Wahrscheinlichkeitsrechnung und Statistik
im Biologie-Bachelorstudiengang der LMU

Übersicht über die besprochenen statistischen Tests

Dirk Metzler

24. März 2026

Inhaltsverzeichnis

1 Welcher Test für welche Daten?

Mittelwert einer "Population" (Grundgesamtheit) $\mu_x = \mu_0$ **?** ein-Stichproben t-Test

Mittelwerte zweier Gruppen gleich $\mu_x = \mu_y$ **?** zwei-Stichproben t-Tests oder Wilcoxon-Test

y_i **von** x_i **unabhängig, also** $b = 0$ **in** $y_i = a + b \cdot x_i + \varepsilon_i$**?** t-Test bei linearer Regression (oder Varianzanalyse für eingebettete Modelle)

Mittelwerte in drei oder mehr Gruppen (Stichproben) gleich $\mu_x = \mu_y = \mu_z = \ldots$ **?** Varianzanalyse (F-Test), Kruskal-Wallis-Test

Lineares Modell mit zusätzlichen Parametern nicht besser als eingebettetes Modell? Varianzanalyse (F-Test)

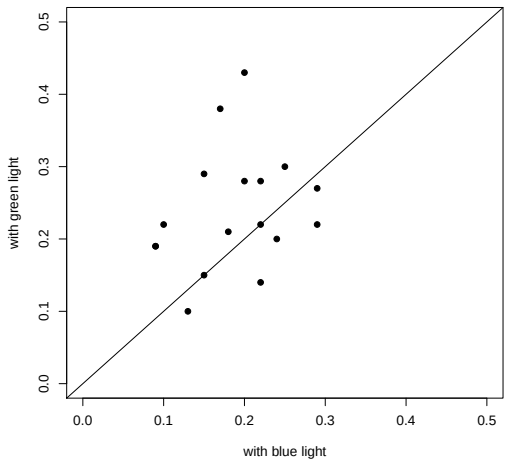
passen beobachtete Anzahlen zu (Annahmen über) Wahrscheinlichkeiten? Chi-Quadrat-Test, Fisher's exakter Test, z-Test

siehe auch uebersicht.pdf

2 t-Tests

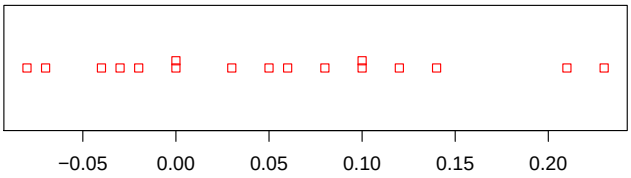
2.1 ein-Stichproben t-Tests und gepaarter t-Test

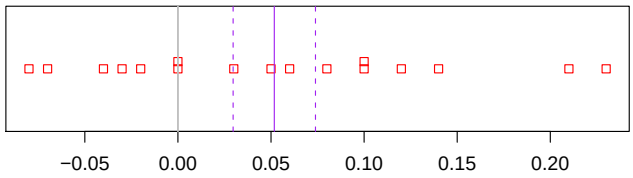
Trauerschnäpper: Länge des Schwerpunktsvektors bei grünem und bei blauem Licht, n=17



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

$x :=$ "Grünwert" $-$ "Blauwert"





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

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

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?

Immer 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.

Allgemein gilt
Sind $X_1, \ldots, X_n$ unabhängig aus einer Normalverteilung mit Mittelwert μ gezogen, so ist

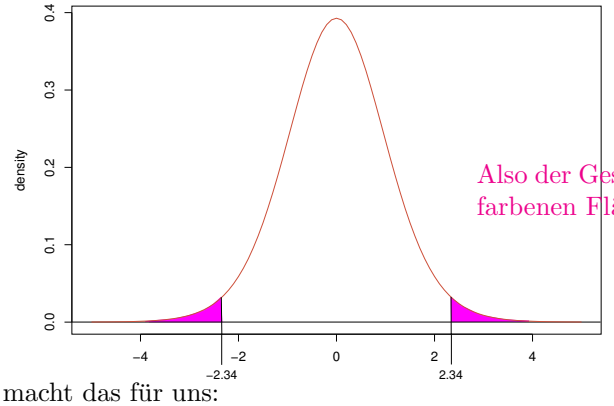
$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

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

Wie (un)wahrscheinlich ist nun eine mindestens so große Abweichung wie 2.34 Standardfehler?

$$\Pr(|T| = 2.34) = 0 \qquad \text{Das bringt nichts!}$$

Zu berechnen ist $\Pr(|T| \geq 2.34)$, der sog. p-Wert.



Also der Gesamtinhalt der magenta-farbenen Flächen.

R macht das für uns:

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

Beachte: `pt(2.34,df=16,lower.tail=FALSE)` ist dasselbe wie `1-pt(2.34,df=16)`, also der `upper tail`.

Zum Vergleich mal mit der Normalverteilung:

```
> pnorm(-2.34)+pnorm(2.34,lower.tail=FALSE)
[1] 0.01928374
```

Vollständiger t-Test mit R

```
> x <- trauerschn$gruen-trauerschn$blau
> 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
```

gepaarter t-Test

Ein gepaarter t-Test für die Werte-Paare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

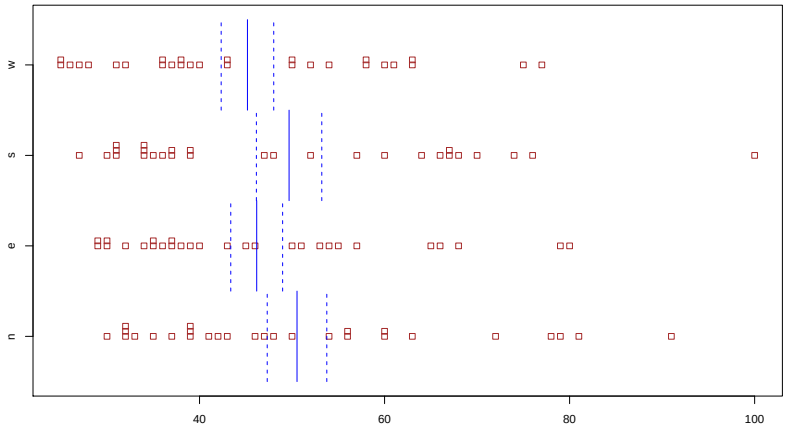
ist ein ein-Stichproben-t-Test mit Nullhypothese $\mu = 0$ angewendet auf die

Differenzen $d_1 = x_1 - y_1,\ d_2 = x_2 - y_2, \ldots,\ d_n = x_n - y_n$.

Wieso gepaart testen, wenn möglich?

Es kann sein, dass es eine Varianz zwischen (x_i, y_i) und (x_j, y_j) gibt, die man herausrechnen sollte. Beispiele:

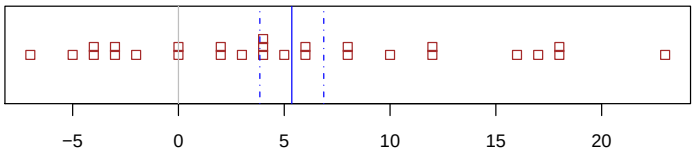
- Verschiedene Vögel sind unterschiedlich stark auf Flugrichtungen festgelegt,
- Große Korkeichen haben generell mehr Kork als kleine.



Kann da was signifikant unterschiedlich sein???

Stripchart der Korkdicken je nach Himmelsrichtung mit Mittelwerten und Mittelwerten ± Standardfehler

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



mit Mittelwert und Mittelwert±Standardfehler

2.2 ungepaarte zwei-Stichproben t-Tests

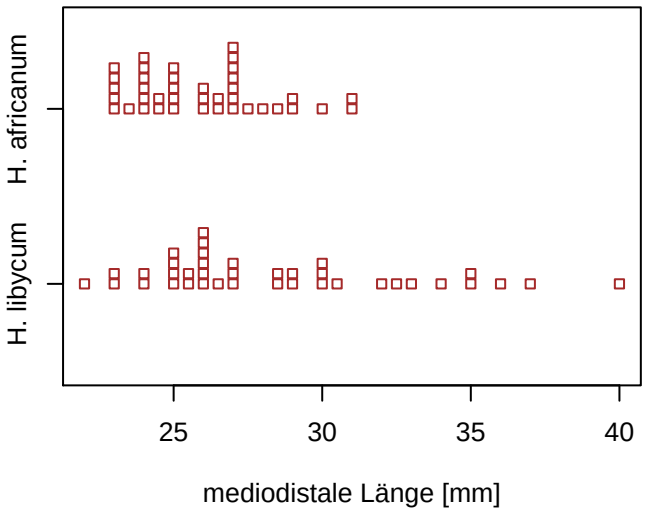
Frage

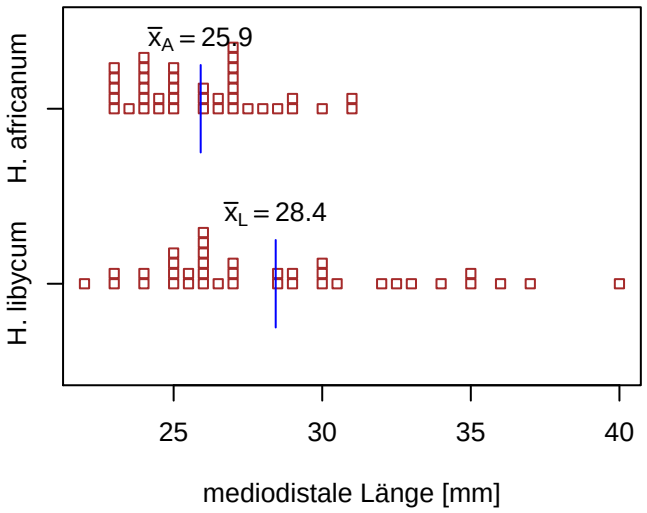
Hipparion:
Laubfresser $\longrightarrow$ Grasfresser

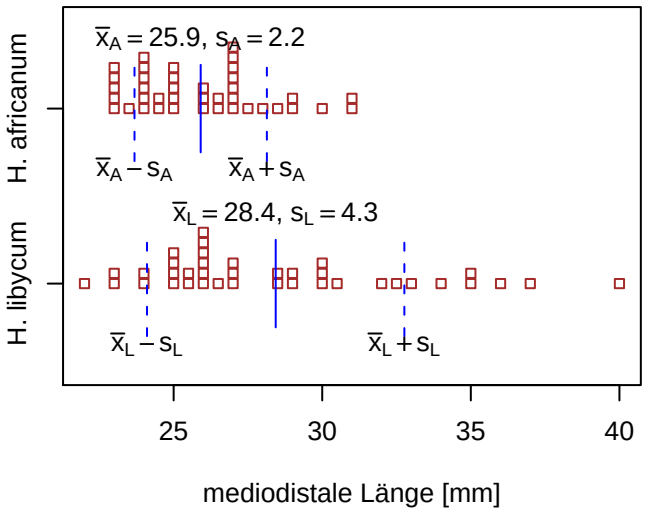
andere Nahrung $\longrightarrow$ andere Zähne?

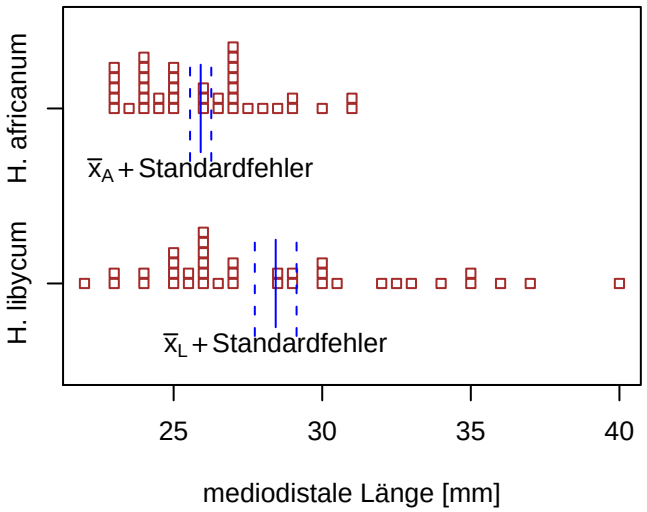
Messungen: mesiodistale Länge

distal = von der Mittellinie weg









Theorem 1 (zwei-Stichproben t-Test, ungepaart mit gleichen Varianzen) *Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängige normalverteilte Zufallsvariablen mit der selben Varianz σ^2. Als* ***gepoolte Stichprobenvarianz*** *definieren wir*

$$s_p^2 = \frac{(n-1)\cdot s_X^2 + (m-1)\cdot s_Y^2}{m+n-2} = \frac{\sum_i (X_i - \overline{X})^2 + \sum_i (Y_i - \overline{Y})^2}{m+n-2}.$$

Unter der Nullhypothese gleicher Erwartungswerte $\mu_X = \mu_y$ folgt die Statistik

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

einer t-Verteilung mit $n + m - 2$ mit Freiheitsgraden.

Theorem 2 (Welch-t-Test, die Varianzen dürfen ungleich sein) *Seien $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_m$ unabhängige normalverteilte Zufallsvariablen mit (möglicherweise verschiedenen) Varianzen $VarX_i = \sigma_X^2$*

und $VarY_i = \sigma_Y^2$. *Seien* s_X *und* s_Y *die aus den Stichproben berechneten korrigierten Standardabweichungen und* $f_x = s_x/\sqrt{n}$ *sowie* $f_y = s_y/\sqrt{m}$ *die Standardfehler. Unter der Nullhypothese gleicher Mittelwerten* $\mathbb{E}X_i = \mathbb{E}Y_j$ *ist die Statistik*

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

ungefähr *t-verteilt mit* $\frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2 \cdot (n-1)} + \frac{s_Y^4}{m^2 \cdot (m-1)}} = \frac{\left(f_x^2 + f_y^2\right)^2}{\frac{f_x^4}{n-1} + \frac{f_y^4}{m-1}}$ *Freiheitsgraden.*

2.3 t-Tests bei der linearen Regression

Beispiel: Rothirsch (*Cervus elaphus*)

Theorie: Hirschkühe können das Geschlecht ihrer Nachkommen beeinflussen.

Unter dem Gesichtspunkt evolutionär stabiler Strategien ist zu erwarten, dass schwache Tiere eher zu weiblichem und starke Tiere eher zu männlichem Nachwuchs tendieren.

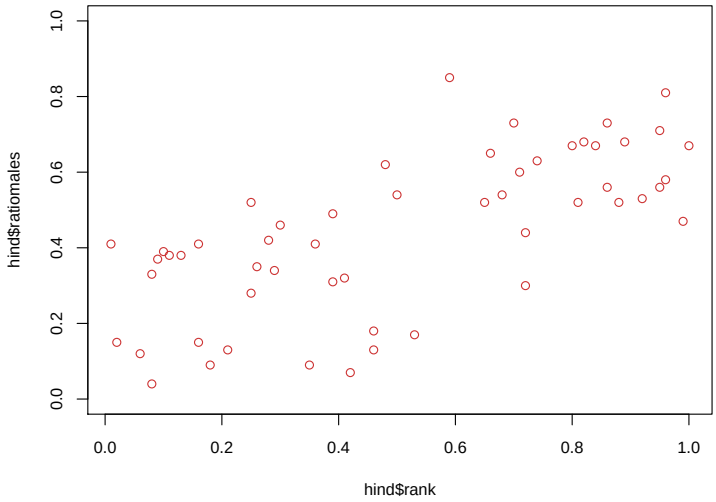
Literatur

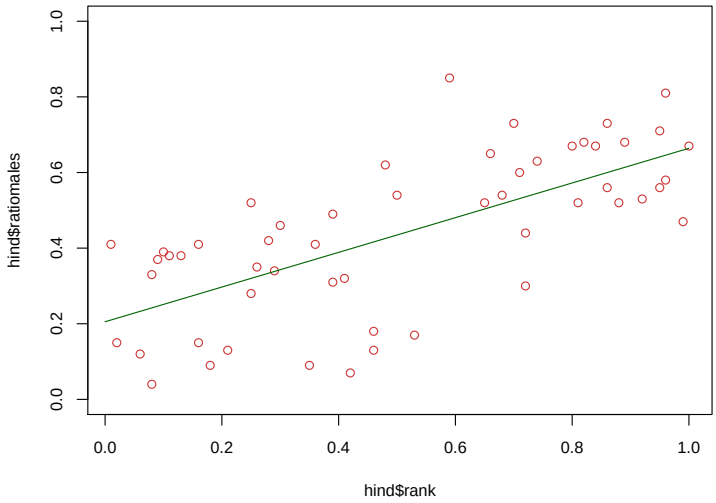
[CAG86] Clutton-Brock, T. H. , Albon, S. D., Guinness, F. E. (1986) Great expectations: dominance, breeding success and offspring sex ratios in red deer. *Anim. Behav.* **34**, 460—471.

```
> hind
   rank ratiomales
1  0.01       0.41
2  0.02       0.15
3  0.06       0.12
4  0.08       0.04
5  0.08       0.33
6  0.09       0.37
.   .          .
.   .          .
.   .          .

52 0.96       0.81
53 0.99       0.47
54 1.00       0.67
```

ACHTUNG: Simulierte Daten, die sich an den Daten aus der Originalpublikation lediglich orientieren.






```
> mod <- lm(ratiomales~rank,data=hind)
> summary(mod)
Call:
lm(formula = ratiomales ~ rank, data = hind)
Residuals:
     Min       1Q   Median       3Q      Max
-0.32798 -0.09396  0.02408  0.11275  0.37403

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20529    0.04011   5.119 4.54e-06 ***
```

```
rank          0.45877    0.06732   6.814 9.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.154 on 52 degrees of freedom
Multiple R-squared: 0.4717,Adjusted R-squared: 0.4616
F-statistic: 46.44 on 1 and 52 DF,  p-value: 9.78e-09
```

Modell:

$$Y = a + b \cdot X + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Wie berechnet man die Signifikanz eines Zusammenhangs zwischen dem *erklärenden Merkmal* X und der *Zielgröße* Y?

Wir haben b durch $\hat{b}$ geschätzt (und $\hat{b} \neq 0$).

Wie testen wir de Nullhypothese $b = 0$?

Wie groß ist der Standardfehler unserer Schätzung $\hat{b}$?

$$y_i = a + b \cdot x_i + \varepsilon \qquad \text{mit } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

nicht zufällig: a, b, x_i, σ^2 zufällig: ε, y_i

$$\text{var}(y_i) = \text{var}(a + b \cdot x_i + \varepsilon) = \text{var}(\varepsilon) = \sigma^2$$

und $y_1, y_2, \ldots, y_n$ sind stochastisch unabhängig.

$$\hat{b} = \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\begin{aligned} \text{var}(\hat{b}) &= \text{var}\left(\frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right) = \frac{\text{var}\left(\sum_i y_i (x_i - \bar{x})\right)}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\ &= \frac{\sum_i \text{var}\left(y_i\right) (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} = \sigma^2 \cdot \frac{\sum_i (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\ &= \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2 \end{aligned}$$

Tatsächlich ist $\hat{b}$ Normalverteilt mit Mittelwert b und

$$\text{var}(\hat{b}) = \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2$$

Problem: Wir kennen σ^2 nicht.

Wir schätzen σ^2 mit Hilfe der beobachten Residuenvarianz durch

$$s^2 := \frac{\sum_i \left(y_i - \hat{a} - \hat{b} \cdot x_i\right)^2}{n - 2}$$

Zu beachten ist, dass durch $n - 2$ geteilt wird. Das hat damit zu tun, dass zwei Modellparameter a und b bereit geschätzt wurden, und somit 2 Freiheitsgrade verloren gegangen sind.

$$\text{var}(\hat{b}) = \sigma^2 \Big/ \sum_i (x_i - \bar{x})^2$$

Schätze σ^2 durch

$$s^2 = \frac{\sum_i \left(y_i - \hat{a} - \hat{b} \cdot x_i\right)^2}{n-2}.$$

Dann ist

$$\frac{\hat{b} - b}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}}$$

Student-t-verteilt mit $n-2$ Freiheitsgraden und wir können den t-Test anwenden, um die Nullhypothese $b = 0$ zu testen.

Verwerfe H_0: „$b = 0$“ zum Signifikanzniveau α, wenn $\left| \frac{\hat{b}}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}} \right| \geq q_{1-\alpha/2}$, wo $q_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Student-Verteilung mit $n-2$ Freiheitsgraden ist.

2.4 Übersicht t-Tests

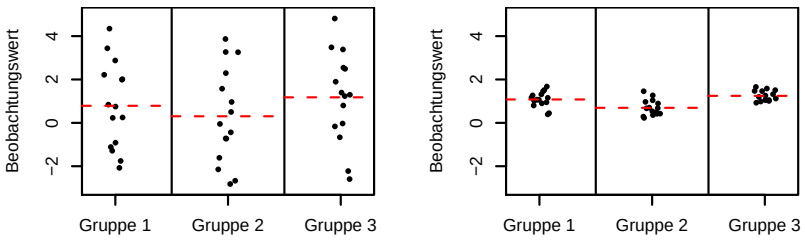
t-Test	t-Statistik	wobei…	Freiheitsgrade
ein-Stichproben-t-Test	$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$	$s = \sqrt{\frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1}}$	$n-1$
gepaarter zwei-Stichproben-t-Test	$t = \frac{\overline{X} - \overline{Y}}{s/\sqrt{n}}$	$s = \sqrt{\frac{\sum_i \left(D_i - \overline{D}\right)^2}{n-1}}$ mit $D_i = X_i - Y_i$	$n-1$
ungepaarter zwei-Stichproben-t-Test mit gleichen Varianzen	$t = \frac{\overline{X} - \overline{Y}}{s_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$	$s_p = \sqrt{\frac{\sum_i \left(X_i - \overline{X}\right)^2 + \sum_i \left(Y_i - \overline{Y}\right)^2}{m+n-2}}$	$n+m-2$
ungepaarter zwei-Stichproben-t-Test mit ungleichen Varianzen	$t = \frac{\overline{X} - \overline{Y}}{\sqrt{f_x^2 + f_y^2}}$	$f_x = \frac{s_x}{\sqrt{n}}, \quad f_y = \frac{s_y}{\sqrt{m}}$	$\frac{\left(f_x^2 + f_y^2\right)^2}{\frac{f_x^4}{n-1} + \frac{f_y^4}{m-1}}$
t-Test für Steigung einer Regressionsgeraden	$\frac{\hat{b} - b}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}}$	$s = \sqrt{\frac{\sum_i \left(y_i - (\hat{a} + \hat{b} \cdot x_i)\right)^2}{n-2}}$	$n-2$

3 Varianzanalysen (ANOVAS)

3.1 ein-Faktor-Anovas

Varianzanalyse

Wir beobachten unterschiedliche Gruppenmittelwerte:



Variabilität innerhalb der Gruppen groß Variabilität innerhalb der Gruppen klein

Sind die beobachteten Unterschiede der Gruppenmittelwerte ernst zu nehmen — oder könnte das alles Zufall sein?

Das hängt vom Verhältnis der Variabilität der Gruppenmittelwerte und der Variabilität der Beobachtungen innerhalb der Gruppen ab: die Varianzanalyse gibt eine (quantitative) Antwort.

Beispiel

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gruppe	Beobachtung							
1	62	60	63	59				
2	63	67	71	64	65	66		
3	68	66	71	67	68	68		
4	56	62	60	61	63	64	63	59

Globalmittelwert $\overline{x}_{..} = 64$,
Gruppenmittelwerte $\overline{x}_{1.} = 61$, $\overline{x}_{2.} = 66$, $\overline{x}_{3.} = 68$, $\overline{x}_{4.} = 61$.

Beispiel

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

Gr.	$\overline{x}_{i.}$	Beobachtung							
1	61	62	60	63	59				
		$(62-61)^2$	$(60-61)^2$	$(63-61)^2$	$(59-61)^2$				
2	66	63	67	71	64	65	66		
		$(63-66)^2$	$(67-66)^2$	$(71-66)^2$	$(64-66)^2$	$(65-66)^2$	$(66-66)^2$		
3	68	68	66	71	67	68	68		
		$(68-68)^2$	$(66-68)^2$	$(71-68)^2$	$(67-68)^2$	$(68-68)^2$	$(68-68)^2$		
4	61	56	62	60	61	63	64	63	59
		$(56-61)^2$	$(62-61)^2$	$(60-61)^2$	$(61-61)^2$	$(63-61)^2$	$(64-61)^2$	$(63-61)^2$	$(59-61)^2$

Globalmittelwert $\overline{x}_{..} = 64$,
Gruppenmittelwerte $\overline{x}_{1.} = 61$, $\overline{x}_{2.} = 66$, $\overline{x}_{3.} = 68$, $\overline{x}_{4.} = 61$.

Die *roten* Werte (ohne die Quadrate) heißen *Residuen*: die „Restvariabilität“ der Beobachtungen, die das Modell nicht erklärt.

Quadratsumme innerhalb der Gruppen: $ss_{\text{innerh}} = 112$, 20 Freiheitsgrade
Quadratsumme zwischen den Gruppen: $ss_{\text{zw}} = 4 \cdot (61-64)^2 + 6 \cdot (66-64)^2 + 6 \cdot (68-64)^2 + 8 \cdot (61-64)^2 = 228$, 3 Freiheitsgrade

$$F = \frac{ss_{\text{zw}}/3}{ss_{\text{innerh}}/20} = \frac{76}{5{,}6} = 13{,}57$$

Beispiel: Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

ANOVA-Tafel („ANalysis Of VAriance“)

	Freiheitsgrade (DF)	Quadratsumme (SS)	mittlere Quadratsumme (SS/DF)	F-Wert
Gruppe	3	228	76	13,57
Residuen	20	112	5,6	

Unter der Hypothese H_0 „die Gruppenmittelwerte sind gleich“ (und einer Normalverteilungsannahme an die Beobachtungen)
ist F Fisher-verteilt mit 3 und 20 Freiheitsgraden, $p = \text{Fisher}_{3,20}([13{,}57, \infty)) \leq 5 \cdot 10^{-5}$.

Wir verwerfen demnach H_0.

95%-Quantil der Fisher-Verteilung in Abhängigkeit der Anzahl Freiheitsgrade (k_1 Zähler-, k_2 Nennerfreiheitsgrade)

$k_2 \backslash k_1$	1	2	3	4	5	6	7	8	9	10	11
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.7
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.6
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.1
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.9	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3	2.91	2.85	2.8	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.6	3.74	3.34	3.11	2.96	2.85	2.76	2.7	2.65	2.6	2.57
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.2	2.96	2.81	2.7	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
19	4.38	3.52	3.13	2.9	2.74	2.63	2.54	2.48	2.42	2.38	2.34
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35	2.31

F-Test

$n = n_1 + n_2 + \cdots + n_I$ Beobachtungen in I Gruppen,
$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$.

Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$,
mit unabhängigen, normalverteilten ε_{ij}, $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

(μ_i ist der „wahre“ Mittelwert innerhalb der i-ten Gruppe.)

$\overline{X}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij}$ (empirisches) „Globalmittel“
$\overline{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ (empirischer) Mittelwert der i-ten Gruppe

$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$ Quadratsumme innerhalb d. Gruppen, $n - I$ Freiheitsgrade

$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$ Quadratsumme zwischen d. Gruppen, $I - 1$ Freiheitsgrade

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

F-Test

$X_{ij} = j$-te Beobachtung in der i-ten Gruppe, $j = 1, \ldots, n_i$,
Modellannahme: $X_{ij} = \mu_i + \varepsilon_{ij}$. $\mathbb{E}[\varepsilon_{ij}] = 0$, $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

$SS_{\text{innerh}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2$ Quadratsumme innerhalb d. Gruppen, $n - I$ Freiheitsgrade

$SS_{\text{zw}} = \sum_{i=1}^{I} n_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$ Quadratsumme zwischen d. Gruppen, $I - 1$ Freiheitsgrade

$$F = \frac{SS_{\text{zw}}/(I-1)}{SS_{\text{innerh}}/(n-I)}$$

Unter der Hypothese $H_0 : \mu_1 = \cdots = \mu_I$ („alle μ_i sind gleich“) ist F Fisher-verteilt mit $I-1$ und $n-I$ Freiheitsgraden

(unabhängig vom tatsächlichen gemeinsamen Wert der μ_i).

F-Test: Wir lehnen H_0 zum Signifikanzniveau α ab, wenn $F \geq q_{1-\alpha}$, wobei $q_{1-\alpha}$ das $(1-\alpha)$-Quantil der Fisher-Verteilung mit $I-1$ und $n-I$ Freiheitsgraden ist.

3.2 Anova für eingebettete Modelle

```
> modell0 <- lm(richness ~ angle2+NAP+grainsize+humus,
+                  data = rikz)
> modell <- lm(richness ~ angle2+NAP+grainsize+humus
+                          +factor(week), data = rikz)
> anova(modell0, modell)
Analysis of Variance Table

Model 1: richness ~ angle2 + NAP + grainsize + humus
Model 2: richness ~ angle2 + NAP + grainsize + humus + factor(week)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     40 531.17
2     37 353.66  3    177.51 6.1902 0.00162 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> lm1 <- lm(Postwt~Prewt+Treat,anorexia)
> lm2 <- lm(Postwt~Prewt*Treat,anorexia)
> anova(lm1,lm2)
Analysis of Variance Table

Model 1: Postwt ~ Prewt + Treat
Model 2: Postwt ~ Prewt * Treat
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     68 3311.3
2     66 2844.8  2     466.5 5.4112 0.006666 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4 Chi-Quadrat-Tests und Fisher's exakter Test

4.1 χ^2-Test für eine feste Verteilung (und z-Test)

- Ein Experiment habe r mögliche Ausgänge (z.B. $r = 6$ beim Werfen eines Würfels).
- Unter der Nullhypothese H_0 habe Ausgang i Wahrscheinlichkeit p_i.
- Unter n unabhängigen Wiederholungen des Experiments beobachten wir B_i mal Ausgang i. Unter H_0 erwarten wir $E_i := \mathbb{E}[B_i] = np_i$ mal Augang i zu beobachten.

Frage: Geben die Beobachtungen Anlass, an der Nullhypothese zu zweifeln?
Erwarte $E_i = np_i$ mal Ausgang i, beobachte B_i mal.
Geben diese Beobachtungen Anlass, an der Nullhypothese zu zweifeln?

Vorgehen:

- Berechne $X^2 = \sum_i \frac{(B_i - E_i)^2}{E_i}$

- X^2 ist unter (approximativ, sofern n genügend groß) χ^2_{r-1}-verteilt („Chi-Quadrat-verteilt mit $r-1$ Freiheitsgraden“)
- Lehne H_0 zum Signifikanzniveau α ab, wenn $X^2 \geq q_{1-\alpha}$, wo $q_{1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $r-1$ Freiheitsgraden ist.

95%-Quantil der χ^2-Verteilung in Abhängigkeit der Anzahl Freiheitsgrade

F.g.	1	2	3	4	5	6	7	8	9	10
Quantil	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31

Beispiel: Unter 12.000 Würfen eines Würfels beobachten wir folgende Häufigkeiten der Augenzahlen:

i	1	2	3	4	5	6
B_i	2014	2000	2017	1925	1998	2046

Ist der Würfel fair (H_0: $p_1 = \cdots = p_6 = 1/6$)?

Es ist $E_1 = \cdots = E_6 = 12.000 \cdot 1/6 = 2000$,

$$\begin{aligned} X^2 &= \frac{(2014-2000)^2}{2000} + \frac{(2000-2000)^2}{2000} + \frac{(2017-2000)^2}{2000} \\ &\quad + \frac{(1925-2000)^2}{2000} + \frac{(1998-2000)^2}{2000} + \frac{(2046-2000)^2}{2000} = 4{,}115. \end{aligned}$$

Das 95%-Quantil der χ^2-Verteilung mit 5 Freiheitsgraden ist 11,07 $>$ 4,115, wir lehnen H_0 nicht ab (zum Signifikanzniveau 5%).

95%-Quantil der χ^2-Verteilung in Abh.keit d. Anz. Freiheitsgrade

F.g.	1	2	3	4	5	6	7	8	9	10
Quantil	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31

Bemerkung: $\chi^2_5([4{,}115, \infty)) = 0{,}533$, d.h. wir finden einen p-Wert von 53%, der Test gibt keinen Anlass zu Zweifel an H_0.

Alternative zum Chi-Quadrat-Anpassungstest, falls es nur zwei Gruppen gibt:
z-Test
Oder auch direkt mit der Binomialverteilung ohne Normalapproximation.

4.2 χ^2-Test auf Unabhängigkeit (oder Homogenität)

Beispiel: 48 Teilnehmer eines Management-Kurses entscheiden über Beförderung:

	Weiblich	Männlich	Summe
Befördern	14	21	35
Ablegen	10	3	13
Summe	24	24	48

Kann das Zufall sein? Testen wir H_0 : „Geschlecht und Beförderungsentscheidung sind unabhängig“.

Anteil Weiblich=24/48=0.5, Anteil befördert=35/48=0.73, also erwartete Zahlen unter H_0:

	Weiblich	Männlich	Summe
Befördern	17.5 $(= 48 \cdot \frac{24}{48} \cdot \frac{35}{48})$	17.5 $(= 48 \cdot \frac{24}{48} \cdot \frac{35}{48})$	35
Ablegen	6.5 $(= 48 \cdot \frac{24}{48} \cdot \frac{13}{48})$	6.5 $(= 48 \cdot \frac{24}{48} \cdot \frac{13}{48})$	13
Summe	24	24	48

H_0 : „Geschlecht und Beförderungsentscheidung sind unabhängig“

Beobachtete Anzahlen:

	Weiblich	Männlich	Summe
Befördern	14	21	35
Ablegen	10	3	13
Summe	24	24	48

Unter H_0 erwartete Anzahlen:

	Weiblich	Männlich	Summe
Befördern	17.5	17.5	35
Ablegen	6.5	6.5	13
Summe	24	24	48

Die X^2-Statistik ist

$$X^2 = \frac{(17.5-14)^2}{17.5} + \frac{(21-17.5)^2}{17.5} + \frac{(10-6.5)^2}{6.5} + \frac{(3-6.5)^2}{6.5} = 5.17.$$

Unter H_0 ist X^2 (approximativ) χ^2-verteilt mit einem Freiheitsgrad ($1 = 4-1-1-1 = (2-1)\cdot(2-1)$): 4 Zellen, ein Freiheitsgrad geht für die feste Gesamtsumme, einer für das (prinzipiell) unbekannte Geschlechterverhältnis und einer für die (prinzipiell) unbekannte Beförderungswahrscheinlichkeit „verloren“.

95%-Quantil der χ^2-Verteilung in Abh.keit d. Anz. Freiheitsgrade

F.g.	1	2	3	4	5	6	7	8	9	10
Quantil	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31

Wir können H_0 zum Signifikanzniveau 5% ablehnen.

(Es ist $\chi^2_1([5.17, \infty)) = 0.023$, d.h. wir finden einen p-Wert von ca. 2%.)

Chi-Quadrat-Test auf Unabhängigkeit, allgemeine Situation:

- 2 Merkmale mit r bzw. s Ausprägungen ($r \times s$-Kontingenztafel), n Beobachtungen
- Bestimme erwartete Anzahlen unter H_0 als Produkt der (normierten) Zeilen- und Spaltensummen
- X^2 ist unter H_0 (approximativ) χ^2-verteilt mit $rs - 1 - (r-1) - (s-1) = (r-1)(s-1)$ Freiheitsgraden.

Bemerkung: Im 2×2-Fall kann man auch Fishers exakten Test verwenden (zumal, wenn n recht klein).

4.3 Fishers exakter Test

Literatur

[McK91] J.H. McDonald, M. Kreitman (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652-654.

	synonym	verändernd
polymorph	43	2
fixiert	17	7

A	B
C	D

- Nullhypothese: $\frac{\mathbb{E}A/\mathbb{E}C}{\mathbb{E}B/\mathbb{E}D} = 1$
- Für 2×2-Tabellen können die p-Werte exakt berechnet werden. (keine Approximation, keine Simulation).

```
> fisher.test(McK)

    Fisher's Exact Test for Count Data

data:  McK
p-value = 0.006653
alternative hypothesis: true odds ratio
                        is not equal to 1
95 percent confidence interval:
  1.437432 92.388001
sample estimates:
odds ratio
  8.540913
```

			$\sum$
	43	2	45
	17	7	24
$\sum$	60	9	69

			$\sum$
	a	b	K
	c	d	M
$\sum$	U	V	N

Unter der Annahme, dass die Zeilen und Spalten unabhängig sind, ist die Wahrscheinlichkeit, dass links oben in der Tabelle der Wert a bzw. oben recht ein $b = K - a$ steht:

$$\Pr(a \text{ oben links}) = \frac{\binom{K}{a}\binom{M}{c}}{\binom{N}{U}} = \Pr(b \text{ oben rechts}) = \frac{\binom{K}{b}\binom{M}{d}}{\binom{N}{V}}$$

"hypergeometrische Verteilung"

			$\sum$
	a	b	45
	c	d	24
$\sum$	60	9	69

b	Pr(b)
0	0.000023
1	0.00058
2	0.00604
3	0.0337
4	0.1117
5	0.2291
6	0.2909
7	0.2210
8	0.0913
9	0.0156

Einseitiger Fisher-Test:
für $b = 2$:
p-Wert=Pr(0) + Pr(1) + Pr(2) = 0.00665313
für $b = 3$:
p-Wert=Pr(0) + Pr(1) + Pr(2) + Pr(3) = 0.04035434

Zweiseitiger Fisher-Test:
Addiere alle Wahrscheinlichkeiten, die kleiner oder gleich Pr(b) sind.
für $b = 2$:
p-Wert=Pr(0) + Pr(1) + Pr(2) = 0.00665313
für $b = 3$:
p-Wert=
Pr(0)+Pr(1)+Pr(2)+Pr(3)+Pr(9) = 0.05599102

4.4 χ^2-Test für allgemeinere Modelle

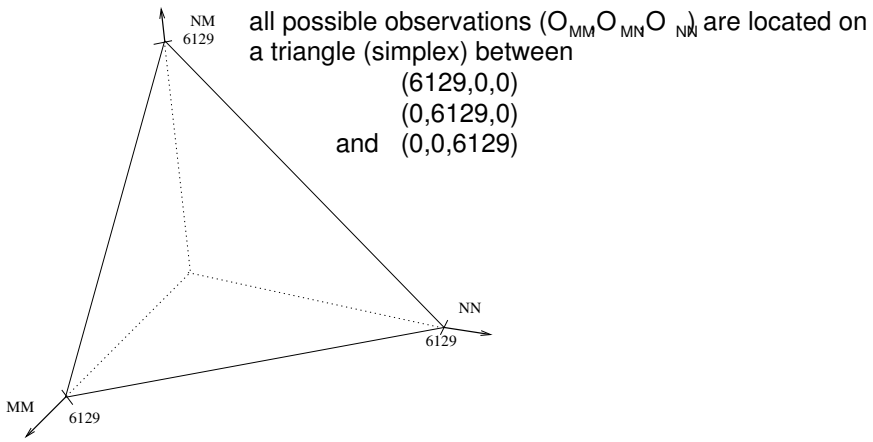
k: Anzahl Gruppen m: Anzahl Modellparameter, die geschätzt werden Anzahl Freiheitsgrade:

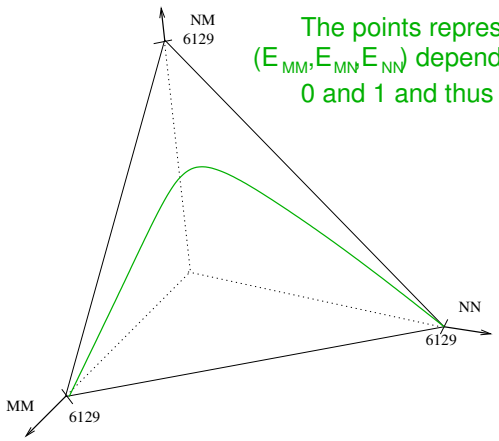
$$\mathsf{df} = k - m - 1$$

Beispiel: Test auf Hardy-Weinberg-Gleichgewicht mit drei Genotypen AA, Aa, aa: $k = 3$ Gruppen, $m = 1$ Parameter wird geschätzt, nämlich der Anteil der Allele vom Typ A.

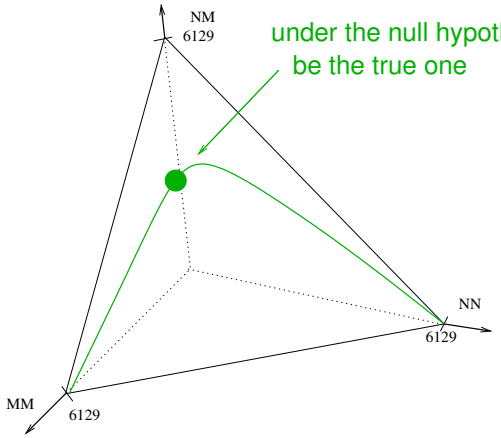
$$\Rightarrow \mathsf{df} = 3 - 1 - 1 = 1$$

5 Freiheitsgrade

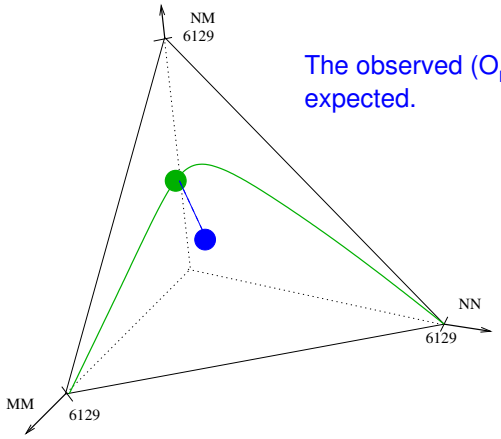




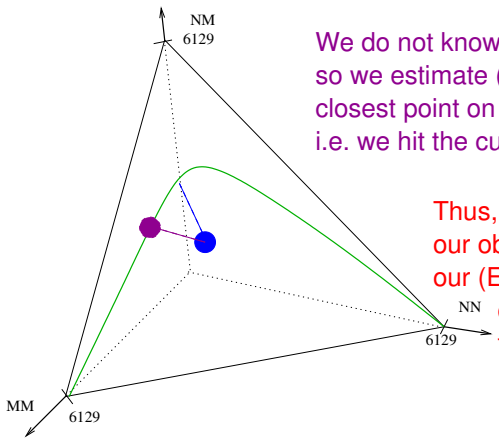
The points representing the Expected Values (E_{MM},E_{MN},E_{NN}) depend on one parameter p between 0 and 1 and thus form a curve in the simplex.



under the null hypothesis, one of these values must be the true one



The observed (O_{MM},O_{NM},O_{NN}) will deviate from the expected.



We do not know the true expectation values so we estimate (E_{MM},E_{MN},E_{NN}) by taking the closest point on the curve of possible values, i.e. we hit the curve in a right angle.

Thus, deviations between our our observations (O_{MM},O_{NM},O_{NN}) and our (E_{MM},E_{NM},E_{NN}) can only be in one dimension: perpendicular to the curve.

Chi-Quadrat-Test	Freiheitsgrade
Chi-Quadrat-Anpassungstest mit n Anzahlen	$n-1$
Chi-Quadrat-Test auf Unabhängigkeit (=Homogenität) mit $n \times m$ Anzahlen	$(n-1)\cdot(m-1)$
Chi-Quadrat-Test bei Nullmodell mit m Parametern und k Anzahlen	$k-m-1$

t-Test	t-Statistik	wobei...	Freiheitsgrade
ein-Stichproben-t-Test	$t=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$	$s=\sqrt{\frac{\sum_i\left(X_i-\overline{X}\right)^2}{n-1}}$	$n-1$
gepaarter zwei-Stichproben-t-Test	$t=\frac{\overline{X}-\overline{Y}}{s/\sqrt{n}}$	$s=\sqrt{\frac{\sum_i\left(D_i-\overline{D}\right)^2}{n-1}}$ mit $D_i=X_i-Y_i$	$n-1$
ungepaarter zwei-Stichproben-t-Test mit gleichen Varianzen	$t=\frac{\overline{X}-\overline{Y}}{s_p\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}$	$s_p=\sqrt{\frac{\sum_i\left(X_i-\overline{X}\right)^2+\sum_i\left(Y_i-\overline{Y}\right)^2}{m+n-2}}$	$n+m-2$
ungepaarter zwei-Stichproben-t-Test mit ungleichen Varianzen	$t=\frac{\overline{X}-\overline{Y}}{\sqrt{f_x^2+f_y^2}}$	$f_x=\frac{s_x}{\sqrt{n}}, \quad f_y=\frac{s_y}{\sqrt{m}}$	$\frac{\left(f_x^2+f_y^2\right)^2}{\frac{f_x^4}{n-1}+\frac{f_y^4}{m-1}}$
t-Test für Steigung einer Regressionsgeraden	$\frac{\hat{b}-b}{s/\sqrt{\sum_i(x_i-\bar{x})^2}}$	$s=\sqrt{\frac{\sum_i\left(y_i-(\hat{a}+\hat{b}\cdot x_i)\right)^2}{n-2}}$	$n-2$

Freiheitsgrade bei der Varianzanalyse mit I Gruppen und n Messungen insgesamt:

$$F=\frac{\sum_{i=1}^{I} n_i(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\cdot})^2/(n-I)}, \qquad \mathsf{df}_1=I-1, \quad \mathsf{df}_2=n-I$$

Freiheitsgrade bei Varianzanalyse für eingebettete Modelle (nested models):
Anzahl zusätzliche Parameter

6 Nichtparamterische Tests und simulationsbasierte Tests

6.1 Wilcoxon's Rangsummentest (Mann-Whitney-U-Test)

Idee der Rangsummentests

$$\begin{gathered}\text{Beobachtungen:}\\ X : x_1, x_2, \ldots, x_m\\ Y : y_1, y_2, \ldots, y_n\end{gathered}$$

- Sortiere alle Beobachtungen der Größe nach.
- Bestimme die Ränge der m X-Werte unter allen $m+n$ Beobachtungen.
- Wenn die Nullhypothese zutrifft, sind die m X-Ränge eine rein zufällige Wahl aus $\{1, 2, \ldots, m+n\}$.
- Berechne die Summe der X-Ränge, prüfe, ob dieser Wert untypisch groß oder klein.

Wilcoxons Rangsummenstatistik

$$\begin{gathered}\text{Beobachtungen:}\\ X : x_1, x_2, \ldots, x_m\\ Y : y_1, y_2, \ldots, y_n\end{gathered}$$

$$\begin{gathered}W = \text{Summe der } X\text{-Ränge} - (1 + 2 + \cdots + m)\\ \text{heißt}\\ \textit{Wilcoxons Rangsummenstatistik}\end{gathered}$$

6.2 Kruskal-Wallis-Test

- Die Abweichung von dieser Erwartung kann man messen mit der Teststatistik

$$S = \sum_{i=1}^{I} J_i \cdot (\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2.$$

- Um aus S einen p-Wert zu erhalten, muss man die Verteilung von S unter der Nullhypothese kennen. Diese kann man für verschiedene I und J_I in Tabellen finden.
- Für $I \geq 3$ und $J_i \geq 5$ sowie $I > 3$ und $J_i \geq 4$ kann man ausnutzen, dass die folgende Skalierung K von S approximativ χ^2-verteilt ist mit $I - 1$ Freiheitsgraden:

$$K \;=\; \frac{12}{N \cdot (N+1)} S \;=\; \frac{12}{N \cdot (N+1)} \cdot \left(\sum_{i=1}^{I} J_i \cdot \overline{R}_{i\cdot}^2 \right) - 3 \cdot (N+1)$$

Kruskal-Wallis-Test mit R

```
> kruskal.test(bgz~beh,data=rat)

Kruskal-Wallis rank sum test

data:  bgz by beh
Kruskal-Wallis chi-squared = 17.0154, df = 3,
                                p-value = 0.0007016
```

6.3 Simulationsbasierte Tests

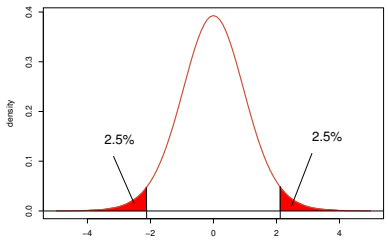
- Beim Chi-Quadrat-Test auf Unabhängigkeit
- Allgemein "Parametrisches Bootstrapping" :
  1. Passe Null-Modell und Alternativmodell an die Daten an
  2. Berechne Statistik S (z.B. likelihood ratio), die zeigt, um wieviel das Alternativmodell besser zu den Daten passt.
  3. wiederhole ganz oft (z.B. 1000 mal):
     (a) simuliere Daten gemäß angepasstem Nullmodell
     (b) passe Nullmodell und Alternativmodel an Daten an
     (c) Berechne Statistik S für Simulierte Daten
  4. Wie oft ist S bei simulierten Daten so extrem wie bei echten?

7 Überblick

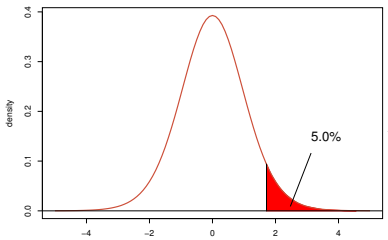
7.1 einseitig oder zweiseitig testen?

Zweiseitig oder einseitig testen?

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



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



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

			$\sum$
	a	b	45
	c	d	24
$\sum$	60	9	69

b	$\Pr(b)$
0	0.000023
1	0.00058
2	0.00604
3	0.0337
4	0.1117
5	0.2291
6	0.2909
7	0.2210
8	0.0913
9	0.0156

Einseitiger Fisher-Test:
für $b = 2$:
p-Wert=$\Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$
für $b = 3$:
p-Wert=$\Pr(0) + \Pr(1) + \Pr(2) + \Pr(3) = 0.04035434$

Zweiseitiger Fisher-Test:
Addiere alle Wahrscheinlichkeiten, die kleiner oder gleich $\Pr(b)$ sind.
für $b = 2$:
p-Wert=$\Pr(0) + \Pr(1) + \Pr(2) = 0.00665313$
für $b = 3$:
p-Wert=
$\Pr(0)+\Pr(1)+\Pr(2)+\Pr(3)+\Pr(9) = 0.05599102$

Die Frage, ob einseitig oder zweiseitig testen, stellt sich bei

- allen t-Tests:
  - ein-Stichprobe
  - zwei-Stichproben
    - gepaart
    - ungepaart mit gleiche Varianzen mit ungleiche Varianzen (Welch)
  - Steigung der Regressionsgeraden
- Wilcoxon-Test,
- Fishers exakter Test,
- z-Test.

Die Frage einseitig/zweiseitig stellt sich nicht bei

- Varianzanalyse (F-Test), Kruskal-Wallis-Test,
- Chi-Quadrat-Tests,

wo ohnehin "nach allen Seiten" getestet wird.

7.2 Nochmal zur Übersicht

Mittelwert einer Gruppe (oder Grundgesamtheit oder Population) $\mu_x = \mu_0$ **?** ein-Stichproben t-Test

Mittelwerte zweier Gruppen gleich $\mu_x = \mu_y$ **?** zwei-Stichproben t-Tests oder Wilcoxon-Test

y_i **von** x_i **unabhängig, also** $b = 0$ **in** $y_i = a + b \cdot x_i + \varepsilon_i$**?** t-Test bei linearer Regression (oder Varianzanalyse für eingebettete Modelle)

Mittelwerte in drei oder mehr Gruppen gleich $\mu_x = \mu_y = \mu_z = \ldots$ **?** Varianzanalyse (F-Test), Kruskal-Wallis-Test

Lineares Modell mit zusätzlichen Parametern nicht besser als eingebettetes Modell? Varianzanalyse (F-Test)

passen beobachtete Anzahlen zu (Annahmen über) Wahrscheinlichkeiten? Chi-Quadrat-Test, Fisher's exakter Test, z-Test