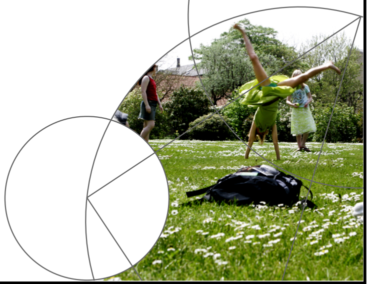
Faculty of Life Sciences

Hypotheses, tests and *p*-values

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



Program

- Example: effect of feed on the hormone concentration
  - statistical model and confidence interval
  - test of hypotheses
  - relation between confidence intervals and tests
- Test theory: concepts and review
- Linear regression (stearic acid and digestibility): test of hypotheses

Hormone concentration: data

The effect of a certain diet on the concentration of a hormone:

- Nine cows have had the diet in a certain period
- The concentration of the hormone was measured before and after

Cow	1	2	3	4	5	6	7	8	9
Initial (μg/ml)	207	196	217	210	202	201	214	223	190
Final (μg/ml)	216	199	256	234	203	214	225	255	182
Difference, y	9	3	39	24	1	13	11	32	-8

Problem: does the diet affect the concentration of the hormone?

Concentration of hormone: statistical model and confidence interval

Consider the differences (after − before), $y_1, \ldots, y_9$.

- Statistical model?
- Parameters? Estimates? Standard error?
- Confidence interval? Interpretation?
- Conclusion: does the diet affect the concentration of the hormone?

Hypothesis

If the diet has no effect, then there is no systematic difference between "before and after" — this implies that $\mu = 0$.

The (null) hypothesis is therefore

$$H_0 : \mu = 0$$

The hypotheses is an extra restriction in the statistical model.

- Under the model: $y_i \sim N(\mu, \sigma^2)$, independent
- If H_0 is true: $y_i \sim N(0, \sigma^2)$, independent

The idea behind a statistical test

Hypothesis $H_0 : \mu = 0$

We have the estimate — "best guess" — $\hat{\mu} = \bar{y}$.

- If $\hat{\mu} = \bar{y}$ is far from zero, it indicates that the hypothesis, H_0, is not correct.
- If $\hat{\mu} = \bar{y}$ is close to zero, it supports the hypothesis.

But what is "far from" and what is "close to"?

- The value $\hat{\mu} = 13.78$ is not sufficient! We need something to compare it to.
- Need to consider the variation in the data!
- Is the mean difference in the sample a real effect or is it due to chance? Imagine you repeated the experiment. Would the difference be reproducible?

The idea behind a statistical test

Far from / close to is measured by the so-called *p*-value resulting from the following reasoning:

- Data are in disagreement with the hypothesis, H_0, if what we have observed would be unlikely if H_0 were true.
- Data are in agreement with the hypothesis if what we have observed would be quite likely if the hypothesis were true.

Thus we need to calculate the following (the *p*-value):

If H_0 really is true ($\mu = 0$) — how likely is it to observe a $\hat{\mu}$ as distant from zero as we actually observed (13.78)?

The *t*-test statistic

Statistical model: $y_i \sim N(\mu, \sigma^2)$.

If the hypothesis $H_0 : \mu = 0$ is true:

- $\hat{\mu} = \bar{y}$ is normally distributed with mean 0 and standard deviation $\sigma / \sqrt{n}$.
- Standardize and replace σ by s:

$$T = \frac{\bar{y} - 0}{\mathrm{SE}(\bar{y})} = \frac{\bar{y} - 0}{s / \sqrt{n}} \sim t_{n-1}$$

We have $\bar{y} = 13.78$ and $s = 15.25$. Hence,

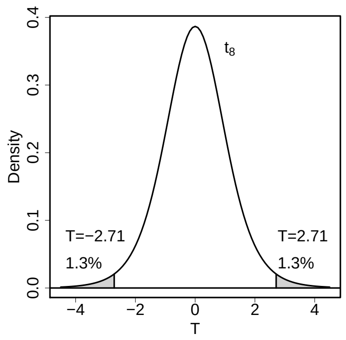
$$T_{\text{obs}} = \frac{13.78 - 0}{15.25/\sqrt{9}} = \frac{13.78}{5.08} = 2.71$$

The *t*-distribution tells us how usual/unusual this is!

The p-value and conclusion from the test

the p-value is the probability of T further away from zero than the observation, T_{obs}.

$$p = P\big(|T| \geq |T_{\text{obs}}|\big) = P\big(|T| \geq 2.71\big) = 2 \cdot P\big(T \geq 2.71\big) = 0.026,$$

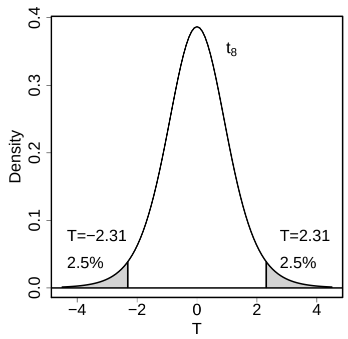



```
> pt(2.71,df=8)
[1] 0.986671
```

If H_0 is true the observations are rather unusual. This suggests that H_0 is false. The p-value is 0.026.

Level of significance

The test is significant at the 5% level of significance, if $|T_{\text{obs}}|$ is larger than the 97.5%-quantile in the t_{n-1}-distribution. This means that the p-value is less than 5%.




```
> qt(0.975, df=8)
[1] 2.306004
```


Hormone concentration: conclusion

The p-value, 0.026, is rather small. There is some evidence that hypothesis is not true, meaning that the feed has some effect on the hormone concentration.

The estimate of the increase in the hormone concentration is 13.78 with 95% confidence interval $(2.06, 25.49)$.

Confidence interval and test

The two methods, confidence interval and test led to the same conclusion:

- Zero is not in the 95%-confidence interval.
- The test is significant at the 5% level (p-value less than 5%).

This is a general relation: the 95%-confidence interval contains exactly those values of μ that are not significant at the 5% level of significance.

Hypothesis testing: concepts and summary

Hypothesis

- Hypothesis: special case of the statistical model (special values of the parameters. Here $H_0 : \mu = 0$.
- Alternative hypothesis. Usually all other cases, here $H_A : \mu \neq 0$.

Test statistic and p-value

- Test statistic: Function of data measuring the agreement between data and hypothesis.

  Here $T = \frac{\hat{\mu} - 0}{\mathrm{SE}(\hat{\mu})}$. Values near zero: good agreement; values far from zero (positive or negative): poor agreement ("critical").
- p-value: the probability of — if H_0 is true — obtaining a test statistic as far out" as the one observed.

  Here:
  $p = P\big(|T| \geq |T_{\mathrm{obs}}|\big) = P\big(|T| \geq |T_{\mathrm{obs}}|\big) = 2 \cdot P\big(T \geq |T_{\mathrm{obs}}\big).$

The scientific conclusion

The p-value summarizes the evidence against the hypothesis. Data are either

in agreement with the hypothesis (large p-value), or

in disagreement with the hypothesis (small p-value).

The experiment cannot tell with certainty what is true. In particular, a large p-value does not tell that the hypothesis is true, only that it agrees with our data.

Fundamental rule: The smaller the p-value, the stronger the evidence against the hypothesis.

Conventional thresholds

From the old days with statistical tables, three "thresholds" have become conventional:

*** $p < 0.001$. Significance at the 0.1% level. Very strong evidence against the hypothesis.

** $p < 0.01$. Significance at the 1% level. Fairly strong evidence against the hypothesis.

* $p < 0.05$. Significance at the 5% level. Some evidence against the hypothesis.

NS $p > 0.05$. Not significant. No trustworthy evidence against the hypothesis.

These thresholds are often used still but have no scientific background. The evidence against the hypothesis is almost the same if the p-value is 5.1% as when it is 4.9%.

Decisions and hypothesis testing

Sometimes a decision has to be made on the basis of a test, for example when

- authorities approve a new drug or not,
- a farmer must decide whether to spray or not.

The decision may be based on a certain level of significance, for example 5%:

- If $p < 0.05$ we reject the hypothesis.
- If $p > 0.05$ we accept the hypothesis.

Accepting/rejecting the hypothesis does not mean that we can decide whether it is true/false. It means that we take action as if it is true/falss.

Error of type I and type II

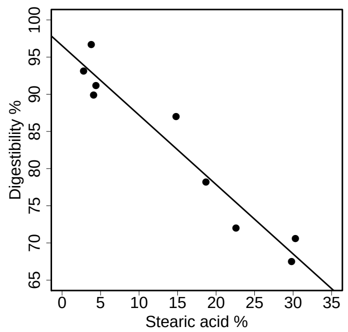
When used to make decision: accept or reject the hypothesis, there are the following four possibilities

	Accept	Reject
H_0 true	OK	type I
H_0 false	type II	OK

If the level of significance used is 5%, then the probability of a type I error is 5%.

Linear regression: stearic acid and digestibility

Statistical model: $y_i = \alpha + \beta \cdot x_i + e_i$ where $e_1, \ldots, e_9 \sim N(0, \sigma^2)$



We want to test the hypothesis that there is no relation between the amount of stearic acid and digestibility.

- What is the hypothesis in terms of the straight line?
- What is the hypothesis, expressed in terms of the parameters in the model?

Linear regression: test for no relation

What er:

- the hypothesis, the alternative hypothesis?
- the test statistic, the *p*-value?
- the conclusion?

```
> model1 <- lm(ford~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
```


Linear regression: test of another hypothesis

According to a (fictive) physiological theory the expected digestibility is 78% when the amount of stearic acid is 20%

Are the data in agreement with this theory?

- Expected digestibility at 20% stearic acid? Estimate?
- What is the hypothesis?
- Test statistic? *p*-value?
- Conclusion?

Review: t-test

- Hypothesis, $H_0 : \theta = \theta_0$ where θ is a parameter or a combination of parameters, and θ_0 is a fixed value.
  - Fx. $\mu = 0$ or $\beta = 0$ or $\alpha + \beta \cdot 20 = 78$.
- Alternative hypothesis, $H_A : \theta \neq \theta_0$
- Test statistic,
$$T = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})} \sim t_{n-p}$$
where p is the number of parameters in the model for the mean
- p-value:
$p = P\big(|T| \geq |T_{\mathrm{obs}}|\big) = P\big(|T| \geq |T_{\mathrm{obs}}|\big) = 2 \cdot P\big(T \geq |T_{\mathrm{obs}}\big)$
- 95%-confidence interval contains exactly those values μ_0 for which the hypothesis $H_0 : \theta \neq \theta_0$ is not significant at the 5% level of significance.
- Remember to quantify the results: $\hat{\theta}$ and 95%-confidence interval.

Lecture summary: main points

- Hypotheses: restriction of parameters in a model
  - Null hypothesis, alternative hypothesis
- How to test an hypothesis?
- Relation between test and confidence interval
- Interpretation of the p-value.