t-Test / Wilcoxon Test
Confidence Interval

Peter Büchel

HSLU W

SA: W 06

t-Test

- So far: Procedure is called z-test
- Tacitly assumed: Standard deviation *known*
- In practice: Never the case
- Why is true standard deviation never known?
  - Take for example bottles with content 500 ml
  - To know true standard deviation: Measure content of *all* bottles
  - Even the ones which haven't been produced yet
- Following *t-test*: Does not assume true standard deviation

- Therefore: t-test much more important than z-test
- Procedure very similar to z test: Only different distribution
- Assumption as before : Data realisations of

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

- But: σ_X is unknown
- Possible to estimate σ_X from data:

$$\widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Additional uncertainty (unknown standard deviation): Change distribution of test statistics

t-distribution

Distribution of test statistics for t-test under null hypothesis

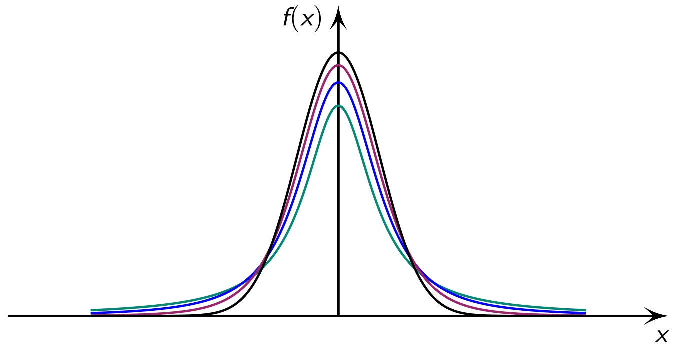
$$H_0 : \quad \mu = \mu_0$$

is given by

$$T = \overline{X}_n \sim t_{n-1}\left(\mu, \frac{\widehat{\sigma}_X^2}{n}\right)$$

where t_{n-1} is a *t-distribution* with $n-1$ degrees of freedom

- Normal distribution is replaced by a t-distribution
- But what is a t-distribution?
- Similar to normal distribution, but flatter, due to greater uncertainty
- Depends on number of observations
- Sketch for $\mu = 0$ and $\sigma \approx 1$ (depends on n):



- Green: $n = 1$, blue: $n = 2$, violet: $n = 5$, black: $\mathcal{N}(0, 1)$
- t_n-distribution symmetric distribution around 0, but flattens out slower than standard normal distribution $\mathcal{N}(0, 1)$
- For large n is t_n similar to $\mathcal{N}(0, 1)$
- t_n tends for $n \to \infty$ to standard normal distribution $\mathcal{N}(0, 1)$
- Important: For t-test use t_{n-1} (technical detail)
- t-distribution: Found by William Gosset (Chief brewer Guiness Brewery) in 1908

R

- All terms from z-test can be used for t-test
- Rejection range: `qt(\ldots)` instead of `qnorm(\ldots)`
- p-value: `pt(\ldots)` instead of `pnorm(\ldots)`
- t-test occurs very often: Whole procedure implemented in `R`
- Enter data in command `t.test(\ldots)` and `R` takes over work
- Rejection zone *not* returned
- But p-value is used for test decision

Example

- Normally distributed data $x_1, \ldots, x_{20}$:

  5.9, 3.4, 6.6, 6.3, 4.2, 2.0, 6.0, 4.8, 4.2, 2.1, 8.7, 4.4, 5.1, 2.7, 8.5, 5.8, 4.9, 5.3, 5.5, 7.9

- Assumption: $x_1, x_2, \ldots, x_{20}$ realisations of

$$X_i \sim \mathcal{N}(5, \sigma_X^2)$$

- σ_X unknown: σ_X thus from data

```
x <- c(5.9, 3.4, 6.6, 6.3, 4.2, 2.0, 6.0, 4.8, 4.2, 2.1,
       8.7, 4.4, 5.1, 2.7, 8.5, 5.8, 4.9, 5.3, 5.5, 7.9)

mean(x)
[1] 5.215

sd(x)
[1] 1.883802
```

- Null hypothesis in this case is:

$$H_0 : \quad \mu_0 = 5$$

- Test whether mean 5.215 matches assumed value μ_0 or not
- Procedure is in itself same as for known standard deviation
- But technically more complicated

- Procedure occurs quite frequently: In R implemented
- Command `t.test(...)` automatically calculates all required quantities (except rejection range):

```
t.test(x, mu = 5)

     One Sample t-test

data:  x
t = 0.51041, df = 19, p-value = 0.6156
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.333353 6.096647
sample estimates:
mean of x
    5.215
```


R Output

- `One Sample t-test`
  A one-sample test is performed (two samples later)
- `data: x`
  Data set `x` that was used
- `t = 0.51041`
  - *t*-value
  - Not interesting in itself
  - "Large" *t*-value: Null hypothesis is rejected
  - *t*-value "close" to 0: Null hypothesis is *not* rejected
  - Important: *p*-value further below
- `df = 19`
  Degree of freedom: Also uninteresting

- `p-value = 0.6156`
  - *p*-value
  - This is *the* crucial value
  - Decides whether null hypothesis is rejected or not
  - Here: Do not reject null hypothesis at significance level 5 %, because *p*-value greater than 0.05
- `alternative hypothesis: true mean is not equal to 5`
  Alternative hypothesis is *noted*
- `95 percent confidence interval: 4.33 6.09`
  Confidence interval (to be introduced soon)
- `mean of x 5.215`
  Average value of `x`

Example: Scale A

- Estimate standard deviation σ_X from data (done by R)
- Assumption: True $\mu = 80$
- t-test at 5 % level of significance
- t-test

```
scaleA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
            79.97, 80.05, 80.03, 80.02, 80.00, 80.02)

t.test(scaleA, mu = 80)

    One Sample t-test

data:  scaleA
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 80.00629 80.03525
sample estimates:
mean of x
 80.02077
```


- p-value: 0.009
- Less than significance level 0.05
- Null hypothesis H_0 is rejected
- Must assume that true mean is statistically significantly *not* 80

Example: Body Height Women

- Newspaper: Average height of adult women in Switzerland is 180 cm
- Suspect: Value too high
- Investigate at a significance level of 5 %
- Randomly select 10 women and measure their height (in cm)
- Measured heights:

  165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4, 163.4, 152.5

- Assume: Average height *less than* 180 cm
- t-test one-sided to the left (left-tailed): `alternative = "less"`:

```
height <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4,
            163.4, 152.5)

t.test(height, mu = 180, alternative = "less")

    One Sample t-test

data:  size
t = -5.4836, df = 9, p-value = 0.0001942
alternative hypothesis: true mean is less than 180
95 percent confidence interval:
    -Inf 169.382
sample estimates:
mean of x
   164.05
```

- p-value: 0.0002: Far below significance level of 0.05
- Null hypothesis

$$H_0: \quad \mu_0 = 180$$

  rejected
- Alternative hypothesis accepted

$$H_A: \quad \mu_0 < 180$$

- Statement of newspaper is therefore statistically significantly *not* true

Confidence Interval

- Point estimate μ of a measurement series: *Single* numerical value
- Don't know how close this estimated mean is to true, but unknown, mean of distribution of observations
- Confidence interval: Interval indicating where, roughly speaking, true mean lies with a certain predefined probability
- Illustration of confidence interval with an example
- See Juypter Notebook `confidence_interval_v2_en.ipynb`
- See also lecture notes

Test Decision with Confidence Interval

- If μ_0 of null hypothesis lies within confidence interval of $\overline{x}_n$, H_0 *not* is rejected
- If μ_0 of null hypothesis does *not* lie within confidence interval of $\overline{x}_n$, H_0 is rejected

- **R** output: Returns confidence interval
- This states that at a significance level of 5 %, true μ lies with probability of 95 % within this interval
- With confidence interval: Make test decision

Example: Scale A

- Null hypothesis
$$H_0: \quad \mu_0 = 80$$
- **R** output: Confidence interval:
$$[80.00629,\ 80.03525]$$
- With probability of 95 %, true μ lies in this interval
- But $\mu_0 = 80$ *not* in this interval
- With 95 % probability, real μ is *not* 80
- Null hypothesis is rejected and alternative hypothesis accepted

Example: Body Height Women

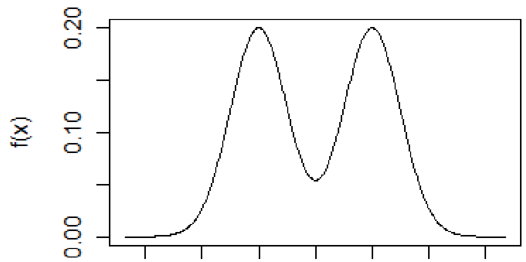
- Null hypothesis:
$$H_0: \quad \mu_0 = 180$$
- **R** output: Confidence interval:
$$(-\infty,\ 169,382]$$
- At 95 % true μ lies in this interval
- $\mu_0 = 180$ *not* in this interval
- With 95 % security is real μ *not* 80
- Reject null hypothesis and accept alternative hypothesis

Remark

- Narrow confidence interval: Know more reliably where true mean is
- Is confidence interval wide, like
$$[10,\ 1000]$$
there is great uncertainty about where real μ lies

Non-Normally Distributed Data: Wilcoxon Test

- Alternative to *t*-test
- Wilcoxon test: Assumes less than *t*-test
- Assume: Distribution under null hypothesis is *symmetrical* with respect to *median* μ_0
- Assume:
$$X_i \sim F \text{ iid}, \quad F \text{ is symmetrical}$$



- A V value (*rank sum*) is calculated: Details are tedious
- Basic idea same as for hypothesis testing so far:
  - V-value *far* away from *median*: Reject null hypothesis
  - V-value close to *median*: Do not reject null hypothesis
  - R calculates p-value

Example: Scale A

- R Output:

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
       79.97, 80.05, 80.03, 80.02, 80.00, 80.02)

wilcox.test(x, mu = 80.00, alternative = "two.sided")

    Wilcoxon signed rank test with continuity correction

data:  x
V = 69, p-value = 0.0195
alternative hypothesis: true location is not equal to 80
```

- Significance level 5 %: Null hypothesis is rejected (p-value < 0.05)
- Significance level 1 %: Null hypothesis is *not* rejected (p-value> 0.01)

Wilcoxon Test versus t-Test

Wilcoxon test versus t-test

Wilcoxon test is in the vast majority of cases preferable to the t-test: It often has much greater power in many situations (probability of correctly rejecting the null hypothesis)

Even in the most extreme cases it is never much worse

Comparing Two Samples: Possible Questions

- Comparison of two measuring methods (measuring device A vs. measuring device B): Is there a significant difference?
- Comparison of two manufacturing processes (A vs. B): Which one has better properties (e.g. regarding brittleness of mobile phone displays)?

Paired Samples

- Example measuring devices: Each test unit is measured with *both* measuring devices
- For each *test unit* two observations: A and B
- So-called *paired samples*
- Both observations are *not* independent, because same experimental unit is measured twice!

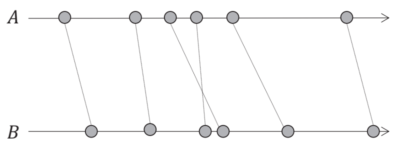
Unpaired (Independent) Samples

- Example manufacturing process: Sample of process A and another sample of process B
- Observations *independent*: There is *nothing* that “connects” them
- So-called *unpaired (or independent) samples*

Distinction Paired versus Unpaired Samples

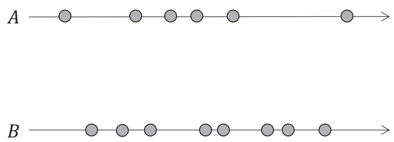
Paired Samples

- Each observation of one group can be clearly assigned to an observation of the other group
- Sample size is inevitably same in both groups



Unpaired Samples

- No assignment of observations possible
- Sample sizes can be different (but do not have to be!)
- Can enlarge one group without enlarging the other



Statistical t-Test for Paired Samples

- *Paired Samples:* Normally distributed data:

$$X_i \sim \mathcal{N}(\mu_X, \sigma_X^2) \qquad \text{and} \qquad Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

- Considering differences:

$$D_i = X_i - Y_i$$

- Performing a t-test
- Normally for null hypothesis:

$$\mathrm{E}(D) = \mu_D = 0$$

- *No* difference!
- If data not normally distributed: Wilcoxon test

- R output:

```
before <- c(25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28)
after <- c(27, 29, 37, 56, 46, 82, 57, 80, 61, 59, 43)

    Paired t-test

data:  after and before
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
  4.91431 15.63114
sample estimates:
mean difference
       10.27273
```

- Null hypothesis is rejected at a significance level of 5 %, since p-value 0.001633 is less than 0.05

- Difference is therefore on 5 % significance level significant, because the p-value is less than 5 %
- 95 % confidence interval: Mean of differences

$$[4.91431, 15.63114]$$

- With 95 % probability is mean of differences between `after` and `before` in this interval
- $\mu_0 = 0$ (no difference) is not in confidence interval: Reject null hypothesis

Statistical t-Test for Unpaired Samples

- *Unpaired samples:* Data X_i and Y_j normally distributed but unpaired
- Example: Scale A and B
- Two sample t-Test for unpaired samples with null hypothesis:

$$\mu_X = \mu_Y \qquad \text{or} \qquad \mu_X - \mu_Y = 0$$

- `R` output:

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
       80.05, 80.03, 80.02, 80.00, 80.02)
y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

t.test(x,
       y,
       alternative = "two.sided",
       mu = 0, paired = FALSE,
       conf.level = 0.95)


    Welch Two Sample t-test

data:  x and y
t = 2.8399, df = 9.3725, p-value = 0.01866
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.008490037 0.073048425
sample estimates:
mean of x mean of y
 80.02077  79.98000
```

- At significance level 5 % null hypothesis is rejected, since p-value 0.01866 is less than 0.05
- But not so at significance level 0.01!
- Difference of averages is therefore on 5 % significance level significant, because p-value is less than 5 %
- 95 % confidence interval: Difference in group mean values:

$$[0.0167, 0.0673]$$

- With 95 % probability is group mean of `x` is a number in this range greater than the group mean of `y`
- $\mu_X - \mu_Y = 0$ (no difference in mean) is not in confidence interval: Reject null hypothesis

Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)

- If data are non-normally distributed
- R output:

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,79.97, 80.05,
       80.03, 80.02, 80.00, 80.02)
y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

wilcox.test(x,
            y,
            alternative = "two.sided",
            mu = 0,
            paired = FALSE,
            conf.level = 0.95)


    Wilcoxon rank sum test with continuity correction

data:  x and y
W = 76.5, p-value = 0.01454
alternative hypothesis: true location shift is not equal to 0
```


Interpreting p-Values

- From: *Hypothesis Testing: An intuitive guide for making data driven decisions* by Jim Frost (with his consent)
- Any time you see a p-value: Looking at results of hypothesis test
- p-values: Determine whether hypothesis test results are statistically significant
- If p-value is less than significance level, reject null hypothesis and conclude that effect or relationship exists
- In other words: Sample evidence is strong enough to determine that effect exists in population
- Statistics use p-values all over the place: t-tests, distribution tests, ANOVA, and regression analysis

- They have become so crucial that they've taken on a life of their own
- Can determine which studies are published, which projects receive funding, and which university faculty members become tenured!
- Ironically, despite being so influential: p-values are misinterpreted very frequently
- What *is* correct interpretation of p-values?
- What do p-values *really* mean?
- p-values are a slippery concept

It's All About the Null Hypothesis

- p-values are directly connected to null hypothesis
- In all hypothesis tests, researchers are testing an effect or relationship of some sort
- Effect can be effectiveness of a new vaccination, durability of a new product, and so on
- There is some benefit or difference that researchers hope to identify
- However, it's possible that there actually is no effect or no difference between experimental groups
- In statistics: Lack of an effect on null hypothesis

- When assessing results of hypothesis test: Can think of null hypothesis as the devil's advocate position, or position you take for sake of argument
- To understand this idea: Imagine a hypothetical study for medication that we know is entirely useless
- In other words: Null hypothesis is true
- There is no difference in patient outcomes at population level between subjects who take medication and subjects who don't
- Despite null being accurate: Likely observe an effect in sample data due to random sampling error
- It is improbable that samples will ever exactly equal null hypothesis value

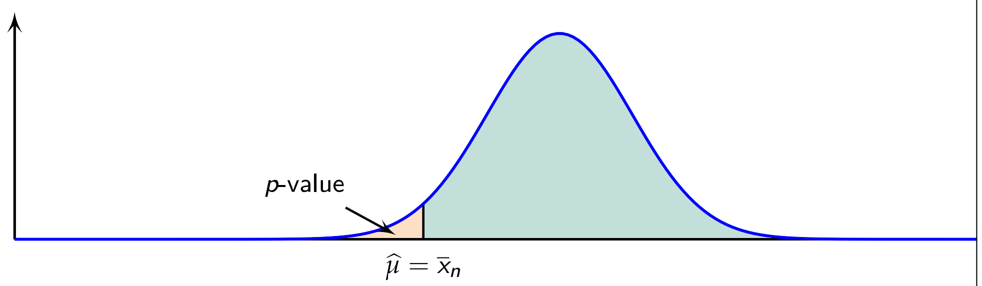
Defining p-Values

- p-value: Indicates believability of devil's advocate case that null hypothesis is correct given sample data
- Gauge how consistent sample statistics are with null hypothesis
- Specifically: If null hypothesis is right, what is probability of obtaining an effect at least as large as the one in sample?
  - High p-values: Sample results are consistent with true null hypothesis
  - Low p-values: Sample results are not consistent with true null hypothesis
- If p-value is small enough: Conclude that sample is so incompatible with null hypothesis that reject null for entire population

- p-values: Integral part of inferential statistics
- Help to use sample to draw conclusions about population
- Technical definition of p-values:

> p-values: Probability of observing a sample statistic that is at least as extreme as sample statistic when assuming that null hypothesis is correct

- One-sided, left-tailed:



- Let's go back to hypothetical medication study
- Suppose hypothesis test generates p-value of 0.03
- Interpret p-value as follows: If medicine has no effect in population, 3 % of studies will obtain effect observed in sample, or larger, because of random sample error
- *Key Point*: How probable are sample data if null hypothesis is correct?
- That's the only question that p-values answer
- This restriction transfers to a persistent and problematic misinterpretation

p-Values Are NOT an Error Rate

- Unfortunately: p-values are frequently misinterpreted
- A common mistake: Represent likelihood of rejecting a null hypothesis that is actually true (Type I error)
- Idea that p-values are probabilities of making a mistake is *WRONG*!
- You can't use p-values to calculate error rate directly for several reasons
- First: p-value calculations assume that null hypothesis is correct
- Thus: From p-value's point of view, null hypothesis is 100 % true
- Remember: p-values assume that null is true, and sampling error caused observed sample effect

- Second: p-values tell how consistent sample data are with true null hypothesis
- However: When data are very inconsistent with null hypothesis, p-values *can't* determine which of the following two possibilities is more probable:
  - Null hypothesis is true, but sample is unusual due to random sampling error
  - Null hypothesis is false
- To figure out which option is right: Apply expert knowledge of study area and, very importantly, assess results of similar studies

- Going back to medication study: Highlight correct and incorrect way to interpret p-value of 0.03:
  - Correct: Assuming medication has zero effect in population, obtain sample effect, or larger, in 3 % of studies because of random sample error
  - Incorrect: There's a 3 % chance of making a mistake by rejecting null hypothesis
- Yes: Incorrect definition seems more straightforward, and that's why it is so common
- Unfortunately: Using this definition gives a false sense of security

What Is True Error Rate?

- Difference between correct and incorrect interpretation is not just a matter of wording
- Fundamental difference in amount of evidence against null hypothesis that each definition implies
- T*p*-value for medication study: 0.03
- If interpreting that *p*-value as a 3 % chance of making a mistake by rejecting null hypothesis: Feel like on pretty safe ground
- However: *p*-values are not an error rate, and can't interpret them this way
- If *p*-value is not error rate for study, what is error rate?
- *Hint*: It's higher!

- Can't directly calculate error rate based on a *p*-value, at least not using the frequentist approach that produces *p*-values
- However: Estimate error rates associated with *p*-values by using Bayesian methodologies and simulation studies
- Sellke et al. have done this
- While exact error rate varies based on different assumptions, values below use middle-of-the-road assumptions

p-value	Probability of rejecting a true null hypothesis
0.05	At least 23 % (and typically close to 50 %)
0.01	At least 7 % (and typically close to 15 %)

- These higher error rates probably surprise you!
- Regrettably: Common misconception that *p*-values are error rate produces false impression of considerably more evidence against null hypothesis than is warranted
- A single study with a *p*-value around 0.05 does not provide substantial evidence that sample effect exists in population
- These estimated error rates emphasize need to have lower *p*-values and replicate studies that confirm initial results before one can safely conclude that an effect exists at the population level
- Additionally, studies with smaller *p*-values have higher reproducibility rates in follow-up studies

p-Values and Reproducibility of Experiments

- At this point: Wouldn't blame you for wondering whether *p*-values are useful
- They are confusing and they don't quite tell us what we most want to know
- Let's do a reality check to see if *p*-values provide any real information!
- Typically, when you perform a study, it's because you aren't sure whether the effect exists
- After all, that's why you're performing the study, right?
- Consequently, when you get your results, whether they are statistically significant or not, you don't know conclusively whether the test results correctly match the underlying reality

Estimating Reproducibility Rate

- Researchers for a study published in August 2015, *Estimating the reproducibility of psychological science*, wanted to estimate reproducibility rate and to identify predictors for successfully reproducing experimental results in psychological studies
- However, there was a shortage of replication studies available to analyze
- Sadly: Lack exists because it is easier for authors to publish new results than to replicate prior studies
- Because of this shortage: Group of 300 researchers first had to conduct their own replication studies
- They identified 100 psychology studies with statistically significant findings that were published in three top psychology journals

- Then: Research group replicated these 100 studies
- After finishing follow-up studies: Calculated reproducibility rate and looked for predictors of success
- To do this: Compared results of each replicate study to corresponding original study
- Researchers: Only 36 of 100 replicate studies were statistically significant
- That's a 36 % reproducibility rate
- This finding sent shock waves through the field of psychology!
- My view of this low reproducibility rate is that science isn't a neat, linear process

- It can be messy
- For science: Take relatively small samples and attempt to model complexities of real world
- Working with samples: False positives are an unavoidable part of process
- Of course, it's going to take repeated experimentation to determine which results represent real findings rather than random noise in data
- You shouldn't expect a single study to prove anything conclusively
- You need to do replication studies