	1	2	3	4	≥ 5
Counts	11	8	2	0	1	0

with the total number of tiles $n = 22$. The Poisson model assumes that the number of gums $X_1, \ldots, X_n$ are independent random variable with the common one-parameter distribution

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots$$

(b) The likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{e^{-n\lambda} \lambda^{n\bar{x}}}{x_1! \cdots x_n!} = \frac{e^{-22\lambda} \lambda^{16}}{2!2!4!}.$$

The log likelihood

$$l(\lambda) = \text{const} - 22\lambda + 16 \log \lambda.$$

The equation $l'(\lambda) = 0$ gives

$$0 = -22 + \frac{16}{\lambda}.$$

The MLE becomes $\hat{\lambda} = \frac{16}{22} = 0.73$, which is $\bar{x}$.

(c) Use the chi-square test of goodness of fit. Combine the cells for 2 and more gums. Compute the expected counts by

$$E_0 = n \cdot e^{-\hat{\lambda}}, \quad E_1 = n \cdot \hat{\lambda} e^{-\hat{\lambda}}, \quad E_2 = n - E_0 - E_1.$$

Then find the test statistic $\mathrm{X}^2 = \sum \frac{(O_k - E_k)^2}{E_k}$ and use the chi-square distribution with $\mathrm{df} = 3 - 1 - 1 = 1$ table to see if the result is significant. For example, if $\sqrt{\mathrm{X}^2} > 1.96$, we reject the Poisson model hypothesis at $\alpha = 5\%$.

(d) Using the Poisson model we estimate p_0 by $\hat{p}_0 = e^{-\hat{\lambda}} = 0.48$, which is close to the sample proportion $\frac{11}{22} = 0.50$.

Answer 24

(a) When the population under investigation has a clear structure it is more effective to use stratified sampling for estimating the overall population mean. 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

$$b_2 = \frac{1}{n s^4} \sum_{i=1}^{n} (x_i - \bar{x})^4,$$

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 $b_2 \approx 3$, 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. Thus a random sample of size 200 may give a decent estimate of the population mean height.

Answer 25

(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

$$F_{2,24} = 3.40, \qquad F_{3,24} = 3.01, \qquad F_{6,24} = 2.51,$$

we reject both null hypotheses on the main factors and do not reject the null hypothesis on interaction.

(c) $s = \sqrt{1.59} = 1.26$.

Answer 26

(a) This is an example of a paired sample, therefore the signed-rank test is appropriate for testing the null hypothesis of no difference.

(b) We use the signed-rank test. The observed test statistics are $W_+ = 39$ and $W_- = 6$.

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

According to Figure 2, the two-sided p-value is larger than 5% because the smaller test statistic $w_- = 6$ is larger than the critical value 5 for $n = 9$. 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.

Answer 27

(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 $x = 77.5$.

Answer 28

(a) The exponential prior with parameter 1 is Gamma$(1, 1)$. Applying the updating rule four times:

$$(1, 1) \to (3, 2) \to (3, 3) \to (5, 4) \to (10, 5),$$

we find the posterior distribution to be Gamma$(10, 5)$. Therefore, $\hat{\theta}_{\text{PME}} = 10/5 = 2$.

(b) The general updating rule for an arbitrary sample $(x_1, \ldots, x_n)$ becomes

- the shape parameter $\alpha' = \alpha + n\bar{x}$,
- the inverse scale parameter $\lambda' = \lambda + n$.

We have $\hat{\theta}_{\text{PME}} = \frac{\alpha+n\bar{x}}{\lambda+n}$. 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 means that the role of the prior is less important with large sample sizes.

Answer 29

(a) We have two independent samples from two distributions: one with parameter p, and the other with parameter q. Using $\text{Bin}(29, p)$ and $\text{Bin}(10, q)$ we compute the likelihood function as

$$L(p, q) = \binom{29}{1} p(1-p)^{28} \binom{10}{4} q^4 (1-q)^6.$$

(b) We test $H_0 : p = q$ against $H_1 : p \neq q$ using the exact Fisher test.

	ECMO	CMT	Total
Died	1	4	5
Alive	28	6	34
Total	29	10	39

The count $y = 1$ is our observed test statistics whose null distribution is $\text{Hg}(39, 29, \frac{5}{39})$. The one-sided p-value is

$$\begin{aligned} \text{P}(Y = 0) + \text{P}(Y = 1) &= \frac{\binom{5}{0}\binom{34}{29}}{\binom{39}{29}} + \frac{\binom{5}{1}\binom{34}{28}}{\binom{39}{29}} = \frac{34!29!10!}{5!29!39!} + \frac{5 \cdot 34!29!10!}{6!28!39!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{39 \cdot 38 \cdot 37 \cdot 36 \cdot 35} + \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 5 \cdot 29}{39 \cdot 38 \cdot 37 \cdot 36 \cdot 35} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot (6 + 5 \cdot 29)}{39 \cdot 38 \cdot 37 \cdot 36 \cdot 35} = 0.011. \end{aligned}$$

The two-sided p-value becomes 2% and we can reject the null hypothesis.

Answer 30

(a) The null distribution of $|\bar{X}|$ is standard normal. The p-value of the test is the probability under the null distribution that $|\bar{X}| > t_{\text{obs}}$. Thus

$$V = \text{P}(|\bar{X}| > t_{\text{obs}} | H_0) = 2(1 - \Phi(t_{\text{obs}})).$$

(b) Different samples will give different observed values $t_{\text{obs}} = |\bar{x}_{\text{obs}}|$, in this sense the p-value

$$V = 2(1 - \Phi(T_{\text{obs}}))$$

is a random variable. We have

$$\text{P}(V \leq 0.05) = \text{P}(1 - \Phi(|\bar{X}_{\text{obs}}|) \leq 0.025) = \text{P}(\Phi(|\bar{X}|) \geq 0.975) = \text{P}(|\bar{X}| > 1.96).$$

(c) If the true value of the population mean is $\mu = 4$, then $\bar{X}$ has distribution $\text{N}(4, 1)$. Using (b) we find

$$\text{P}(V \leq 0.05) \approx \text{P}(\bar{X} > 2) = 1 - \Phi(2 - 4) = \Phi(2) \approx 0.975.$$

(d) We see from (c) that even with such a big separation between the null-hypothesis and the true parameter values, there is a probability of 2.5% that the p-value will exceed 5%. One has to be aware of this variability while interpreting the p-value produced by your statistical analysis.

Answer 31

(a) Stratified sample mean $\bar{x}_{\mathrm{s}} = 0.3 \cdot 6.3 + 0.2 \cdot 5.6 + 0.5 \cdot 6.0 = 6.01$.

(b) We are in the one-way Anova setting with $I = 3$ and $J = 13$. The 95% Bonferroni simultaneous confidence intervals for the three differences $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$ are computed as

$$B_{\mu_u - \mu_v} = \bar{x}_u - \bar{x}_v \pm t_{36}(0.05/6) s_p \sqrt{2/13},$$

with the pooled sample variance given by

$$s_p^2 = \frac{12 \cdot s_1^2 + 12 \cdot s_2^2 + 12 \cdot s_3^2}{36} = \frac{2.14^2 + 2.47^2 + 3.27^2}{3} = 2.67^2.$$

This yields

$$B_{\mu_u - \mu_v} = \bar{x}_u - \bar{x}_v \pm 2.5 \cdot 2.67 \cdot 0.39 = \bar{x}_u - \bar{x}_v \pm 2.62.$$

(c) We would not reject the null hypothesis of equality $\mu_1 = \mu_2 = \mu_3$, since for all three pairwise differences the confidence intervals contain zero:

$$0.7 \pm 2.62, \quad 0.3 \pm 2.62, \quad 0.4 \pm 2.62.$$

Answer 32

(a) This is another example of the Simpson paradox. The confounding factor here is the difficulty to enter the 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!

Answer 33

Let $X \sim \mathrm{Bin}(n, p)$. Two observed counts $(x, n - x)$ are used to test $H_0 : p = p_0$. The corresponding chi-square test statistic

$$\sum_{j=1}^{2} \frac{(O_j - E_j)^2}{E_j} = \frac{(x - np_0)^2}{np_0} + \frac{(n - x - n(1-p_0))^2}{n(1-p_0)} = \frac{(x - np_0)^2}{np_0(1-p_0)} = z^2,$$

where

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

is the test statistic for the large sample test for a proportion.