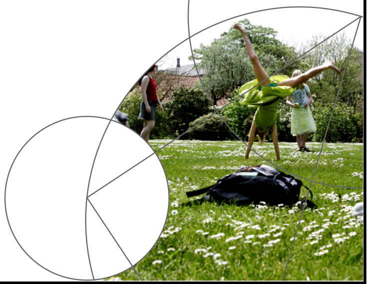
Faculty of Life Sciences

Statistical inference

Statistical models, estimation and confidence intervals

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



Program

- Distribution of a sample mean
- Statistical inference for a single sample
  - statistical model
  - estimation and precision of estimates
  - the t-distribution
  - confidence intervals
- Statistical inference for linear regression

The sample mean

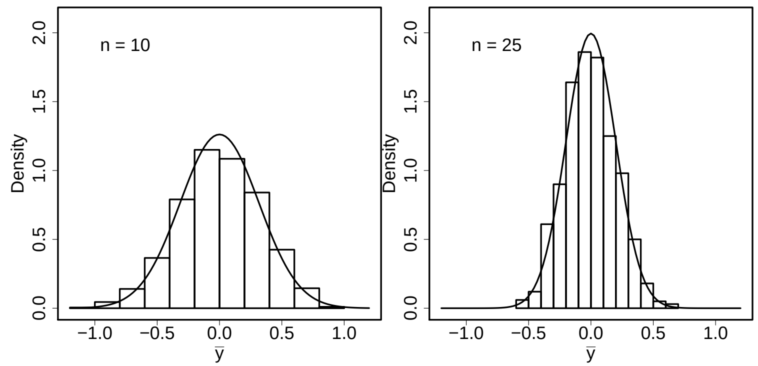
Weights of crabs:

- Wanted: the mean weight in the population — μ
- We have: a sample of $n = 162$ weights: $y_1, \ldots, y_{162}$.
- Sample statistics, $\bar{y} = 12.76$ and $s = 2.25$.
- Estimate of μ is $\hat{\mu} = \bar{y} = 12.76$
- But how precise is it?
  How large can we expect $\hat{\mu} - \mu$ to be?

To answer this we make a confidence interval for μ. This requires a statistical model.

Distribution of a sample mean

Histograms of the sample mean of n independent $N(0,1)$ variables.



Mean? — Standard deviation? — distribution?

Distribution of a sample mean

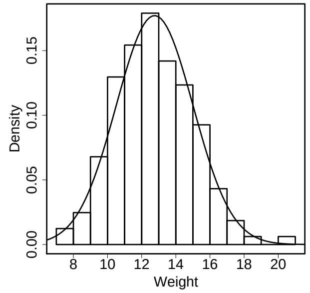
In practice we only observe one sample mean, so how can we find its distribution?

- Answer: Mathematical computation!
- Because a mean of n independent $N(\mu, \sigma^2)$-variables is normal with mean μ and standard deviation $\sigma/\sqrt{n}$

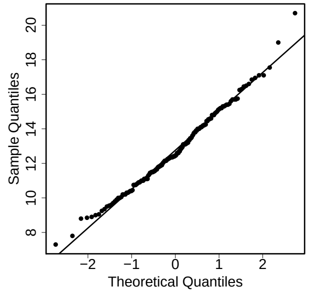
...and σ can be estimated from the sample.

Statistical model

Histogram and N-density



QQ-plot



Statistical model:
$y_1, \dots, y_{162}$ are independent and $y_i \sim N(\mu, \sigma^2)$
In words, the observations are normally distributed, have the same mean, the same standard deviation and are independent.

Estimation

Statistical model:

$$y_1, \dots, y_{162} \sim N(\mu, \sigma^2) \text{ independent}$$

Parameters in the model

- mean μ — in the population
- standard deviation σ — in the population

Estimation: The population parameters are estimated as the sample statistics:

- $\hat{\mu} = \bar{y}$
- $\hat{\sigma} = s$

Precision of $\hat{\mu}$

The estimate $\hat{\mu}$ tells nothing about the precision. But we know that

- $\mathrm{sd}(\bar{y}) = \sigma/\sqrt{n}$
- $\bar{y}$ is within $\mu \pm 1.96 \cdot \sigma/\sqrt{n}$ with 95% probability.

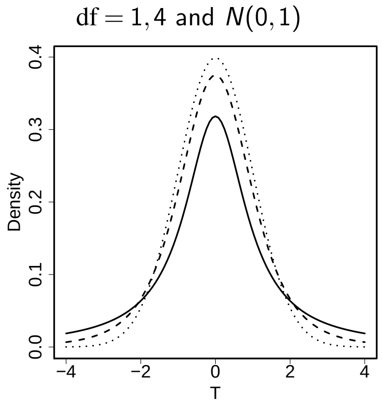
But we don't know σ, just the estimate (s).

- Standard error of $\bar{y}$ — estimated standard deviation:

$$\mathrm{SE}(\bar{y}) = s/\sqrt{n}$$

- $\bar{y}$ is within $\mu \pm ??? \cdot s/\sqrt{n}$ with probability 95%.

The t-distribution



Standardization

$$z = \frac{\sqrt{n}(\bar{y}-\mu)}{\sigma} \sim N(0,1),$$

When the estimate, s, of σ is inserted the distribution is changed from a normal distribution to a t-distribution:

$$T = \frac{\sqrt{n}(\bar{y}-\mu)}{s} \sim t_{n-1}$$

The t-distribution with $n-1$ degrees of freedom.

- Thicker tails than $N(0,1)$
- Resembles $N(0,1)$ more and more as df increases.

Confidence interval for μ

If $t_{0.975,n-1}$ is the 97.5%-quantile in the t_{n-1}-distribution:

$$P\left(-t_{n-1,0.975} < \frac{\sqrt{n}(\bar{y}-\mu)}{s} < t_{n-1,0.975}\right) = 0.95.$$

These two inequalities can be rearranged to give two inequalities for μ:

$$P\left(\bar{y} - t_{n-1,0.975}\cdot\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{n-1,0.975}\cdot\frac{s}{\sqrt{n}}\ \right) = 0.95$$

This interval contains the population mean, μ, with probability 95%.

The interval is called a 95% confidence interval for μ.

Confidence intervals: weights of crabs

Recall: $n = 162$, $\bar{y} = 12.75$ and $s = 2.25$.

Quantiles:

```
> qt(0.975,161)
[1] 1.974808
> qt(0.95,161)
[1] 1.654373
```

Compute

- Standard error, $\mathrm{SE}(\hat{\mu})$?
- 95% confidence interval?
- 90% confidence interval?

Confidence intervals: interpretation

95%-confidence interval for μ

$$\bar{y} \pm t_{n-1,0.975}\cdot\frac{s}{\sqrt{n}} = \hat{\mu} \pm t_{n-1,0.975}\cdot\mathrm{SE}(\hat{\mu})$$

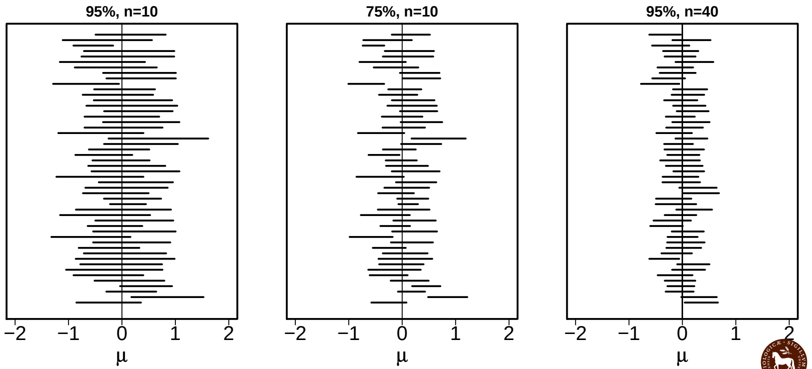
Interpretation: with probability 95%, the interval contains the population mean, μ.

What happens when the sample size, n, increases? Does the 95% confidence interval become wider or narrower?

Confidence intervals: interpretation

If we repeated the experiment, then in the long run 95% of the confidence intervals would contain the population mean.

Confidence intervals for 50 data sets from $N(0,1)$.



The central limit theorem

The main reason that the normal distribution is so important.

The central limit theorem

Assume that $Y_1, \ldots, Y_n$ are independent random variables with the same distribution with mean μ and standard deviation σ. Then their mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \sigma^2/n),$$

has a distribution which approaches the normal distribution as n increases. More precisely,

$$P\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi(z)$$

Hence, the confidence interval for the mean may be OK, even if the population is not normal.

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: $\mathrm{SE}(\hat{\mu}) = s/\sqrt{n}$
- 95%-confidence interval:
  $\bar{y} \pm t_{n-1,0.975} \cdot \frac{s}{\sqrt{n}} = \hat{\mu} \pm t_{n-1,0.975} \cdot \mathrm{SE}(\hat{\mu})$

Statistical model and parameters

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

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

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.

Main points from this lecture

- Statistical model and parameters
- Estimates, distribution of estimates, standard error
- Confidence intervals: $\text{estimate} \pm t\text{-fraktil} \cdot \text{SE}(\text{estimate})$ and interpretation