Statistics for LMU Bio Master's programs
t-tests and non-parametric alternatives

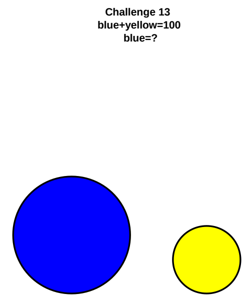
Dirk Metzler

May 1, 2021

Contents

1 The one-sample t-Test

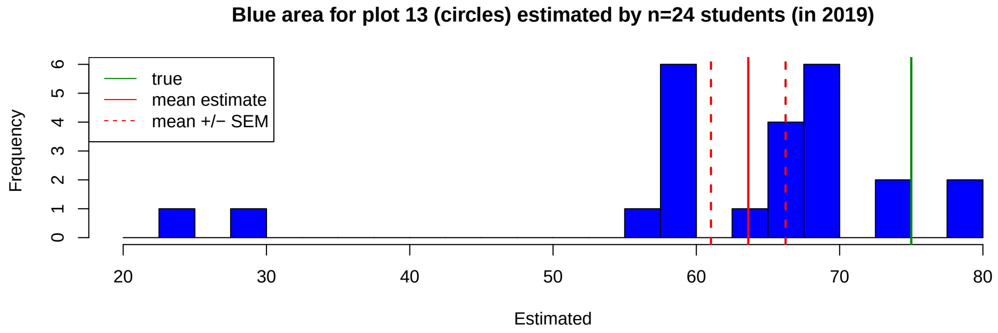
Are large circle areas over- or underestimated?



true: 75% vs 25%

Overestimated because volume could be taken instead of area?

Underestimated because diameter could be taken instead of area?



Null hypothesis: estimation X is unbiased. That is, the mean of the estimated values is

$$\mu := \mathbb{E}X = 75$$

.

Let $x_1, \dots, x_n$ be the observed values.
$|\overline{x} - \mu| = |63.6 - 75| = 11.4$ is 4.37 standard errors ($s_x/\sqrt{n}$).

$$t \ := \ \frac{\overline{x} - \mu}{s_x/\sqrt{n}} \ = \ 4.37$$

The t value is the test statistic in Student's (that is, Gosset's) one-sample t-test.
If the null hypothesis is true, what is the distribution of t?
Assume

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

$$\begin{aligned} \Rightarrow \overline{X} &\sim \mathcal{N}(\mu, \sigma^2/n) \\ \Rightarrow \overline{X} - \mu &\sim \mathcal{N}(0, \sigma^2/n) \\ \Rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} &\sim \mathcal{N}(0, 1) \end{aligned}$$

But σ is replaced by s in t, which adds some variation.

General Rule 1. *If $X_1, \dots, X_n$ are independently drawn from a normal distributed with mean μ and*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

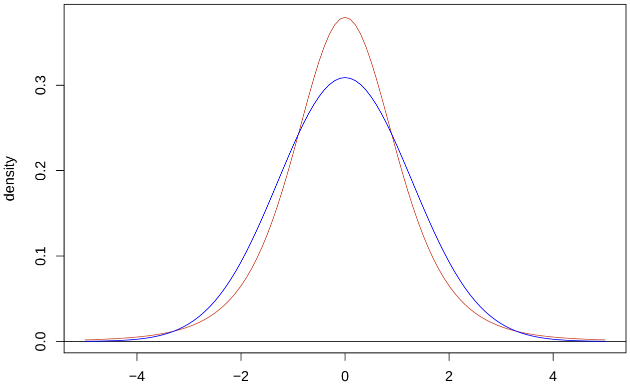
then

$$t \ = \ \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

is t-distributed with $n-1$ degrees of freedom (df).

The t-distribution is also called *Student-distribution* since Gosset published it using this synonym.

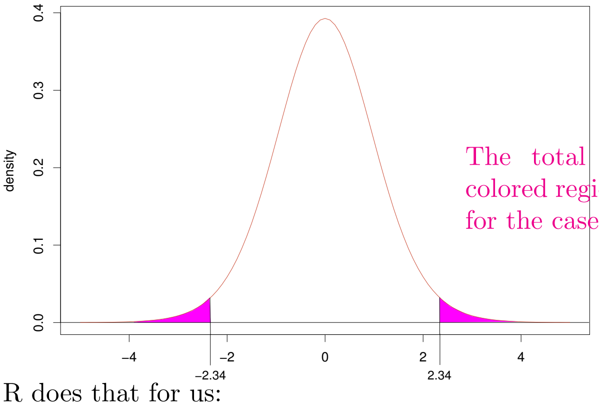
Density of the t-distribution



Density of t-distribution with 5 df Density of normal distribution with $\mu = 0$ and $\sigma^2 = \frac{5}{3}$
How (im)probable is a deviation by at least 4.37 standard errors?

$$\Pr(T = 4.37) = 0 \qquad \text{does not help!}$$

Compute $\Pr(|T| \geq 4.37)$, which is the p-value.



The total area of the magenta-colored regions would be the p-value for the case of $|t| = 2.34$.

R does that for us:

```
> pt(-4.37,df=23)+pt(4.37,df=23,lower.tail=FALSE)
[1] 0.0002238728
```

t-Test with R

```
> t.test(ba$estimated[ba$X==13],mu=mean(ba$true[ba$X==13]))

One Sample t-test

data:  ba$estimated[ba$X == 13]
t = -4.3668, df = 23, p-value = 0.0002257
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 58.23633 69.01367
sample estimates:
mean of x
   63.625
```

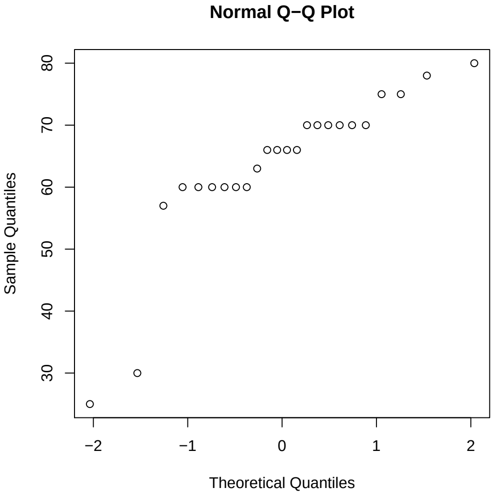
We note:

$$p - \text{value} = 0.0002257$$

i.e.: If the **Null hypothesis** "just random", in our case $\mu = 75$, holds, any deviation like the observed one or even more is very improbable.

If we always reject the null hypothesis if and only if the p-value is below a level of significance of 0.05, then we can say following holds:

If the null hypothesis holds, then the probability that we reject it, is only 0.05.



Why normal q-q-plot should roughly show a straigth line.

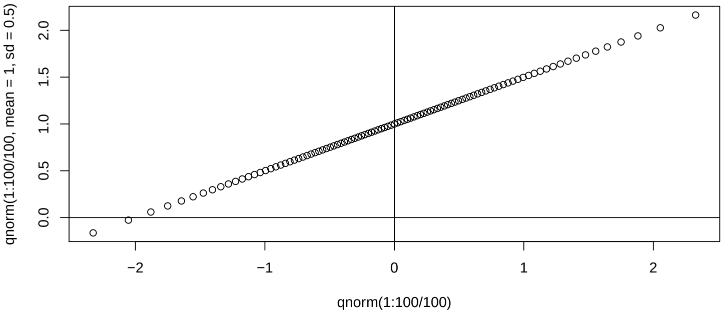
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

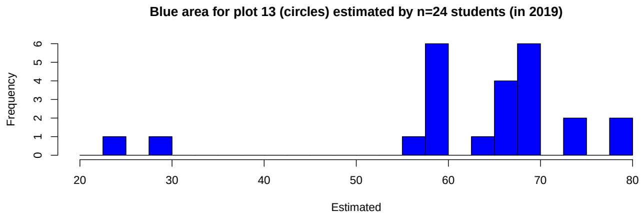
$\Rightarrow$

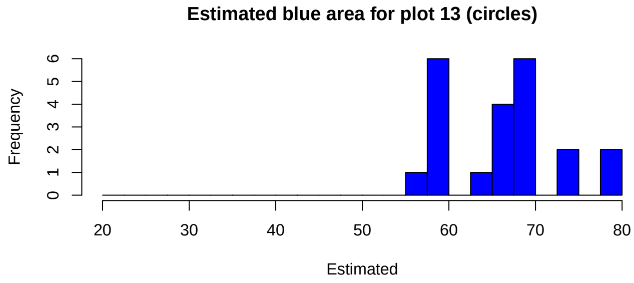
$$\Pr\left(\frac{X-\mu}{\sigma} < q\right) = \Pr(X < \sigma \cdot q + \mu)$$

This means that if q is the p-quantile of $\mathcal{N}(0,1)$, then $\sigma \cdot q + \mu$ is the p-quantile of X.

With **exact** quantiles of $\mathcal{N}\left(\mu = 1, \sigma^2 = 0.5^2\right)$:



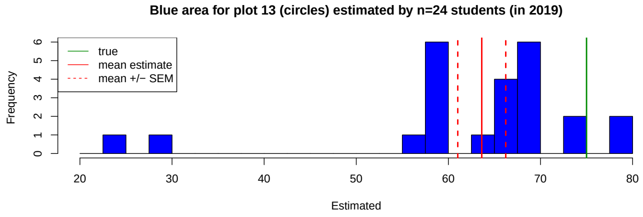


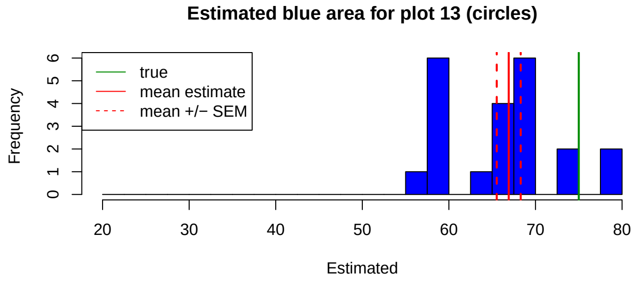


```
> t.test(ba$estimated[ba$X==13 & ba$estimated>50],
         mu=mean(ba$true[ba$X==13]))

One Sample t-test

data:  ba$estimated[ba$X == 13 & ba$estimated > 50]
t = -5.8431, df = 21, p-value = 8.443e-06
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 64.02948 69.78870
sample estimates:
mean of x
 66.90909
```





But is it really well justfied to remove outliers?

Perhaps rather not.

In any case the decision must be documented and reasoned.

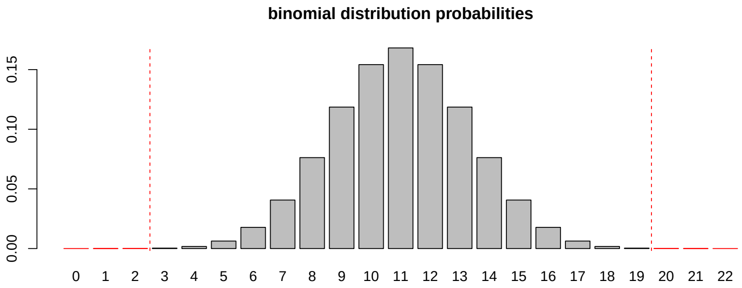
2 Reducing the data even further

For our example data we can also try just using the following information:

$n = 22$ students did not give the correct value of 75

$k = 20$ of them underestimated the blue area

Null hypothesis is now: overestimation and underestimation have the same probability.As always we further assume independence among the estimations. Thus, k is a realization of $K \sim \text{bin}(n, 1/2)$.



With two-sided testing, the p-value is the sum of all probabilities that are smaller or equal to $\Pr(K = k)$.

For $p = 1/2$ the binomial distribution is symmetric, such that the p-value is

$$\Pr(K \leq n - k) + \Pr(K \geq k) = \Pr(K \leq 2) + \Pr(K \geq 20) \approx 0.00012.$$

```
> (s <- dbinom(0:n,n,0.5)<=dbinom(k,n,0.5))
 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
> sum(dbinom(0:n,n,0.5)[s])
[1] 0.0001211166
```

3 Non-parametric alternative: Wilcoxon's signed rank test

Wilcoxon's signed rank test

signed rank: take rank of absolute value, equipped with sign of original value

example:

values:	-2.6	-2.5	-2.3	-1.3	-0.6	1.6	2.2	6.1
absolute values:	2.6	2.5	2.3	1.3	0.6	1.6	2.2	6.1
ranks:	7	6	5	2	1	3	4	8
signed ranks:	-7	-6	-5	-2	-1	3	4	8

Test statistic: $V = 3 + 4 + 8 = 15$ (sum of positive ranks)

```
> wilcox.test(c(-2.6,-2.5,-2.3,-1.3,-0.6,1.6,2.2,6.1))

Wilcoxon signed rank test
```

```
data:  c(-2.6, -2.5, -2.3, -1.3, -0.6, 1.6, 2.2, 6.1)
V = 15, p-value = 0.7422
alternative hypothesis: true location is not equal to 0
```

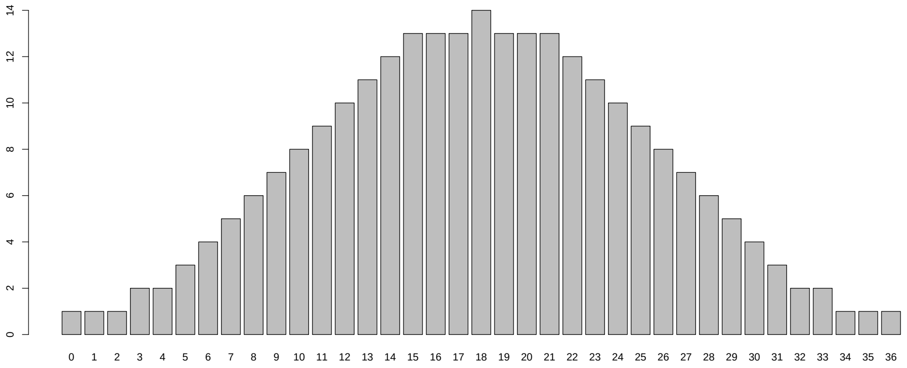
How many combinations of ranks lead to a certain value of V?

$$\begin{array}{rcrl}
-1-2-3-4-5-6-7-8 & = & -36 & V=0 \\
+1\ -2-3-4-5-6-7-8 & = & -34 & V=1 \\
-1\ +2\ -3-4-5-6-7-8 & = & -32 & V=2 \\
-1-2\ +3\ -4-5-6-7-8 & = & -30 & V=3 \\
+1+2\ -3-4-5-6-7-8 & = & -30 & V=1+2=3 \\
\vdots & \vdots & \vdots & \\
-1+2+3+4+5+6+7+8 & = & 34 & V=2+3+\cdots+8=35 \\
1+2+3+4+5+6+7+8 & = & 36 & V=1+2+\cdots+8=36
\end{array}$$

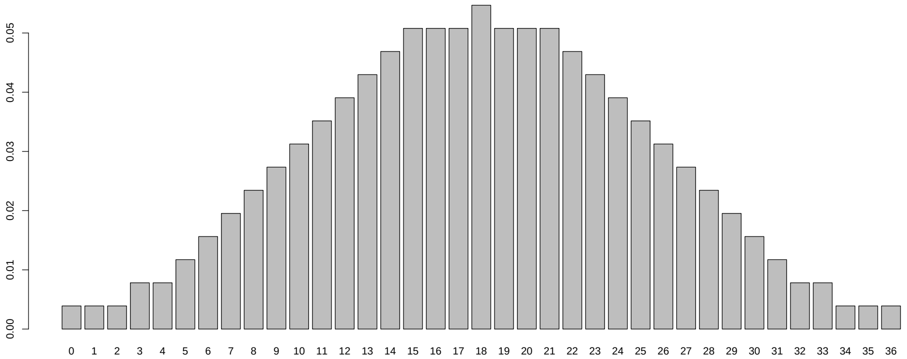
The null hypothesis implies that for each rank the sign is purely random.Thus, each line above has a probability of $1/2^n$.There are e.g. two ways of getting a $V=3$ and two of getting $V=33$.

$$\Pr(V=3)=\Pr(V=33)=\frac{2}{2^8}=0.0078$$

Numbers of possibilities for each possible value of V for $n=8$



Probabilities under H_0 of possible values of V for $n=8$



example

values:	-2.6	-2.5	-2.3	-1.3	-0.6	-0.2	0.1	0.4
absolute values:	2.6	2.5	2.3	1.3	0.6	0.2	0.1	0.4
ranks:	8	7	6	5	4	2	1	3
signed ranks:	-8	-7	-6	-5	-4	-2	1	3

p-value for $V = 4$ (with $n = 8$):

$$\Pr(V \in \{0, 1, 2, 3, 4, 32, 33, 34, 35, 36\}) = \frac{1+1+1+2+2+2+2+1+1+1}{2^8} = 0.054$$

Wilcoxon's signed rank test

```
> wilcox.test(c(-2.6,-2.5,-2.3,-1.3,-0.6,-0.2,0.1,0.5))

Wilcoxon signed rank test

data:  c(-2.6, -2.5, -2.3, -1.3, -0.6, -0.2, 0.1, 0.5)
V = 4, p-value = 0.05469
alternative hypothesis: true location is not equal to 0
```

Wilcoxon's signed rank test

```
> wilcox.test(ba$estimated[ba$X==13],
              mu=mean(ba$true[ba$X==13]))

Wilcoxon signed rank test with continuity correction

data:  ba$estimated[ba$X == 13]
V = 6, p-value = 8.779e-05
alternative hypothesis: true location is not equal to 75

Warnmeldungen:
1: In wilcox.test.default(ba$estimated[ba$X == 13], mu = mean(ba$true[ba$X ==  :
  cannot compute exact p value due to bindings
2: In wilcox.test.default(ba$estimated[ba$X == 13], mu = mean(ba$true[ba$X ==  :
  cannot compute exact p value due to bindings
```

4 General principles of statistical testing

- We want to argue that some deviation in the data is not just random.
- To this end we first specify a null hypothesis H_0, i.e. we define, what “just random” means.
- Then we try to show: If H_0 is true, then a deviation that is at least at large as the observed one, is very improbable.
- If we can do this, we reject H_0.
- How we measure deviation, must be clear *before* we see the data.

Statistical Testing: Important terms

null hypothesis H_0 : says that what we want to substantiate is not true and anything that looks like evidence in the data is just random. We try to reject H_0.

significance level α : If H_0 is true, the probability to falsely reject it, must be $\leq \alpha$ (often $\alpha = 0.05$).

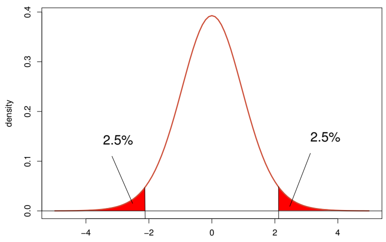
test statistic : measures how far the data deviates from what H_0 predicts into the direction of our alternative hypothesis.

p **value** : Probability that, if H_0 is true, a dataset leads to a test statistic value that is as least as extreme as the observed one.

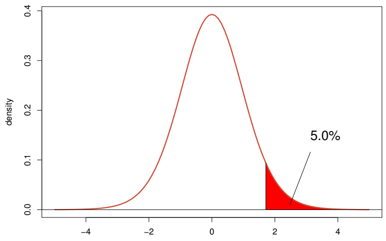
- We reject the null hypothesis H_0 if the p value is smaller than α.
- Thus, if H_0 is true, the probability to (falsely) reject it is α (not the p value).
- This entails that a researcher who performs many tests with $\alpha = 0.05$ on complete random data (i.e. where H_0 is always true), will falsely reject H_0 in 5% of the tests.
- Therefore it is a severe violation of academic soundness to perform tests until one shows significance, and to publish only the latter.

Testing two-sided or one-sided?

We observe a value of x that is much larger than the H_0 expectation value μ.



p-value=$\Pr_{H_0}(|X-\mu| \geq |x-\mu|)$



p-value=$\Pr_{H_0}(X \geq x)$

The pure teachings of statistical testing

- Specify a null hypothesis H_0, e.g. $\mu = 0$.
- Specify level of significance α, e.g. $\alpha = 0.05$.
- Specify an event $\mathcal{A}$ such that
  $$\Pr_{H_0}(\mathcal{A}) = \alpha$$
  (or at least $\Pr_{H_0}(\mathcal{A}) \leq \alpha$). e.g. $\mathcal{A} = \{\overline{X} > q\}$ or $\mathcal{A} = \{|\overline{X} - \mu| > r\}$ in general: $\mathcal{A} = \{p\text{-value} \leq \alpha\}$
- AND AFTER THAT: Look at the data and check if if $\mathcal{A}$ occurs.
- Then, the probability that H_0 is rejected in the case that H_0 is actually true ("Type I error") is just α.

Violations against the pure teachings

"The two-sided test gave me a p-value of 0.06. Therefore, I tested one-sided and this worked out nicely."

is as bad as:

"At first glance I saw that $\overline{x}$ is larger than μ_{H_0}. So, I immediately applied the one-sided test."

Important
The decision between one-sided and two-sided must not depend on the concrete data that are used in the test. More generally: If $\mathcal{A}$ is the event that will lead to the rejection of H_0, (if it occurs) then $\mathcal{A}$ must be defined without being influenced by the data that is used for testing.

This means: Use separate data sets for exploratory data analysis and for testing.
In some fields these rules are followed quite strictly, e.g. testing new pharmaceuticals for accreditation.
In some other fields the practical approach is more common: Just inform the reader about the p-values of different null-hypotheses.

Severe violations against scientific standards

HARKing: Hypothesize After Results Known

p-hacking: try out different tests and different preprocessing methods until you obtain significance

If H_0 is rejected on the 5%-level, which of the following statements is true?

- ~~The null hypothesis is wrong.~~
- ~~H_0 is wrong with a probability of 95%.~~
- If H_0 is true, you will see such an extreme event only in 5% of the data sets. ✓

If the test did not reject H_0, which of the following statements are true?

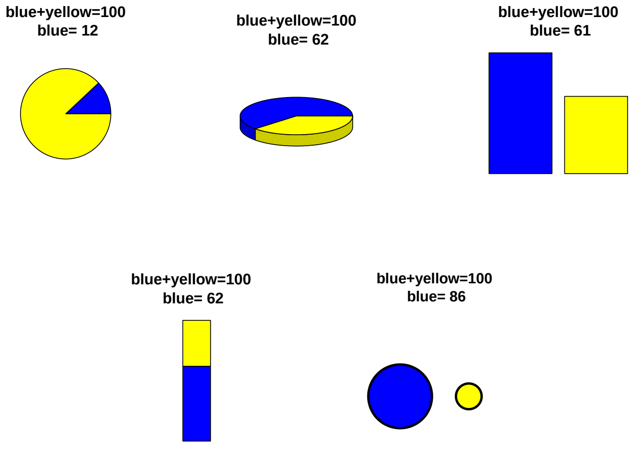
- ~~We have to reject the alternative H_1.~~
- ~~H_0 is true~~
- ~~H_0 is probably true.~~
- ~~It is safe to assume that H_0 was true.~~
- Even if H_0 is true, it is not so unlikely that our test statistic takes a value that is as extreme as the one we observed.✓
- With this respect, H_0 is compatible with the data.✓

Some of what you should be able to explain

- Structure of t-statistic and df for one-sample t-test
- what is a qqnorm plot and how can it be used to check normality
- Rationale of Wilcoxon's signed rank test and how to compute its test statistic
- principles of statistical testing and **exact** meaning of
  - p-value
  - significance level α
  - null hypothesis
- how to report significance and non-significance

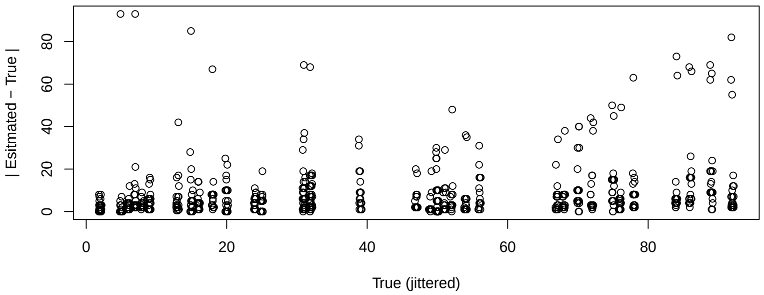
5 two-sample t-tests

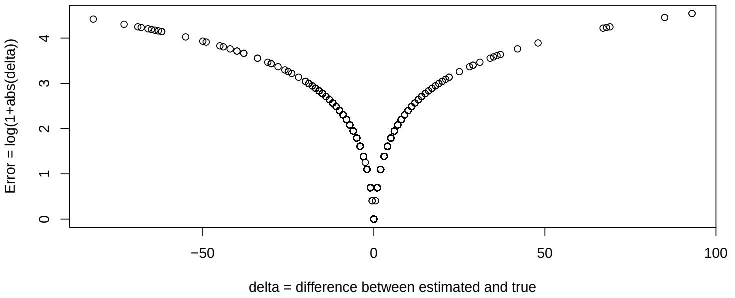
Do different visualization types lead to different error rates?

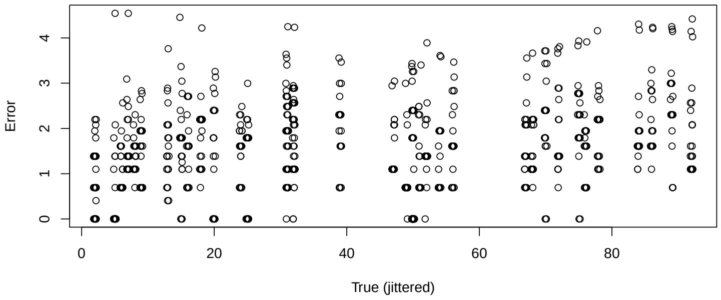


How to measure error rate?

Error$= \log(1 + |\text{estimated} - \text{true}|)$







Make sure later that assuptions for tests (e.g. normal distribution) are fulfilled; otherwise try a different error function.

Using multiple measurements per student would lead to dependencies in the data that would violate independence assumtions in our tests.

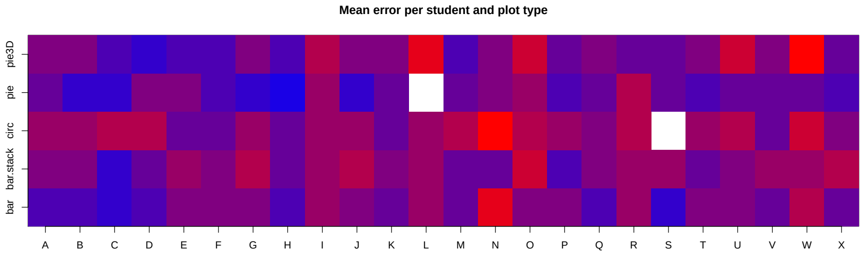
One possible solution: Take for each student and each plot type only the mean value of the measurements (the measurements are the estimation errors in the case).

Average errors of each student for each plot type

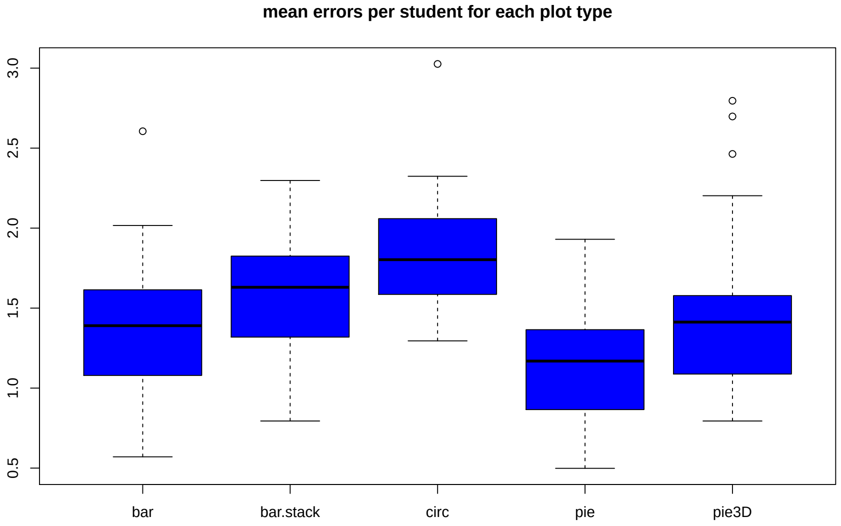
```
                 A         B         C         D        E         F        G
bar       1.065645 0.9642789 0.7630992 0.9450101 1.546024 1.5283079 1.395527
bar.stack 1.394888 1.6230186 0.7945135 1.3210294 1.711141 1.6088525 2.028028
circ      1.678805 1.7038884 2.1449357 2.0660967 1.295430 1.2954301 1.802887
pie       1.353468 0.7864461 0.6212267 1.4907948 1.376918 0.9744558 0.735763
pie3D     1.424133 1.4940845 1.0316535 0.7945135 1.023586 0.8451966 1.426942
```

```
                 H        I        J        K        L         M        N
bar       1.014962 1.655458 1.395527 1.165459 1.798220 1.3727371 2.605710
bar.stack 1.195617 1.682711 2.028028 1.637742 1.758303 1.1670114 1.316792
circ      1.346113 1.697049 1.802887 1.298239 1.845993 2.0249917 3.026309
pie       0.498623 1.654818 0.735763 1.319477       NA 1.3519713 1.497177
pie3D     1.001792 2.025146 1.426942 1.393275 2.698072 0.8730895 1.592161
                 O         P        Q        R         S         T        U
bar       1.599032 1.6300899 1.092981 1.859423 0.5705435 1.5607468 1.384883
bar.stack 2.297891 1.0595464 1.532854 1.892707 1.7193107 1.1750788 1.623019
circ      2.158900 1.8075584 1.561459 2.128278        NA 1.9099707 2.052395
pie       1.755138 0.9678003 1.133020 1.930243 1.1178076 0.8730895 1.263201
pie3D     2.463760 1.1968729 1.564268 1.231577 1.1442213 1.4176781 2.202529
                 V        W         X
bar       1.194348 2.016658 1.2799950
bar.stack 1.891898 1.724770 2.1487236
circ      1.340298 2.324150 1.6102338
pie       1.168980 1.344205 0.8583667
pie3D     1.407673 2.796155 1.2931365
```

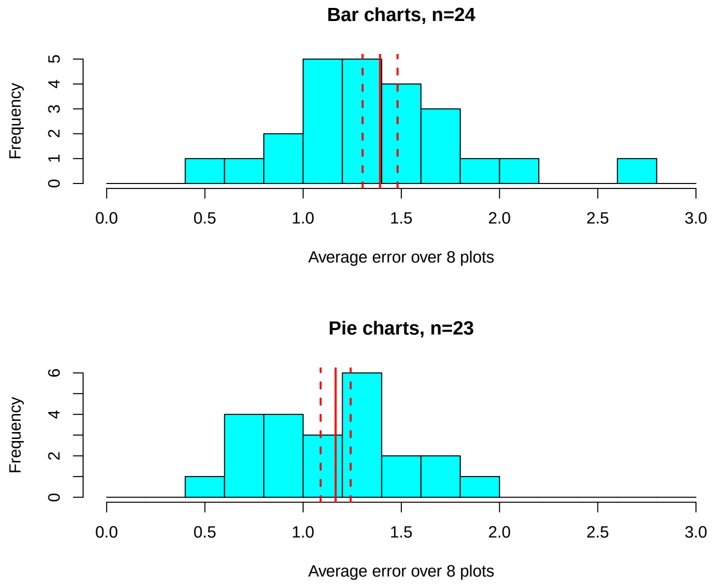
Average errors of each student for each plot type

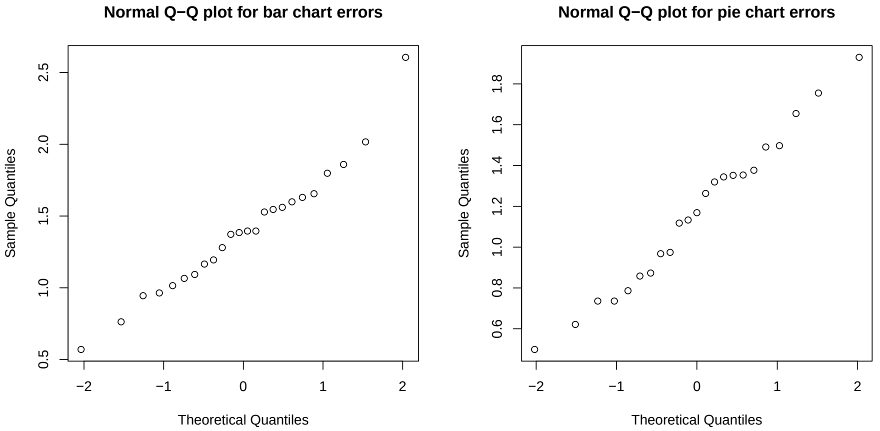


blue: small error,
red: large error,
white: missing data



Let's test whether errors (averaged for each student) differ significantly between **bar** charts and **pie** charts.





```
> t.test(M["bar",],M["pie",])

Welch Two Sample t-test

data:  M["bar", ] and M["pie", ]
t = 1.9106, df = 44.466, p-value = 0.06252
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01233732  0.46486341
sample estimates:
mean of x mean of y
 1.391861  1.165598
```

Theorem 1 (Welch's t-test)**.** *Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent and normally distributed random variables with $\mathbb{E}X_i = \mathbb{E}Y_j$ and potentially different variances $VarX_i = \sigma_X^2$ and $VarY_i = \sigma_Y^2$. Let s_X^2 and s_Y^2 be the sample variances. The statistic*

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

approximately *follows a* t *distribution with*

$$\frac{\left(\frac{s_X^2}{n}+\frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2\cdot(n-1)}+\frac{s_Y^4}{m^2\cdot(m-1)}}$$

degrees of freedom.

```
> t.test(M["bar",],M["pie",],var.equal=TRUE)

Two Sample t-test

data:  M["bar", ] and M["pie", ]
t = 1.9043, df = 45, p-value = 0.06328
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01305144  0.46557752
sample estimates:
mean of x mean of y
 1.391861  1.165598
```

Theorem 2 (Student's two-sample t-test, unpaired with equal variances)**.** *Suppose that* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$ *are independent and normally distributed random variables with the same mean* μ *and the same variance* σ^2*. Define the* ***pooled sample variance*** *to be*

$$s_p^2 = \frac{(n-1)\cdot s_X^2 + (m-1)\cdot s_Y^2}{m+n-2}.$$

The statistic

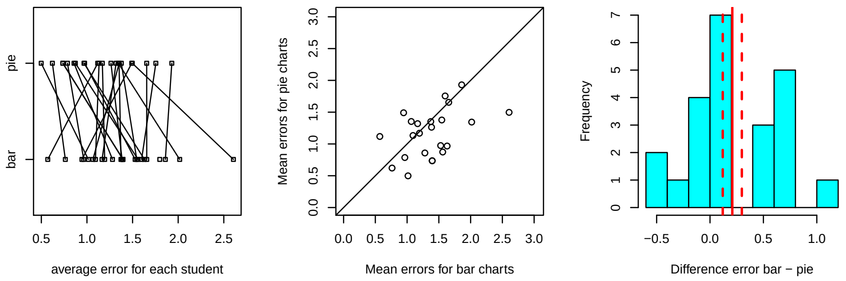
$$t = \frac{\overline{X}-\overline{Y}}{s_p\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

follows a t *distribution with* $n+m-2$ *degrees of freedom.*

Have we missed some relevant information?

Maybe there was variation in error among the students.

We can remove this variation by **paired** testing, that is, apply one-sample t-test to differences in error per student.

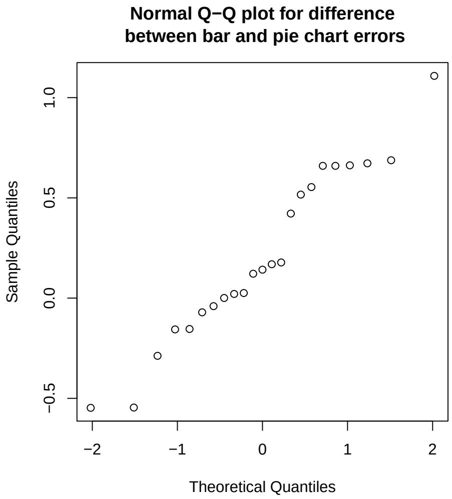


```
> t.test(M["bar",],M["pie",],paired=TRUE)

Paired t-test

data:  M["bar", ] and M["pie", ]
t = 2.3331, df = 22, p-value = 0.02918
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02317965 0.39401089
sample estimates:
mean of the differences
              0.2085953
```



Note: what we did here looks like we tried out several tests until a test indicated significance.

This would be a severe violation of the principles of statistical testing and of good scientific practice!

The various tests were applied above only to show the differences for teaching purposes. (So do as I say, not as I apparently did.)

Correct: before looking at the data, anticipate that between-student variation is possible and decide to apply paired t-test.

What is the conclusion from our test?

Let's assume the 24 test persons were chosen randomly among all LMU biology students (which was actually not the case).

Can we say

"LMU biology students could guess fractions significantly better from pie charts (of this kind) than from bar charts (of this kind)."

or can we only say

"LMU biology students could guess fractions significantly better from *these* 8 pie charts than from *these* 8 bar charts."

???

- The true value of each plot may influence the error distribution.

- In fact, the pie and bar charts were randomly generated from a class of charts for which they are representative
  - blue and yellow colour
  - blue area is a purely randonm value from $\{1, 2, \ldots, 99\}$.
- Thus, each single plot had the same chance to have a difficult or an easy true value.
- But the values used for the t-test were not the independent observations of each plot but averages over 8 plots of each type.
- Thus, each of the eight plots was used in all 23 (or 24) values, and the t-test (standard errors in denominator of t) could not account for the variation between plots.

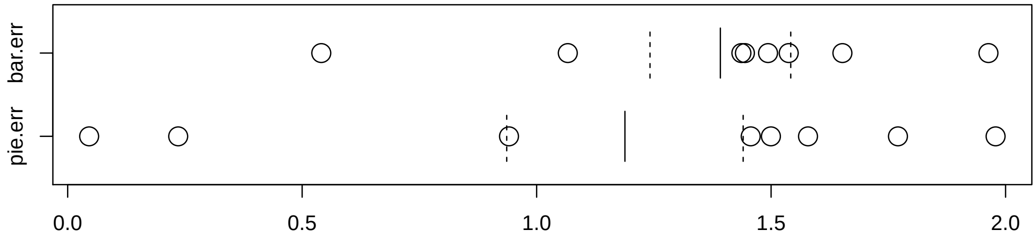
This implies that the results could be representative for LMU biology students (if the test persons had been sampled randomly) but not for a class of pie/bar plots.

How can we apply the t-test to independent values for each plot?

Average for each plot the errors from all students:

pie.err	0.24	0.05	1.46	0.94	1.77	1.58	1.98	1.50
bar.err	1.07	1.44	1.96	0.54	1.49	1.54	1.65	1.44

(for one of the pie charts we averaged over only 23 student errors as one was NA.)



```
> t.test(pie.err,bar.err)

Welch Two Sample t-test

data:  pie.err and bar.err
t = -0.69367, df = 11.407, p-value = 0.5018
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8464499  0.4394227
sample estimates:
mean of x mean of y
 1.188347  1.391861
```

So, if we use the errors values for strip charts and bar charts in a way such that they are representative for charts of their class, the differences in error values are not significant any more.

But were the students representative for other students in this analysis?

No, because we averaged over the students, such that the t-test could not account for variation among the students.

Anyway, we could not find significant differences (maybe because of the small number of plots of each type?), but even if we had found significant differences, the study design would not allow to draw conclusions about other students than the ones involved in the experiment.

How could we draw conclusions about a class of pie charts and bar charts and about a larger population of persons?

With different tests, e.g. nested anova / mixed-effects models.

or:

With a different study design, e.g. many students, each sees only one bar chart and one pie chart; always new ones.

6 Wilcoxon's rank sum test

Wilcoxon's rank sum test (or equivalently Mann-Whitney U test) is a non-parametric alternative for the unpaired two-sample t-test with equal variances. (But not for Welch's t-test!)

References

[1] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1**:8083.

[2] Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**:5060.

Observations: Two samples

$$X : x_1, x_2, \ldots, x_m$$
$$Y : y_1, y_2, \ldots, y_n$$

Test the *null hypothesis*: that X and Y come from the same population

alternative: X "typically larger" than Y or Y "typically larger" than X

Idea

Observations:

$$X : x_1, x_2, \ldots, x_m$$
$$Y : y_1, y_2, \ldots, y_n$$

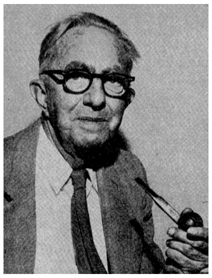
- Sort all observations by size.
- Determine the *ranks* of the m X-values among all $m+n$ observations.

- If the null hypothesis is true, than the m X-ranks are randomly chosen from $\{1, 2, \ldots, m+n\}$.
- Compute the sum of the X-ranks and check if it is untypically small or large compared to sum of random ranks.

Wilcoxon's rank-sum statistic

Observation:

$$X : x_1, x_2, \ldots, x_m$$
$$Y : y_1, y_2, \ldots, y_n$$



Frank Wilcoxon,
1892-1965

$$W = \text{Sum of the } X\text{-ranks} - (1 + 2 + \cdots + m)$$

is called

Wilcoxon's rank-sum statistic

Wilcoxon's rank-sum statistic

Note:

$$W = \text{Sum of the } X\text{-ranks} - (1 + 2 + \cdots + m)$$

We could also use the sum of the Y-ranks, because

$$\begin{aligned} &\text{Sum of the } X\text{-ranks} + \text{Sum of the } Y\text{-ranks} \\ &= \text{Sum of all ranks} \\ &= 1 + 2 + \cdots + (m+n) = \frac{(m+n)(m+n+1)}{2} \end{aligned}$$

A *small* example

- Observations:

$$\begin{aligned} X &: 1.5, 5.6, 35.2 \\ Y &: 7.9, 38.1, 41.0, 56.7, 112.1, 197.4, 381.8 \end{aligned}$$

- Pool observations and sort: 1.5, 5.6, 7.9, 35.2, 38.1, 41.0, 56.7, 112.1, 197.4, 381.8
- Determine ranks: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- rank-sum: $W = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Significance

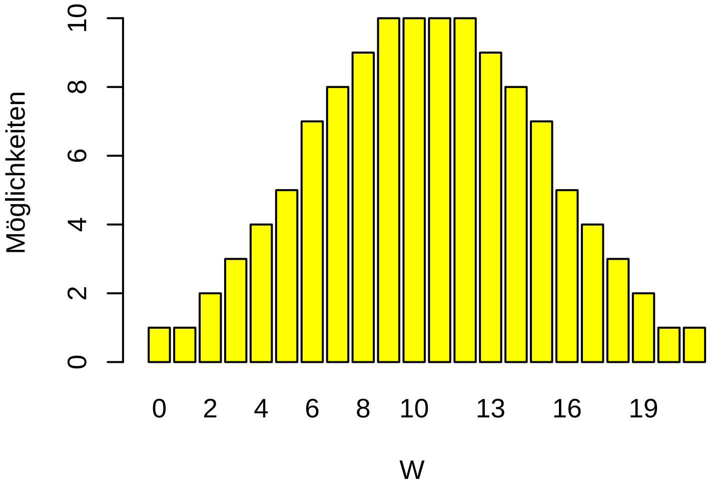
Null hypothesis:
X-sample and Y-sample were taken from the same distribution

The 3 ranks of the X-sample 1 2 3 4 5 6 7 8 9 10
could just as well have been any 3 ranks 1 2 3 4 5 6 7 8 9 10

There are $\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$ possibilities.

(In general: $\frac{(m+n)(m+n-1)\cdots(n+1)}{m(m-1)\cdots 1}) = \frac{(m+n)!}{n!m!} = \binom{m+n}{m}$ possibilities)

Distribution or the Wilcoxon statistic $(m = 3, n = 7)$[1ex]



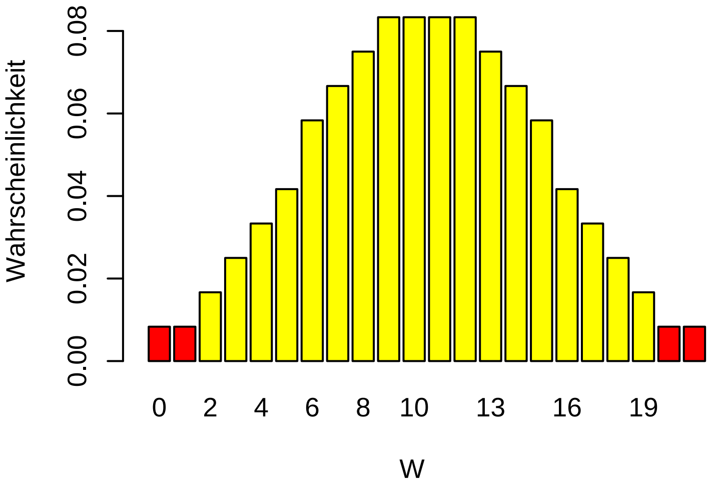
Under the null hypothesis all rank configurations are equally likely, thus

$$\mathbb{P}(W = w) = \frac{\text{number of possibilities with rank-sum statistic } w}{120}$$

We see in our example: 1.5, 5.6, 7.9, 35.2, 38.1, 41.0, 56.7, 112.1, 197.4, 381.8 $W = 1$

$\mathbb{P}(W \leq 1) + \mathbb{P}(W \geq 20) \;= \mathbb{P}(W = 0) + \mathbb{P}(W = 1) + \mathbb{P}(W = 20) + \mathbb{P}(W = 21) = \frac{1+1+1+1}{120} \doteq 0.033$

Distribution of the Wilcoxon statistic $(m = 3, n = 7)$[1ex]



$$\text{For our example } (W = 1):$$

$$p\text{-value} = \mathbb{P}(\text{such an extreme } W) = 4/120 = 0.033$$

We *reject* the *null hypothesis*, that the distributions of X and Y were equal, on the 5%-level.

Wilcoxon test in R with `wilcox.test`:

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> wilcox.test(x,y)

      Wilcoxon rank sum test

data:  x and y
W = 1, p-value = 0.03333
alternative hypothesis: true location shift is
not equal to 0
```

IMPORTANT!

Can Wilcoxon's rank sum test replace Welch's t-test? Not in general, because its null hypothesis is that the data come from the same distribution, not just that the means are equal. If we want to test whether the means are different but allow the standard deviations to be different (like in the assumptions of Welch's t-test), the Wilcoxon test cannot be applied!

Some of what you should be able to explain

- Structure of t-statistic and df for
  - one-sample t-test
  - paired two-sample t-test
  - unpaired two-sample t-test
    - with equal variances
    - Welch's t-test
- when and why to use the different t-test variants
- summarizing measurements in a statistic (here: our Error function) that fulfills basic assumptions of the test
- summarize data to avoid dependencies; e.g. average over multiple measurements
- Assumptions of Wilcoxon's rank sum test and when and how to apply it.

see also the topics listed on page 11