Biostatistik, WS 2010/2011

Der t-Test

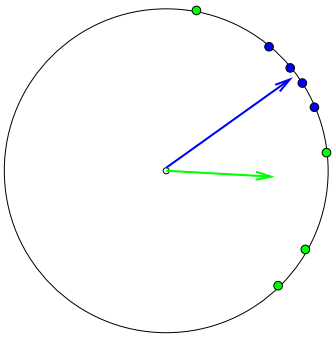
(Zwei gepaarte Stichproben bzw. eine Stichprobe)

Matthias Birkner

http://www.mathematik.uni-mainz.de/~birkner/Biostatistik1011/

10.12.2010

JOHANNES GUTENBERG UNIVERSITÄT MAINZ



Pfeilspitze: Schwerpunkt der Austrittspunkte bei grünem Licht. Dasselbe für die “blauen” Austrittspunkte.

Je variabler die Richtungen, desto kürzer der Pfeil!

Fragestellung

Hat die Farbe der monochromatischen Beleuchtung einen Einfluss auf die Orientierung?
Experiment: Bei 17 Vögeln wurde die Länge des Schwerpunktsvektors sowohl bei blauem als auch bei grünem Licht bestimmt.

Wir verwenden im Folgenden simulierte Daten, die sich an der Literatur orientieren.

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.

Wiltschko, R.; Ritz, T.; Stapput, K.; Thalau, P.; Wiltschko, W.

Two different types of light-dependent responses to magnetic fields in birds.

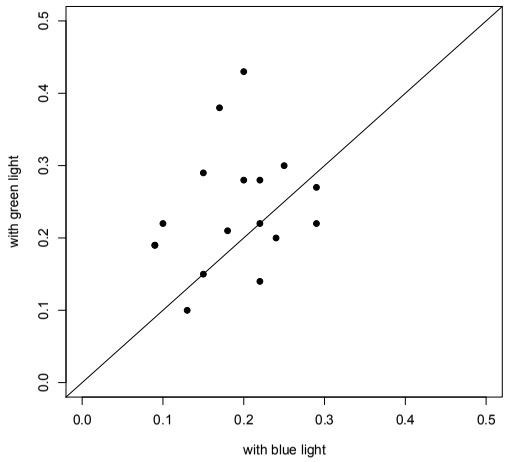
Curr Biol 15.16, pp. 1518-23, 2005.

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.

Trauerschnäpper:
Länge des Schwerpunktsvektors
bei grünem und bei blauem Licht, n=17

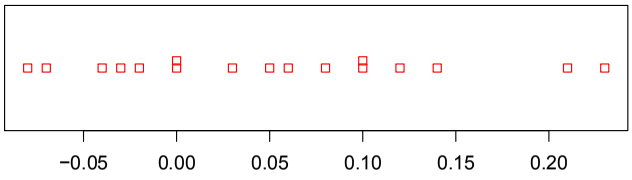


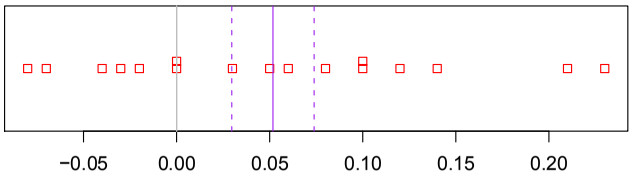
Wie kann ich
statistisch testen,
ob die Farbe
einen Einfluss hat?

Wir berechnen nun für jeden Vogel den Abstand des Punktes von der Diagonale,

d.h.

$$x := \text{“Grünwert”} - \text{“Blauwert”}$$





Kann der wahre Mittelwert $\mu = 0$ sein?

$$\begin{aligned} \overline{x} &= 0.0518 \\ s &= 0.0912 \\ \text{Standardfehler} &= \frac{s}{\sqrt{n}} = \frac{0.912}{\sqrt{17}} = 0.022 \end{aligned}$$

Ist $|\overline{x} - \mu| \approx 0.0518$ eine große Abweichung?

Groß? Groß im Vergleich zu was?

In welcher Vergleichseinheit soll $|\overline{x} - \mu|$ gemessen werden?

Im Vergleich zum
Standardfehler!

$|\overline{x} - \mu|$
gemessen in der Einheit 'Standardfehler'
heißt t-Statistik

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

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

$t = 1$ bedeutet
1 Standardfehler von μ entfernt
(kommt häufig vor)

$t = 3$ bedeutet
3 Standardfehler von μ entfernt
(kommt selten vor)

In unserem Fall:

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

Also: $\overline{x}$ ist mehr als 2.3 Standardfehler von $\mu = 0$ entfernt.

Wie wahrscheinlich ist das, wenn 0 der wahre Mittelwert ist?

anders gefragt:

Ist diese Abweichung signifikant?

Für die Antwort benötigen wir die Verteilung der t-Statistik.

Wir wissen:

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

ist (für große n) ungefähr standardnormalverteilt (dies ist die Grundlage des z-Tests).

Aber:
Die t-Statistik ist jedoch mit s an Stelle von σ definiert (und nicht normalverteilt), die Approximation mit der Normalverteilung ist (speziell für kleine und „mittelgroße“ Stichprobengrößen n) häufig zu grob.

Allgemein gilt

Sind $X_1, \ldots, X_n$ unabhängig aus einer Normalverteilung mit Mittelwert μ gezogen (und beliebiger Varianz $\sigma^2 > 0$), so ist

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

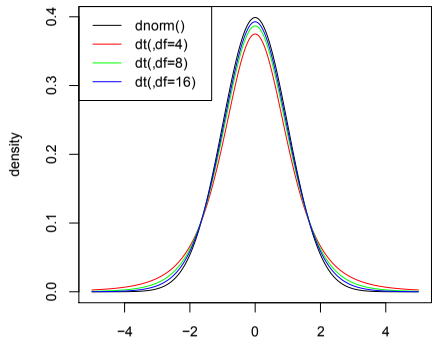
t-verteilt mit $n - 1$ Freiheitsgraden (df=*degrees of freedom*).

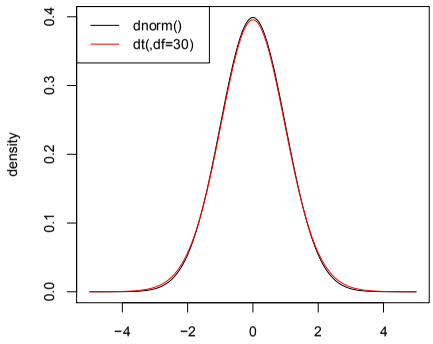
Eine t-verteilte Zufallsvariable bezeichnen wir meist mit T.

Die t-Verteilung heißt auch Student-Verteilung. Sie wurde von William Gosset erforscht und von ihm 1908 veröffentlicht, während er in einer Guinness-Brauerei arbeitete. Da sein Arbeitgeber die Veröffentlichung nicht gestattete, veröffentlichte Gosset sie unter dem Pseudonym *Student*.

W.S. Gossett, 1876–1937

Dichte der t-Verteilung





Freiheitsgrade?

Beispiel: Es gibt 5 Freiheitsgrade im Vektor

$$x = (x_1, x_2, x_3, x_4, x_5)$$

da 5 Werte frei wählbar sind. Der Vektor

$$v := x - \overline{x} := (x_1 - \overline{x}, x_2 - \overline{x}, x_3 - \overline{x}, x_4 - \overline{x}, x_5 - \overline{x})$$

hat 4 Freiheitsgrade, denn nach Wahl von v_1, v_2, v_3, v_4 ist v_5 festgelegt wegen $v_1 + \cdots + v_4 + v_5 = 0$.

Die Bezeichnung „Student-Verteilung mit $n-1$ Freiheitsgraden“ ist motiviert durch die Tatsache, dass $t = (\overline{x} - \mu)/(s/\sqrt{n})$, wo $s = \sqrt{s^2} = \frac{1}{\sqrt{n-1}}\left((x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right)^{1/2}$ (bis auf Normierung) die Länge eines „Vektors mit $n-1$ Freiheitsgraden“ ist.

Wir meinen:
Die Farbe der Beleuchtung
hat einen Einfluss auf die Orientierung

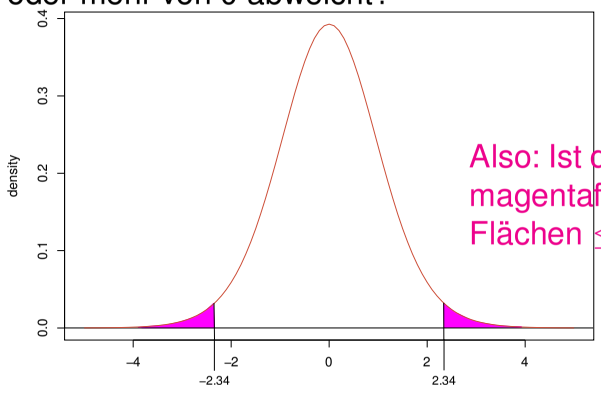
Ein Skeptiker würde erwidern:
Alles nur Zufall

Wir wollen nun zeigen:
Unter der Annahme 'Kein Einfluss'
ist die Beobachtung sehr unwahrscheinlich

Nullhypothese: $\mu = 0$

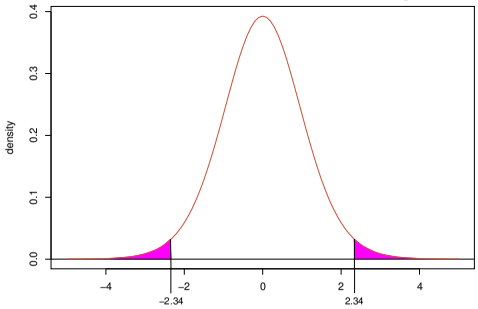
Wie (un)wahrscheinlich ist nun eine mindestens
so große Abweichung wie 2.34 Standardfehler?
Idee des statistischen Testens: Wir legen ein Signifikanzniveau $\alpha \in (0, 1)$, z.B. $\alpha = 5\%$, fest und fragen:

Wenn das wahre $\mu = 0$ ist (d.h. unter der Nullhypothese), wie wahrscheinlich ist es, einen t-Wert zu beobachten, der um 2,34 oder mehr von 0 abweicht?



Also: Ist der Gesamtinhalt der
magentafarbenen
Flächen $\leq \alpha$?

Beobachtet: $t = 2{,}34$
Ist der Gesamtinhalt der magentafarbenen Flächen $\leq \alpha$?



Das 97,5%-Quantil der Student-Verteilung mit 16 Freiheitsgraden ist ca. 2,12 $\leq$ 2,34, also: Ja.

Genauer: Es gilt $\mathbb{P}(|T| \geq 2.34) \approx 0{,}0325$ ($\leq 0{,}05$).
Man nennt $\mathbb{P}(|T| \geq t)$ den p-Wert (bei beobachtetem Wert t).

Wir halten fest:

$$p - \text{Wert} = 0.03254 \ (\leq 0.05)$$

Wenn die **Nullhypothese** “alles nur Zufall” (hier $\mu = 0$) gilt, dann ist eine mindestens so große Abweichung sehr unwahrscheinlich.

Sprechweise:

Wir verwerfen die Nullhypothese
auf dem 5%-Signifikanzniveau.

Oder:

Die Differenz zwischen grün und blau
ist auf dem 5%-Niveau signifikant.

Die Nullhypothese wurde also auf dem 5%-Niveau verworfen.
Welche Aussagen sind wahr/sinnvoll?

- Die Nullhypothese ist falsch. ~~Die Nullhypothese ist falsch.~~
- Die Nullhypothese ist mit 95%-iger Ws falsch.
  ~~Die Nullhypothese ist mit 95%-iger Ws falsch.~~
- Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in höchstens 5% der Fälle. Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in höchstens 5% der Fälle. ✓
- Die Orientierung der Vögel ist bei blau und grün verschieden.
  ~~Die Orientierung der Vögel ist bei blau und grün verschieden.~~
- Die Orientierung bei grün und blau war in dem Experiment auf dem 5%-Niveau signifikant verschieden. Die Orientierung bei grün und blau war in dem Experiment auf dem 5%-Niveau signifikant verschieden.✓

Man könnte auch ein anderes Signifikanzniveau α wählen.
Dann müsste man zeigen, dass der p-Wert kleiner als α ist.

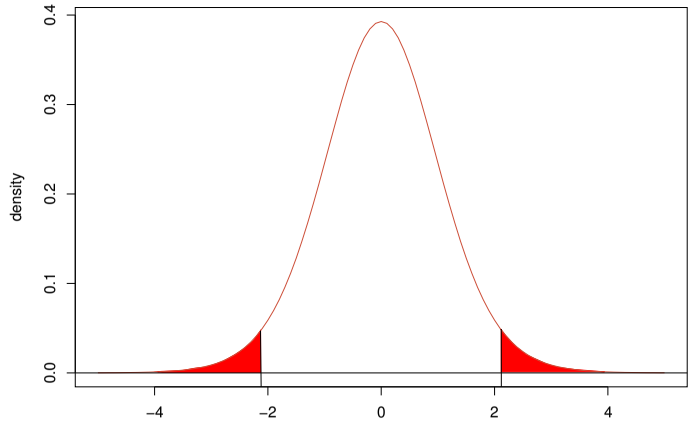
Wichtig: Wähle zuerst das Signifikanzniveau und ermittle erst dann den p-Wert! Das Signifikanzniveau je nach p-Wert zu wählen ist geschummelt.

In der Literatur wird üblicherweise
5% als Signifikanzniveau gewählt.

Beachte:

Falls die Nullhypothese zutrifft,
ist die Wahrscheinlichkeit,
dass wir sie zu Unrecht auf dem 5%-Niveau verwerfen,
höchstens 5%.

Wir verwerfen also die Nullhypothese auf 5%-Niveau, wenn der Wert der t-Statistik in den roten Bereich fällt:



(hier am Beispiel der $t-$Verteilung mit df$=$ 16 Freiheitsgraden)

Welche t-Werte sind auf dem 5%-Niveau signifikant?

Anzahl Freiheitsgrade	$\lvert t\rvert \geq \ldots$
5	2.57
10	2.23
20	2.09
30	2.04
100	1.98
„∞“	1.96

Diese sog. kritischen Werte (für das zweiseitige 5%-Signifikanzniveau beim t-Test) sind die 97,5%-Quantile der t-Verteilung mit der jeweils angegebenen Anzahl Freiheitsgrade.

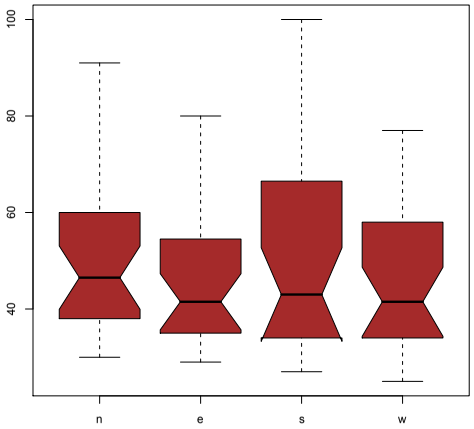
In der Praxis entnimmt man sie einem Statistik-Computerprogramm oder einer Tabelle.

Bei $n = 28$ Bäumen wurden die Korkdicken [mm] in den vier Himmelsrichtungen gemessen:

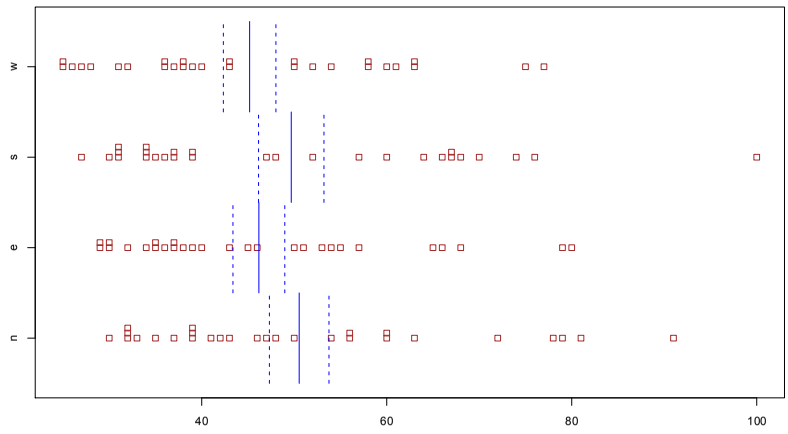
n	o	s	w
72	66	76	77
60	53	66	63
5	57	64	58
41	29	36	38
32	32	35	36
30	35	34	26
39	39	31	27
.	.	.	.
.	.	.	.

(Wir verwenden wieder simulierte Daten, die aber Daten aus echten Studien nachempfunden sind, auch im Ergebnis.)

Korkdicken nach Himmelsrichtung getrennt



Kann da was signifikant unterschiedlich sein???



Stripchart der Korkdicken je nach Himmelsrichtung und Mittelwerten $\pm$ Standardfehler

Kann da was signifikant unterschiedlich sein???

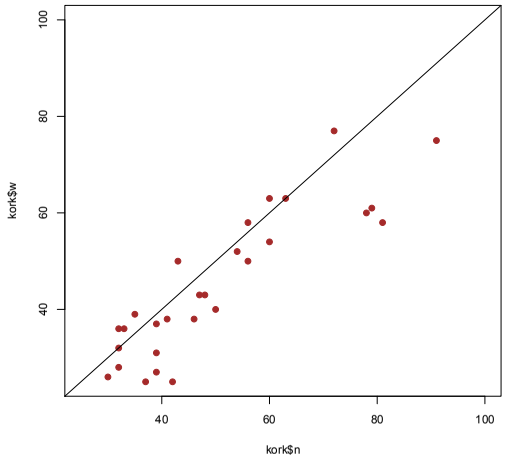
Haben wir irgend etwas übersehen?

Wir haben bisher vernachlässigt,
welche Werte von demselben Baum kommen!
Die Bäume unterscheiden sich sehr in ihrer Größe und Dicke.

Vergleiche also jeweils Paare von Korkdicken,
die von demselben Baum kommen!
($\rightsquigarrow$ gepaarter t-Test)

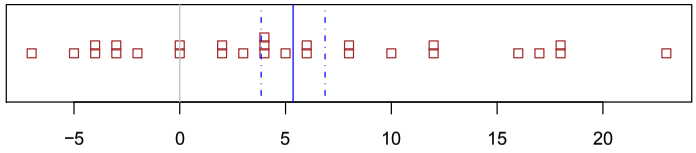
Korkdicken [mm] bei $n = 28$ Bäumen

Korkdicke an der Westseite



Korkdicke an der Nordseite

Differenz der Korkdicken an der Nord- und der Westseite für $n = 28$ Bäume



und Mittelwert±Standardfehler

Ist die Differenz signifikant von 0 verschieden?

$$
\begin{aligned}
x \;:=\; &(\text{Korkdicke Nordseite}) \;-\; (\text{Korkdicke Westseite}) \\
\overline{x} &\approx 5{,}36 \\
s_x &\approx 7{,}99 \\
\frac{s_x}{\sqrt{n}} &\approx 1{,}51 \\
t-\text{Wert} \;=\; \frac{\overline{x}}{s_x/\sqrt{n}} &\approx 3{,}547 \\
\text{Anzahl Freiheitsgrade: df} \;=\; n-1 \;&=\; 27
\end{aligned}
$$

Das 97,5%-Quantil der Student-t-Verteilung mit 27 Freiheitsgraden ist $\approx 2{,}05$,

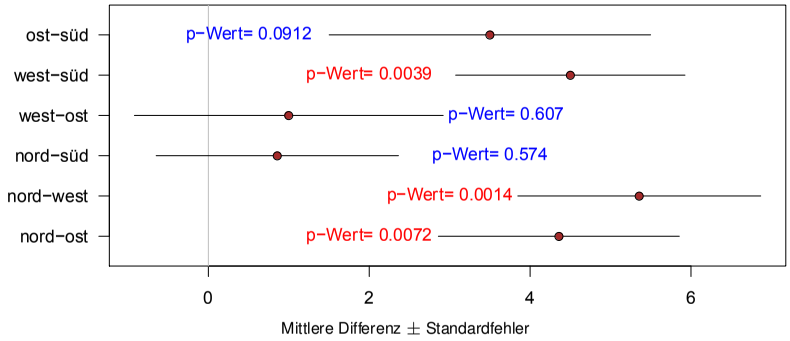
$$|3{,}547| > 2{,}05.$$

Die Differenz der Korkdicke an Nord- und Westseite ist auf dem 5%-Niveau signifikant verschieden.

Bemerkung: Die 6 möglichen Paarungen im Kork-Beispiel

Paarung	N-O	N-W	N-S	W-O	W-S	O-S
Mittlere Differenz $\overline{x}$	4,36	5,36	0,86	1,00	4,50	3,50
Standardfehler $s/\sqrt{n}$	1,50	1,51	1,51	1,92	1,43	2,00
$t = \frac{\overline{x}-0}{s/\sqrt{n}}$	2,91	3,55	0,57	0,52	3,16	1,75
$\lvert t\rvert \geq 2{,}052$?	ja	ja	nein	nein	ja	nein

(Das 97,5%-Quantil der Student-Verteilung mit 27 Freiheitsgraden ist 2,052.)



Zusammenfassung gepaarter t-Test

Gegeben: gepaarte Beobachtungen

$$(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$$

Nullhypothese H_0: $\mu_Y = \mu_Z$
Signifikanzniveau: α (meist $\alpha = 5\%$)
Test: gepaarter t-Test (genauer: zweiseitiger gepaarter t-Test)

Berechne Differenz $X := Y - Z$
Berechne Teststatistik

$$t := \frac{\overline{X}}{s(X)/\sqrt{n}}$$

Verwirf Nullhypothese, falls $|t| \geq (1 - \alpha/2)$-Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden
(d.h. wenn der p-Wert $\mathbb{P}(|T_{n-1}| \geq |t|) \leq \alpha$ ist)

Zusammenfassung Ein-Stichproben t-Test

Gegeben: Beobachtungen

$$X_1, X_2, \ldots, X_n$$

Nullhypothese H_0: $\mu_X = c$ (Den Wert c kennt man, oft $c = 0$)
Signifikanzniveau: α (meist $\alpha = 5\%$)
Test: t-Test

Berechne Teststatistik

$$t := \frac{\overline{X} - c}{s(X)/\sqrt{n}}$$

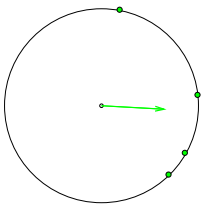
Verwirf Nullhypothese, falls $|t| \geq (1 - \alpha/2)$-Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden
(d.h. wenn der p-Wert $\mathbb{P}(|T_{n-1}| \geq |t|) \leq \alpha$ ist)

Beispiel: Codon Bias

- Wir beobachten 101844 mal CCT und 106159 mal CCA
- Wenn beide eigentlich gleich wahrscheinlich sind, erwarten wir 104001.5 von jedem.
- Die Beobachtung weicht um 2156 von diesem Erwartungswert ab
- z-Test: Die Wahrscheilichkeit einer mindestens so großen Abweichung ist kleiner als 10^{-20}
- Also sind CCT und CCA wohl nicht gleich wahrscheinlich.

Beispiel: Zugvogelorientierung

- Wie variabel ist die Abflugrichtung bei grünem und bei blauem Licht.
- Wir messen die Variabilität durch die Länge des Schwerpunktsvektors.

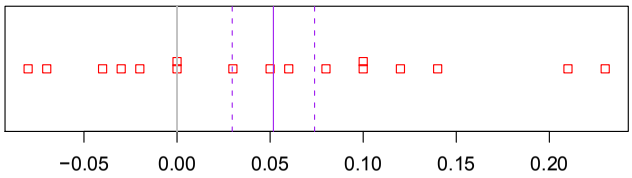


- Quantifiziere Unterschied durch $X =$(Länge grün)$-$ (Länge blau).
- Wenn das Licht keinen Einfluss hat, gilt $\mathbb{E}X = 0$.

Beispiel: Zugvogelorientierung

$X =$(Länge grün)$-$ (Länge blau)

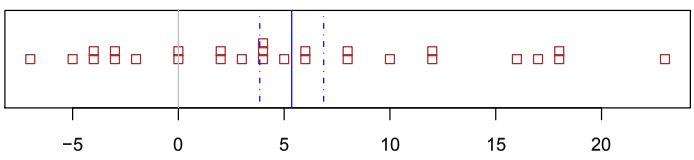
- Wenn das Licht keinen Einfluss hat, gilt $\mathbb{E}X = 0$.
- Wir beobachten aber $\overline{X} = 0.0518$ und SEM=0.022



- t-Test: p-Wert dieser Abweichung ist ca. 3.3%.
- Vermutlich hat die Lichtfarbe also doch einen Einfluss

Beispiel: Dicke des Korks

- X=(Korkdicke an der Nordseite)$-$ (Korkdicke an der Westseite)
- Wenn die Seite keine Rolle spielt, ist $\mathbb{E}X = 0$.
- Wir sehen aber $\overline{X} = 5.36$ und SEM$= 1.51$



- t-Test: p-Wert dieser Abweichung ist ca. 0.14%.
- Also hat die Himmelsrichtung wohl doch einen Einfluss.

Prinzip des statistischen Testens

- Wir möchten belegen, dass eine Abweichung in den Daten vermutlich nicht allein auf Zufallsschwankung beruht.
- Dazu spezifizieren wir zunächst eine Nullhypothese H_0, d.h. wir konkretisieren, was “allein auf Zufall beruhen” bedeutet.
- Dann versuchen wir zu zeigen: Wenn H_0 gilt, dann sind Abweichungen, die mindestens so groß sind wie die beobachtete, sehr unwahrscheinlich.
- Wenn uns das gelingt, verwerfen wir H_0.
- Was wir als Abweichung auffassen, sollte klar sein, bevor wir die Daten sehen.

Nullhypothesen

- H_0 bei Codon-Bias: CCT und CCA haben jeweils W’keit $\frac{1}{2}$ Außerdem: alle Positionen entscheiden unabhängig zwischen CCT und CCA
- H_0 bei Vogelorientierung und Korkdicken: $\mathbb{E}X = 0$. Außerdem: X normalverteilt, X_i unabhängig.

Abweichungen und p-Werte

- Codon Bias: Anzahl CCT weicht um 2156 vom Mittelwert ab.
  Wegen der Binomialverteilungsannahme gehen wir von festem σ aus und berechnen mit dem z-Test den p-Wert: Die Wahrscheinlichkeit, dass eine $\mathrm{bin}(n, \frac{1}{2})$-verteilte Zufallsgröße um mindestens 2156 von $n/2$ abweicht.
- Vogelorientierung und Korkdicke:

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

  p-Wert: W'keit, dass t-Wert bei $n-1$ mindestens so stark von 0 abweicht wie beobachtet.

Zweiseitig oder einseitig testen?

In den meisten Fällen will man testen, ob zwei Stichproben sich signifikant unterscheiden.
$\rightsquigarrow$ zweiseitiger Test

In manchen Fällen

- kann man von vornherein ausschließen, dass die erste Stichprobe kleinere Werte als die zweite Stichprobe hat. Dann will man testen, ob die erste Stichprobe signifikant größer ist.
- will man nur testen, ob die erste Stichprobe signifikant größer ist.
- will man nur testen, ob die erste Stichprobe signifikant kleiner ist.

$\rightsquigarrow$ einseitiger Test

Beispiel für einseitigen Test:

Man will zeigen,
dass ein Wachstumhormon wirkt,
also kein Placebo ist.

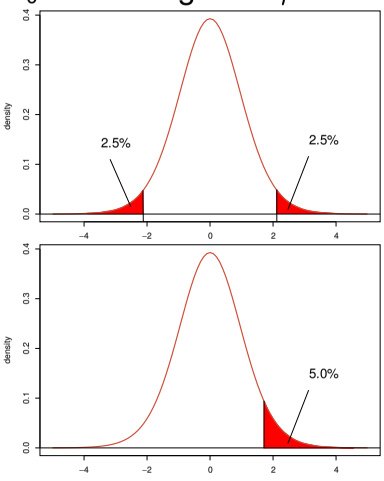
Dazu müssen die Größen Y in der behandelten Gruppe
signifikant größer sein
als die Größen Z in der Kontrollgruppe.
Die zu entkräftende Nullhypothese wäre hier:

Nullhypothese $\mu_Y \leq \mu_Z$

Definiere die Differenz $X := Y - Z$.

Zweiseitig oder einseitig testen?

Wir beobachten einen Wert x, der deutlich größer als der H_0-Erwartungswert μ ist.



p-Wert=$\mathbb{P}_{H_0}(|X - \mu| \geq |x - \mu|)$

p-Wert=$\mathbb{P}_{H_0}(X - \mu \geq x - \mu)$

Reine Lehre des statistischen Testens

- Formuliere eine Nullhypothese H_0, z.B. $\mu = 0$.
- Lege ein Signifikanzniveau α fest; üblich ist $\alpha = 0.05$.
- Lege ein Ereignis $\mathcal{A}$ fest, so dass

$$\Pr_{H_0}(\mathcal{A}) = \alpha$$

(oder zumindest $\mathbb{P}_{H_0}(\mathcal{A}) \leq \alpha$).
z.B. $\mathcal{A} = \{\overline{X} > q\}$ oder $\mathcal{A} = \{|\overline{X} - \mu| > r\}$

- ERST DANN: Betrachte die Daten und überprüfe, ob $\mathcal{A}$ eintritt.
- Dann ist die Wahrscheinlichkeit, dass H_0 verworfen wird, wenn H_0 eigentlich richtig ist ("Fehler erster Art") , lediglich α.

Verstöße gegen die reine Lehre

"Beim zweiseitigen Testen kam ein *p*-Wert von 0.06 raus. Also hab ich einseitig getestet, da hat's dann funktioniert."

genauso problematisch:

"Beim ersten Blick auf die Daten habe ich sofort gesehen, dass $\overline{x}$ größer ist als μ_{H_0}. Also habe ich gleich einseitig getestet"

Wichtig

Die Entscheidung, ob einseitig oder zweiseitig getestet wird, darf nicht von den konkreten Daten abhängen, die zum Test verwendet werden.
Allgemeiner: Ist $\mathcal{A}$ das Ereignis, dass zum Verwerfen von H_0 führt (falls es eintritt), so muss die Festlegung von H_0 stattfinden bevor man die Daten betrachtet hat.

Die Wahl von $\mathcal{A}$ sollte von der Alternative H_1 abhängen, also davon, was wir eigentlich zeigen wollen, indem wir H_0 durch einen Test verwerfen. Es muss gelten:

$$\mathbb{P}_{H_0}(\mathcal{A}) \leq \alpha$$

und

$$\mathbb{P}_{H_1}(\mathcal{A}) = \text{möglichst groß},$$

damit die W'keit eines Fehlers zweiter Art, dass also H_0 nicht verworfen wird, obwohl H_1 zutrifft, möglichst klein ist.

Beispiele

- Wenn wir von Anfang an unsere Vermutung belegen wollten, dass sich die Trauerschnäpper bei grünem Licht stärker auf eine Richtung konzentrieren als bei blauem, dürfen wir einseitig testen.
- Wenn dann aber noch so deutlich herauskommt, dass die Richtungswahl bei blauem Licht enger konzentriert war, so ist das dann strenggenommen nicht als signifikant zu betrachten.
- Wenn wir von Anfang an die Vermutung belegen wollten, dass der Kork an der Nordseite des Baumes dicker war, dürfen wir einseitig testen.
- Wenn dann aber noch so deutlich herauskommt, dass der Kork im Westen dicker ist, ist das strenggenommen nicht signifikant.

Angenommen, H_0 wird auf dem 5%-Niveau verworfen. Welche Aussage gilt dann?

- Die Nullhypothese ist falsch. ~~Die Nullhypothese ist falsch.~~
- H_0 ist mit 95%-iger Wahrscheinlichkeit falsch. ~~H_0 ist mit 95%-iger Wahrscheinlichkeit falsch.~~
- Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in 5% der Fälle. Falls die Nullhypothese wahr ist, beobachtet man ein so extremes Ergebnis nur in 5% der Fälle. ✓

Angenommen, H_0 konnte durch den Test nicht verworfen werden. Welche Aussagen sind dann richtig?

- Wir müssen die Alternative H_1 verwerfen.
  ~~Wir müssen die Alternative H_1 verwerfen.~~
- H_0 ist wahr. ~~H_0 ist wahr.~~
- H_0 ist wahrscheinlich wahr. ~~H_0 ist wahrscheinlich wahr.~~
- Es ist ungefährlich, davon auzugehen, dass H_0 zutrifft.
  ~~Es ist ungefährlich, davon auzugehen, dass H_0 zutrifft.~~
- Auch wenn H_0 zutrifft, ist es nicht sehr unwahrscheinlich, dass unsere Teststatistik einen so extrem erscheinenden Wert annimmt. Auch wenn H_0 zutrifft, ist es nicht sehr unwahrscheinlich, dass unsere Teststatistik einen so extrem erscheinenden Wert annimmt. ✓
- Die Nullhypothese ist in dieser Hinsicht mit den Daten verträglich. Die Nullhypothese ist in dieser Hinsicht mit den Daten verträglich. ✓