Slides 1: Random sampling

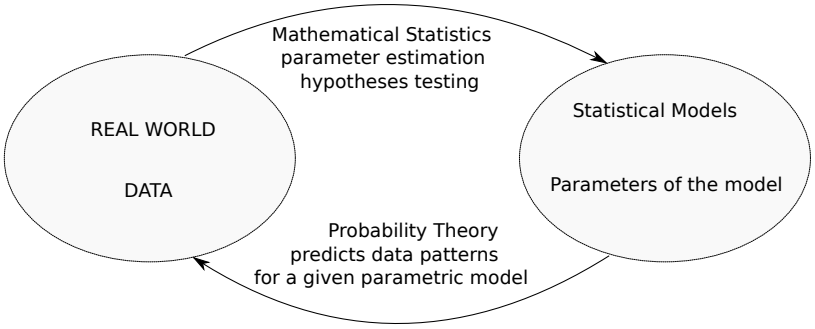
- Probability theory and Mathematical statistics
- Randomisation
- Sample mean and sample variance
- Approximate confidence intervals
- Finite population correction factor

Course compendium can be viewed from here:

http://www.math.chalmers.se/∼serik/Statistical Inference.pdf

Probability theory and Mathematical statistics

Statistical analysis : collection and summarising data, making inferences.



The main focus of this course is on the issues of parameter estimation and hypothesis testing using properly collected, relatively small data sets.

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.

Question. How data should be sampled, so that probability theory can be applied to analyse the data?

Population distribution

What is the population distribution of heights? If you pick an individual at random, the individual height can be $x_1 = 182$ cm. Another random measurement might give $x_2 = 173$ cm.

It is convenient to think of a population distribution in terms of a random variable X generated by taking a measurement of a randomly chosen individual. The population mean and standard deviation

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

give a center of the distribution and its spread around the mean.

Recall that for a random variable X with mean μ, the difference $(X - \mu)$ is called deviation and the variance of X is defined

$$\sigma^2 = \mathrm{E}[(X - \mu)^2]$$

as the mean of the squared deviation. Show that $\sigma^2 = \mathrm{E}(X^2) - \mu^2$.

Question. A company producing clothes is interested in the population distribution of heights. What it would like to know beyond μ and σ?

A random sample of n observations $(x_1, \ldots, x_n)$ is a single realisation of a random vector $(X_1, \ldots, X_n)$, whose components are independent random variables, with each X_i having the same population distribution.

Randomisation protects against investigator's biases even unconscious

The population distribution is unknown and we would like to estimate the unknown population parameters (μ, σ) using the random sample $(x_1, \ldots, x_n)$ at hand.

Example: 24 heights. A random sample of $n = 24$ male students was drawn from a certain population. Their heights in cm $(x_1, \ldots, x_n)$ are given below in an ascending order:

$$170, 175, 176, 176, 177, 178, 178, 179, 179,$$

$$180, 180, 180, 180, 180, 181, 181, 182, 183, 184, 186, 187, 192, 192, 199.$$

Question. What is your best guess on the population mean μ for the heights?

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

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

The summary statistics $\bar{x}$ and s^2 are realisations of the random variables

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

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

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

Question. Why defining s^2 do we divide by $n-1$ and not by n?

Unbiased estimates have no systematic error: if you produce k different samples and compute their sample means and variances, then

$$\frac{\bar{x}_1 + \ldots + \bar{x}_k}{k} \to \mu, \quad \frac{s_1^2 + \ldots + s_k^2}{k} \to \sigma^2, \quad k \to \infty$$

The random errors are captured by the variances

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

Since the variances tend to zero as $n \to \infty$, both $\bar{x}$ and s^2 are consistent estimators of μ and σ^2 in that

$$\bar{x} \to \mu, \quad s \to \sigma^2, \quad n \to \infty$$

The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$. The standard error of $\bar{x}$ is obtained after replacing the unknown σ by its estimate s

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Question. Is s a consistent estimate of σ? Is it an unbiased estimate?

The normal distribution $\mathrm{N}(\mu, \sigma)$ is uniquely determined by its mean μ and standard deviation σ. If $Z \sim \mathrm{N}(0, 1)$, then

$$\mathrm{P}(|Z| < 1.645) = 0.90, \quad \mathrm{P}(|Z| < 1.96) = 0.95, \quad \mathrm{P}(|Z| < 2.58) = 0.99.$$

Central Limit Theorem (CLT). *A sum of n independent random variables is asymptotically normally distributed as $n \to \infty$, provided that none of the summands dominates the others.*

By CLT, the distribution of the sample mean is approximately normal

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

This yields the following formulas of approximate 90%, 95%, and 99% confidence interval for μ:

$$I_\mu^{90} \approx \bar{x} \pm 1.645 \cdot s_{\bar{x}}, \quad I_\mu^{95} \approx \bar{x} \pm 1.96 \cdot s_{\bar{x}}, \quad I_\mu^{99} \approx \bar{x} \pm 2.58 \cdot s_{\bar{x}}$$

- the larger is n the narrower is I_μ, since $s_{\bar{x}} = \frac{s}{\sqrt{n}}$,
- the higher is confidence level the wider is the confidence interval.

Example: 24 heights

The population mean height μ is estimated by the sample mean

$$\bar{x} = 181.46 \text{ cm}$$

Furhtermore, we compute

$$\overline{x^2} = 32964.2, \quad \overline{x^2} - \bar{x}^2 = 37.08, \quad s^2 = \frac{24}{23} \cdot 37.08 = 38.69$$

and get

$$s = 6.22, \quad s_{\bar{x}} = \frac{s}{\sqrt{n}} = 1.27.$$

This brings the following 95% confidence interval

$$I_\mu = 181.46 \pm 1.96 \cdot 1.27 = 181.46 \pm 2.49 \approx (179, 184)$$

The exact meaning of the confidence level 95% is tricky, because $\{179 < \mu < 184\}$ is not a random event. Randomness in the sampling procedure and $(179, 184)$ is just one possible outcome of a random trial.

Question. If we collect 10 samples and then compute 10 confidence intervals I_μ, how many of them will cover the true value of μ?

A population of size N can be viewed as an urn containing N balls labeled by values $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$. Population mean and population variance in this case are computed as

$$\mu = \frac{x^{(1)} + \ldots + x^{(N)}}{N}$$

$$\sigma^2 = \frac{(x^{(1)} - \mu)^2 + \ldots + (x^{(N)} - \mu)^2}{N}$$

There are two basic ways of random sampling:

drawing with replacement,

drawing without replacement.

Drawing n balls with replacement produces what we call random sample. Drawing n balls without replacement results in (negatively) dependent observations $(x_1, \ldots, x_n)$, which will be called a simple random sample.

Question. If $\frac{n}{N}$ is small, then the two approaches are almost indistinguishable. Why?

Finite population correction factor

For sampling wout replacment, $\bar{x}$ is still an unbiased and consistent estimate of μ, despite the dependence between observations

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

where N is the finite population size. However, the sample variance s^2 is a biased estimate of σ^2 because

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

An unbiased estimate of σ^2 given by $s^2 \cdot \frac{N-1}{N}$, leads to an unbiased estimate of $\mathrm{Var}(\bar{X})$

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

Thus, for the sampling without replacement, the formula for the estimated standard error takes a new form

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

The extra term $\sqrt{1 - \frac{n}{N}}$ is called the finite population correction factor.