Statistik für Informatiker, SS 2018

2. Ideen aus der Statistik

2.4 Weitere Tests

Matthias Birkner

http://www.staff.uni-mainz.de/birkner/StatInfo18/

2.7.2018

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

Beispiel. Wir vermuten, dass ein gegebener sechsseitiger Würfel unfair ist.

Bei 120-maligem Würfeln finden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
h_i	13	12	20	18	26	31

Wenn der Würfel fair wäre, würden wir jeden möglichen Ausgang $1, 2, \ldots, 6$ im Mittel jeweils 20-mal erwarten.

Sind die beobachteten Abweichungen durch „reine Zufallsschwankungen“ plausibel erklärbar?

χ^2-Test für feste Gewichte: Die allgemeine Situation

Ein Experiment mit s möglichen Ausgängen werde n mal (unabhängig) wiederholt, Ausgang i habe die (unbekannte) Wahrscheinlichkeit ϑ_i, $i = 1, \ldots, s$.

Angenommen, wir beobachten h_i-mal Ausgang i für $i = 1, \ldots, s$.

Passt dies zur (Null-)Hypothese

$$H_0 : \vartheta = (\vartheta_1, \ldots, \vartheta_s) = (\rho_1, \ldots, \rho_s) = \rho$$

für einen vorgegebenen Vektor ρ von Wahrscheinlichkeitsgewichten (auf $\{1, \ldots, s\}$)?

Wenn die Nullhypothese gilt, so ist der Vektor der beobachteten Häufigkeiten multinomialverteilt (vgl. Beispiel 1.18): $(H_1^{(n)}, \ldots, H_s^{(n)}) \sim \mathrm{Mult}_{n;\rho_1,\ldots,\rho_s}$, d.h.

$$P_\rho\big(H_1^{(n)} = k_1, \ldots, H_s^{(n)} = k_s\big) = \binom{n}{k_1, k_2, \ldots, k_s} \rho_1^{k_1} \rho_2^{k_2} \cdots \rho_s^{k_s}$$

und insbesondere ist $\mathbb{E}_\rho[H_i^{(n)}] = n\rho_i$ für $i = 1, \ldots, s$

(beachte: unter H_0 ist $H_i^{(n)} \sim \mathrm{Bin}_{n,\rho_i}$)

Wir bilden

$$D := \sum_{i=1}^{s} \frac{(H_i^{(n)} - n\rho_i)^2}{n\rho_i}$$

um die typischen Abweichungen vom Erwartungswert zu quantifizieren.

Satz. Sei $\rho \in \Delta_s := \{(\vartheta_1, \ldots, \vartheta_s) \in [0,1]^s : \vartheta_1 + \cdots + \vartheta_s = 1\}$,

$$(H_1^{(n)}, \ldots, H_s^{(n)}) \sim \mathrm{Mult}_{n;\rho_1,\ldots,\rho_s},$$

dann gilt

$$\sum_{i=1}^{s} \frac{(H_i^{(n)} - n\rho_i)^2}{n\rho_i} \xrightarrow[n\to\infty]{d} \chi^2_{s-1}$$

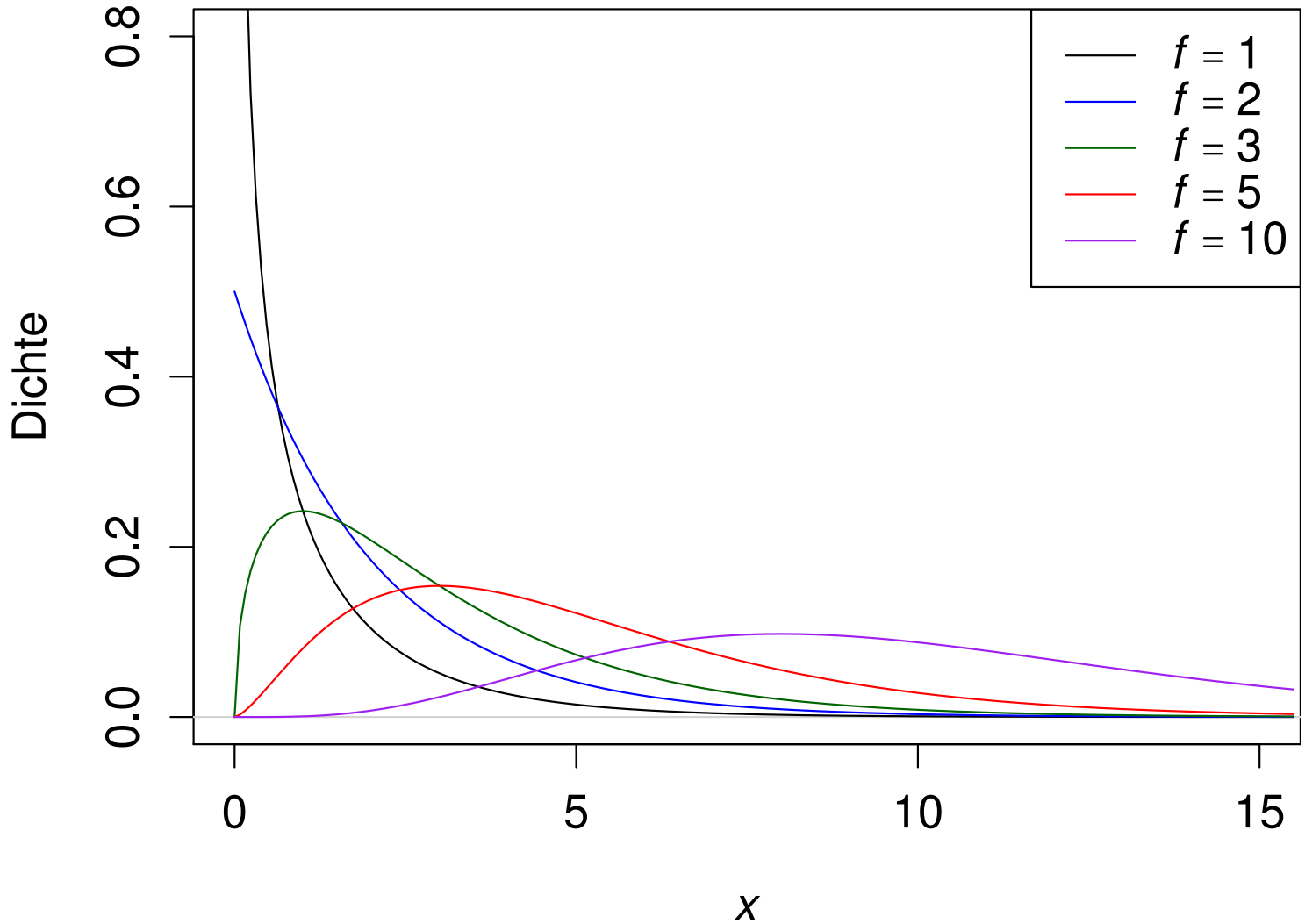
Für $m \in \mathbb{N}$ ist die χ^2_m-Verteilung („Chi-Quadrat-Verteilung mit m Freiheitsgraden“) die Verteilung der Summe

$$Z_1^2 + \cdots + Z_m^2 \sim \chi^2_m$$

wobei $Z_1, \ldots, Z_m$ u.i.v. $\sim \mathcal{N}_{0,1}$

Die χ^2_m-Verteilung besitzt die Dichtefunktion

$$\frac{1}{\Gamma(m/2)} 2^{-m/2} x^{\frac{m}{2}-1} e^{-x/2} \mathbf{1}_{[0,\infty)}(x)$$


Dichte der χ^2_f-Verteilung für verschiedene f
Dichte
0.0
0.2
0.4
0.6
0.8
0
5
10
15
x
$f = 1$
$f = 2$
$f = 3$
$f = 5$
$f = 10$

Zum theoretischen Hintergrund

Wo kommen hier die Quadrate von Normalverteilten her?

Es steckt eine Version des multivariaten zentralen Grenzwertsatzes dahinter:

Sei $u_\rho := (\sqrt{\rho_1}, \ldots, \sqrt{\rho_s}) \in \mathbb{R}^s$ (beachte: $||u_\rho|| = 1$),
$\mathbb{H}_\rho := \{x \in \mathbb{R}^s : x \cdot u_\rho = 0\}$
(die Hyperebene durch den Ursprung mit Normale u_ρ)
und $\Pi_\rho : \mathbb{R}^s \to \mathbb{H}_\rho$, $\Pi_\rho(x) = x - (x \cdot u_\rho)u_\rho$
(die orthogonale Projektion auf $\mathbb{H}_\rho$)

dann gilt

$$\left(\frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}\right)_{i=1,\ldots,s} \xrightarrow[n\to\infty]{d} \Pi_\rho(Z)$$

mit $Z = (Z_1, \ldots, Z_s)$ s-dim. Standard-normal

$$\left(\frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}\right)_{i=1,\dots,s} \xrightarrow[n\to\infty]{d} \Pi_\rho(Z)$$

mit $Z = (Z_1, \dots, Z_s)$ s-dim. Standard-normal

Man „verliert“ gewissermaßen einen Freiheitsgrad durch die Projektion auf den $(s-1)$-dimensionalen Teilraum Π_ρ

Für Details siehe die Literatur, z.B. [Georgii, Kap. 11.1–11.2], wir beobachten hier nur:

$$\mathrm{Cov}[H_i^{(n)}, H_j^{(n)}] = \begin{cases} n\rho_i(1-\rho_i), & i = j, \\ -n\rho_i\rho_j, & i \neq j \end{cases}$$

somit für $\widetilde{H}_i^{(n)} = \frac{H_i^{(n)} - n\rho_i}{\sqrt{n\rho_i}}$: $\mathrm{Cov}[\widetilde{H}_i^{(n)}, \widetilde{H}_j^{(n)}] = \begin{cases} 1-\rho_i^2, & i = j, \\ -\rho_i\rho_j, & i \neq j \end{cases}$

d.h. die Kovarianzmatrix $C = \left(\mathrm{Cov}[H_i^{(n)}, H_j^{(n)}]\right)_{i,j=1,\dots,s}$ ist $C = I - (\rho_i\rho_j)_{i,j=1,\dots,s}$ (mit $I = s \times s$-Einheitsmatrix) und dies ist die Kovarianzmatrix von $\Pi_\rho(Z)$

χ^2-Test für feste Gewichte (auch χ^2-Anpassungstest genannt)

Sei $\vartheta \in \Theta = \Delta_s := \{(\vartheta_1, \ldots, \vartheta_s) \in [0,1]^s : \vartheta_1 + \cdots + \vartheta_s = 1\}$,
unter P_ϑ sei $(H_1, \ldots, H_s) \sim \mathrm{Mult}_{n;\vartheta_1,\ldots,\vartheta_s}$.

Sei $\rho \in \Delta_s$,

$$D := \sum_{i=1}^{s} \frac{(H_i - n\rho_i)^2}{n\rho_i},$$

$\alpha \in (0,1)$, q das $(1-\alpha)$-Quantil der χ^2_{s-1}-Verteilung.

Der Test von $H_0 : \{\vartheta = \rho\}$ gegen $H_1 : \{\vartheta \neq \rho\}$ mit Ablehnungsbereich $\{D > q\}$ hat (asymptotisches) Niveau α.

(Dies folgt aus obigem Satz.)

Die Quantile der χ^2-Verteilungen findet man (traditionell) in Quantiltabellen,

`R` kennt `[d|p|q|r]chisq`
(z.B. `qchisq(`β`, df=`m`)` berechnet das β-Quantil von χ^2_m)

Zurück zu unserem Beispiel:

Bei 120-maligem Würfeln fanden wir folgende Häufigkeiten:

i	1	2	3	4	5	6
h_i	13	12	20	18	26	31

`R` kennt den χ^2-Test: `chisq.test`

```
> w <- c(13,12,20,18,26,31)
> chisq.test(w,p=c(1/6,1/6,1/6,1/6,1/6,1/6))

Chi-squared test for given probabilities

data:  w
X-squared = 13.7, df = 5, p-value = 0.01763
```

Eine Alternative: *p*-Wert per Simulation

Die asymptotische Aussage obigen Satzes sagt nichts darüber, wie groß n sein sollte, damit die Approximation plausibel ist.

Eine oft zitierte Faustregel (für die Gültigkeit der χ^2-Approximation) ist $n\rho_i \geq 5$ für alle i.

Wir können den *p*-Wert auch näherungsweise bestimmen, indem wir D sehr oft unter der Nullhypothese simulieren:

$D_1, \ldots, D_M$ simulierte Werte, wir haben $D = d$ beobachtet: Lehne H_0 zum Niveau α ab, wenn

$$\frac{1}{M}\#\{1 \leq i \leq M : D_i \geq d\} \leq \alpha$$

gilt.

Eine Alternative: *p*-Wert per Simulation

Lassen wir für das Beispiel R den *p*-Wert via Simulation bestimmen:

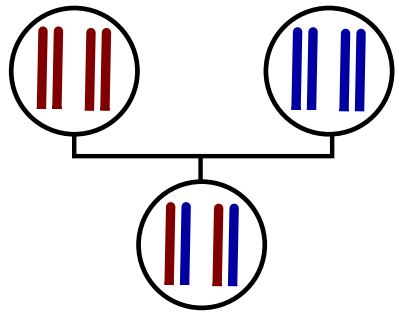
```
> w <- c(13,12,20,18,26,31)
> chisq.test(w, p=c(1/6,1/6,1/6,1/6,1/6,1/6),
             simulate.p.value=TRUE)

Chi-squared test for given probabilities with
        simulated p-value
        (based on 2000 replicates)

data:  w
X-squared = 13.7, df = NA, p-value = 0.01799
```

Beispiel (Mendels Erbsenexperimente[1])**.**

Beim Kreuzen von Doppelhybriden erwarten wir folgende Phänotypwahrscheinlichkeiten unter Mendel'scher Segregation („rund“ und „gelb“ sind jeweils dominant, $n = 556$ Versuche):



Mendels Beobachtungen:

Typ	rund/gelb	rund/grün	kantig/gelb	rund/gelb
Anteil	9/16	3/16	3/16	1/16
Erwartete Anz.	315	104,25	104,25	34,75
beobachtet	315	108	101	32

[1] Gregor Mendel, 1822–1884; G. Mendel, Versuche über Pflanzenhybriden, Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen: 3–47, (1866).

Analysieren wir Mendels Daten mit R:

```
> x <- c(315, 108, 101, 32)
> chisq.test(x, p=c(9/16,3/16,3/16,1/16))

Chi-squared test for given probabilities

data:  x
X-squared = 0.47, df = 3, p-value = 0.9254
```

Insoweit passen die Daten sehr gut zu den theoretischen Häufigkeiten.

(Manche haben später argumentiert: „fast zu gut“)

χ^2-Test auf Homogenität

auch χ^2-Test auf Unabhängigkeit oder
Pearsons χ^2-Test genannt

(nach Karl Pearson, 1857–1936)

In einem Experiment werden zwei „Merkmale“ beobachtet,

und es geht grob gesprochen darum, ob die Verteilungen von Merkmal 1 und von Merkmal 2 unabhängig sind.

χ^2-Test auf Homogenität: Situation (abstrakt)

In einem Experiment werden zwei „Merkmale“ beobachtet, wobei das erste Merkmal a und das zweite Merkmal b viele Ausprägungen besitzt (also insgesamt $s = a \cdot b$ mögliche Ausgänge).

Unter n u.a. Wiederholungen werde h_{ij} mal Ausgang (i, j) beobachtet ($i \in \{1, 2, \ldots, a\}$, $j \in \{1, 2, \ldots, b\}$), man fasst die Beobachtungen in einer $a \times b$-Kontingenztafel zusammen:

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

mit Zeilensummen $h_{i\cdot} = \sum_{j=1}^{b} h_{ij}$,
Spaltensummen $h_{\cdot j} = \sum_{i=1}^{a} h_{ij}$
und Gesamtsumme $h_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} h_{ij} = n$

i \ j	1	2	3	
1	h_{11}	h_{12}	h_{13}	$h_{1\cdot}$
2	h_{21}	h_{22}	h_{23}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$h_{\cdot 3}$	$h_{\cdot\cdot} = n$

Wir fassen die beobachteten Häufigkeiten als Realisierungen einer multinomial$(n, (\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b})$-verteilten ZV $(H_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$ auf, wobei

$(\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b}$ ein $a \cdot b$-dimensionaler Vektor von Wahrscheinlichkeitsgewichten ist.

Passen die Beobachtungen zur Nullhypothese, dass

$$\vartheta_{ij} = \eta_i \cdot \rho_j, \quad \text{für } i = 1, \ldots, a,\ j = 1, \ldots, b$$

mit $(\eta_i)_{i=1,\ldots,a}$, $(\rho_j)_{j=1,\ldots,b}$ gewissen a- bzw. b-dimensionalen Vektoren von Wahrscheinlichkeitsgewichten?

Wir bilden

$$\widehat{\vartheta}_{i\cdot} = \frac{H_{i\cdot}}{n}, \quad \widehat{\vartheta}_{\cdot j} = \frac{H_{\cdot j}}{n}$$

(dies sind die ML-Schätzer für η_i bzw. für ρ_j)

und die Teststatistik

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j})^2}{n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j}}$$

Unter

$$H_0 : \text{„}(\vartheta_{ij})_{i=1,\ldots,a;j=1,\ldots,b} \text{ hat Produktform“}$$

ist D (approximativ) $\chi^2_{(a-1)(b-1)}$-verteilt.

χ^2-Test auf Homogenität

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j})^2}{n\widehat{\vartheta}_{i\cdot}\widehat{\vartheta}_{\cdot j}}$$

mit

$$\widehat{\vartheta}_{i\cdot} = \frac{H_{i\cdot}}{n}, \quad \widehat{\vartheta}_{\cdot j} = \frac{H_{\cdot j}}{n}$$

Lehne H_0 zum Niveau α ab, falls der beobachtete Wert größer ist als das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $(a-1)(b-1)$ Freiheitsgraden.

Dahinter steckt wieder eine Projektion einer (hochdimensionalen) Normalverteilung, Intuition für die Anzahl Freiheitsgrade:
Es gibt im „freien Modell“ (ohne Unabhängigkeitsannahme) $a \cdot b - 1$ Parameter, wir haben $(a-1)+(b-1)$ Parameter geschätzt, es bleiben $a \cdot b - 1 - (a-1) - (b-1) = (a-1)(b-1)$ „Freiheitsgrade“.

Beispiel

Der Kuhstärling ist ein Brutparasit des Oropendola.

photo (c) by J. Oldenettel

- Normalerweise entfernen Oropendolas alles aus ihrem Nest, was nicht genau nach ihren Eiern aussieht.
- In einigen Gegenden sind Kuhstärling-Eier gut von Oropendola-Eiern zu unterscheiden und werden trotzdem nicht aus den Nestern entfernt.
- Mögliche Erklärung: Nester mit Kuhstärling-Eiern sind eventuell besser vor Befall durch Dasselfliegenlarven geschützt.

(vgl. N.G. Smith, The advantage of being parasitized.
Nature 219(5155):690-4, (1968))

Anzahlen von Nestern, die von Dasselfliegenlarven befallen sind

Anzahl Kuhstärling-Eier	0	1	2
befallen	16	2	1
nicht befallen	2	11	16

In Prozent:

Anzahl Kuhstärling-Eier	0	1	2
befallen	89%	15%	6%
nicht befallen	11%	85%	94%

- Anscheinend ist der Befall mit Dasselfliegenlarven reduziert, wenn die Nester Kuhstärlingeier enthalten. Statistisch signifikant?
- Nullhypothese: Die Wahrscheinlichkeit eines Nests, mit Dasselfliegenlarven befallen zu sein hängt nicht davon ab, ob oder wieviele Kuhstärlingeier in dem Nest liegen.

Anzahlen der von Dasselfliegenlarven befallenen Nester

Anzahl Kuhstärling-Eier	0	1	2	$\sum$
befallen	16	2	1	19
nicht befallen	2	11	16	29
$\sum$	18	13	17	48

Welche Anzahlen würden wir unter der Nullhypothese erwarten?

Das selbe Verhältnis 19/48 in jeder Gruppe.

Erwartete Anzahlen von Dasselfliegenlarven befallener Nester, bedingt auf die Zeilen- und Spaltensummen:

Anzahl Kuhstärling-Eier	0	1	2	$\sum$
befallen	7.13	5.15	6.72	19
nicht befallen	10.87	7.85	10.28	29
$\sum$	18	13	17	48

$$18 \cdot \frac{19}{48} = 7.13 \qquad 13 \cdot \frac{19}{48} = 5.15$$

Alle anderen Werte sind nun festgelegt durch die Summen.

beobachtet (O, observed):

befallen	16	2	1	19
nicht befallen	2	11	16	29
$\sum$	18	13	17	48

erwartet: (E):

befallen	7.13	5.15	6.72	19
nicht befallen	10.87	7.85	10.28	29
$\sum$	18	13	17	48

O-E:

befallen	8.87	-3.15	-5.72	0
nicht befallen	-8.87	-3.15	5.72	0
$\sum$	0	0	0	0

$$D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(H_{ij} - n\widehat{\vartheta}_{i.}\widehat{\vartheta}_{.j})^2}{n\widehat{\vartheta}_{i.}\widehat{\vartheta}_{.j}} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 29.5$$

Beachte:

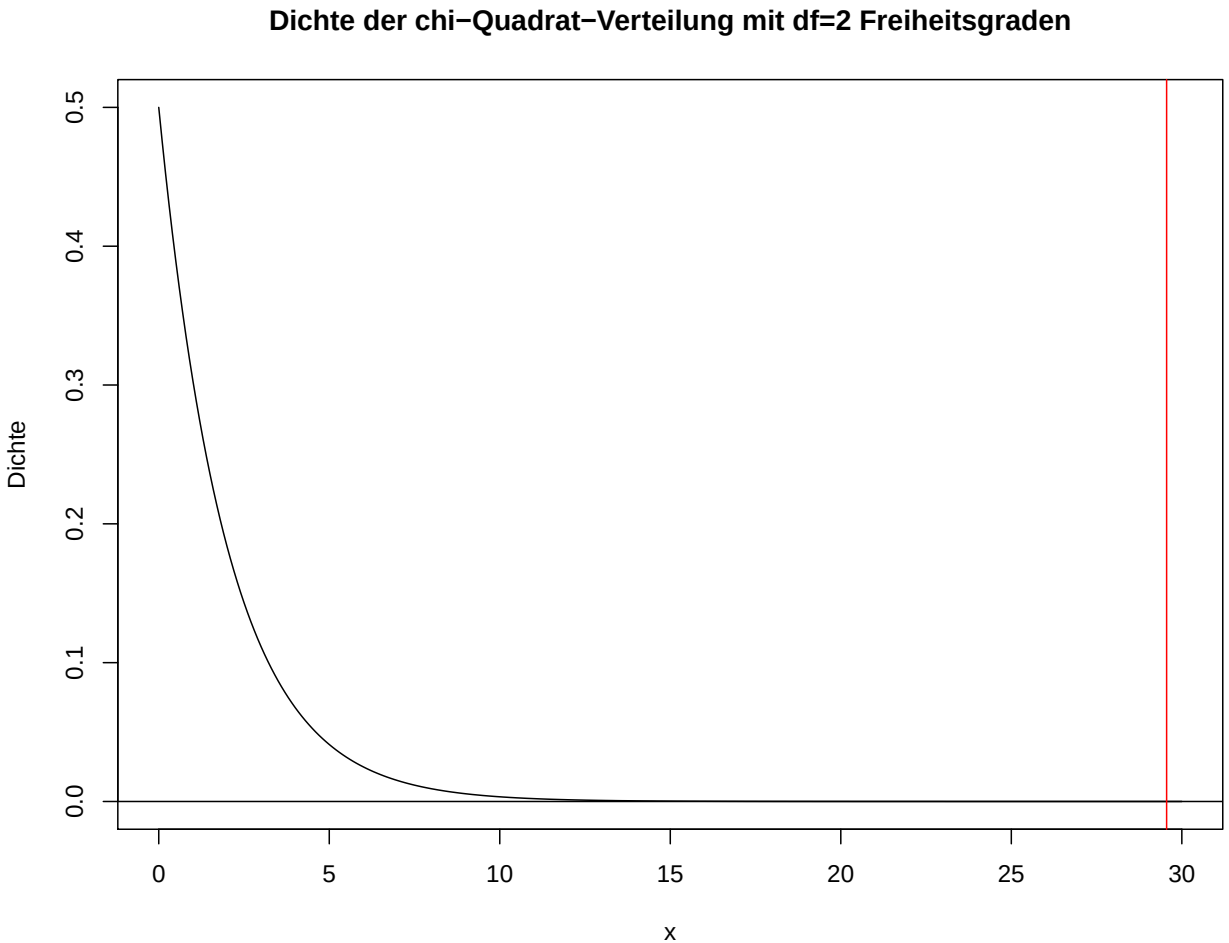
- Wenn die Zeilen- und Spaltensummen gegeben sind, bestimmen bereits 2 Werte in der Tabelle alle anderen Werte
- Insbesondere df=2 für Kontingenztafeln mit zwei Zeilen und drei Spalten.

Wir haben den Wert $\chi^2 = 29.5$ beobachtet.

Unter der Nullhypothese „die Wahrscheinlichkeit, mit der ein Nest von Dasselfliegenlarven befallen wird, hängt nicht von der Anzahl Kuhstärling-Eier ab“ ist die Teststatistik (approximativ) χ^2-verteilt mit $2 = (2-1)\cdot(3-1)$ Freiheitsgraden.

Das 99%-Quantil der χ^2-Verteilung mit df=2 ist 9.21 (<29.5), wir können also die Nullhypothese zum Signifikanzniveau 1% ablehnen.
(Denn wenn die Nullhypothese zutrifft, so würden wir in weniger als 1% der Fälle einen so extremen Wert der χ^2-Statistik beobachten.)

(Siehe die folgenden Folien für die mit dem Computer bestimmten exakten *p*-Werte.)



Bemerkung 1: Genauere Rechnung ergibt: Für ein χ^2_2-verteiltes X gilt $\mathbb{P}(X \geq 29.6) = 3.74 \cdot 10^{-7}$ (was hier wörtlich der p-Wert des χ^2-Tests auf Unabhängigkeit wäre, in dieser Genauigkeit für statistische Zwecke allerdings sinnlos ist).

Bemerkung 2: Um die Gültigkeit der χ^2-Approximation (und der Faustregel) in diesem Beispiel einzuschätzen, könnten wir wieder den Computer beauftragen, durch vielfach wiederholte Simulation den *p*-Wert zu schätzen. Mit **R** funktioniert das folgendermaßen:

```
> M <- matrix(c(16,2,2,11,1,16),nrow=2)
> M
     [,1] [,2] [,3]
[1,]   16    2    1
[2,]    2   11   16
> chisq.test(M,simulate.p.value=TRUE,B=50000)

  Pearson's Chi-squared test with simulated p-value
  (based on 50000 replicates)

data:  M
X-squared = 29.5544, df = NA, p-value = 2e-05
```

Wir sehen: Der empirisch geschätze *p*-Wert $2 \cdot 10^{-5}$ stimmt zwar nicht mit dem aus der χ^2-Approximation überein, aber beide sind hochsignifikant klein (und in einem Bereich, in dem der exakte Wert sowieso statistisch „sinnlos“ ist). Insoweit ist die Faustregel hier bestätigt.

Simpson-Paradoxon

Durch Zusammenfassen von Gruppen können sich (scheinbare) statistische Trends in ihr Gegenteil verkehren.

Dieses Phänomen heißt Simpson-Paradoxon oder Yule-Simpson-Effekt.

(nach Edward H. Simpson, *1922 und George Udny Yule, 1871–1951)

Simpson-Paradoxon

Beispiel: Zulassungsstatistik der UC Berkeley 1973

Im Herbst 1973 haben sich an der Universität Berkeley 12763 Kandidaten für ein Studium beworben, davon 8442 Männer und 4321 Frauen. Es kam zu folgenden Zulassungszahlen:

	Aufgenommen	Abgelehnt
Männer	3738	4704
Frauen	1494	2827

Demnach betrug die Zulassungsquote
bei den Männern $\frac{3738}{8442} \approx 44\%$, bei den Frauen nur $\frac{1494}{4321} \approx 35\%$.

Ein χ^2-Test auf Homogenität (z.B. mit R) zeigt, dass eine solche Unverhältnismäßigkeit nur mit verschwindend kleiner Wahrscheinlichkeit durch „reinen Zufall“ entsteht:

```
> berkeley <- matrix(c(3738,1494,4704,2827),
                                          nrow=2)
> berkeley
     [,1] [,2]
[1,] 3738 4704
[2,] 1494 2827
> chisq.test(berkeley,correct=FALSE)

Pearson's Chi-squared test

data:  berkeley
X-squared = 111.2497, df = 1, p-value < 2.2e-16
```

Dieser Fall hat einiges Aufsehen erregt, s.a. P.J. Bickel, E.A. Hammel, J.W. O'Connell, Sex Bias in Graduate Admissions: Data from Berkeley, *Science*, **187**, no. 4175, 398–404 (1975).

Das Ungleichgewicht verschwindet, wenn man die Zulassungszahlen nach Departments aufspaltet:

Es stellt sich heraus, dass innerhalb der Departments die Aufnahmewahrscheinlichkeiten nicht signifikant vom Geschlecht abhängen, aber sich Frauen häufiger bei Departments mit (absolut) niedriger Aufnahmequote beworben haben als Männer – dies ist ein Beispiel für das *Simpson-Paradox*.

Die genauen nach Departments aufgeschlüsselten Bewerber- und Zulassungszahlen sind leider nicht öffentlich zugänglich (siehe aber Abb. 1 in Bickel et. al, loc. cit., für eine grafische Aufbereitung der Daten, die den Simpson-Effekt zeigt).

Bickel et. al demonstrieren das Phänomen mittels eines hypothetischen Beispiels:

	Aufgenommen	Abgelehnt
	Department of machismathics	
Männer	200	200
Frauen	100	100
	Department of social warfare	
Männer	50	100
Frauen	150	300
	Gesamt	
Männer	250	300
Frauen	250	400

Erinnerung

Nehmen wir an, wir haben zufällige Stichproben aus 2 Gruppen:

$x_1, x_2, \ldots, x_{n_1}$ n_1 Beobachtungswerte aus Population 1,

$y_1, y_2, \ldots, y_{n_2}$ n_2 Beobachtungswerte aus Population 2

(beispielsweise die Länge von Backenzähnen für zwei Stichproben von zwei verschiedenen Urpferdchen-Arten).

Der (uns unbekannte) wahre Populationsmittelwert ist

μ_1 in Population 1, μ_2 in Population 2.

Frage Ist (angesichts der Beobachtungen) die Annahme

$\mu_1 = \mu_2$ plausibel?

Erinnerung (ungepaarter t-Test)

Gegeben

$x_1, x_2, \ldots, x_{n_1}$ n_1 Beobachtungswerte aus Population 1,

$y_1, y_2, \ldots, y_{n_2}$ n_2 Beobachtungswerte aus Population 2

Um die Nullhypothese

$H_0 : \mu_1 = \mu_2$ d.h. Mittelwerte in beiden Populationen gleich

zu püfen, können wir den ungepaarten t-Test verwenden.

Erinnerung

(zweiseitiger, ungepaarter t-Test, Ann. gleicher Varianzen)

Mit $$\overline{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i,\ \overline{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i,$$

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\overline{x})^2,\ s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i-\overline{y})^2$$

(Stichprobenmittelwerte und korrigierte Stichprobenvarianzen),

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

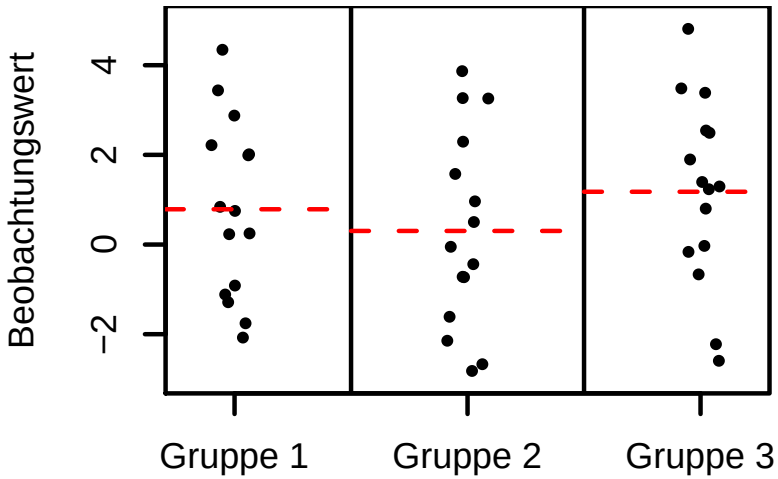
(gepoolte Stichprobenvarianz) berechne $t = \frac{\overline{x}_1-\overline{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, lehne $H_0 : \mu_1 = \mu_2$ zum Signifikanzniveau α ab, wenn

$|t| >$ $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung mit n_1+n_2-2 Freiheitsgraden.

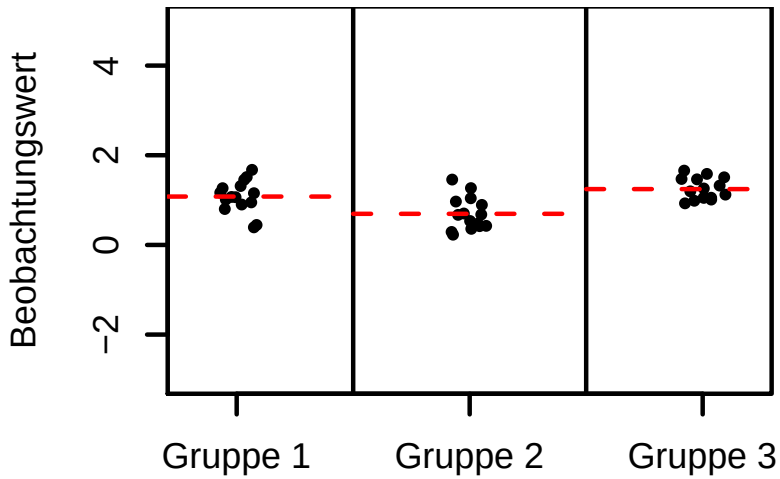
Frage Was tun, wenn mehr als zwei Gruppen vorliegen?

Grundidee der 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: Blutgerinnungszeiten

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

Bemerkung: Der Globalmittelwert ist in diesem Beispiel auch der Mittelwert der Gruppenmittelwerte. Das muss aber nicht immer so sein!

Beispiel

Blutgerinnungszeit bei Ratten unter 4 versch. Behandlungen

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

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“)

	Freiheits-grade (DF)	Quadrat-summe (SS)	mittlere Quadrat-summe (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,
das 95%-Quantil der $\mathrm{Fisher}_{3,20}$-Verteilung ist 3,098 ($< 13{,}57$).

Wir können demnach H_0 zum Signifikanzniveau 5% ablehnen.

(Der p-Wert ist $\mathrm{Fisher}_{3,20}([13{,}57, \infty)) \leq 5 \cdot 10^{-5}$.)

F-Test, allgemein

$n = n_1 + n_2 + \cdots + n_I$ 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

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

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 $F \geq q_\alpha$, wobei q_α das $(1-\alpha)$-Quantil der Fisher-Verteilung mit $I - 1$ und $n - I$ Freiheitsgraden ist.

Zur Theorie:

Seien $m, n \in \mathbb{N}$, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ unabhängig, $\sim \mathcal{N}_{0,1}$.

$$F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2} \text{ hat Dichte}$$

$$f_{m,n}(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{(m+n)}{2}}} \mathbf{1}_{(0,\infty)}(x).$$

$\mathscr{L}(F_{m,n})$ heißt *Fisher-Verteilung*[2] mit m und n Freiheitsgraden (präziser: mit m Zähler- und n Nenner-Freiheitsgraden).

[2] Nach Ronald Aylmer Fisher, 1890–1962

Tabelle der 95%-Quantile der F-Verteilung

Die folgende Tabelle zeigt (auf 2 Nachkommastellen gerundet) das 95%-Quantil der Fisher-Verteilung mit k_1 und k_2 Freiheitsgraden (k_1 Zähler- und k_2 Nennerfreiheitsgrade)

k_2 \ 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

Bemerkung: F-Test mit 2 Gruppen $\widehat{=}$ t-Test

Für $I = 2$ Gruppen ist $\overline{X}_{..} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij} = \frac{n_1}{n_1+n_2}\overline{X}_{1.} + \frac{n_2}{n_1+n_2}\overline{X}_{2.}$

und somit

$\overline{X}_{1.} - \overline{X}_{..} = \frac{n_2}{n_1+n_2}(\overline{X}_{1.} - \overline{X}_{2.})$, $\overline{X}_{2.} - \overline{X}_{..} = \frac{n_1}{n_1+n_2}(\overline{X}_{2.} - \overline{X}_{1.})$, d.h.

$$SS_{\text{zw}} = n_1(\overline{X}_{1.} - \overline{X}_{..})^2 + n_2(\overline{X}_{2.} - \overline{X}_{..})^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{X}_{1.} - \overline{X}_{2.}\right)^2.$$

Weiter ist $SS_{\text{innerh}} = \sum_{j=1}^{n_1}(X_{1j} - \overline{X}_{1.})^2 + \sum_{j=1}^{n_2}(X_{2j} - \overline{X}_{2.})^2 = (n_1 + n_2 - 2)s^2$

($s^2 = \frac{n_1-1}{n_1+n_2-2}s_1^2 + \frac{n_2-1}{n_1+n_2-2}s_2^2$ ist die gepoolte Stichprobenvarianz)

Insgesamt:

$$F = \frac{SS_{\text{zw}}/1}{SS_{\text{innerh}}/(n_1 + n_2 - 2)} = \frac{(\overline{X}_{1.} - \overline{X}_{2.})^2}{s^2(\frac{1}{n_1} + \frac{1}{n_2})} = T^2$$

Berechnung der Signifikanz mit R

Wie muss man q wählen, damit $\mathbb{P}(F \leq q) = 0.95$ für Fisher(6,63)-verteiltes F?

```
> qf(0.95,df1=6,df2=63)
[1] 2.246408
```

p-Wert-Berechnung: Wie wahrscheinlich ist es, dass eine Fisher(3,20)-verteilte Zufallsgröße einen Wert ≥ 13.57 annimmt?

```
> pf(13.57, df1=3, df2=20, lower.tail=FALSE)
[1] 4.66169e-05
```

Varianzanalyse komplett in R

Die Text-Datei gerinnung.txt enthält eine Spalte “bgz” mit den Blutgerinnungszeiten und eine Spalte “beh” mit der Behandlung (A,B,C,D).

```
> rat<-read.table("gerinnung.txt",header=TRUE)
> rat.aov <- aov(bgz~beh,data=rat)
> summary(rat.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
beh          3    228    76.0  13.571 4.658e-05 *
Residuals   20    112     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
```

Mediantest

Modell: $X_1, \ldots, X_n$ u.i.v. reellwertig, $\sim \rho$ mit Median $m(\rho)$
(und ρ habe stetige Verteilungsfunktion)

Wir möchten anhand von Beobachtungen $x_1, \ldots, x_n$ prüfen, ob $m(\rho) = m_0$ plausibel ist, d.h. wir testen

$$H_0 : m(\rho) = m_0 \quad \text{gegen} \quad H_1 : m(\rho) \neq m_0$$

Sei

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)} \quad \text{die Ordnungsstatistik,} \quad \alpha \in (0, 1)$$

wähle k (möglichst groß) mit $\mathrm{Bin}_{n,1/2}(\{0, \ldots, k-1\}) \leq \frac{\alpha}{2}$,

wenn $[x_{(k)}, x_{(n-k+1)}] \ni m_0$, so nehme H_0 an, sonst lehne H_0 ab zugunsten von H_1.

Dieser Test hält das Niveau α ein.

Mediantest, einseitig

Analog im einseitigen Fall:

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)} \quad \text{die Ordnungsstatistik,} \quad \alpha \in (0,1)$$

wähle k' möglichst groß, so dass $Bin_{n,1/2}(\{0,\ldots,k'-1\}) \leq \alpha$ gilt.

Wir testen

$$H_0 : m(\rho) \leq m_0 \quad \text{gegen} \quad H_1 : m(\rho) > m_0$$

folgendermaßen:

Lehne H_0 ab, wenn $x_{(k')} > m_0$.

Dieser Test hält das Niveau α ein.

Mediantest: Theoretische Begründung

Sei ρ eine Verteilung auf $\mathbb{R}$ mit stetiger Verteilungsfunktion und Median $m(\rho) = m_0$, $X_1, \ldots, X_n$ u.i.v., $\sim \rho$

$$\begin{aligned} P_\rho\big(X_{(k)} > m(\rho)\big) &= P_\rho\big(|\{1 \le i \le n : X_i \le m(\rho)\}| \le k-1\big) \\ &= \mathrm{Bin}_{n,1/2}(\{0,\ldots,k-1\}) \le \frac{\alpha}{2}, \end{aligned}$$

analog ist

$$P_\rho\big(X_{(n-k-1)} < m(\rho)\big) = P_\rho\big(|\{1 \le i \le n : X_i \ge m(\rho)\}| \le k-1\big) \le \frac{\alpha}{2},$$

somit

$$\begin{aligned} &P_\rho\Big(\big[X_{(k)}, X_{(n-k+1)}\big] \not\ni m(\rho)\Big) \\ &\quad \le P_\rho\big(X_{(k)} > m(\rho)\big) + P_\rho\big(X_{(n-k-1)} < m(\rho)\big) \le \alpha. \end{aligned}$$

(Da ρ stetige Verteilungsfunktion hat, gilt $P_\rho(X_i = X_j) = 0$ für $i \neq j$ und $P_\rho(X_i = m(\rho)) = 0$.)

Wilcoxons Rangsummen-Test

Ein „verteilungsfreier“ Test,
mit dem man die Lage zweier Verteilungen
zueinander testen kann.

Beobachtungen: Zwei Stichproben

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

Wir möchten die Nullhypothese:
X und Y haben diesselbe Verteilung
testen

gegen die Alternative:
Die beiden Verteilungen sind gegeneinander verschoben.

(Die Situation ist ähnlich zum zwei-Stichproben-t-Test, aber wir möchten *nicht* die implizite Annahme treffen, dass es sich dabei (wenigstens ungefähr) um Normalverteilungen handelt.)

Idee

Beobachtungen:

$$X : x_1, x_2, \ldots, x_m \qquad \text{und} \qquad Y : y_1, y_2, \ldots, y_n$$

- 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

Beobachtungen:

$X : x_1, x_2, \ldots, x_m$

$Y : y_1, y_2, \ldots, y_n$

W = Summe der X-Ränge $- (1 + 2 + \cdots + m)$

heißt

Wilcoxons Rangsummenstatistik

Die Normierung ist so gewählt, dass $0 \leq W \leq m\,n$.

Wilcoxons Rangsummenstatistik

Bemerkung 1:

$$W = \text{Summe der } X\text{-Ränge} - (1 + 2 + \cdots + m)$$

Wir könnten auch die Summe der Y-Ränge benutzen, denn

$$\begin{aligned} &\text{Summe der } X\text{-Ränge} + \text{Summe der } Y\text{-Ränge} \\ = \;& \text{Summe aller Ränge} \\ = \;& 1 + 2 + \cdots + (m+n) = \frac{(m+n)(m+n+1)}{2} \end{aligned}$$

Bemerkung 2:

Der Wilcoxon-Test heißt auch Mann-Whitney-Test, die Rangsummenstatistik auch Mann-Whitney Statistik U, sie unterscheidet sich (je nach Definition) von W um eine Konstante.

(In der Literatur sind beide Bezeichnungen üblich, man prüfe vor Verwendung von Tabellen, etc. die verwendete Konvention.)

Ein kleines Beispiel

- Beobachtungen:

$$X : 1{,}5;\ 5{,}6;\ 35{,}2$$
$$Y : 7{,}9;\ 38{,}1;\ 41{,}0;\ 56{,}7;\ 112{,}1;\ 197{,}4;\ 381{,}8$$

- Lege Beobachtungen zusammen und sortiere:
  $1{,}5;\ 5{,}6;\ 7{,}9;\ 35{,}2;\ 38{,}1;\ 41{,}0;\ 56{,}7;\ 112{,}1;\ 197{,}4;\ 381{,}8$
- Bestimme Ränge:
  $1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10$
- Rangsummenstatistik hier: $W = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Interpretation von W

X-Population kleiner $\Longrightarrow$ W klein:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 0$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 1$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 2$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 2$

X-Population größer $\Longrightarrow$ W groß:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 21$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 20$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 19$

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 $W = 19$

Signifikanz

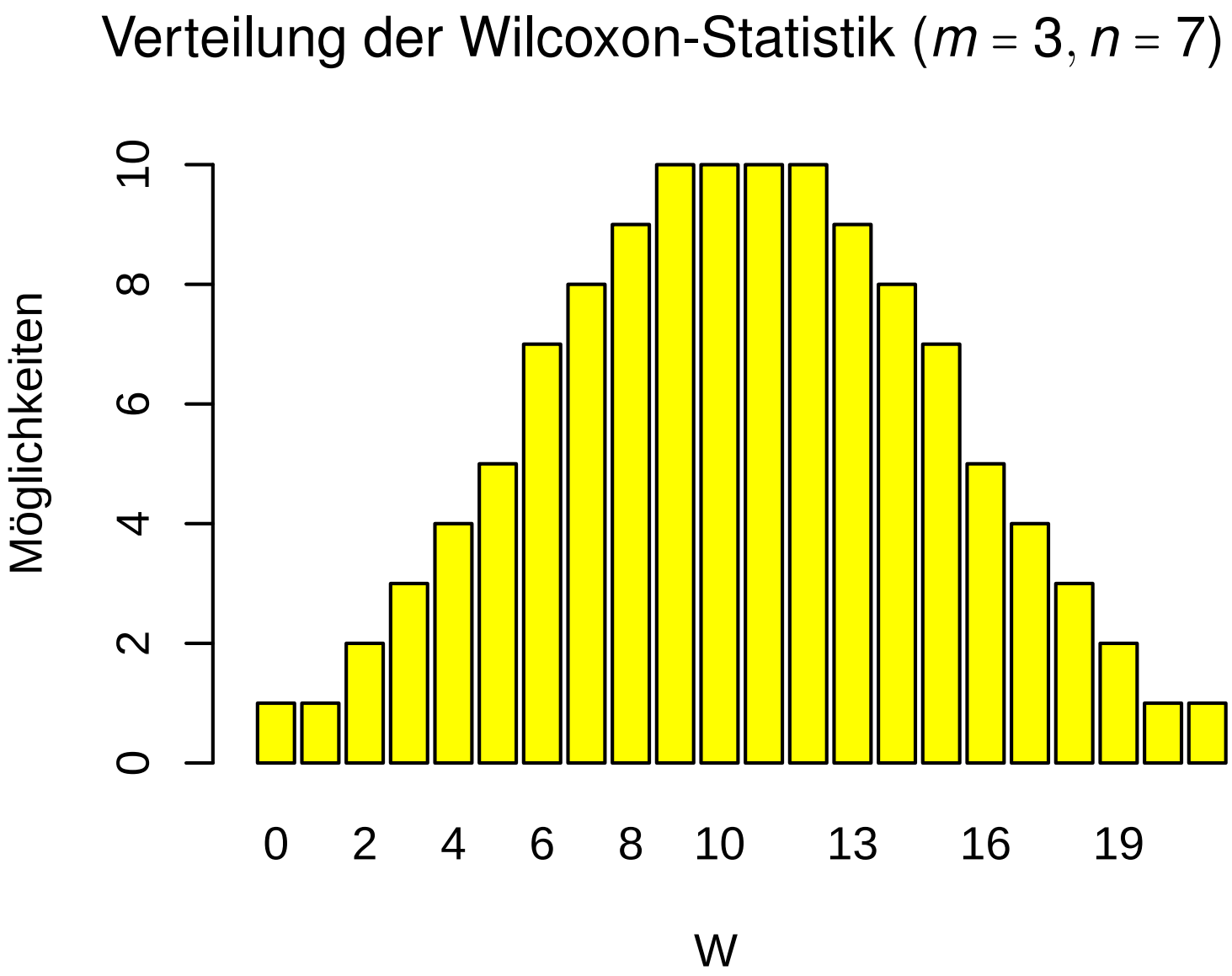
Nullhypothese:
X-Stichprobe und Y-Stichprobe
stammen aus
derselben Verteilung

Die 3 Ränge der X-Stichprobe
1 2 3 4 5 6 7 8 9 10

hätten genausogut irgendwelche 3 Ränge
1 2 3 4 5 6 7 8 9 10
sein können.

Es gibt $\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$ Möglichkeiten.

(Allgemein: $\frac{(m+n)(m+n-1)\cdots(n+1)}{m(m-1)\cdots 1}) = \frac{(m+n)!}{n!m!} = \binom{m+n}{m}$ Möglichkeiten)


Verteilung der Wilcoxon-Statistik ($m = 3, n = 7$)
Möglichkeiten
0
2
4
6
8
10
0
2
4
6
8
10
13
16
19
W

Unter der Nullhypothese sind alle Rangbelegungen gleich wahrscheinlich, also

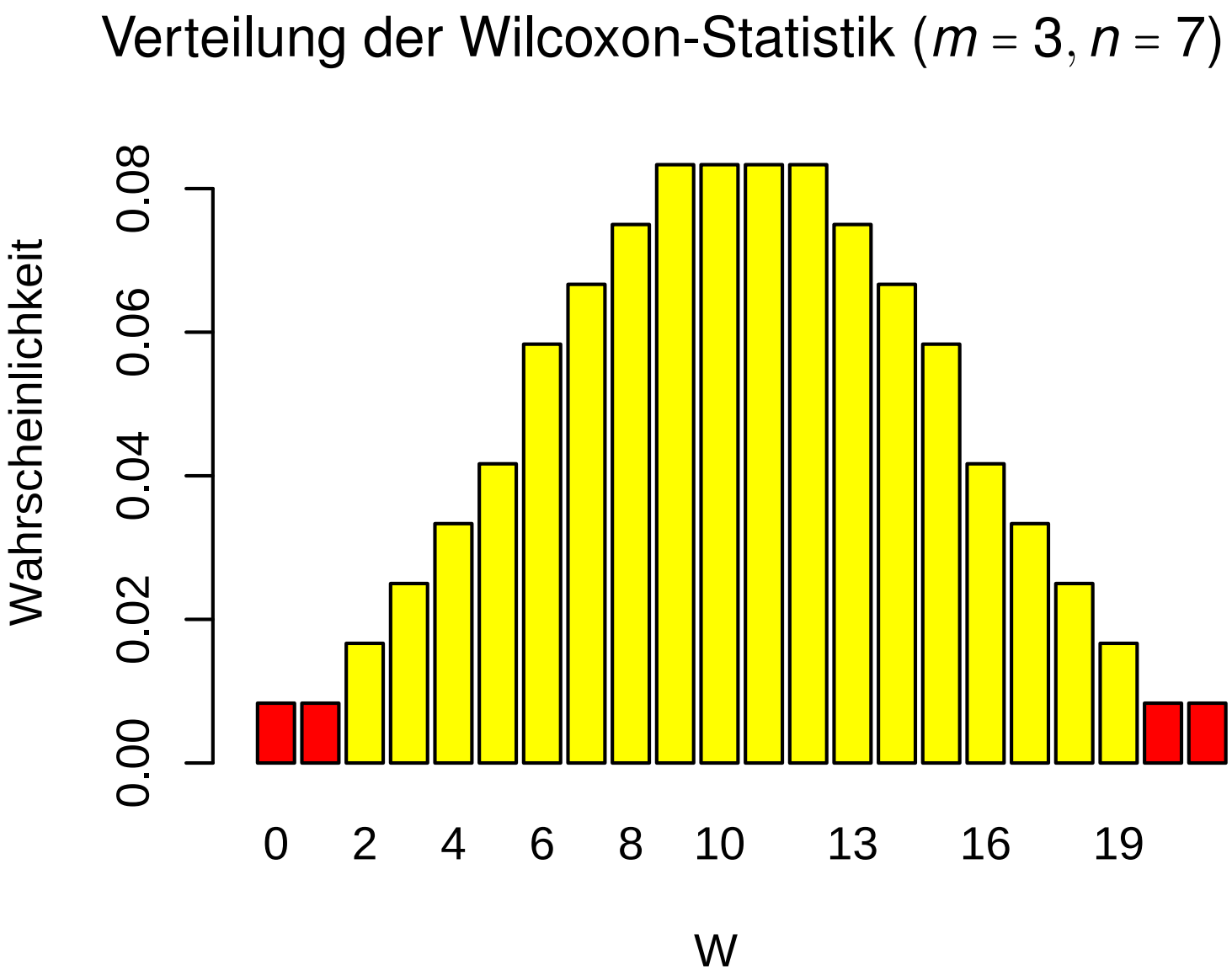
$$\mathbb{P}(W = w) = \frac{\text{Anz. Möglichkeiten mit Rangsummenstatistik } w}{120}$$

Wir beobachten in unserem Beispiel:

1,5, 5,6; 7,9; 35,2; 38,1; 41,0; 56,7; 112,1; 197,4; 381,8

somit $W = 1$

$$\begin{aligned} &\mathbb{P}(W \leq 1) + \mathbb{P}(W \geq 20) \\ &= \mathbb{P}(W = 0) + \mathbb{P}(W = 1) + \mathbb{P}(W = 20) + \mathbb{P}(W = 21) \\ &= \tfrac{1+1+1+1}{120} \doteq 0{,}033 \end{aligned}$$


Verteilung der Wilcoxon-Statistik ($m = 3, n = 7$)
Wahrscheinlichkeit
0.00
0.02
0.04
0.06
0.08
0
2
4
6
8
10
13
16
19
W

Prüfen wir in unserem Beispiel die Nullhypothese, dass die Verteilungen von X und Y identisch sind, auf dem 5%-Niveau:

Wir haben $W = 1$ beobachtet, also

$$p\text{-Wert} = P(\text{ein so extremes } W) = 4/120 = 0{,}033$$

Wir lehnen die Nullhypothese auf dem 5%-Niveau ab.

Bem.: Die Verteilungsgewichte von W kann man mittels einer Rekursionsformel explizit bestimmen (was für „mittelgroße“ m und n praktikabel ist),
bei großem m und n verwendet man eine Normalapproximation.

R kennt den Wilcoxon-Test mittels `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
```

Vergleich von *t*-Test und Wilcoxon-Test

Beachte:

Sowohl der *t*-Test als auch
der Wilcoxon-Test können verwendet werden,
um eine vermutete Verschiebung der Verteilung zu stützen.

Der *t*-Test testet „nur“ auf Gleichheit der Erwartungswerte.
Der Wilcoxon-Test dagegen testet auf Gleichheit
der gesamten Verteilungen.
(Der Wilcoxon-Test kann beispielsweise Signifikanz anzeigen,
selbst wenn die Stichproben-Mittelwerte übereinstimmen)

In besonderen Fällen

- Verteilungen sind asymmetrisch
- Stichprobenlänge ist klein

hat der Wilcoxon-Test eine höhere Testpower.

Vergleichen wir (spaßeshalber) mit dem *t*-Test (mit R ausgeführt):

```
> x
[1]  1.5  5.6 35.2
> y
[1]   7.9  38.1  41.0  56.7 112.1 197.4 381.8
> t.test(x,y,var.equal=TRUE)

Two Sample t-test

data:  x and y
t = -1.3319, df = 8, p-value = 0.2196
alternative hypothesis: true difference in means
95 percent confidence interval:
 -287.30632   76.93489
sample estimates:
mean of x mean of y
  14.1000  119.2857
```

Kolmogorov-Smirnov-Test

(Eine Erinnerung an / Einordnung von Übungsaufgabe 9.4)

Wir möchten prüfen, ob unabhängige, identisch verteilte $X_1, \ldots, X_n$ eine gewisse Verteilung ρ besitzen.
(und ρ habe stetige Verteilungsfunktion F_ρ)

Idee: Wenn $X_i \sim \rho$ gilt, so auch

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq t\}} \xrightarrow[n\to\infty]{} F_\rho(t) = \rho\big((-\infty, t]\big)$$

zumindest für jedes $t \in \mathbb{R}$ (mit dem Gesetz der großen Zahlen)

und die Fluktuationen sind von der Ordnung $O(1/\sqrt{n})$:

$$\sqrt{n} \cdot \sup_{t\in\mathbb{R}} \left|\widehat{F}_n(t) - F_\rho(t)\right| \xrightarrow[n\to\infty]{} K$$

wo K die *Kolmogorov-Verteilung besitzt*.

Sei $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n$ eine geordnete Stichprobe x,

$$\widehat{F}_x(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[x_i, \infty)}(t), \quad t \in \mathbb{R}.$$

die empirische Verteilungsfunktion.

Die Kolmogorov-Smirnov-Statistik ist

$$D^\rho(x) = \sup_t |\widehat{F}_x(t) - F_P(t)| = \max\{d_1, d_2, \ldots, d_n\}$$

mit $d_i = \max\{|\widehat{F}_x(x_i) - F_\rho(x_i)|, |\widehat{F}_x(x_i) - F_\rho(x_{i+1})|\}$ für $i = 1, 2, \ldots, n-1$ und $d_n = |1 - F_\rho(x_n)|$

Kolmogorov-Smirnov-Test

Lehne H_0 : Die x_i stammen aus Verteilung ρ zum (asymptotischen) Niveau α ab, wenn $\sqrt{n} D^\rho(x) >$ $(1-\alpha)$-Quantil von K.

Bemerkung:

K ist die Verteilung des Maximums des Betrags der sogenannten Brown'schen Brücke, es gilt

$$P(K > z) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-k^2 z^2}$$

R-Befehl: `ks.test`

Zur „reinen Lehre“ des statistischen Testens

Nehmen wir an, wir möchten eine gewisse Aussage anhand experimenteller oder empirischer Daten statistisch prüfen. Das korrekte („lehrbuchmäßige“) Vorgehen sieht folgendermaßen aus:

1. Statistisches Modell formulieren, Nullhypothese und Alternative angeben
(was die Nullhypothese ist, hängt von der konkreten Anwendungsfrage ab, oft ernennt man „das Gegenteil dessen, was man erhärten möchte“ zur Nullhypothese).

2. Dann einen Test (einschließlich gewünschtem Niveau) festlegen.

3. Dann erst: Daten erheben (bzw. Daten anschauen), Test-Entscheidung fällen.

Zur „reinen Lehre“ des statistischen Testens

Die Kontrolle der Fehlerwahrscheinlichkeiten, die die Theorie des statistischen Testens liefert, bezieht sich auf dieses Vorgehen.

Wenn man die Reihenfolge herumdreht, also zuerst die Daten anschaut und dann einen Test wählt, verfälscht man strenggenommen zumindest das Signifikanzniveau, möglicherweise bis ins Unsinnige.

Beispiel: zuerst den empirischen Mittelwert bestimmen, dann je nachdem, ob er links oder rechts von ϑ_0 liegt, entscheiden, ob man eine rechts- oder eine linksseitige Alternative wählt, ist offenbar „geschummelt“.

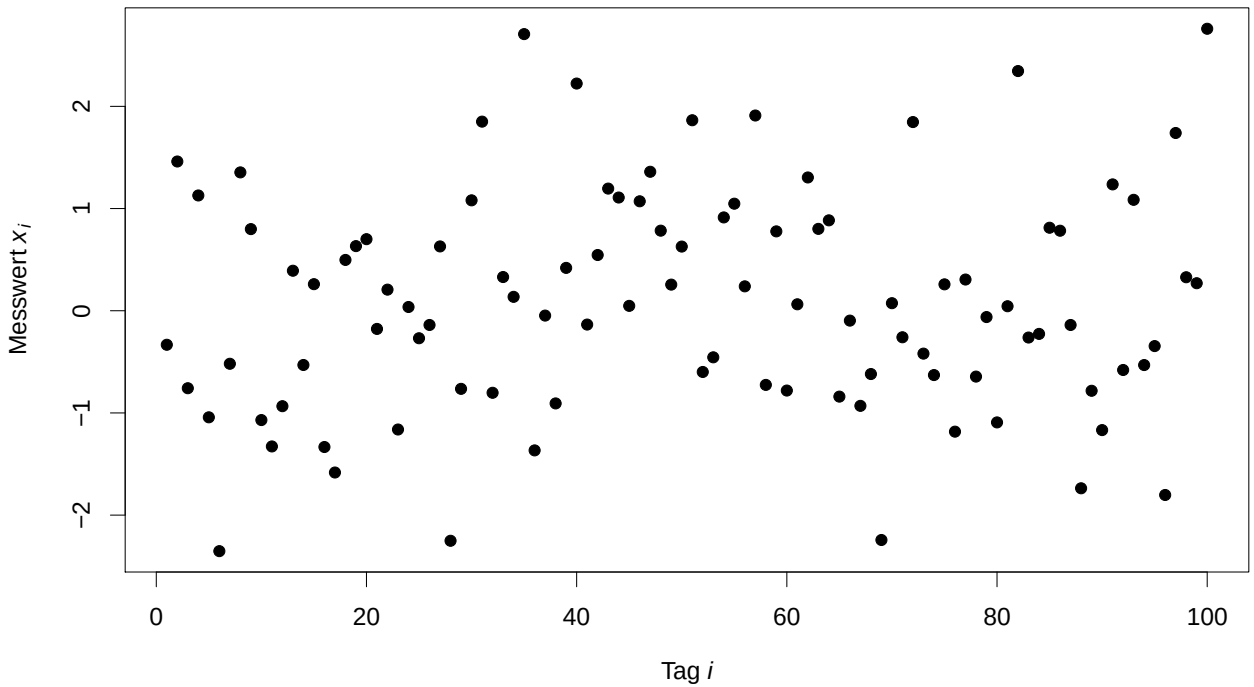
Zur „reinen Lehre“ des statistischen Testens

Man sollte dieselben Daten nicht für explorative Statistik (d.h. Beobachtungen, die zu neuen Hypothesen führen [sollen]) und schließende Statistik (d.h. Beobachtungen, anhand denen eine Hypothese getestet werden soll) zugleich verwenden.

Wir betrachten zum Abschluss ein (simuliertes) Beispiel, das zeigt, was sonst schief gehen kann ...

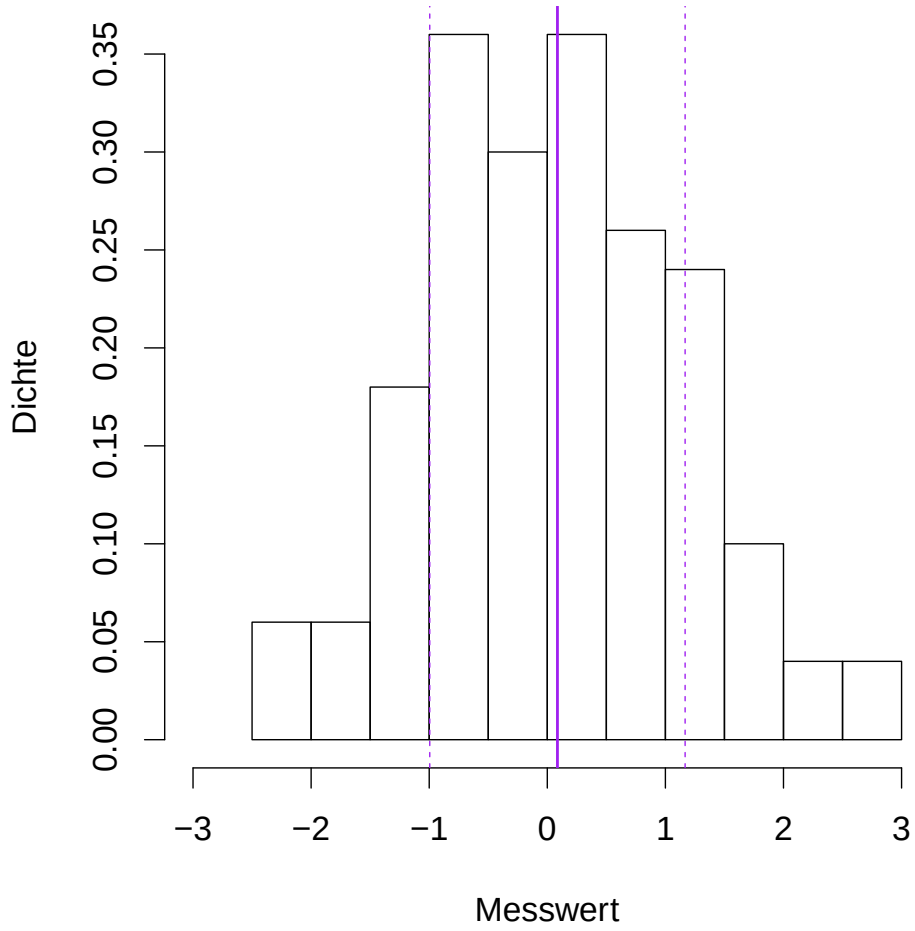
Ein simuliertes Experiment

Ein Versuch werde an $n = 100$ aufeinanderfolgenden Tagen unabhängig unter identischen Bedingungen wiederholt,
x_i = Messergebnis am i-ten Tag



(unter der Nullhypothese $\mu = 0$ simulierte Daten, d.h. es gibt in Wirklichkeit keinen Effekt)

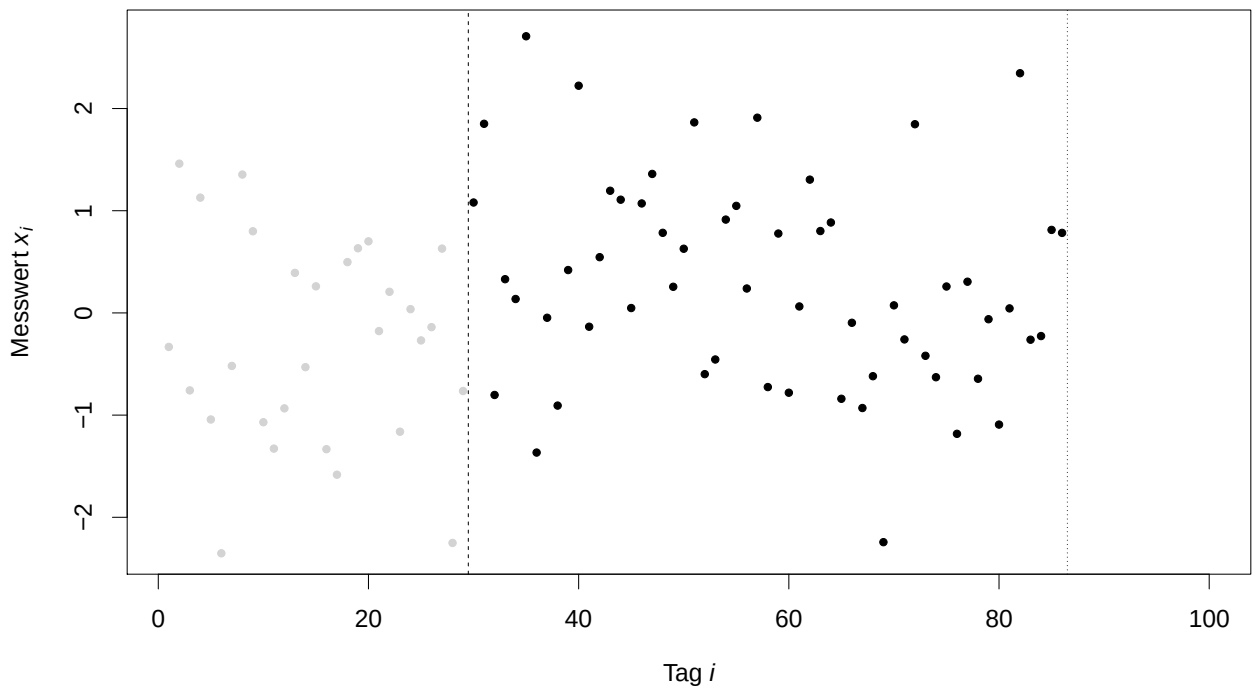
Ein simuliertes Experiment



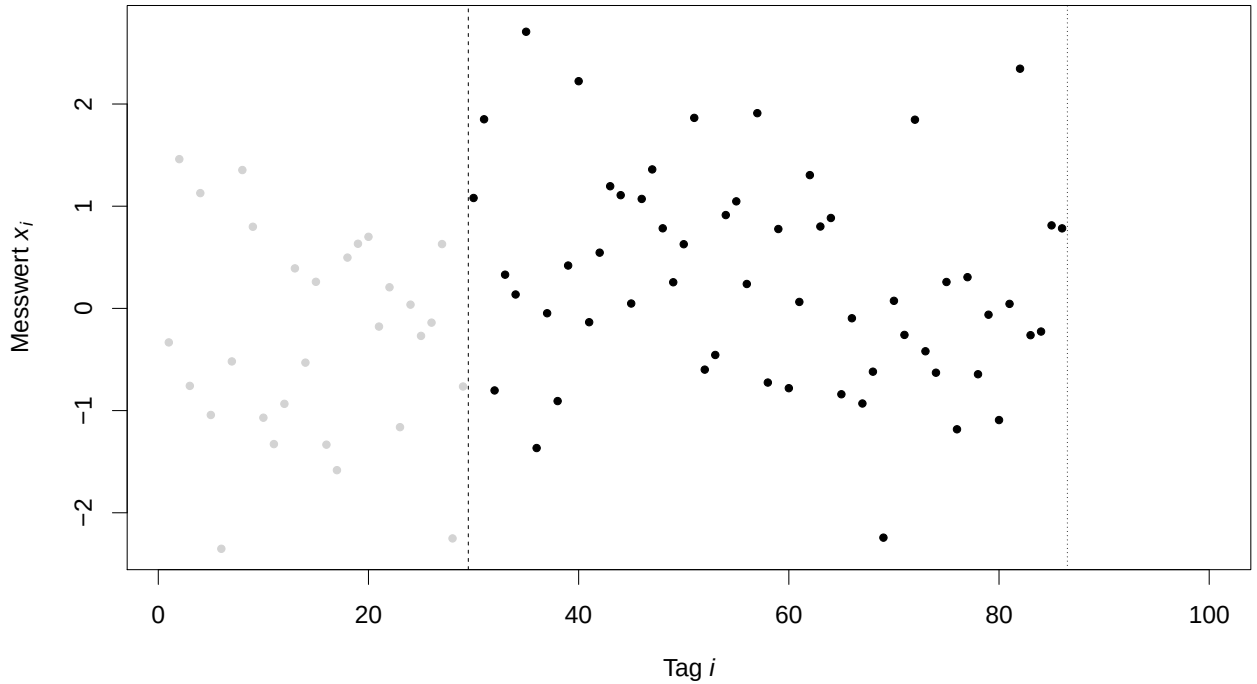
$n = 100$, $\overline{x} = 0.086$, $s/\sqrt{n} = 0.108$, $t = 0.794$, p-Wert ist 0.43
(zweiseitiger t-Test)

„Aufhören, wenn es gut aussieht“

Der Experimentator überlegt am Tag 86:
Der Monat erste war noch eine Übungs- und Kalibrierungsphase,
ich lasse einmal die ersten 29 Beobachtungen weg und schaue, was ich dann bis jetzt so habe (57 Beobachtungen)



„Aufhören, wenn es gut aussieht“

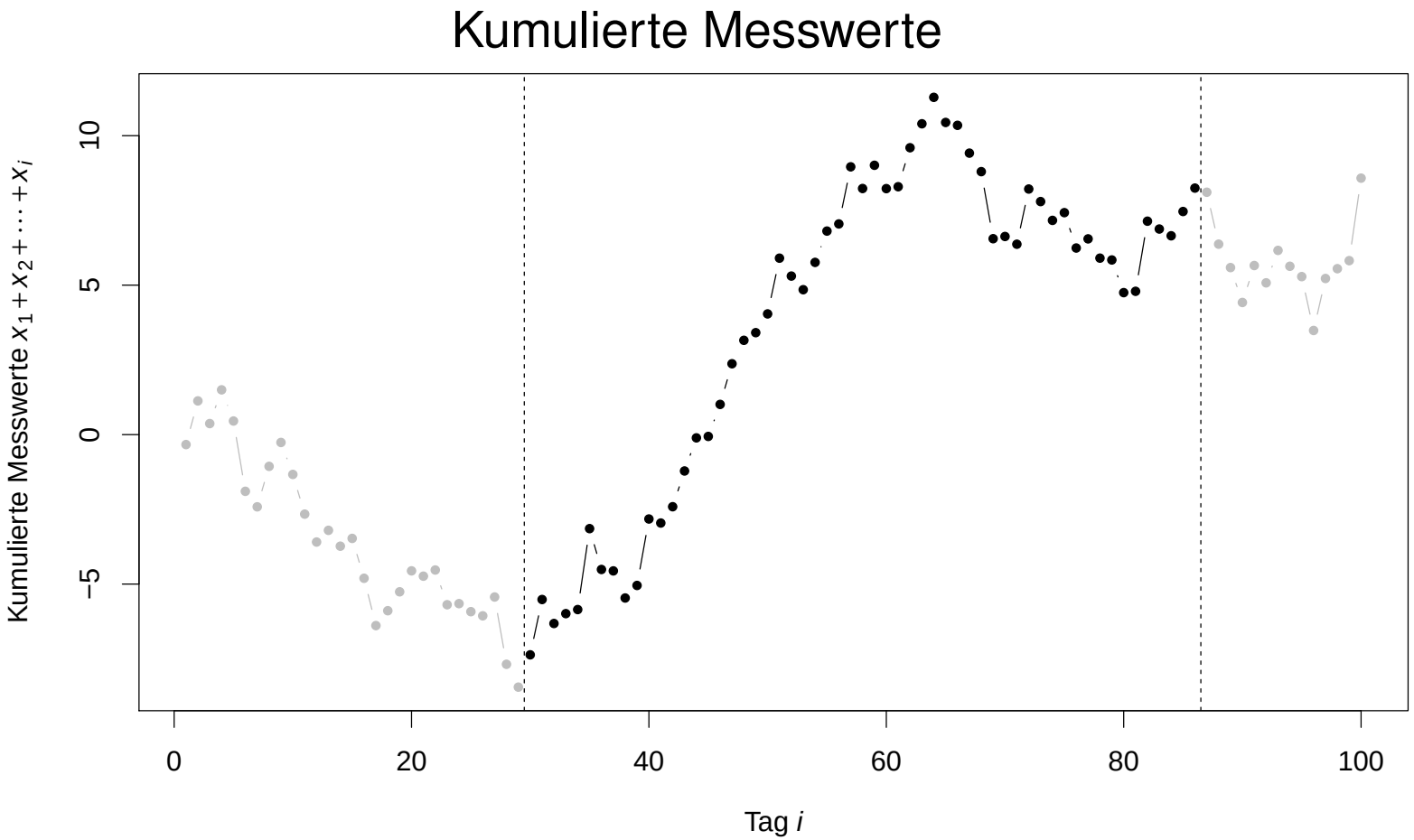


Mit den $n = 57$ Messwerten $x_{30}, x_{31}, \ldots, x_{85}, x_{86}$ ergibt sich
$\overline{x} = 0.293$, $s = 1.021$, $s/\sqrt{n} = 0.135$, $t = 2.167$, p-Wert ist 0.035
(zweiseitiger t-Test)

Demnach: Wir sehen scheinbar eine signifikante Abweichung von der 0?

Was ist hier passiert?

„Aufhören, wenn es gut aussieht?!“



Problem des multiplen Testens

Wenn wir den Beginn und die Länge der „richtigen“ Versuchsreihe nicht im vorhinein festlegen, haben wir ein multiples Testproblem vorliegen:

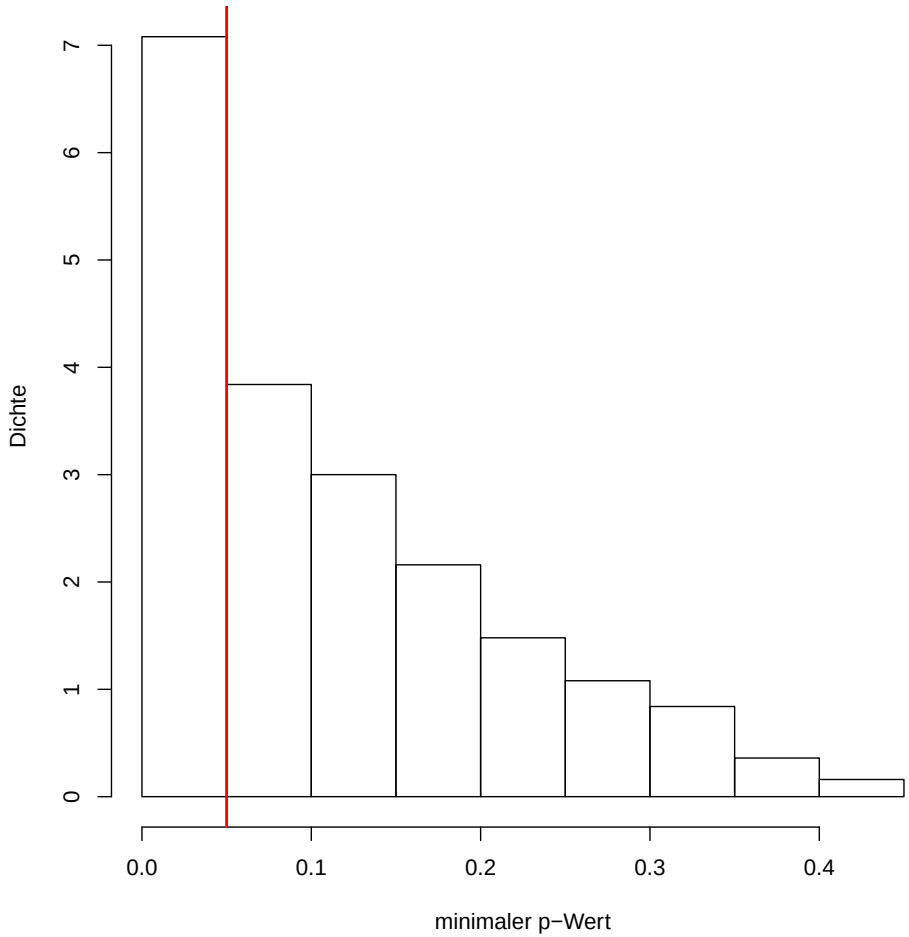
Angenommen, an jedem Tag $i = 50, 51, \ldots, 100$ geht der Experimentator die $i - 50 + 1$ möglichen Messreihen

$$
\begin{aligned}
&x_1, x_2, \ldots, x_{i-1}, x_i \\
&x_2, x_3, \ldots, x_{i-1}, x_i \\
&\vdots \\
&x_{i-50+1}, x_{i-50+2} \ldots, x_{i-1}, x_i
\end{aligned}
$$

der Länge ≥ 50, die mit dem heutigen Tag enden, durch und führt mit jeder davon einen (zweiseitigen ein-Stichproben) t-Test zur Nullhypothese $\mu = 0$ aus.

Problem des multiplen Testens

Dann wurden insgesamt $1 + 2 + \cdots + 51 = \frac{51 \cdot 52}{2} = 1326$ Tests ausgeführt. Wie wahrscheinlich ist es, dass mindestens einer einen p-Wert < 0.05 liefert?



500 simulierte Versuchsreihen

W’keit, dass mindestens einer der Tests anschlägt ≈ 0.35.