Faculty of Life Sciences

Chapter 5 overview

Statistical models, estimation and confidence intervals

Ib Skovgaard & Claus Ekstrøm
E-mail: ims@life.ku.dk

Program

Concepts and methods

- statistical model
- parameters
- estimates
- standard errors of the estimates
- confidence intervals
  - degrees of freedom

Models

- a single sample
- linear regression
- two samples: paired and unpaired
- one-way ANOVA

Summary-matrix

Name	One sample	Lin. reg.	1-way ANOVA
Model	$y_i = \mu + e_i$	$y_i = \alpha + \beta x_i + e_i$	$y_i = \alpha_g + e_i$
Parameters	μ, σ	α, β, σ	α_1, α_2, ..., σ
Estimates	$\bar{y}$, s	$\hat{\alpha}$, $\hat{\beta}$, s (p. 109)	$\bar{y}_g$, s
SE(est.)	$\text{SE}(\hat{\mu}) = \sigma/\sqrt{n}$	$\text{SE}(\hat{\alpha})$, $\text{SE}(\hat{\beta})$	$\text{SE}(\hat{\alpha}_1)$, $\text{SE}(\hat{\alpha}_2)$, ...
95% CI	$\hat{\mu} \pm t_{...} s/\sqrt{n}$	p. 109	separate

The estimate of σ is always the residual s defined as

$$s = \sqrt{\text{SS}_e/\text{df}_e}$$

where SS_e is the sum of squared residuals, and df_e is the corresponding degrees of freedom.

Summary: a single sample

- Statistical model: $y_1, \ldots, y_{162}$ independent and $y_i \sim N(\mu, \sigma^2)$
- Parameters, μ and σ: mean and standard deviation in the population.
- Estimates: $\hat{\mu} = \bar{y}$ and $\hat{\sigma} = s$
- Distribution of the estimate: $\hat{\mu}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n}$
- Standard error is an estimate of the standard deviation of an estimate: $\text{SE}(\hat{\mu}) = s/\sqrt{n}$
- 95%-confidence interval: $\bar{y} \pm t_{n-1,0.975} \cdot \frac{s}{\sqrt{n}} = \hat{\mu} \pm t_{n-1,0.975} \cdot \text{SE}(\hat{\mu})$

Linear regression

Statistical model: the deviations from the straight line are normally distributed and independent

$$y_i = \alpha + \beta \cdot x_i + e_i, \quad e_1, \ldots, e_n \sim N(0, \sigma^2) \text{ independent}$$

In words: The mean of y_i is $\alpha + \beta \cdot x_i$ and the remainders (or residuals) are normal and independent with the same standard deviation.

Parameters (population constants)

- Intercept α and slope β
- Standard deviation σ for the deviations from the line

Estimates and distribution of the estimates

Estimates $\hat\beta$ and $\hat\alpha$ shown earlier (Chapter 2).

Estimate of the residual standard deviation:

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat\alpha - \hat\beta \cdot x_i)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} r_i^2}$$

$\hat\beta$ and $\hat\alpha$ are normally distributed:

$$\hat\beta \sim N\left(\beta, \frac{\sigma^2}{\mathrm{SS}_x}\right), \quad \hat\alpha \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}\right)\right), \quad \mathrm{SS}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The statistical experiment is an instrument that "measures" the values α and β with a precision given by the standard errors.

Standard errors and confidence intervals

Distributions:

$$\hat\beta \sim N\left(\beta, \frac{\sigma^2}{\mathrm{SS}_x}\right), \quad \hat\alpha \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}\right)\right)$$

Standard errors — estimates of standard deviations

$$\mathrm{SE}(\hat\beta) = \frac{s}{\sqrt{\mathrm{SS}_x}}, \quad \mathrm{SE}(\hat\alpha) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}}$$

95% confidence intervals:

$$\hat\beta \pm t_{0.975,n-2} \cdot \mathrm{SE}(\hat\beta), \quad \hat\alpha \pm t_{0.975,n-2} \cdot \mathrm{SE}(\hat\alpha)$$

Note: t-distribution with $n-2$ degrees of freedom is used.

Stearic acid example

```
> model1 = lm(digest~st.acid}
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.53336    1.67518   57.63 1.24e-10 ***
st.acid     -0.93374    0.09262  -10.08 2.03e-05 ***

Residual standard error: 2.97 on 7 degrees of freedom
```

- Statistical model? Interpretation of models?
- Estimates? Confidence intervals?

Reflection: What is a statistical model?

- A statistical model describes the probability distribution of the population from which our sample is drawn.
- But how can we know that?
- We can't, but a model is just a rough picture displaying the important features.
- Some of these features are not known. This is why we measure a sample.
- Therefore a statistical model is not complete; some aspects have to be estimated from the sample.
- These aspects may be given as a number of parameters such as mean and standard deviation.
- The remaining part of the model is assumed and should be validated as well as possible.

Without a model we have no basis for probability calculations.

A typical statistical model

Many statistical models consist of two parts:

$$\begin{aligned}\text{observation} &= \text{fixed part} + \text{random part}\\ &= \text{predictable part} + \text{unpredictable part}\end{aligned}$$

Predictable means that it depends on factors we know (type of antibiotics, amount of stearic acid, age, treatment, etc.).
The random part is defined by the equation above as the remainder (or residual)

$$\text{random part} = \text{observation} - \text{fixed part}$$

The random part is often assumed to be normally distributed.