Statistics for LMU Bio Master’s programs
More biological examples for t-tests

Dirk Metzler

April 7, 2026

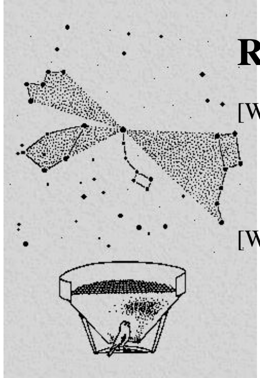
Contents

1 t-Test for paired samples

1.1 Example: Orientation of pied flycatchers

References

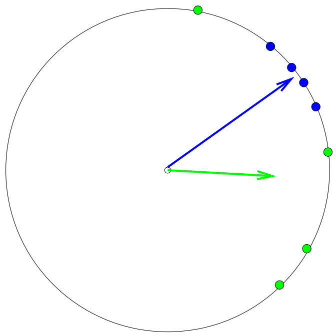
[WGS+04] Wiltschko, W.; Gesson, M.; Stapput, K.; Wiltschko, R. Light-dependent magnetoreception in birds: interaction of at least two different receptors. *Naturwissenschaften* 91.3, pp. 130-4, 2004.



References

[WRS+05] Wiltschko, R.; Ritz, T.; Stapput, K.; Thalau, P.; Wiltschko, W. Two different types of light-dependent responses to magnetic fields in birds. *Curr Biol* 15.16, pp. 1518-23, 2005.

[WSB+07] Wiltschko, R.; Stapput, K.; Bischof, H. J.; Wiltschko, W. Light-dependent magnetoreception in birds: increasing intensity of monochromatic light changes the nature of the response. *Front Zool*, 4, 2007.



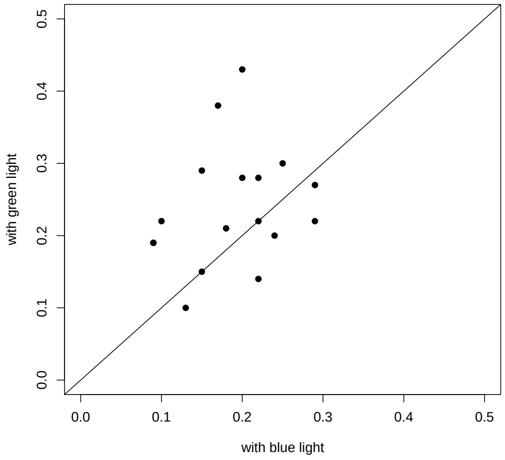
Direction of flight with blue light.
Another flight of the same bird with blue light.
All flight's directions of this bird in blue light.
corresponding exit points
Directions of this bird's flights in green light. the more variable the directions
corresponding exit points
arrow heads: barycenter of exit points for green light
the same for "blue" directions

the shorter the arrows!

Question of interest

Does the color of monochromatic light influence the orientation?
Experiment: For 17 birds compare barycenter vector lengths for blue and green light.

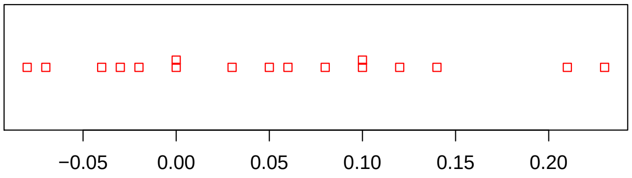
Pies flycathchers : barycenter vector lengths for green light and for blue light b, n=17

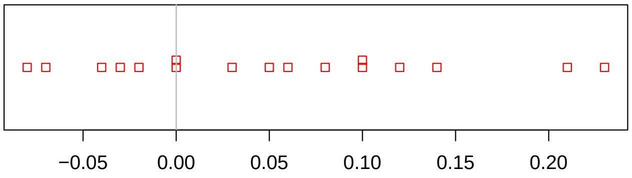


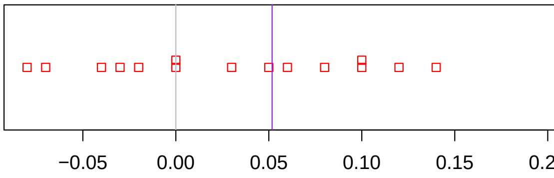
Compute for each bird the distance to the diagonal,

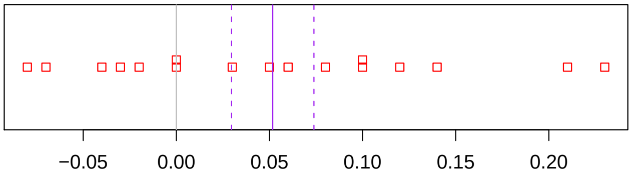
i.e.

$$x := \text{“green length”} - \text{“blue length”}$$









Can the theoretical mean be $\mu = 0$? (i.e. the expected value for a randomly chosen bird)

$$\begin{aligned} \overline{x} &= 0.0518 \\ s &= 0.0912 \\ \text{SEM} &= \frac{s}{\sqrt{n}} = \frac{0.0912}{\sqrt{17}} = 0.022 \end{aligned}$$

Is $|\overline{x} - \mu| \approx 0.0518$ a large deviation?

Large? Large compared to what?

Large compared the the standard error?

$$\frac{|\overline{x} - \mu|}{s/\sqrt{n}} \approx \frac{0.0518}{0.022} \approx 2.34$$

So, $\overline{x}$ is more than 2.3 standard deviations away from $\mu = 0$.

How probable is this when 0 is the true theoretical mean? in other words:

Is this deviation significant?

We know:

$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

is asymptotically (for large n) standard-normally distributed.

σ is the theoretical standard deviation, but we do not know it. Instead of σ we use the sample standard deviation s. We have applied the "2/3 rule of thumb" which is based on the normal distribution. For questions about the significance this approximation is too imprecise if σ is replaced by s (and n is not extremely large).

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

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

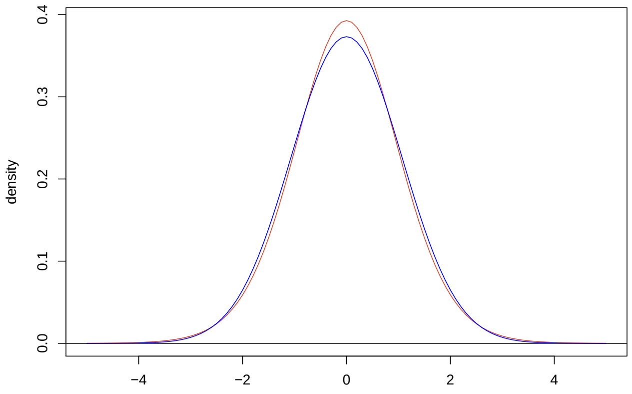
then

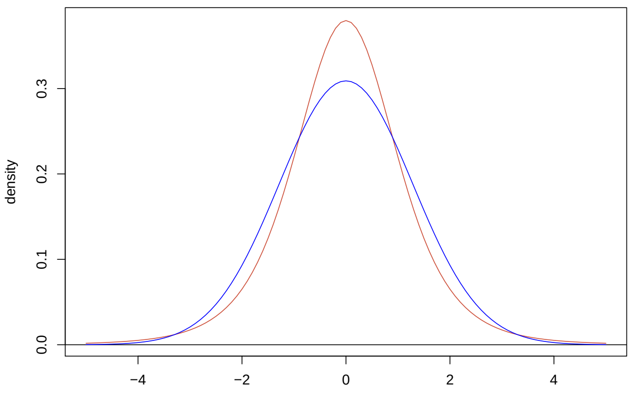
$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

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



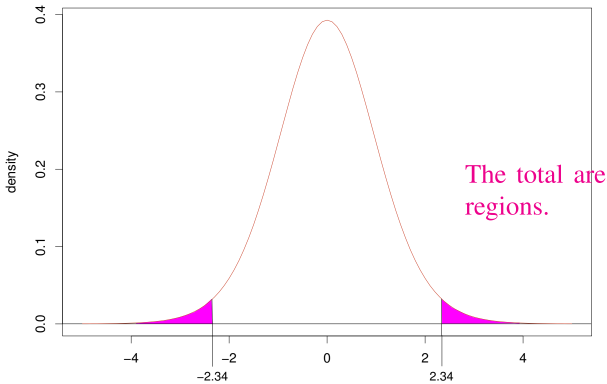


Density of t-distribution with 16 df Density of standard normal distribution Density of standard normal distribution $\mu = 0$ and $\sigma^2 = \frac{16}{14}$ Density of t-distribution with 5 df Density of normal distribution with $\mu = 0$ and $\sigma^2 = \frac{5}{3}$

How (im)probable is a deviation by at least 2.35 standard errors?

$$\Pr(T = 2.34) = 0 \qquad \text{does not help!}$$

Compute $\Pr(T \geq 2.34)$, the so-called p-value.



The total area of the magenta-colored regions.

R does that for us:

```
> pt(-2.34,df=16)+pt(2.34,df=16,lower.tail=FALSE)
[1] 0.03257345
```

Comparison with normal distribution:

```
> pnorm(-2.34)+pnorm(2.34,lower.tail=FALSE)
[1] 0.01928374
> pnorm(-2.34,sd=16/14)+
+ pnorm(2.34,sd=16/14,lower.tail=FALSE)
[1] 0.04060902
```

t-Test with R

```
> x <- length$green-length$blue
> t.test(x)

One Sample t-test

data:  x
t = 2.3405, df = 16, p-value = 0.03254
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.004879627 0.098649784
sample estimates:
 mean of x
0.05176471
```

We note:

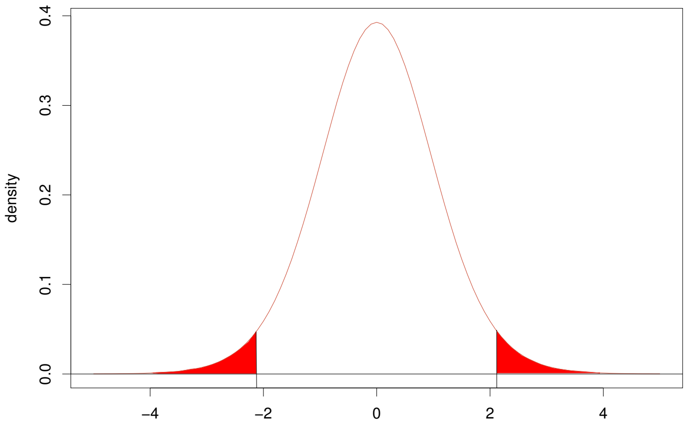
$$p\text{-value} = 0.03254$$

i.e.: If the **Null hypothesis** “just random”, in our case $\mu = 0$ 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.

If we set the level of significance to $\alpha = 0.05$, we reject the null hypothesis H_0 if and only if the t-value is in one of the red ranges:



(example of t-distribution with df$= 16$)

1.2 Example: direction-dependent thickness of cork

Simulated data!

The data in the following example are simulated. The computer simulations are inspired by a classical example data set.

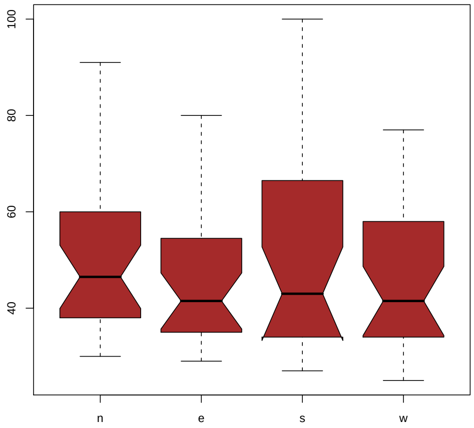
References

[RAO48e] Rao, C.R. Tests of significance in multivariate analysis. *Biometrika* 35, pp. 58-79, 1948.

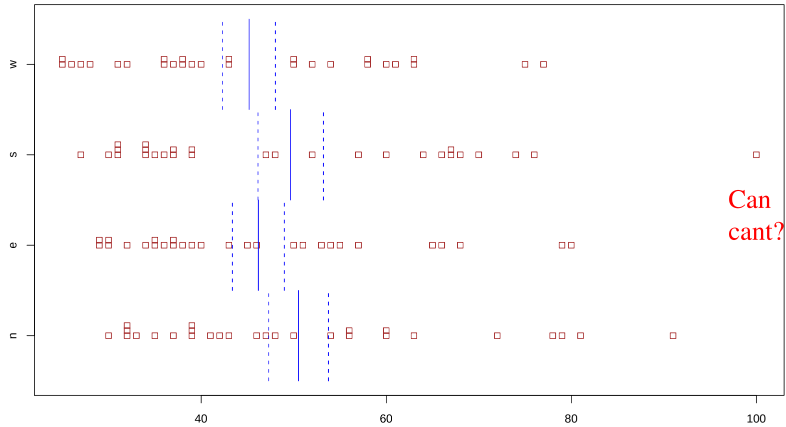
For $n = 28$ the thickness of cork [mm] was measured in each direction:

```
n        e         s         w
72      66        76        77
60      53        66        63
 5      57        64        58
41      29        36        38
32      32        35        36
30      35        34        26
39      39        31        27
 .       .         .         .
 .       .         .         .
```

cork thickness for each direction



Can some of these differences be significant???



Can some of these differences be significant???

Stripchart of cork thickness for each direction with sample means and sample means ± standard errors

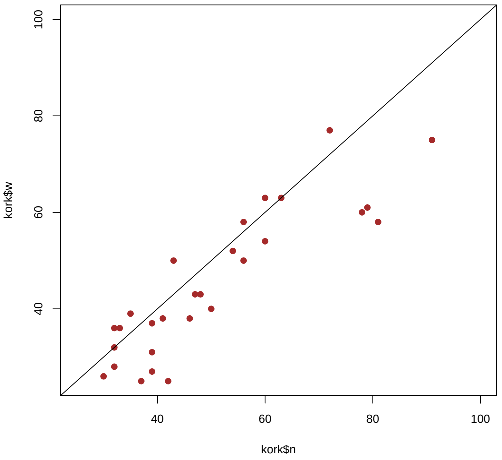
Have we missed something?

We have neglected from which measurements came from the same tree!
The trees differ in their size.

We must compare values that come from the same tree! ($\rightsquigarrow$ paired t-test)

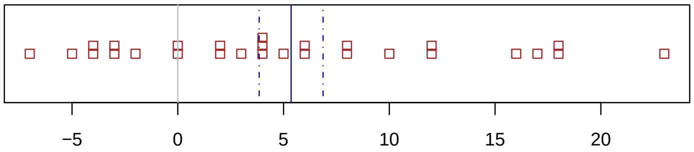
Cork thickness [mm] for $n = 28$ trees

Cork thickness on west aspect



cork thickness on north aspect

Differences in cork thickness between west and north aspect for $n = 28$ trees



with sample mean and sample mean±standard error

Is the difference significantly different from 0?

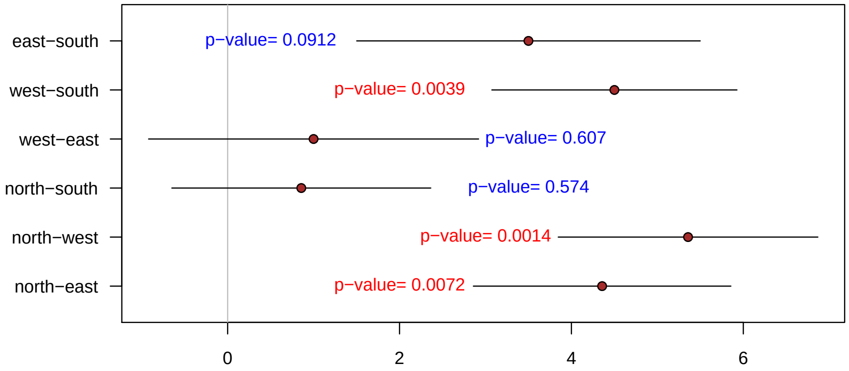
$$
\begin{aligned}
x \;:=\; (\text{thickness north}) \;&-\; (\text{thickness west}) \\
\overline{x} \;&\approx\; 5.36 \\
s_x \;&\approx\; 7.99 \\
\frac{s_x}{\sqrt{n}} \;&\approx\; 1.51 \\
t-\text{value} \;=\; \frac{\overline{x}}{s_x/\sqrt{n}} \;&\approx\; 3.547 \\
\text{degrees of freedom: df} \;&=\; n-1 \;=\; 27
\end{aligned}
$$

```
> t.test(cork$n-cork$w)

        One Sample t-test

data:  cork$n - cork$w
t = 3.5471, df = 27, p-value = 0.001447
```

```
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.258274 8.456012
sample estimates:
mean of x
 5.357143
```



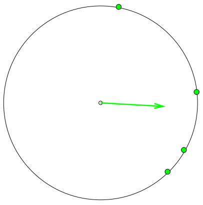
1.3 Principle of statistical testing

Example: Codon Bias

- We see 16710 times CCT an 18895 times CCC as a codon for prolin in a large sample of human genes.
- If both are equally probable, we expect 17802.5 of each.
- The observation differs from this value by 1092.5.
- z-Test: The probability of a deviation like this or more is smaller than 10^{-30}.
- We conclude that CCT and CCC do not seem to be equally probable.

Example: bird orientation

- Does the variability of flight direction depend on ambient light color?
- Measure variability by length of barycenter vector length.

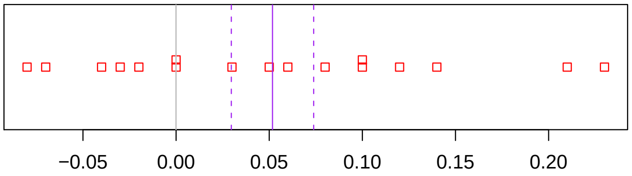


- Quantify difference by X =(length green)− (length blue).
- If light color has no influence, then $\mathbb{E}X = 0$.

Example: bird orientation

X =(Length green)$-$ (Length blue)

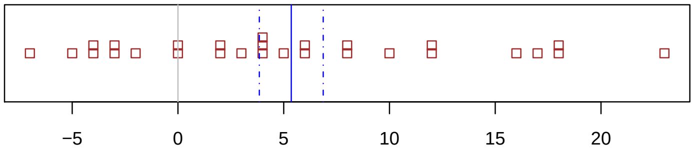
- If the light color has no influence, then $\mathbb{E}X = 0$.



- But we observe that $\overline{X} = 0.0518$ and SEM=0.022
- t-test: p-value of this deviation is ca. 3.3%.
- It seems likely that the color of the light has some influence.

Example: Thickness of cork

- X=(thickness on north aspect)$-$ (thickness on west aspect)
- If the direction has no effect, then $\mathbb{E}X = 0$.
- But we see $\overline{X} = 5.36$ and SEM$= 1.51$



- t-test: p-value of this deviation is ca. 0.14%.
- Thus, it seems that the direction has some effect.
- 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.

Null hypotheses

- H_0 for Codon-Bias: CCT and CCA have the same probability of $\frac{1}{2}$ Moreover: all positions decide independently between CCT and CCA
- H_0 for bird orientation and cork thickness: $\mathbb{E}X = 0$. Moreover: X normally distributed, X_i independent.

Deviations and p-values

- Codon Bias: Number of CCT deviates by 1092.5 from H_0-expected. binomial distribution model gives us $\sigma = \sqrt{n/4}$ an we can apply the z-test to compute the p-value: The probability that a bin$(n, \frac{1}{2})$-distributed random variable deviates from $n/2$ by 1092.5 or more.
- bird orientation and cork thickness:

$$t\text{-value} = \frac{\overline{X} - \mathbb{E}X}{s/\sqrt{n}}$$

  p-value: Probability, that a t-distributed value with $n-1$ df deviates from 0 at least as much as the t-value calculated from the observed data.

Statistical Testing: Imporatant 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 $\alpha = 0.05$).

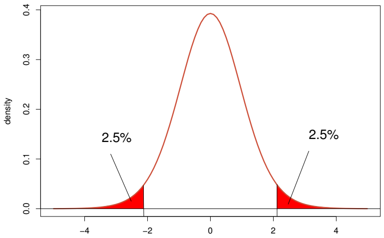
test statistic : measures how far the data deviates from what H_0 predicts into the direction of our alternative hypothesis.

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.

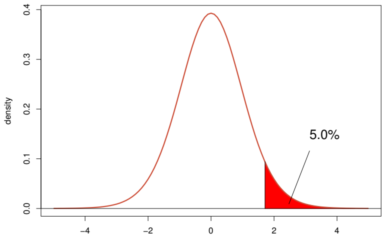
- We reject the null hypothesis H_0 if the p value is smaller than α.
- Thus, if H_0 is true, the probability to (falsely) reject it is α (not the p value).
- This entails that a researcher who performs many tests with $\alpha = 0.05$ on complete random data (i.e. where H_0 is always true), will falsely 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 expected value μ.



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



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

Examples

- If our original intention is to show that the flight direction of pied flycatchers varies more in blue light than in green light, then we can use a one-sided test.
- Even if the data then shows quite clearly that the opposite is true, we can not reject the null hypothesis that the color of the light has no effect (at least not if we really stick to the pure teachings).
- If we initially want to show that the cork on the northern aspect of the tree is thicker than on the western, we can use a one-sided test.
- If the data then clearly show that the opposite is true, we may, strictly speaking, not call this significant.

2 t-test for unpaired samples

2.1 H_0: means and variances are equal

`http://en.wikipedia.org/wiki/File:Tetranychus-urticae.jpg`
photo (c) by J. Holopainen

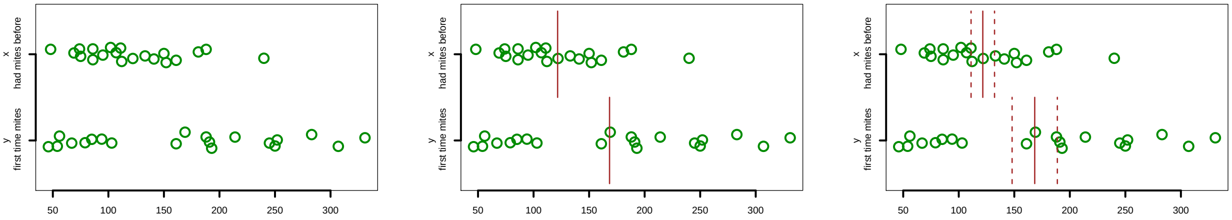
Example: Do mites favor plants that have not been infected with mites before?

Infect cotton plants with mites (*Tetranychus urticae*) and count mites on plants that had have mites before and plants that are infected for the first time.

Data for mites-cotton example are artificially generated by computer simulation inspired by actual experiments, cf. e.g.

References

[1] S. Harrison, R. Karban: Behavioral response of spider mites (Tetranychus urticae) to induced resistance of cotton plants *Ecological Entomology* **11**:181-188, 1986.



$$\mu(y) = 168.4$$
$$sd(y) = 91.09763$$
$$sd(y)/\sqrt{20} = 20.37005$$
$$\mu(x) = 121.65$$
$$sd(x) = 47.24547$$
$$sd(x)/\sqrt{20} = 10.56441$$

Our null hypothesis H_0:

All data values are independent samples from the same normal distribution.

This H_0 includes that the theoretical means in the two groups are equal, but also that the theoretical standard deviations are equal.

Theorem 1 (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

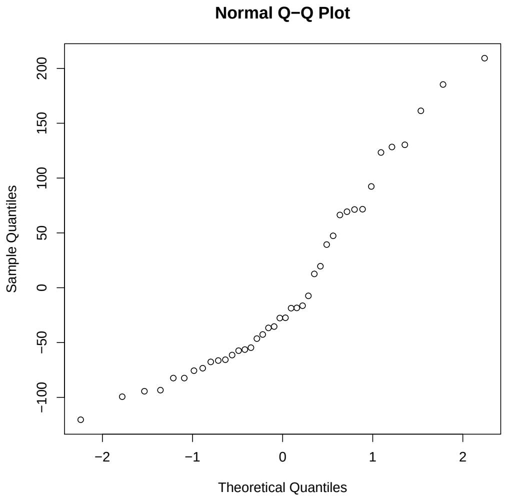
$$t = \frac{\overline{X} - \overline{Y}}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

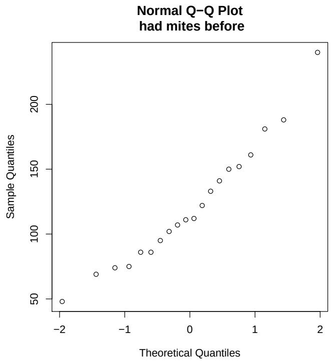
follows a t distribution with $n + m - 2$ *degrees of freedom.*

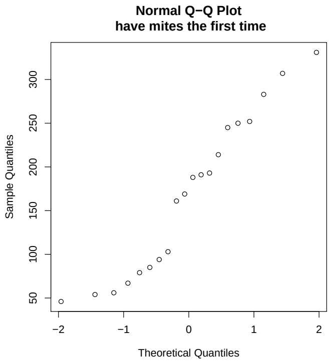
```
> t.test(y,x,var.equal=TRUE)

Two Sample t-test

data:  y and x
t = 2.0373, df = 38, p-value = 0.04862
alternative hypothesis: true difference in
 means is not equal to 0
95 percent confidence interval:
  0.2970719 93.2029281
sample estimates:
mean of x mean of y
   168.40    121.65
```







permutation test

1. let t_{orig} be the original t-value from comparison between plants that were infected for the first time and plants that had been infected before.
2. set counter k to 0
3. repeat the following steps e.g. 100000 times
   - randomly shuffle lables “first time infected” and “had mites before”
   - compute t-statistic t between these random groups
   - if $|t| \geq |t_{\text{orig}}|$ increase counter k by 1.
4. simulated p-value is $k/100000$, the fraction of simulated t-values that were more extereme than t_{orig}.

permutation test

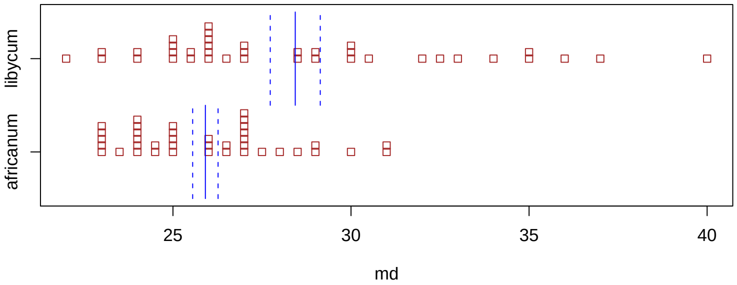
```
> t.orig <- t.test(cont,ex,var.equal=TRUE)$statistic
> count <- 0
> for( i in 1:100000) {
+   random.group <- sample(c(rep("cont",20),rep("ex",20)))
+   y <- all[random.group=="cont"]
+   x <- all[random.group=="ex"]
+   t <- t.test(x,y,var.equal=TRUE)$statistic
+   if( abs(t)>=abs(t.orig) ) {
+     count <- count+1
+   }
+ }
> count/100000
> [1] 0.04895
```

2.2 H_0: mean are equal, variances may be different

Hipparion dataset

- Hipparion is an extinct genus of horse.
- *Hipparion africanum* lived 4 mio years ago.

- *Hipparion libycum* lived 2.5 mio years ago.
- 2.8 mio years ago climate became cooler and dryer in east Africa.
- Hipparion ate less leaves and more grass.
- can we see this in fossil teeth?
- data: mesiodistal lengths from 77 back teeth which were found in Chiwondo Beds, Malawi, and are now in the Landesmuseum Darmstadt.



Theorem 2 (Welch's t-test)**.** *Suppose that $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ are independent and normally distributed random variables with $\mathbb{E}X_i = \mathbb{E}Y_j$ and potentially different variances $\mathrm{Var}X_i = \sigma_X^2$ and $\mathrm{Var}Y_i = \sigma_Y^2$. 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*

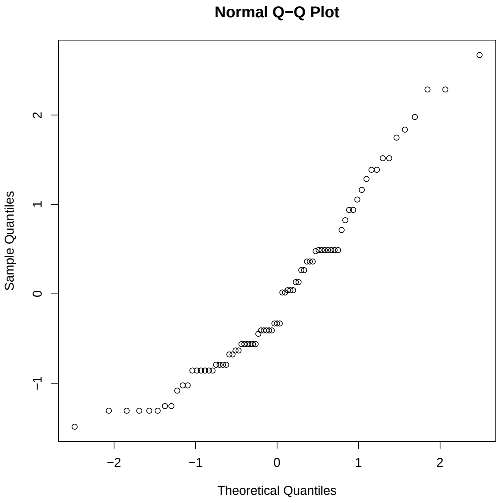
$$\frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2 \cdot (n-1)} + \frac{s_Y^4}{m^2 \cdot (m-1)}}$$

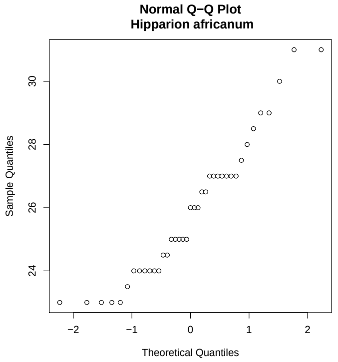
degrees of freedom.

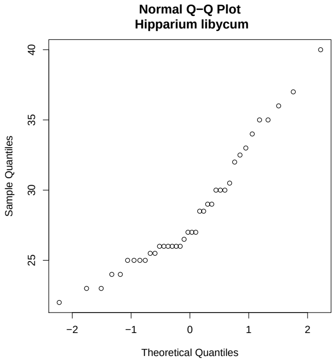
```
> t.test(md[Species=="africanum"],md[Species=="libycum"])

Welch Two Sample t-test

data:  md[Species == "africanum"] and
md[Species == "libycum"]
t = -3.2043, df = 54.975, p-value = 0.002255
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -4.1025338 -0.9453745
sample estimates:
mean of x mean of y
 25.91026  28.43421
```







- qqnorm plots show deviation from normality assumption
- not a priori clear if t-test is applicable
- permutation test: difficult because H_0 allows that standard deviations are different
- examine robustness of t-test by simulation studies
- or rescale data to make residuals normally distributed
- is non-parametric test applicable?

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 paired t-test

- summarizing measurements in a statistic
- what is a qqnorm plot and when and how should it be used to check normality
- the idea of permutation tests