Hypothesis test, error types and p-values

Søren Højsgaard

Department of Mathematical Sciences, Aalborg University

(updated: 2019-04-19)

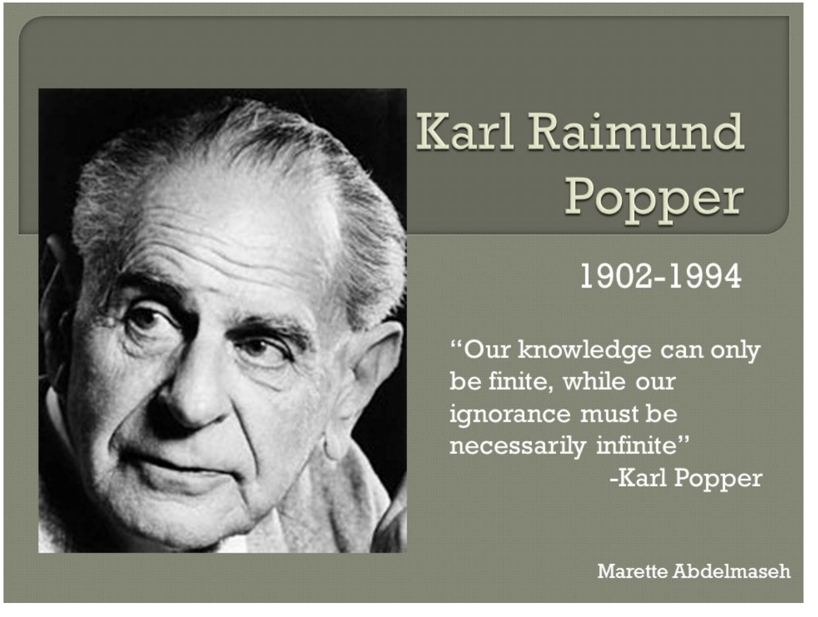
Statistical test

- A statistical test is a confrontation of the real world (data) with a theory (model).
- Conducted with the aim of falsifying the model.
- That is: You use data to prove that you are "not right"
- NB: In Danish: I statistik hedder det ET test; ikke EN test!

Karl Popper (1902-1994)

Fits well into Karl Poppers (1902-1994) theory of science:

- You can not empirically verify scientific theories, you can only falsify them.
- Scientific progress is made by "subscribing to a theory until it is falsified"
- See "Conjectures and Refutations" and "The Logic of Scientific Discovery"



The interpretation rule of statistics

In statistics we employ the following rule of interpretation:

> **Unlikely things do not happen**

- That is, if you observe data which - if the model is true - are very unlikely, then you reject the model.
- Such a rule is necessary because otherwise you can never recognize anything via statistics! The reason being that you would always be able claim that the dataset at hand simply is an *unfortunate* outcome, an outcome which is unlikely but nonetheless also possible.
- Example: To see 20 heads out of 20 tosses with a fair coin happens with probability $\approx 10^{-6}$; that is about 1 out of $1.000.000$ times. Therefore, 20 heads is *possible* but not very *probable* if the coin is fair. Therefore we would be inclined to say that the coin is not fair - that is, the model is wrong.

Is the number of newborn boys and girls the same?

Over some years these data were collected (at a London hospital) - the LARGE dataset:

	boys	**girls**	**total**
counts	6.389	6.135	12.524
proportion	0,51	0,49	1,00

Later on, we shall use a smaller dataset - the SMALL data set (10% of the large dataset):

	boys	**girls**	**total**
counts	639	614	1.253
proportion	0,51	0,49	1,00

	boys	**girls**	**total**
counts	6.389	6.135	12.524
proportion	0,51	0,49	1,00

- Is there a 50-50 chance for a boy and a girl?
- Clearly not - in this dataset
- But what about in the *population*? After all, 51\% is not far from 50\% and the deviation could well be a coincidence.
- The question is: Is the deviation so large that it can not be attributed a coincidence?

Model for data:

To make progress we need a *model* - a mechanism that could have generated data:

We shall assume the following:

- All women have the same probability θ for giving birth to a boy.
- The outcome of all pregnancies are independent (also different pregnancies for the same woman and also for different pregnancies with the same father).

Are these assumptions reasonable? Well - perhaps - and: no assumptions, no conclusions

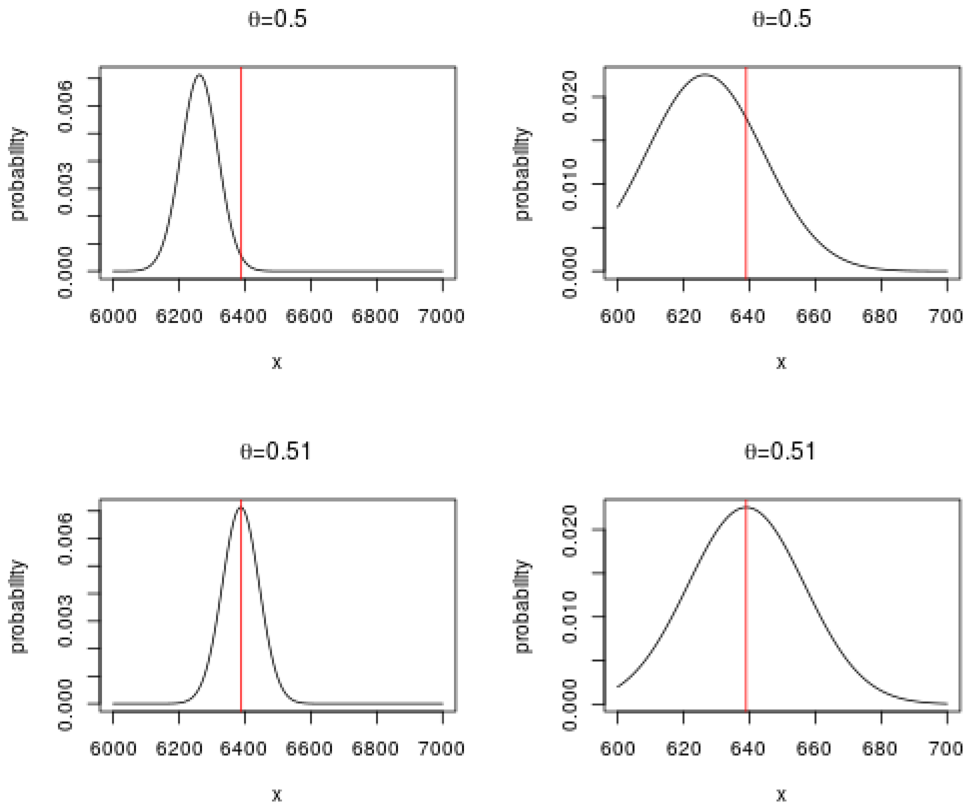
Leads to that the number of boys X is binomial distributed

$$X \sim bin(N, \theta), \quad N = 12524$$

That is, the probability of observing x boys in N pregnancies where there each time is probability θ for a boy is er

$$Pr(X = x; \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

Binomial densities



We can assume that the probability θ is 50% and then look at what data *could have looked like*:

boys	girls	total	boys	girls	total
6263	6261	12524	6252	6272	12524
6223	6301	12524	6271	6253	12524
6243	6281	12524	6298	6226	12524
6263	6261	12524	6324	6200	12524
6403	6121	12524	6215	6309	12524

Compare with observed data:

boys	girls	total
6389	6135	12524

NB: One simulated dataset have the same number or more boys than we have observed.

The SMALL dataset

boys	girls	total	boys	girls	total
627	626	1253	625	628	1253
615	638	1253	631	622	1253
623	630	1253	639	614	1253
627	626	1253	613	640	1253
670	583	1253	600	653	1253

Compare with observed data:

boys	girls	total
639	614	1253

NB: Two simulated datasets have the same number or more boys than we have observed.

Hypothesis test

- A "scientific" / subject matter questions: Is the number of newborn boys and girls the same?
- Translated to a statistical question: Is θ (probability of a boy) equal to 0.5?
- Usually formulate questions as hypotheses:
  - Testing the null-hypothesis: $H_0 : \theta = \theta_0$, where $\theta_0 = 1/2$ mod den
  - Alternative hypothesis: $H_A : \theta \neq \theta_0$.
- We can make one of two decisions: Reject or accept H_0.
- Poppers line of thinking: To reject a hypothesis is the **strong conclusion**
- NB: Perhaps "accept" should be replaced with "not reject" - but "accept" is a established terminology that can not be changed.

Classical procedure

- Let x denote data.
- Choose a function $t(x)$ with the property that $t(x)$ is (numerically) large if data does not fit to the model and small otherwise.
- We call $t(x)$ a *test statistic*
- For example, we can take

$$t(x) = |x/N - \theta_0| = |x/N - 1/2|$$

- The observed test statistic is $t_{obs} = t(x) = t(6389) = 0.0101$
- Question: Is t_{obs} a "large" or a "small" number?
- One answer: What is the probability of observing values of $t(x)$ that are larger or equal to t_{obs} if the model is true, i.e. if $\theta = \theta_0$?

The idea:

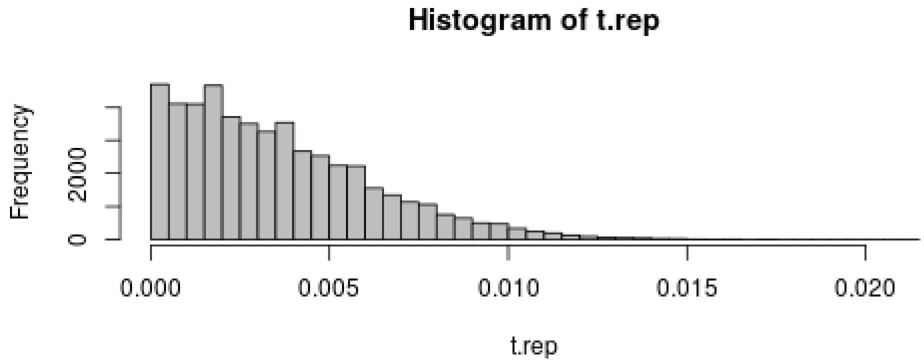
- Suppose there is some remote corner of the world where we (for some reason) know that the hypothesis is true, i.e. $\theta = \theta_0 = 1/2$.

- In this remote corner we repeat the study M times where the study consists in:

  1. Wait until $N = 12524$ babies are born and
  2. Note the number of boys x^j for $j = 1, \ldots, M$.

- Compute $t(x^j)$ for each x^j and draw a histogram of the $t(x^j)$'s.

- Good news: We do not need look after this remote corner of the world: The computer has been invented and we can do the studies by simulation ("in silica trial"):



- Our task is to make a decision: **Accept** H_0 or **reject** H_0.
- To do so, we create a decision rule: Reject H_0 if $t(x)$ is "large";
- To be specific:

> **reject** H_0 **if** $t(x) \geq c$

- The value c is called the **critical value**.
- So far we have no clue about how to choose c - but it will come soon.

- There are two types of errors we can make:
    - Reject H_0 even though H_0 is true; is called a **type-I** error
    - Accept H_0 even though H_0 is false; is called a **type-II** error.
- One often decides on beforehand that the probability of making at type-I error must be smaller than a small number α, e.g.\ $\alpha = 0.05$.

$$Pr_{\theta_0}(\text{reject } H_0) \leq \alpha$$

where $Pr_{\theta_0}()$ indicates that the probability is computed for $\theta = \theta_0$.

- If the decision rule is "Reject H_0 if $t(x) \geq c$ then we can find c from:

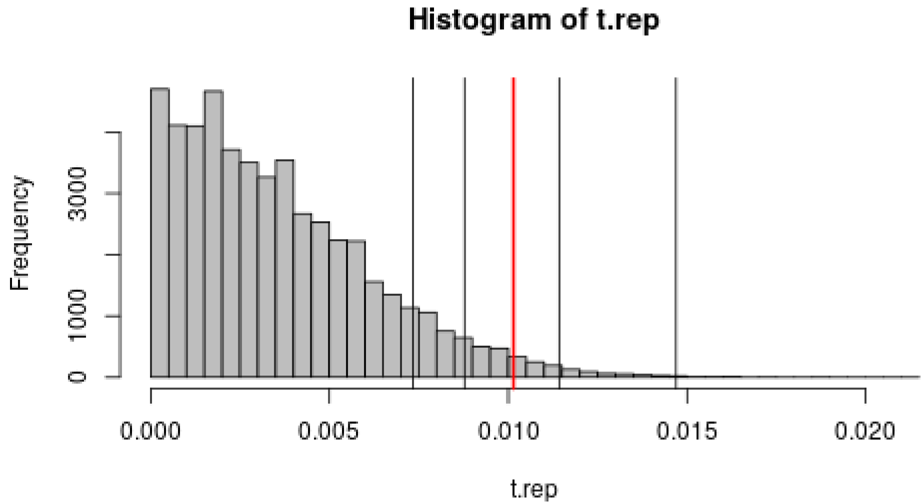
$$Pr_{\theta_0}(t(X) \geq c) \leq \alpha$$

- If $t_{obs} \geq c$ we say that **the test is significant at level** α.

- For each α we can find the critical value c_α:

```
##      0.1     0.05     0.01    0.001
## 0.007346 0.008783 0.011418 0.014692
```

Compare with t_{obs} = 0.0101



We say that the test is **signifikant at level niveau** 5% (but not significant at level 1%).

Often one use the *significance levels* 0.10, 0.05, 0.01 og 0.001 -- but there is nothing divine to the values; they have historical reasons.

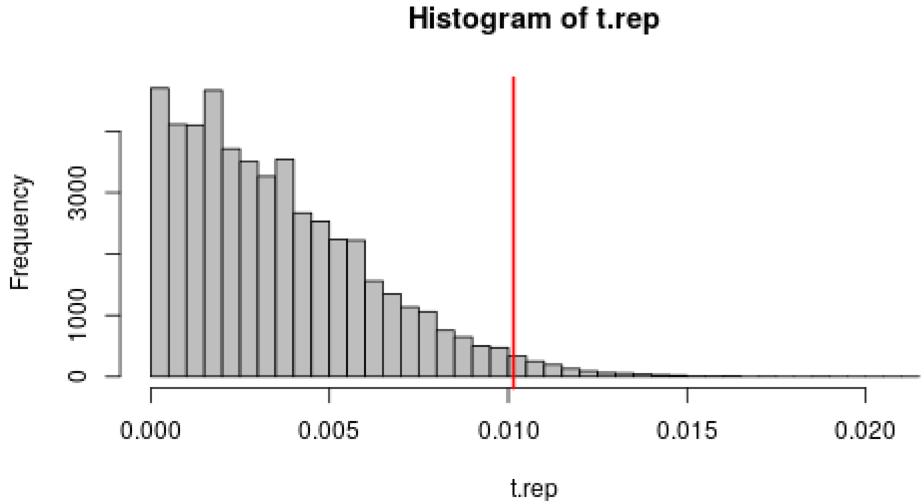
The p-value

A slightly different approach is: Compute the p-values defined as

$$p = Pr_{\theta_0}(t(X) \geq t_{obs})$$

That is; the probability of observing a test statistic $t()$ which is larger or equal to the observed value t_{obs}.

We find that the p-value is 0.0235



- Thus one can say that the p-value is a measure of "the degree of evidence against a hypothesis".

- In some settings, this approach makes much more sense than by creating a decision problem.

Interpretation of p-values

- Back to the original question is there at a 50-50 chance for boys and girls?
- A p-value can be regarded as a measure of evidence against a hypothesis: A small p-value indicates strong evidence against the hypothesis.
- Here the p-value is small so we doubt the hypothesis.
- Can we from this conclude that the null-hypothesis $H_0 : \theta = \theta_0 = 1/2$ is false? Have we "proven" that $\theta \neq 1/2$.
- No! If $\theta = 1/2$, the probability of observing 6389 boys in 12524 pregnancies is 0.00054 or about 1 in every 2000 times we have seen 12524 pregnancies . It is a small probability, sure, but the data are definitely possible even if the hypothesis is true.
- However, many studies indicate that more boys than girls are born.

- Sometimes p-values are erroneously interpreted along these lines:

> **the p--value is the probability that the hypothesis is true**.

- This is wrong. Probabilities are numbers we assign to random phenomena - phenomena where there is uncertainty about the outcome (e.g. toss a coin or a die).

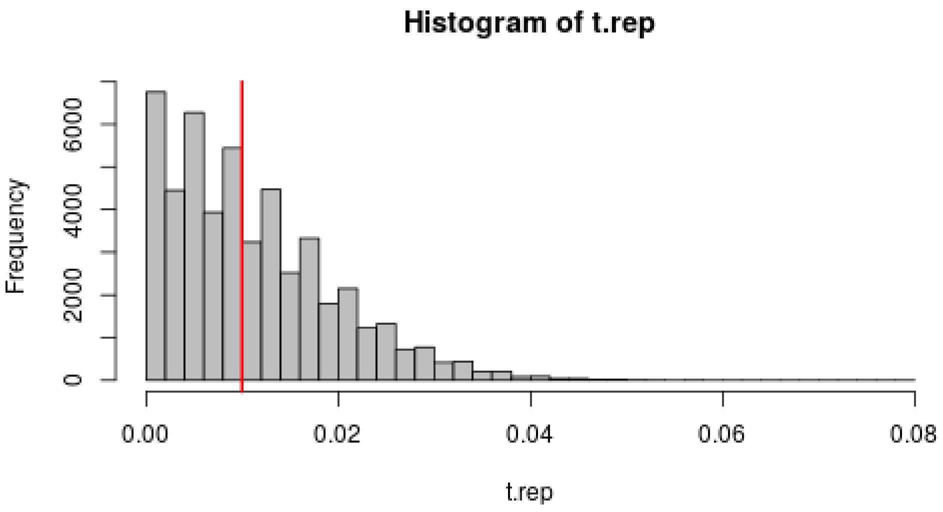
- There is not randomness related to the hypothesis: The hypothesis is either true or false (we just do not know what it is because we have no divine insight).

The effect of sample size

The SMALL dataset:

	drenge	**girls**	**total**
counts	639	614	1.253
proportion	0,51	0,49	1,00

We only have 10\% of data, but the proption of boys is still 0.51.


Histogram of t.rep
Frequency
0
2000
4000
6000
0.00
0.02
0.04
0.06
0.08
t.rep

What can we conclude?

t_{obs} is the same in the small and large dataset (apart from a few decimals): 0.01 but

- With 12524 births 6389 boys there is strong evidence against the hypothesis $\theta = \frac{1}{2}$: We get the that the p-value is 2%.
- With 1253 births and 639 boys there is very little evidence against the hypothesis We get that the p-value is 0.4975
- In both cases the proportion of boys is 0.51. What to make of this?

- We establish a hypothesis about "the true state of the world" and then we "ask data" if there is evidence **against** the hypothesis.
    - If there is no evidence against the hypothesis is data it can be because the hypothesis is true, or
    - Because there insufficient data (information) to provide this evidence (that is to "prove" that the hypothesis is wrong).

Again: Think of p-value as a measure of evidence against the hypothesis

- The p-value reflects the "distance" between data and model (between $\hat{\theta} = 0.51$ and $\theta_0 = 0.5$)
- The p-value ALSO reflects the amount of data.

More poetically:

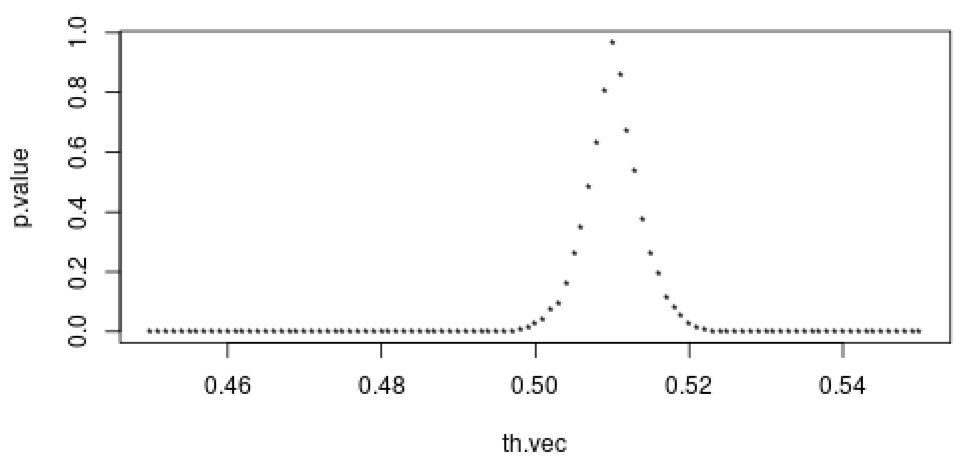
"Absence of evidence (of an effect) is NOT the same as evidence of absence (of an effect)."

Test and confidence intervals -- two sides of the same coin

Above, we tested the hypothesis $\theta = \theta_0$ where $\theta_0 = 1/2$.

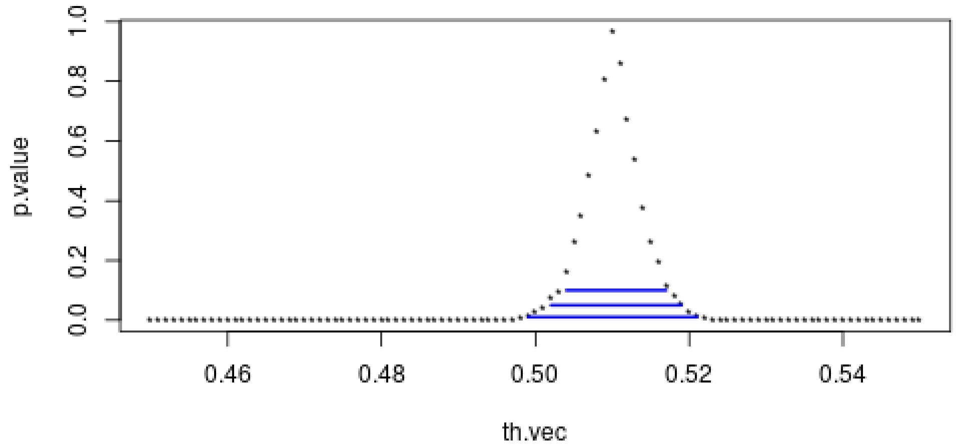
We could have tested the hypothesis for many other values of θ_0.

For each value of θ_0 will be compute the p-value and plot it against θ_0



Remember: Small p-values are evidence against the hypothesis.

Add intervals indicating where the p-value is larger than 0.01, 0.05 and 0.10.

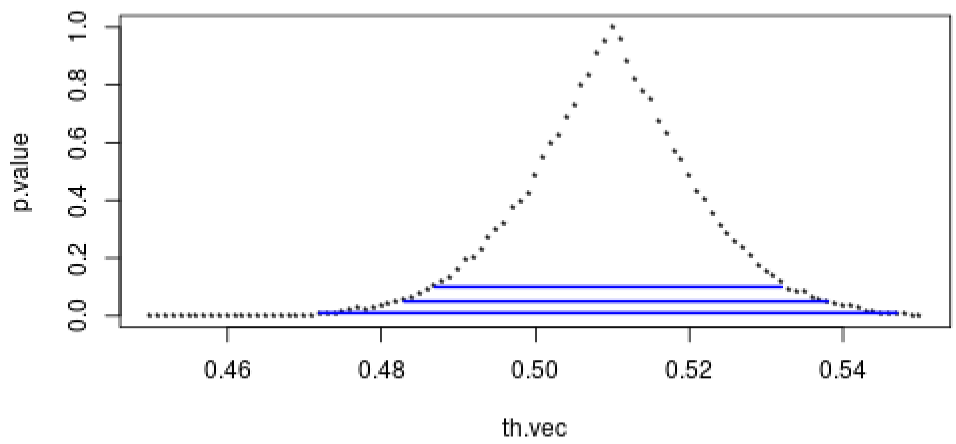


These intervals are 99%, 95% og 90% **confidence intervals**:

99% confidence interval: [0.499; 0.521]

95% confidence interval: [0.502; 0.519]

90% confidence interval: [0.504; 0.517]



These intervals are 99%, 95% og 90% **confidence intervals**:

99% confidence interval: [0.475; 0.547]

95% confidence interval: [0.483; 0.538]

90% confidence interval: [0.487; 0.532]

Statistical significance, practical significance, clinical significance...

The origin is the latin **significantia** which means **importance**

When we find a statistically significant "effect" then this means that the effect is so large that we can not reasonably attribute it to being a coincidence.

Many studies indicate that 50.5 boys and 49.5% are born.

But when you expect at child you thing that there is a 50--50 chance either gender.

Hence, the statistical significance does not necessarily mean to much in practice.

You find the same in the health science: A statistically significant effect can be so small that it is not *clinically relevant* for the patient.