Biostatistik

Der zwei-Stichproben-t-Test

(t-Test für ungepaarte Stichproben)

Matthias Birkner

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

1. t-Test für ungepaarte Stichproben
   - Beispiel: Backenzähne von Hipparions
   - Allgemein: ungepaarter *t*-Test mit Annahme gleicher Varianzen
   - Bericht: *t*-Test ohne Annahme gleicher Varianz
   - Power eines Tests
   - Vergleich: gepaarter *t*-Test und ungepaarter *t*-Test

Beispiel: Backenzähne von Hipparions

(c): public domain

Die Daten

77 Backenzähne

gefunden in den Chiwondo Beds, Malawi,

jetzt in den Sammlungen des
Hessischen Landesmuseums, Darmstadt

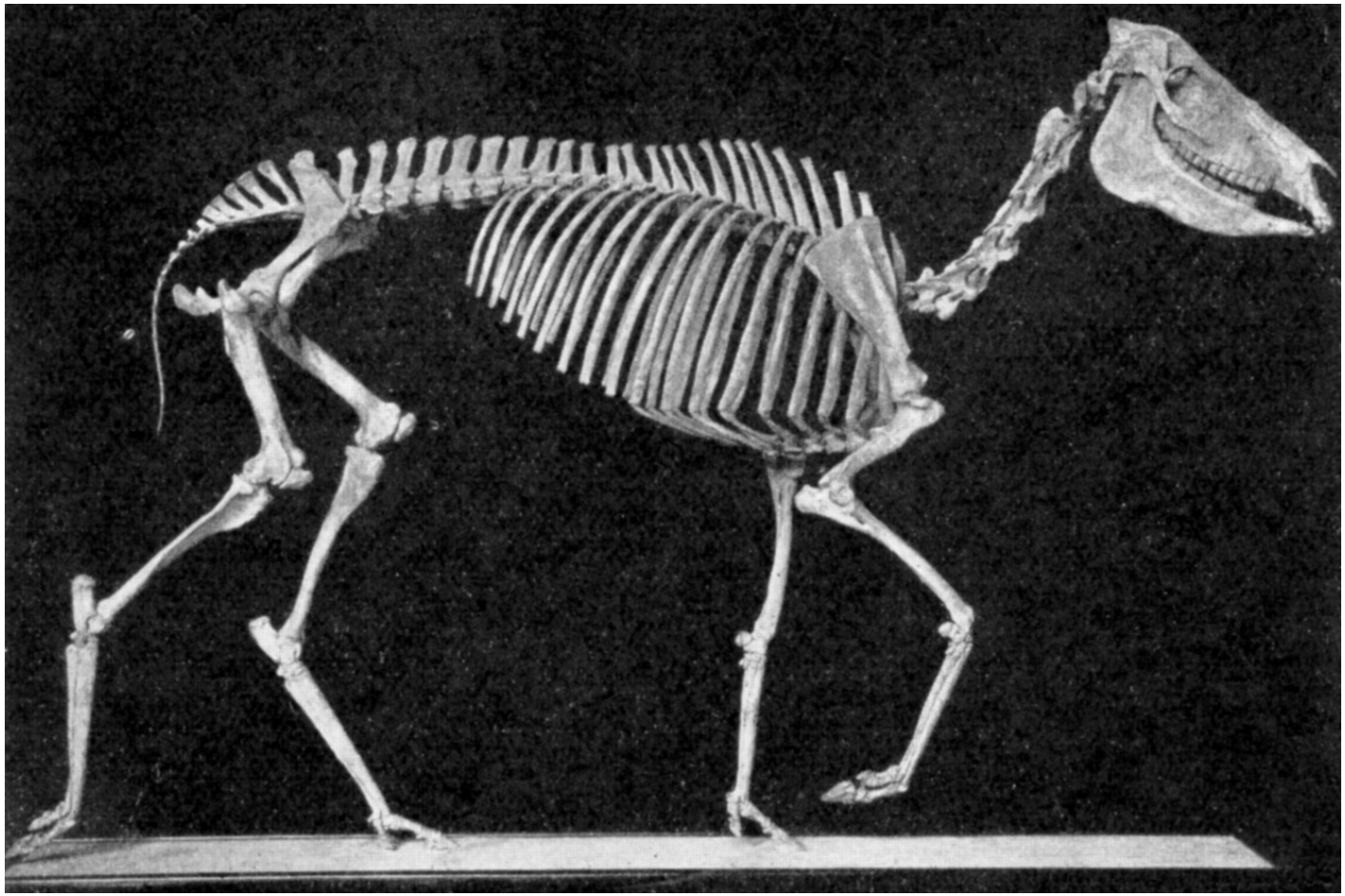
(c): Rei-artur

Zuordnung

Die Zähne wurden zwei Arten zugeordnet:

Hipparion africanum
$\approx$ 4 Mio. Jahre

Hipparion libycum
$\approx$ 2,5 Mio. Jahre



(c): public domain

Geologischer Hintergrund

Vor 2,8 Mio. Jahren kühlte sich das Klima weltweit ab.

Das Klima in Ostafrika:

warm-feucht $\longrightarrow$ kühl-trocken

Hipparion:

Laubfresser $\longrightarrow$ Grasfresser

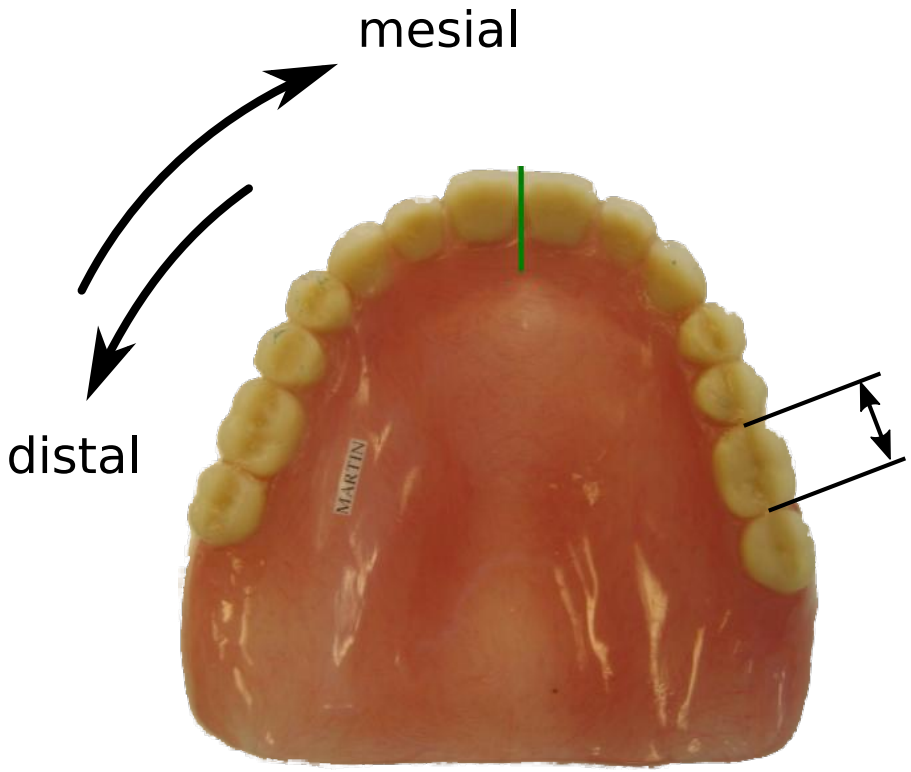
Frage

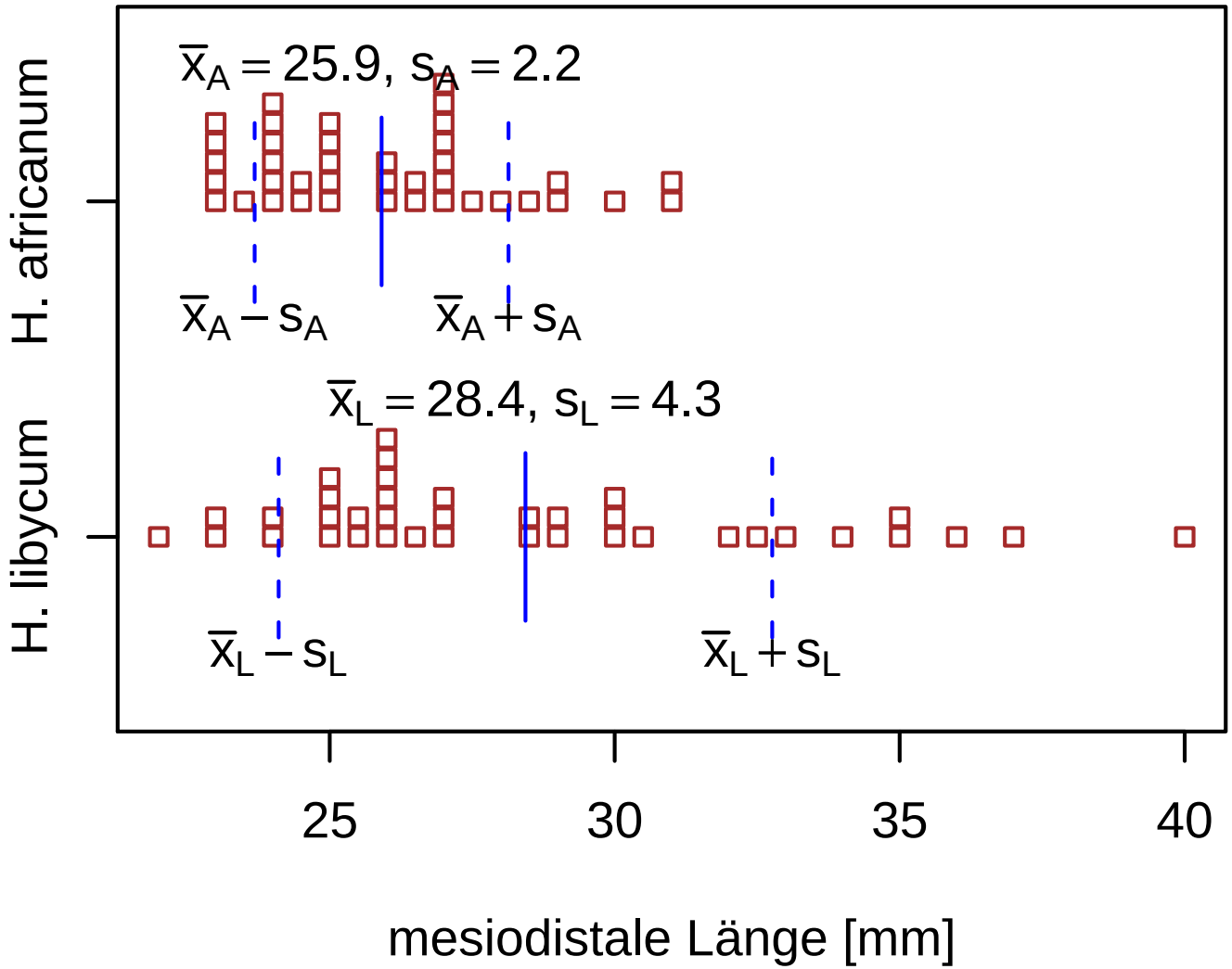
Hipparion:

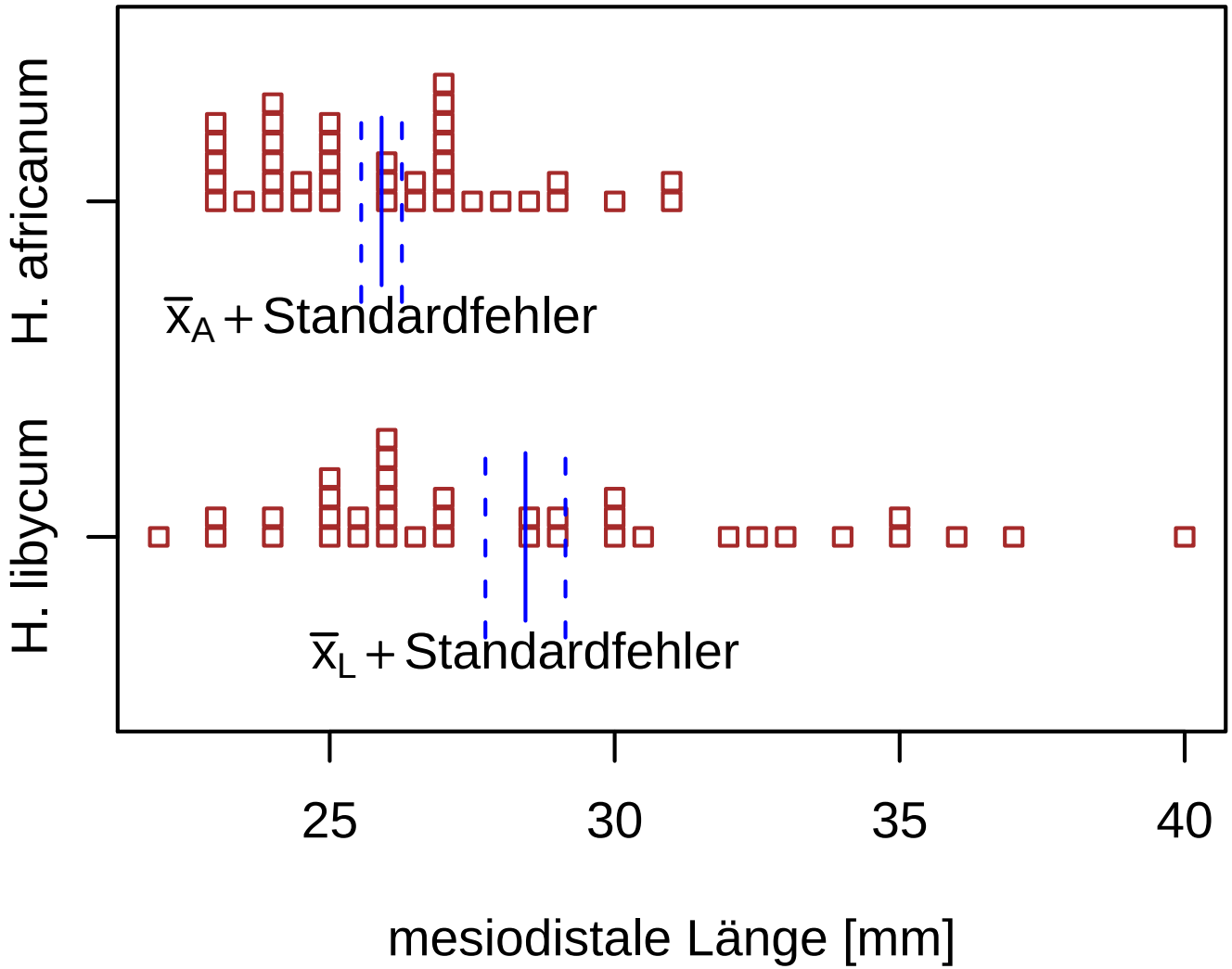
Laubfresser $\longrightarrow$ Grasfresser

andere Nahrung $\longrightarrow$ andere Zähne?

Messungen: mesiodistale Länge


mesial
distal
MARTIN


$\bar{x}_A = 25.9$, $s_A = 2.2$
$\bar{x}_A - s_A$
$\bar{x}_A + s_A$
$\bar{x}_L = 28.4$, $s_L = 4.3$
$\bar{x}_L - s_L$
$\bar{x}_L + s_L$
H. africanum
H. libycum
25
30
35
40
mesiodistale Länge [mm]


H. africanum
H. libycum
$\overline{x}_A$ + Standardfehler
$\overline{x}_L$ + Standardfehler
25
30
35
40
mesiodistale Länge [mm]

Wir beobachten ($n_A = 39$, $n_L = 38$):

$$\overline{x}_A = 25{,}9,\ s_A = 2{,}2$$

(unser Schätzwert für die Streung von $\overline{x}_A$ ist also $f_A = \frac{s_A}{\sqrt{n_A}} = 2{,}2/\sqrt{n_A} = 0{,}36$ (Standardfehler)),

$$\overline{x}_L = 28{,}4,\ s_L = 4{,}3$$

(unser Schätzwert für die Streung von $\overline{x}_L$ ist also $f_L = \frac{s_L}{\sqrt{n_L}} = 4{,}3/\sqrt{n_L} = 0{,}70$).

Ist die beobachtete Abweichung $\overline{x}_L - \overline{x}_A = 2{,}5$ mit der Nullhypothese verträglich, dass $\mu_L = \mu_A$?

Die „Bilderbuchsituation“

Annahme: Wir haben zwei unabhängige Stichproben $x_{1,1}, \ldots, x_{1,n_1}$ und $x_{2,1}, \ldots, x_{2,n_2}$.

Die $x_{1,i}$ stammen aus einer Normalverteilung mit (unbekanntem) Mittelwert μ_1 und (unbekannter) Varianz $\sigma^2 > 0$, die $x_{2,i}$ aus einer Normalverteilung mit (unbekanntem) Mittelwert μ_2 und derselben Varianz σ^2.

Seien

$$\overline{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1,i}, \ \overline{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2,i}$$

die jeweiligen Stichprobenmittelwerte,

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1,i} - \overline{x}_1)^2, \ s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2,i} - \overline{x}_2)^2,$$

die (korrigierten) Stichprobenvarianzen.

Wir möchten die Hypothese H_0 : „$\mu_1 = \mu_2$“ prüfen.

Wenn $\mu_1 = \mu_2$ gilt, so sollte $\overline{x}_1 = \overline{x}_2$ „bis auf Zufallsschwankungen“ gelten, denn $\mathbb{E}[\overline{X}_1] = \mu_1$, $\mathbb{E}[\overline{X}_2] = \mu_2$.

Was ist die Skala der typischen Schwankungen von $\overline{x}_1 - \overline{x}_2$?

$$\mathrm{Var}\left[\overline{X}_1 - \overline{X}_2\right] = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Problem (wie bereits im ein-Stichproben-Fall): Wir kennen σ^2 nicht.

Wir schätzen es im zwei-Stichproben-Fall durch die gepoolte Stichprobenvarianz

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(= \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1} (x_{1,i} - \overline{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2,i} - \overline{x}_2)^2 \right) \right)$$

und bilden die Teststatistik

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Es gilt dann: Wenn $\mu_1 = \mu_2$ gilt, so ist

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Student-verteilt mit $n_1 + n_2 - 2$ Freiheitsgraden.

Verfahren des (zweiseitigen) zwei-Stichproben t-Tests (mit Annahme gleicher Varianzen) also:

Lehne die Nullhypothese $\mu_1 = \mu_2$ zum Signifikanzniveau α ab, wenn der beobachtete Wert von $|t|$ größer ist als das $(1 - \alpha/2)$-Quantil der Student-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden.

Sei $\alpha = 0{,}01$, das 99,5%-Quantil der Student-Vert. mit 75 Freiheitsgraden ist 2,64.

Im Hipparion-Beispiel war $\overline{x}_L = 28{,}4$, $\overline{x}_A = 25{,}9$, $s_A = 2{,}2$, $s_A = 4{,}3$.

Wir finden $t = 3{,}2$.

Wir können die Nullhypothese „die mittlere mesiodistale Länge bei *H. lybicum* und bei *H. africanum* sind gleich“ zum Signifikanzniveau 1% ablehnen.

Mögliche Formulierung:

„Die mittlere mesiodistale Länge
war signifikant größer (28,4 mm) bei *H. libycum*
als bei *H. africanum* (25,9 mm)
(zweiseitiger t-Test, $\alpha = 0{,}01$).“

t-Statistik ohne Annahme gleicher Varianz

Es gibt auch eine Version des zwei-Stichproben-t-Tests, der die Annahme gleicher Varianzen nicht trifft (Welchs t-Test):

Ist die beobachtete Abweichung $\overline{x}_L - \overline{x}_A = 2{,}5$ mit der Nullhypothese verträglich, dass $\mu_L = \mu_A$?

Wir schätzen die Streuung von $\overline{x}_L - \overline{x}_A$ durch f, wo

$$f^2 = f_L^2 + f_A^2$$

(mit $f_L = \frac{s_L}{\sqrt{n_L}}$, $f_A = \frac{s_A}{\sqrt{n_A}}$ den zugehörigen Standardfehlern)

und bilden $t = \dfrac{\overline{x}_L - \overline{x}_A}{f}$.

Wenn die Nullhypothese zutrifft, ist t „Student-verteilt mit g Freiheitsgraden."

(wobei g aus den Daten geschätzt wird, $g = \dfrac{\left(\frac{s_A^2}{n_A} + \frac{s_L^2}{n_L}\right)^2}{\frac{s_A^4}{n_A^2(n_A-1)} + \frac{s_L^4}{n_L^2(n_L-1)}}$)

Zwei-Stichproben-*t*-Test mit R

```
> A <- md[Art=="africanum"]
> L <- md[Art=="libycum"]
> t.test(L,A)

Welch Two Sample t-test

data:  L and A
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:
 0.9453745 4.1025338
sample estimates:
mean of x mean of y
 28.43421  25.91026
```

Mögliche Formulierung:

„Die mittlere mesiodistale Länge
war signifikant größer (28,4 mm) bei *H. libycum*
als bei *H. africanum* (25,9 mm)
(zweiseitiger *t*-Test, $p = 0{,}002$).“

Wir haben keine Annahme an die Varianzen der beiden Stichproben gemacht (und dann verwendet R standardmäßig den Welch-*t*-Test). R bietet auch die Variante des *t*-Tests mit der zusätzlichen Annahme $\sigma_A^2 = \sigma_L^2$:

```
> t.test(L,A,var.equal=TRUE)

Two Sample t-test

data:  L and A
t = 3.2289, df = 75, p-value = 0.001845
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 0.9667634 4.0811448
sample estimates:
mean of x mean of y
 28.43421  25.91026
```

Testpower bzw. Testmacht

Salopp gesprochen ist die
Power oder Macht eines Tests
die Wahrscheinlichkeit, die Nullhypothese abzulehnen,
falls die Alternative zutrifft.

Bei einer einelementigen Alternative
ist dies leicht zu formulieren: $H_0 : \mu = 0$
$H_1 : \mu = m_1$

Die Testpower (oder auch Testmacht) ist dann definiert als
$\mathbb{P}_{H_1}$(Nullhypothese wird abgelehnt)

Warum interessiert uns die Testmacht?

Im Extremfall ist die Testmacht gleich 0,
dann wird die Nullhypothese nie abgelehnt.
Somit können wir unsere Vermutung nicht stützen.

Je größer die Testmacht,
desto wahrscheinlicher wird die Nullhypothese abgelehnt.
Beachte: Die Testmacht hängt stark
von der Stichprobenlänge ab.

In der Praxis muss man sich bereits vor Versuchsbeginn
Gedanken machen, wie groß die Stichprobenlänge sein muss,
damit man die Vermutung stützen kann.

Wann gepaarter *t*-Test und wann ungepaarter *t*-Test?

Wenn die Stichprobenlänge unterschiedlich ist,
ergibt „gepaart“ keinen Sinn.

Wenn die Stichprobenlänge gleich ist:

- Sind die Stichproben unabhängig voneinander?
  Falls ja, dann ungepaart testen. Ein gepaarter Test würde sinnlose Abhängigkeiten unterstellen und hätte wegen der geringeren Anzahl Freiheitsgrade auch eine geringere Macht.
- Sind die Stichproben voneinander abhängig?
  (z.B. Messungen von denselben Individuen bzw. Objekten)
  Falls ja, dann ist ein gepaarter Test angemessen. Bei starker Abhängigkeitsstruktur hat der gepaarte *t*-Test höhere Macht (da der Test von Variabilität zwischen den Individuen bereinigt ist), zudem würde der ungepaarte Test die Varianz der Differenz der Mittelwerte „falsch“ schätzen und damit das Niveau verzerren.