Biostatistik

Ergänzungen: Rangsummen-Test, Faktorielle Varianzanalyse und *F*-Test, sowie etwas zu multiplen Tests

Matthias Birkner

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Inhalt

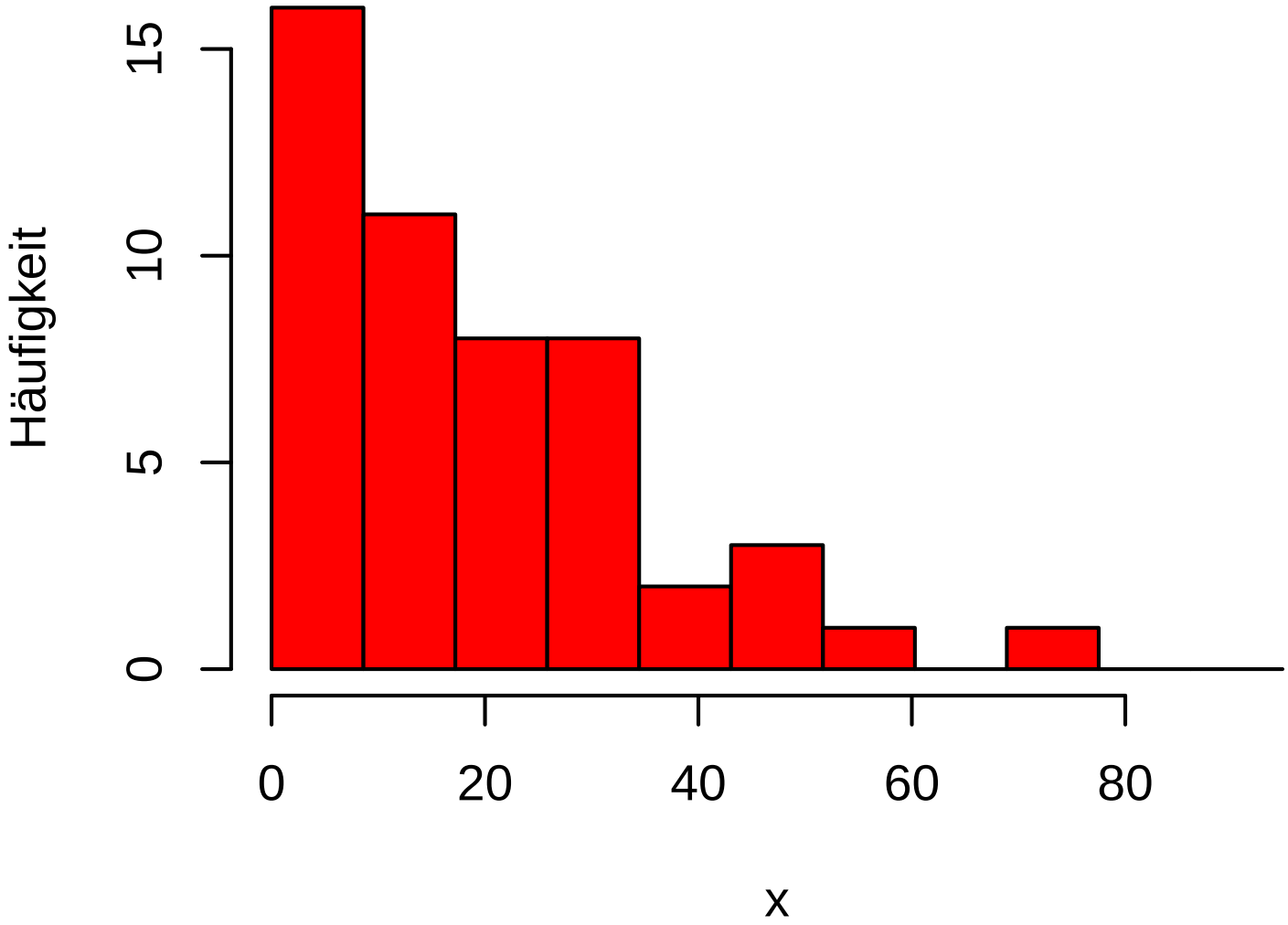
1. Wilcoxons Rangsummentest
   - Motivation
   - Wilcoxon-Test für unabhängige Stichproben
2. Einfaktorielle Varianzanalyse und F-Test
   - Exkurs: Varianzanalyse mit `R`
3. Über paarweise Vergleiche und multiples Testen
   - Exkurs: Bonferroni-Holm-Methode
4. Problematisch: „Aufhören, wenn es gut aussieht“

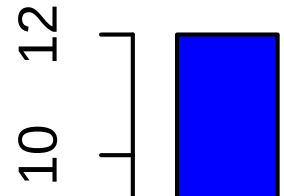
Bei (ungefähr) glockenförmigen und
symmetrisch verteilten Beobachtungen
oder wenn die Stichprobenumfänge genügend groß sind
können wir den t-Test benutzen,
um die Nullhypothese $\mu_1 = \mu_2$ zu testen:

Die t-Statistik ist (annährend) Student-verteilt.

Besonders bei sehr asymmetrischen und
langschwänzigen Verteilungen
kann das anders sein

Nehmen wir an, wir sollten folgende Verteilungen vergleichen:





Beispiele

- Wartezeiten
- Ausbreitungsentfernungen
- Zelltypenhäufigkeiten
- Genexpressionsniveaus

Gesucht:

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

(Eine Aufgabe der *nicht-parametrischen Statistik*.)

Beobachtungen: Zwei Stichproben

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

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

Wir möchten die Nullhypothese:
X und Y aus derselben Population
(X und Y haben diesselbe Verteilung)
testen

gegen die Alternative:
Die beiden Verteilungen sind gegeneinander verschoben.

Wir sind also in einer Situation, die wir schon beim t-Test getroffen haben: Die zwei Verteilungen sind möglicherweise gegeneinander verschoben (haben insbesondere möglicherweise unterschiedliche Mittelwerte), aber wir möchten *nicht* die implizite Annahme treffen, dass es sich dabei (wenigstens ungefähr) um Normalverteilungen handelt.

Idee

Beobachtungen:

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

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

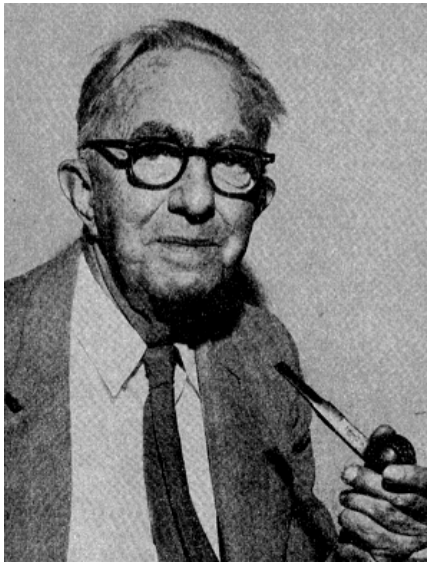
- Sortiere alle Beobachtungen der Größe nach.
- Bestimme die Ränge der m X-Werte unter allen $m + n$ Beobachtungen.
- Wenn die Nullhypothese zutrifft, sind die m X-Ränge eine rein zufällige Wahl aus $\{1, 2, \ldots, m + n\}$.
- Berechne die Summe der X-Ränge, prüfe, ob dieser Wert untypisch groß oder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

Wilcoxons Rangsummenstatistik

Beobachtungen:

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

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



Frank Wilcoxon,
1892–1965

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

heißt

Wilcoxons Rangsummenstatistik

(Der minimal mögliche Wert ist 0, der maximal mögliche Wert ist $m \cdot n$.)

Wilcoxons Rangsummenstatistik

Bemerkung:

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

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

Der Wilcoxon-Test heißt auch Mann-Whitney-Test.

In der Literatur sind verschiedene Zentrierungen der Rangsumme gebräuchlich, ggfs. prüfen, bevor Sie eine Formel aus einem Buch verwenden.

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:
  $1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10$
- Rangsumme: $U = 1 + 2 + 4 - (1 + 2 + 3) = 1$

Interpretation von U

X-Population kleiner $\Longrightarrow$ U klein:

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 0$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 1$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 2$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 2$

X-Population größer $\Longrightarrow$ U groß:

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 21$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 20$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 20$

$1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \quad U = 19$

Signifikanz

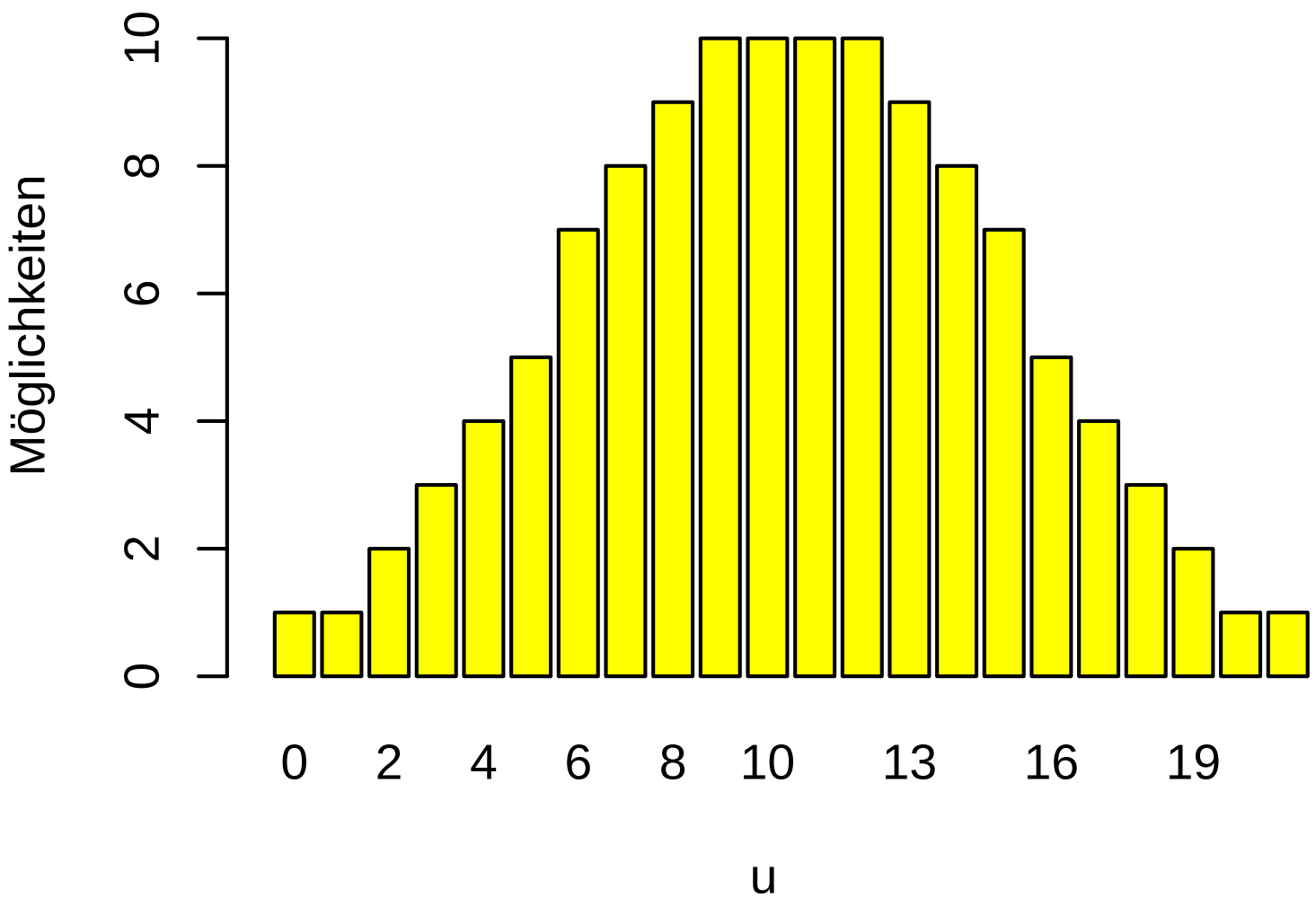
Nullhypothese:
X-Stichprobe und Y-Stichprobe
stammen aus
derselben Verteilung

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
u

Unter der Nullhypothese sind alle Rangbelegungen gleich wahrscheinlich, also

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

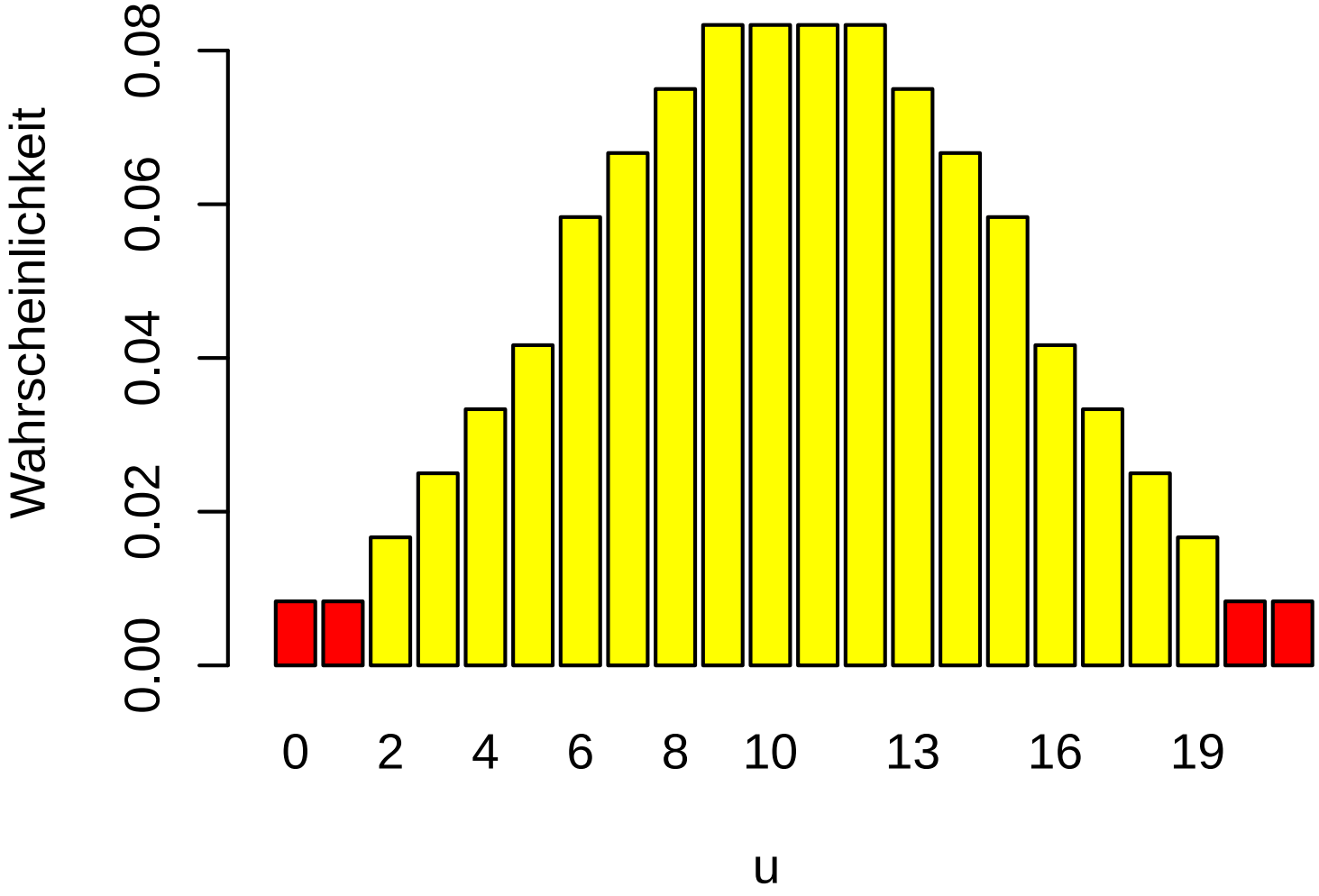
Wir beobachten in unserem Beispiel:
1,5, 5,6; 7,9; 35,2; 38,1; 41,0; 56,7; 112,1; 197,4; 381,8
somit $U = 1$

$$\mathbb{P}(U \leq 1) + \mathbb{P}(U \geq 20)$$

$$= \mathbb{P}(U = 0) + \mathbb{P}(U = 1) + \mathbb{P}(U = 20) + \mathbb{P}(U = 21)$$

$$= \tfrac{1+1+1+1}{120} \doteq 0{,}033$$

Verteilung der Wilcoxon-Statistik ($m = 3, n = 7$)



Für unser Beispiel ($U = 1$) also:

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

Wir lehnen die Nullhypothese,
dass die Verteilungen
von X und Y
identisch sind,
auf dem 5%-Niveau ab.

R kennt den Wilcoxon-Test mittels `wilcox.test`:

```
> x <- c(1.5,  5.6, 35.2)
> y <- c(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
```

Wilcoxon-Test: allgemeine Theorie

Gegeben zwei Stichproben $x_1, x_2, \ldots, x_m$ (m x-Werte) und $y_1, \ldots, y_n$ (n y-Werte)

bilde (normierte) Rangsumme

$$U = \text{Summe aller } x\text{-Ränge} - (1 + 2 + \cdots + m)$$
$$= \sum_{i=1}^{m} \#\{j \leq n : x_i > y_j\} =: \sum_{i=1}^{m} U_i$$

(mit möglichen Werten in $0, 1, \ldots, m \cdot n$).

Unter H_0 : die x-Werte und die y-Werte sind unabhängige Stichproben aus derselben Verteilung

hängt die Verteilung von U nicht von der tatsächlichen Verteilung ab.

(Strenggenommen: Sofern die Verteilung „stetig“ ist, also kein Wert mehrmals vorkommen kann)

Wilcoxon-Test: allgemeine Theorie, 2

normierte Rangsumme $U = U_{m,n} = \sum_{i=1}^{m} \#\{j \leq n : x_i > y_j\}$

H_0 : x-Werte und y-Werte stammen aus derselben Verteilung

Die Verteilung von $U_{m,n}$ unter H_0 ist (prinzipiell) ein rein kombinatorisches Problem:

$$\mathbb{P}_{H_0}(U_{m,n} = u) = \frac{\#\begin{pmatrix}\text{Zuordnungen von } m \text{ Rängen an } x\text{-Werte,} \\ \text{die Rangsumme } u \text{ ergeben}\end{pmatrix}}{\binom{m+n}{m}}$$

Für sehr kleine Stichproben kann man dies durch Auszählen bestimmen (wie vorhin), für größere Stichproben verwendet man eine Rekursionsformel (in R implementiert) oder eine Normalapproximation.

Wilcoxon-Test: allgemeine Theorie, 3

$$\mathbb{P}_{H_0}(U_{m,n} = u) = \frac{\#\begin{pmatrix}\text{Zuordnungen von } m \text{ Rängen an } x\text{-Werte,} \\ \text{die Rangsumme } u \text{ ergeben}\end{pmatrix}}{\binom{m+n}{m}}$$

(in `R`: `dwilcox(u,m,n)`, z.B. `dwilcox(1,3,7)=0.00833`, `R` kennt auch die Verteilungsfunktion `pwilcox` und die Quantilfunktion `qwilcox`.)

Quantile $u_{m,n,\alpha}$ mit

$$\mathbb{P}_{H_0}(U_{m,n} \leq u_{m,n,\alpha}) = \alpha$$

für $\alpha \in (0, 1)$ findet man in Tabellen

(oder mit `R`: `qwilcox(`α`,m,n)`)

Zweiseitiger Wilcoxon-Rangsummen-Test

Gegeben m x-Werte $x_1, \ldots, x_m$, n y-Werte $y_1, \ldots, y_n$, Signifikanzniveau $\alpha \in (0, 1)$.

Bestimme (normierte) Rangsumme $U = \sum_{i=1}^{m} \#\{j \leq n : x_i > y_j\}$, $u_{m,n,1-\alpha/2}$ mit $\mathbb{P}_{H_0}(U_{m,n} \leq u_{m,n,1-\alpha/2}) = 1 - \alpha/2$ (aus Tabelle oder mit R)

Lehne H_0: die x-Werte und die y-Werte sind unabhängige Stichproben aus derselben Verteilung

ab zugunsten von

H_1: die Verteilung der x-Werte ist verschoben im Vergleich zu der der y-Werte

ab, falls

$$U > u_{m,n,1-\alpha/2} \quad \text{oder} \quad U < m \cdot n - u_{m,n,1-\alpha/2}$$

Einseitiger Wilcoxon-Rangsummen-Test

Gegeben m x-Werte $x_1, \ldots, x_m$, n y-Werte $y_1, \ldots, y_n$, Signifikanzniveau $\alpha \in (0, 1)$.

Bestimme (normierte) Rangsumme $U = \sum_{i=1}^{m} \#\{j \leq n : x_i > y_j\}$,

$u_{m,n,1-\alpha}$ mit $\mathbb{P}_{H_0}(U_{m,n} \leq u_{m,n,1-\alpha}) = 1 - \alpha$ (aus Tabelle oder mit R)

Lehne H_0 ab zugunsten von

H_1: die Verteilung der x-Werte hat mehr Gewicht auf größen Werten im Vergleich zu der der y-Werte

ab, falls

$$U > u_{m,n,1-\alpha}$$

(analog für H_1 : „Verteilung der x-Werte nach links verschoben im Vgl. zu y-Werten“, falls $U < m \cdot n - u_{m,n,1-\alpha}$, oder vertausche Rollen von x und y.)

Wilcoxon-Test: Normalapproximation

Unter H_0 : x-Werte und y-Werte stammen aus derselben Verteilung

ist

$$\frac{U_{m,n} - \frac{m \cdot n}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \approx \text{Standard-normalverteilt}$$

Damit lautet der (approximative) zweiseitige Test zum Signifikanzniveau α bei beobachter Rangsumme U:

$$\text{lehne } H_0 \text{ ab, wenn } \left| U - \frac{m \cdot n}{2} \right| > z_{1-\alpha/2} \sqrt{\frac{mn(m+n+1)}{12}}$$

wobei $z_{1-\alpha/2} = (1 - \alpha/2)$-Quantil der Standard-Normalverteilung

Analog einseitig (H_1 = „Verteilung der x-Werte nach rechts verschoben im Vgl. zu y-Werten“):

Lehne H_0 ab, falls $U > \frac{m \cdot n}{2} + z_{1-\alpha} \sqrt{\frac{mn(m+n+1)}{12}}$.

Wilcoxon-Test, Fall mit Bindungen

Bei diskreten Verteilungen (und in der Praxis!) kann es vorkommen, dass einzelne Werte in den Daten x_i und y_j mehrfach vorkommen. Dies nennt man Bindungen. In diesem Fall wird der Rang von x_i in den $y_1, \ldots, y_n$ berechnet als

$$U_i = \text{Anzahl der } j \text{ mit } y_j < x_i + \frac{1}{2}\text{Anzahl der } j \text{ mit } y_j = x_i$$

und die Rangsumme als

$$U = \sum_{i=1}^{m} U_i.$$

In diesem Fall hält der Wilcoxon-Test das geforderte Niveau nicht exakt ein, meistens aber doch recht gut. `R` gibt dann eine Warnmeldung.

Beachte:

Wenn der Wilcoxon-Test Signifikanz anzeigt,
so kann das daran liegen, dass die zugrunde
liegenden Verteilungen verschiedene Formen haben.

Der Wilcoxon-Test kann beispielsweise Signifikanz anzeigen,
selbst wenn die Stichproben-Mittelwerte übereinstimmen!

Vergleich von t-Test und Wilcoxon-Test

Sowohl der (2-Stichproben-)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.

In den meisten Fällen liefern beide Tests dasselbe Ergebnis.
Sofern die zugrundeliegenden Verteilungen einigermaßen
glockenförmig sind, ist der t-Test empfehlenswerter.

In besonderen Fällen

- Verteilungen sind asymmetrisch
- Stichprobenlänge ist klein

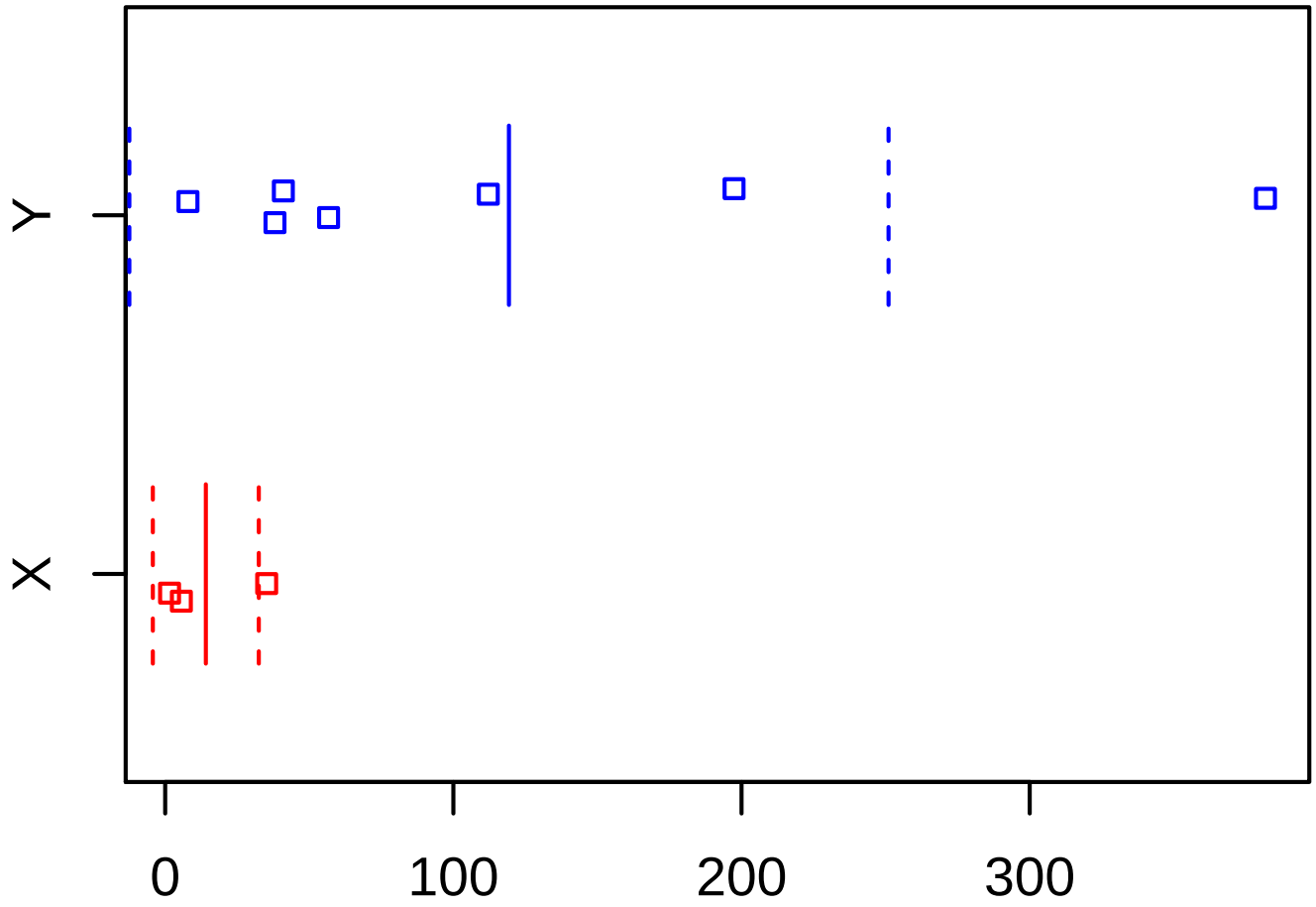
hat der Wilcoxon-Test eine höhere Testpower.

Vergleichen wir (spaßeshalber) mit dem *t*-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
> t.test(x,y)

Welch Two Sample t-test

data:  x and y
t = -2.0662, df = 6.518, p-value = 0.08061
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -227.39182   17.02039
sample estimates:
mean of x mean of y
  14.1000  119.2857
```


Y
X
0
100
200
300

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),

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

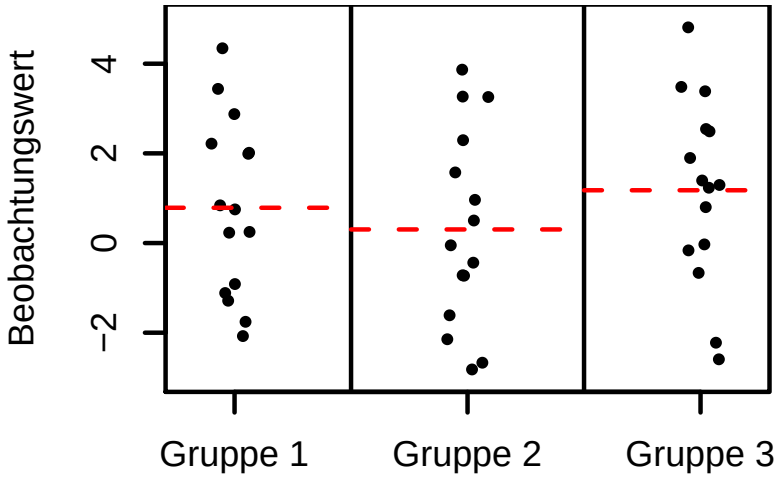
(gepoolte Stichprobenvarianz) berechne $t = \frac{\overline{x}-\overline{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, lehne $H_0 : \mu_1 = \mu_2$ zum Signifikanzniveau α ab, wenn

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

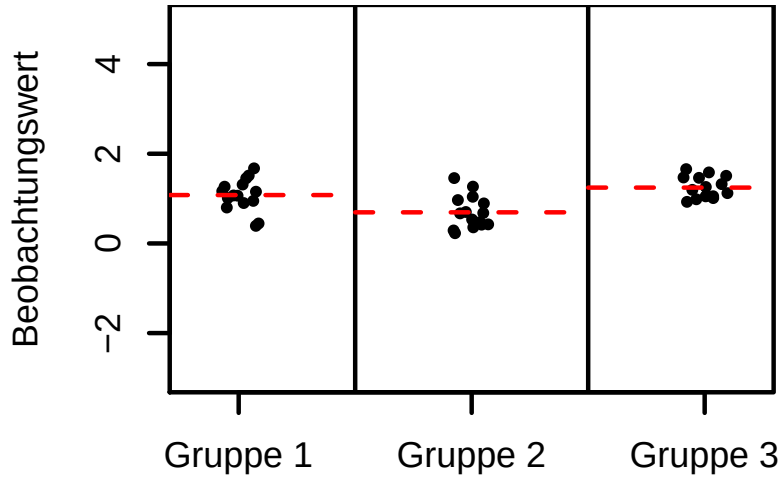
Frage Was tun, wenn mehr als zwei Gruppen vorliegen?

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

Sir Ronald Aylmer Fisher,
1890–1962

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

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

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

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.

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

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).

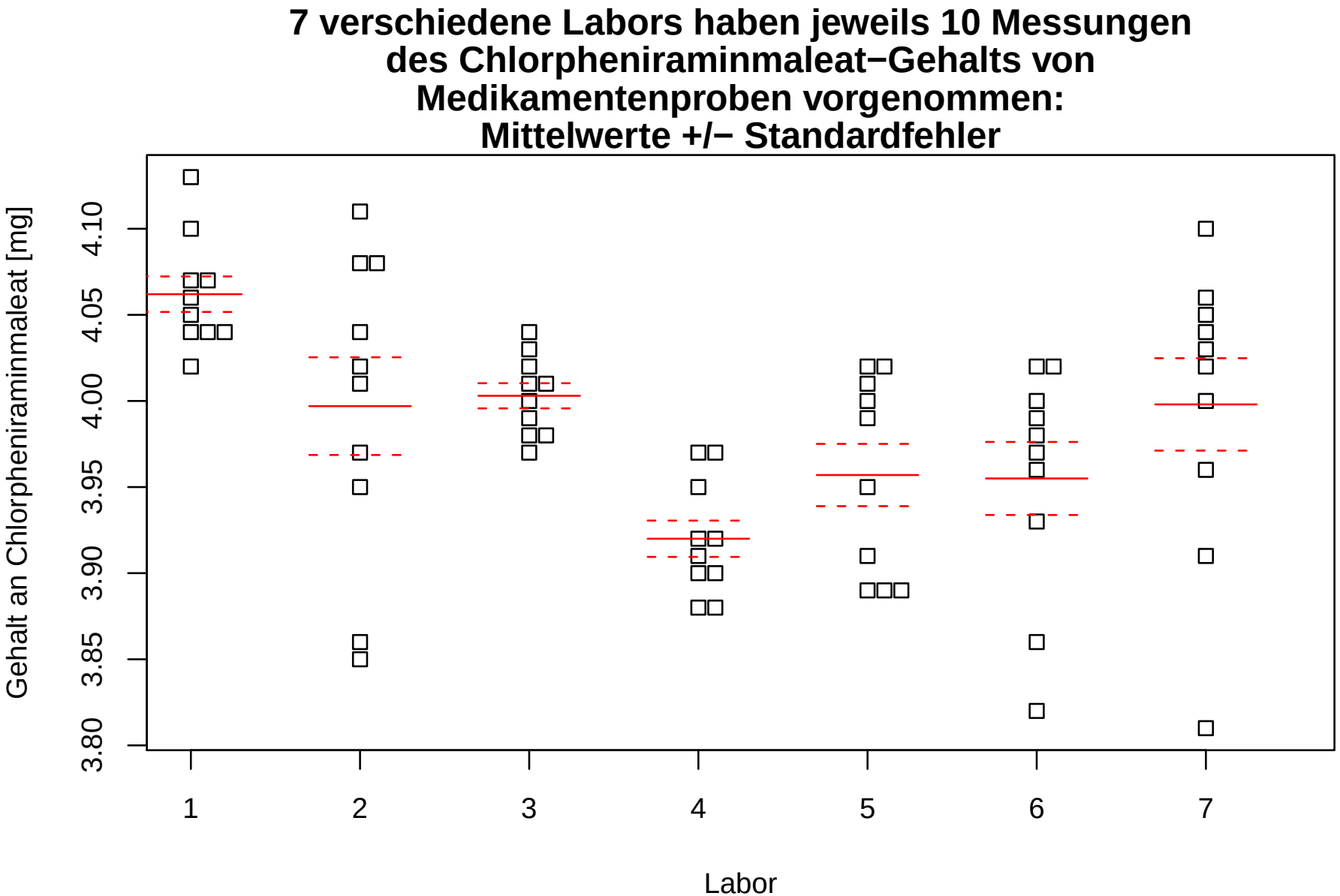
```
> rat<-read.table("gerinnung.txt",header=TRUE)
> rat.aov <- aov(bgz~beh,data=rat)
> summary(rat.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
beh          3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Ein weiteres Beispiel

7 verschiedene Labors haben jeweils 10 Messungen des Chlorpheniraminmaleat-Gehalts von Medikamentenproben vorgenommen.

Die Daten liegen in der Datei `chlorpheniraminmaleat.txt` als Tabelle vor:

```
  Gehalt Labor
1 4.13 1
2 4.07 1
3 4.04 1
4 4.07 1
5 4.05 1
6 4.04 1
7 4.02 1
8 4.06 1
9 4.1 1
10 4.04 1
11 3.86 2
12 3.85 2
13 4.08 2
14 4.11 2
15 4.08 2
16 4.01 2
17 4.02 2
18 4.04 2
...
```



Daten aus R.D. Kirchhoefer, Semiautomated method for the analysis of chlorpheniramine maleate tablets: collaborative study, *J. Assoc. Off. Anal. Chem.* 62(6):1197-1201 (1979),
zitiert nach John A. Rice, *Mathematical statistics and data analysis*, 2nd ed., Wadsworth, 1995

Beachte: Die Labore sind sind mit Zahlen nummeriert. Damit R das nicht als numerische Werte sondern als Nummern der Labore auffasst, müssen wir die Variable “Labor” in einen sog. Factor umwandeln:

```
> chlor <- read.table("chlorpheniraminmaleat.txt")
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4
 $ Labor : int  1 1 1 1 1 1 1 1 1 1 ...
> chlor$Labor <- as.factor(chlor$Labor)
> str(chlor)
'data.frame': 70 obs. of  2 variables:
 $ Gehalt: num  4.13 4.07 4.04 4.07 4.05 4.04 4.02 4.06 4.1 4
 $ Labor : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1
```

Nun können wir die Varianzanalyse durchführen:

```
> chlor.aov <- aov(Gehalt~Labor,data=chlor)
> summary(chlor.aov)
            Df  Sum Sq  Mean Sq F value    Pr(>F)
Labor        6 0.12474 0.020789  5.6601 9.453e-05 ***
Residuals   63 0.23140 0.003673
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

In vorigem Beispiel:

Sei μ_i der (uns unbekannte, wahre Populations-)Mittelwert der Messungen aus Labor i, für $i = 1, \ldots, 7$.

Die Varianzanalyse zeigte, dass es signifikante Unterschiede zwischen den Laboren gibt.

Aber welche Labore unterscheiden sich signifikant?

Wir könnten dazu für jedes Paar i, j von Labors jeweils einen (zwei-Stichproben-)t-Test durchführen, um die Nullhypothese

$$H_{0,(i,j)} : \mu_i = \mu_j$$

(zu einem vorgegebenen Signifikanzniveau α, sagen wir $\alpha = 5\%$) zu testen.

Welche Labore unterscheiden sich signifikant?

Wert der t-Statistik aus paarweisen Vergleichen mittels t-Tests:
[(zweiseitiger) zwei-Stichproben t-Tests mit Annahme gleicher Varianzen]

	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
Lab1	2.154	4.669	9.632	5.046	4.539	2.227
Lab2		-0.205	2.545	1.189	1.186	-0.026
Lab3			6.470	2.359	2.140	0.180
Lab4				-1.768	-1.478	-2.706
Lab5					0.072	-1.268
Lab6						-1.258

Das 97,5%-Quantil der
t-Verteilung mit 18 Freiheitsgraden ist 2.101,
also würde für die rot markierten Paare (jeweils für sich betrachtet)
ein t-Test $H_{0,(i,j)}$ zum Signifikanzniveau 5% ablehnen.

Welche Labore unterscheiden sich signifikant?

Alternative Darstellung:

p-Werte aus paarweisen Vergleichen mittels *t*-Tests:

[(zweiseitiger) zwei-Stichproben *t*-Tests mit Annahme gleicher Varianzen]

	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
Lab1	0.04506	0.00019	0.00000	0.00008	0.00025	0.03894
Lab2		0.84000	0.02033	0.24980	0.25103	0.97985
Lab3			0.00000	0.02982	0.04626	0.85929
Lab4				0.09398	0.15662	0.01446
Lab5					0.94356	0.22113
Lab6						0.22459

Erinnerung:

p-Wert = W'keit (unter der Nullhypothese) einen mindestens so extremen Wert der *t*-Statistik wie den beobachteten zu erhalten

[Hier: $2(1 - F_{Student(18)}(|t|))$, mit $F_{Student(18)}$ Verteilungsfunktion der Student-Verteilung mit 18 Freiheitsgraden]

Problem des multiplen Testens

Wir haben $7 \cdot 6 \cdot \frac{1}{2} = 21$ paarweise Vergleiche; auf dem 5%-Niveau zeigen einige davon Signifikanz an.

Wenn die Nullhypothese(n) („alles nur Zufallsschwankungen“) stimmt/en, verwirft man im Schnitt bei 5% der Tests die Nullhypothese zu Unrecht.

Testet man mehr als 20 mal und gelten jeweils die Nullhypothesen, wird man also im Schnitt mehr als eine Nullhypothese zu Unrecht verwerfen.

Dieses Phänomen müssen wir bei multiplen Tests berücksichtigen

(und mit entsprechend angepassten Tests bzw. mit korrigierten *p*-Werten arbeiten).

Eine ganz allgemeine Korrektur für multiples Testen ist die **Bonferroni[1]-Methode**:

Wenn m Tests zum *multiplen Signifikanzniveau* $\alpha \in (0, 1)$ durchgeführt werden sollen,

so führe jeden Test für sich zum *lokalen* Signifikanzniveau $\frac{\alpha}{m}$ durch.

Alternativ bedeutet dies: Multipliziere jeden (individuellen) p-Wert mit der Anzahl m der durchgeführten Tests.

[denn wenn die jeweilige Nullhypthese zutrifft, so ist der p-Wert uniform verteilt in $[0, 1]$]

Dann gilt: Die Wahrscheinlichkeit, dass *irgendeine zutreffende* Nullhypothese zu Unrecht ablehnt wird,

beträgt höchstens α.

[1] Carlo Emilio Bonferroni, 1892–1960

Labor-Vergleichs-Beispiel mit Bonferroni-Korrektur

Wert der t-Statistik aus paarweisen Vergleichen mittels t-Tests:

	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
Lab1	2.154	4.669	9.632	5.046	4.539	2.227
Lab2		-0.205	2.545	1.189	1.186	-0.026
Lab3			6.470	2.359	2.140	0.180
Lab4				-1.768	-1.478	-2.706
Lab5					0.072	-1.268
Lab6						-1.258

Betrachte $\alpha = 5\%$. Hier $m = 21$, das $(1 - \frac{1}{2}\frac{\alpha}{m})$-Quantil $(1 - \frac{1}{2}\frac{\alpha}{m} = 0.99881)$ der t-Verteilung mit 18 Freiheitsgraden ist 3.532,

also können wir für die rot markierten Paare $H_{0,(i,j)}$ zum multiplen Signifikanzniveau 5% ablehnen.

Labor-Vergleichs-Beispiel mit Bonferroni-Korrektur

Alternativ: $21 \times$ (p-Wert aus paarweisem t-Test)

	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7
Lab1	0.94626	0.00399	0.00000	0.00168	0.00525	0.8177
Lab2		17.64000	0.42693	5.24580	5.27163	20.576
Lab3			0.00000	0.62622	0.97146	18.045
Lab4				1.97358	3.28902	0.3036
Lab5					19.81476	4.6437
Lab6						4.7163

Für die rot markierten Paare ist der korrigierte p-Wert < 0.05.

Bonferroni-Korrektur: Theoretischer Hintergrund

Sei $\alpha \in (0,1)$, es seien m Nullhypothesen $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ und m Tests $\mathcal{T}_1, \ldots, \mathcal{T}_m$ gegeben mit

$$\mathbb{P}_{H_{0,i}}(\mathcal{T}_i \text{ lehnt } H_{0,i} \text{ ab}) \leq \frac{\alpha}{m}, \quad i = 1, \ldots, m$$

(d.h. wenn $H_{0,i}$ [und ggfs. noch irgendwelche anderen $H_{0,j}$] zutrifft, so wird sie von $\mathcal{T}_i$ nur mit W'keit $\leq \alpha/m$ zu Unrecht abgelehnt).

Eine gewisse Teilmenge $W \subset \{1, 2, \ldots, m\}$ der Nullhypothesen sei wahr. Dann ist

$$\mathbb{P}_{\cap_{i \in W} H_{0,i}}\Big(\text{es gibt ein } j \in W, \text{ so dass } H_{0,j} \text{ von } \mathcal{T}_j \text{ abgelehnt wird}\Big)$$

$$\leq \sum_{j \in W} \mathbb{P}_{H_{0,j}}(\mathcal{T}_j \text{ lehnt } H_{0,j} \text{ ab}) \leq \sum_{j \in W} \frac{\alpha}{m} = |W| \frac{\alpha}{m} \leq \alpha.$$

Nochmal das Ratten-Beispiel: Paarweise Vergleiche (mittels t-Test) für die Blutgerinnungszeiten bei vier verschiedenen Behandlungen, zunächst ohne Korrektur für multiples Testen:

```
  B       C         D
A 0.0147  0.00024   1.00000
B         0.16689   0.00509
C                   0.00010
```

Nun mit Bonferroni-Korrektur (alle Werte mit $\binom{4}{2} = 6$ multiplizieren):

```
  B       C         D
A 0.0882  0.00144   6.00000
B         1.00134   0.03054
C                   0.00060
```

Nach Bonferroni-Korrektur führen folgende Paare von Behandlungen zu jeweils signifikant unterschiedlichen Ergebnissen: A/C, B/D sowie C/D. (Der Bonferroni-korrigierte p-Wert von 6.0 für den Vergleich der Behandlungen A und D ist natürlich nicht als echter p-Wert zu interpretieren.)

Die Bonferroni-Methode ist sehr *konservativ*, d.h. um auf der sicheren Seite zu sein, lässt man sich lieber die eine oder andere Signifikanz entgehen.

Eine Verbesserung der Bonferroni-Methode ist die **Bonferroni-Holm-Methode**:
Ist m die Anzahl der Tests, so multipliziere den kleinsten p-Wert mit m, den zweitkleinsten mit $m-1$, den drittkleinsten mit $m-2$ usw.,

lehne all die Nullhypothesen ab,

deren so korrigierter p-Wert $< \alpha$ ist.

Dies ist ein Test aller m Nullhypothesen gleichzeitig zum multiplen Signifikanzniveau α.

Im Ratten-Beispiel:

Unkorrigierte *p*-Werte (aus paarweisen *t*-Tests)

```
  B       C        D
A 0.0147  0.00024  1.00000
B         0.16689  0.00509
C                  0.00010
```

$$0.00010 < 0.00024 < 0.00509 < 0.01470 < 0.16689 < 1.00000$$

p-Werte nach Bonferroni-Holm-Korrektur

	B	C	D
A	$0.0147 \cdot 3 = 0.0441$	$0.00024 \cdot 5 = 0.0012$	$1.0 \cdot 1 = 1.0$
B		$0.16689 \cdot 2 = 0.33378$	$0.00509 \cdot 4 = 0.02036$
C			$0.0001 \cdot 6 = 0.0006$

Wir sehen: Nun sind auf multiplen 5%-Niveau die Paare A/B, A/C, B/D und C/D signifikant verschieden.

Übrigens:

In R gibt des den Befehl `p.adjust`, der *p*-Werte für multiples Testen korrigiert und dabei defaultmäßig die Bonferroni-Holm-Korrektur verwendet:

```
> pwerte <- c(0.01470, 0.00024, 0.16689, 1.00000,
+  0.00509, 0.00010)
> pwerte
[1] 0.01470 0.00024 0.16689 1.00000 0.00509 0.00010

> p.adjust(pwerte)
[1] 0.04410 0.00120 0.33378 1.00000 0.02036 0.00060
> p.adjust(pwerte, method="bonferroni")
[1] 0.08820 0.00144 1.00000 1.00000 0.03054 0.00060
```

Übrigens, 2:

Für paarweise *t*-Tests gibt es ebenfalls eine R-Funktion, die per default die Bonferroni-Holm-Korrektur verwendet:

```
> pairwise.t.test(rat$bgz,rat$beh,
+                 pool.sd=FALSE,var.equal=TRUE)

Pairwise comparisons using t tests with non-pooled SD

data:  rat$bgz and rat$beh

  A       B       C
B 0.04410 -       -
C 0.00121 0.33378 -
D 1.00000 0.02036 0.00059

P value adjustment method: holm
```

Übrigens, $2\frac{1}{2}$:

Wenn man keine *p*-Wert-Korrektur wünscht, kann man sie im R-Befehl `pairwise.t.test` mit dem Zusatzparameter `p.adjust.method='none'` explizit ausschalten.

```
> pairwise.t.test(rat$bgz,rat$beh,
+  pool.sd=FALSE,var.equal=TRUE, p.adjust.method="none")

Pairwise comparisons using t tests with non-pooled SD

data:  rat$bgz and rat$beh

  A       B       C
B 0.01470 -       -
C 0.00024 0.16689 -
D 1.00000 0.00509 9.9e-05

P value adjustment method: none
```

Bonferroni-Holm-Korrektur: Theoretischer Hintergrund

Gegeben m Nullhypothesen $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$
und m Tests $\mathcal{T}_1, \ldots, \mathcal{T}_m$, P_i sei der p-Wert aus dem i-ten Test
($\mathcal{T}_i$ ist ein gültiger Test für $H_{0,i}$, d.h. $\mathbb{P}_{H_{0,i}}(P_i \leq u) \leq u$ für $u \in [0,1]$).

Seien

$$P_{(1)} < P_{(2)} < \cdots < P_{(m)}$$

die der Größe nach sortierten p-Werte und
$H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(m)}$ die entsprechend umsortierten Hypothesen, $\alpha \in (0,1)$.

Wenn

$$mP_{(1)}, (m-1)P_{(2)}, \ldots, (m-\ell-1)P_{(\ell)} < \alpha \leq (m-\ell)P_{(\ell+1)}$$

gilt, so lehne $H_{0,(1)}, H_{0,(2)}, \ldots, H_{0,(\ell)}$ (zum multiplen Niveau α ab)
(und behalte $H_{0,(\ell+1)}, \ldots, H_{0,(m)}$ bei).

Bonferroni-Holm-Korrektur: Theoretischer Hintergrund

Sei $W \subset \{1, \ldots, m\}$ (mit $|W| = k$, sagen wir) und die Nullhypothesen $H_{0,i}$, $i \in W$ seinen wahr.

Es gilt $\bigcap_{i \in W} \left\{ P_i > \frac{\alpha}{k} \right\} \subset \left\{ P_{(\, m-(k-1)\,)} > \frac{\alpha}{k} \right\}$,

insbesondere stoppt das Verfahren dann in Schritt $\ell \leq m - k + 1$ und alle $H_{0,i}$, $i \in W$ werden akzeptiert.

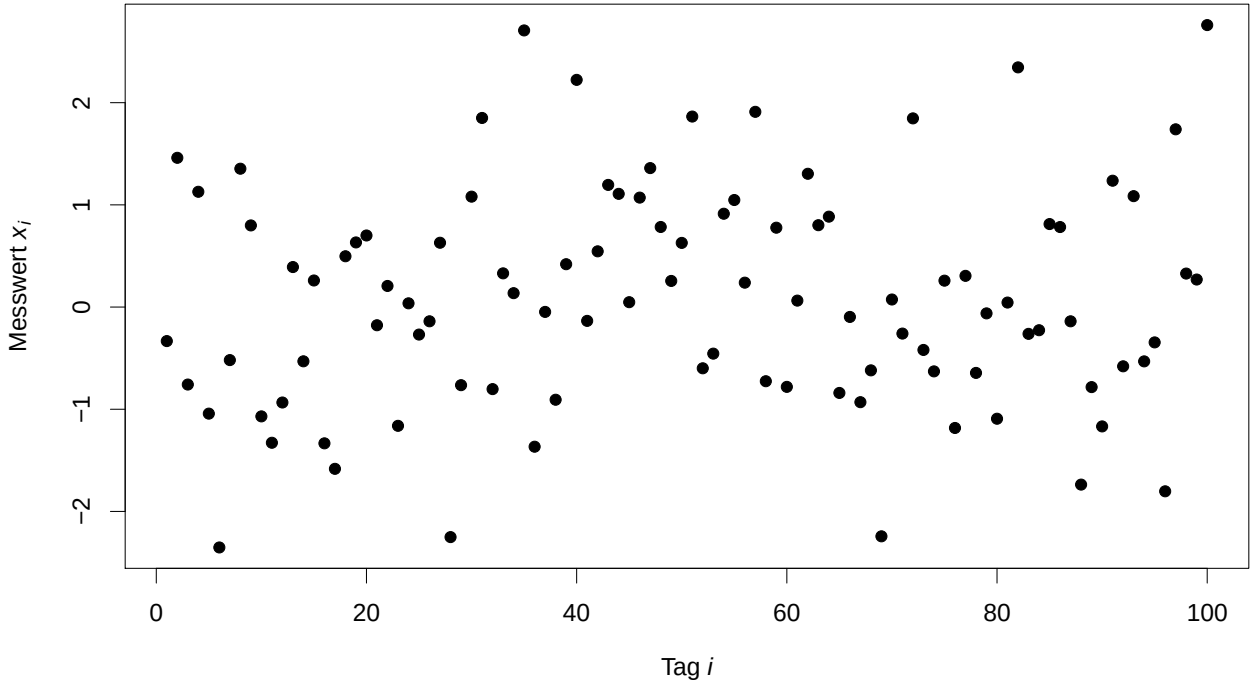
Weiter ist

$$\mathbb{P}_{\cap_{i \in W} H_{0,i}} \left(\bigcap_{i \in W} \left\{ P_i > \tfrac{\alpha}{k} \right\} \right) = 1 - \mathbb{P}_{\cap_{i \in W} H_{0,i}} \left(\bigcup_{i \in W} \left\{ P_i \leq \tfrac{\alpha}{k} \right\} \right)$$

$$\geq 1 - \sum_{i \in W} \mathbb{P}_{H_{0,i}} \left(P_i \leq \tfrac{\alpha}{k} \right) \geq 1 - \sum_{i \in W} \frac{\alpha}{k} = 1 - \alpha.$$

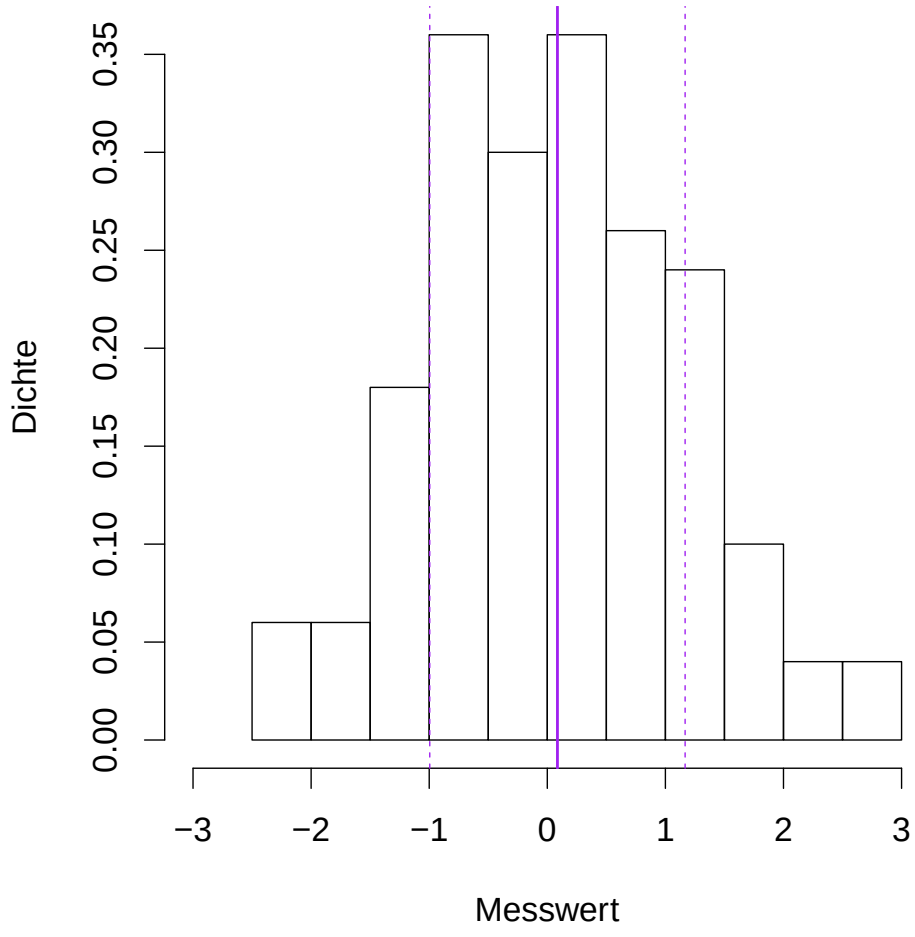
Ein simuliertes Experiment

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)

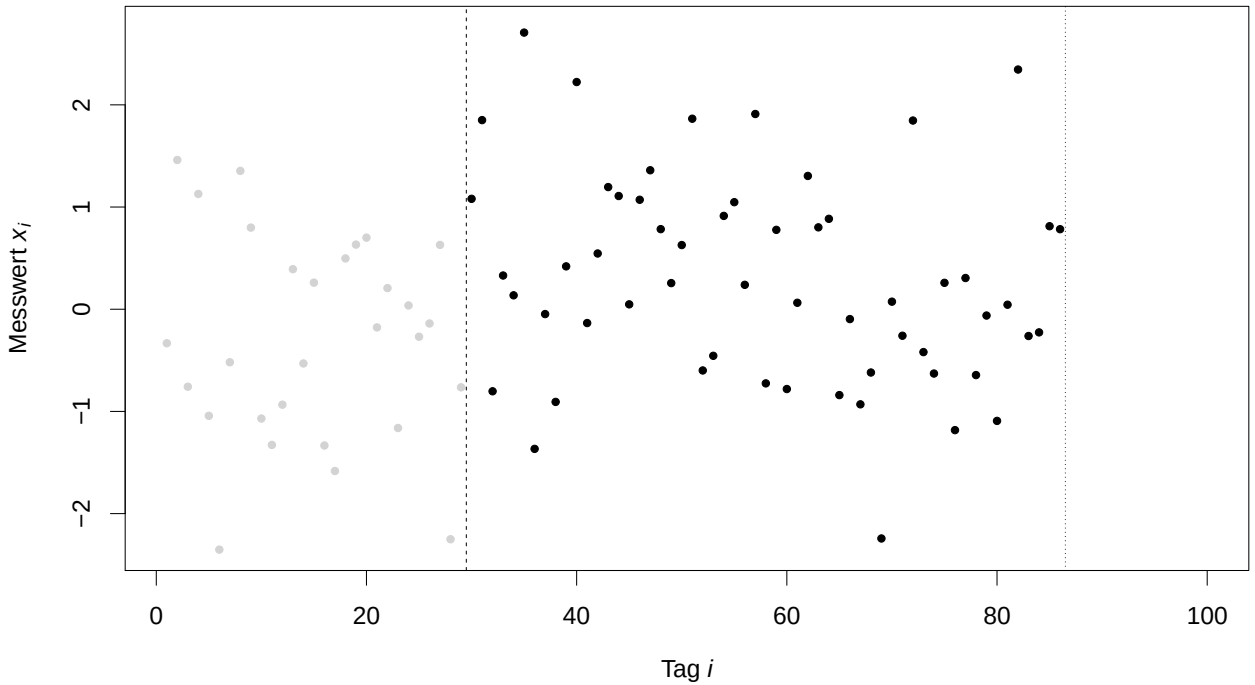
Ein simuliertes Experiment



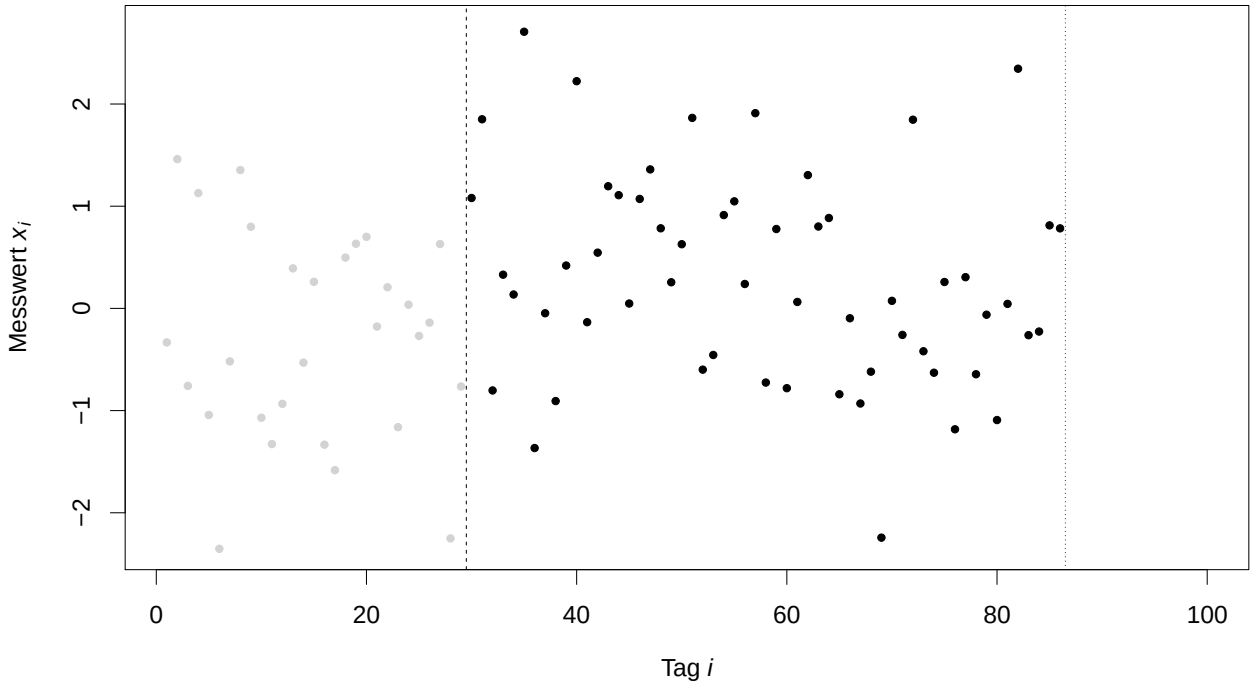
$n = 100$, $\overline{x} = 0.086$, $s/\sqrt{n} = 0.108$, $t = 0.794$, p-Wert ist 0.43 (zweiseitiger t-Test)

„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“

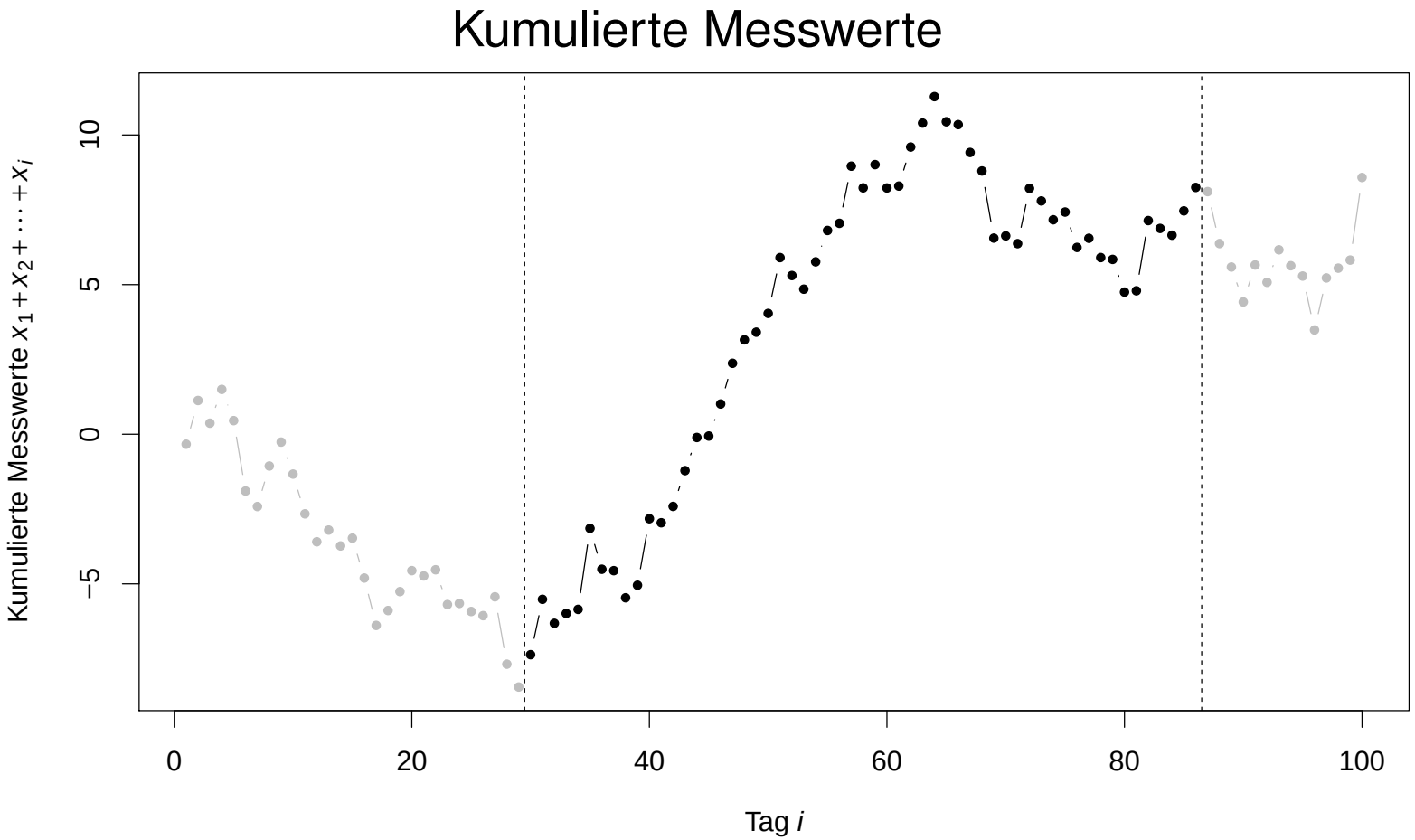


Mit den $n = 57$ Messwerten $x_{30}, x_{31}, \ldots, x_{85}, x_{86}$ ergibt sich $\overline{x} = 0.293$, $s = 1.021$, $s/\sqrt{n} = 0.135$, $t = 2.167$, p-Wert ist 0.035 (zweiseitiger t-Test)

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:

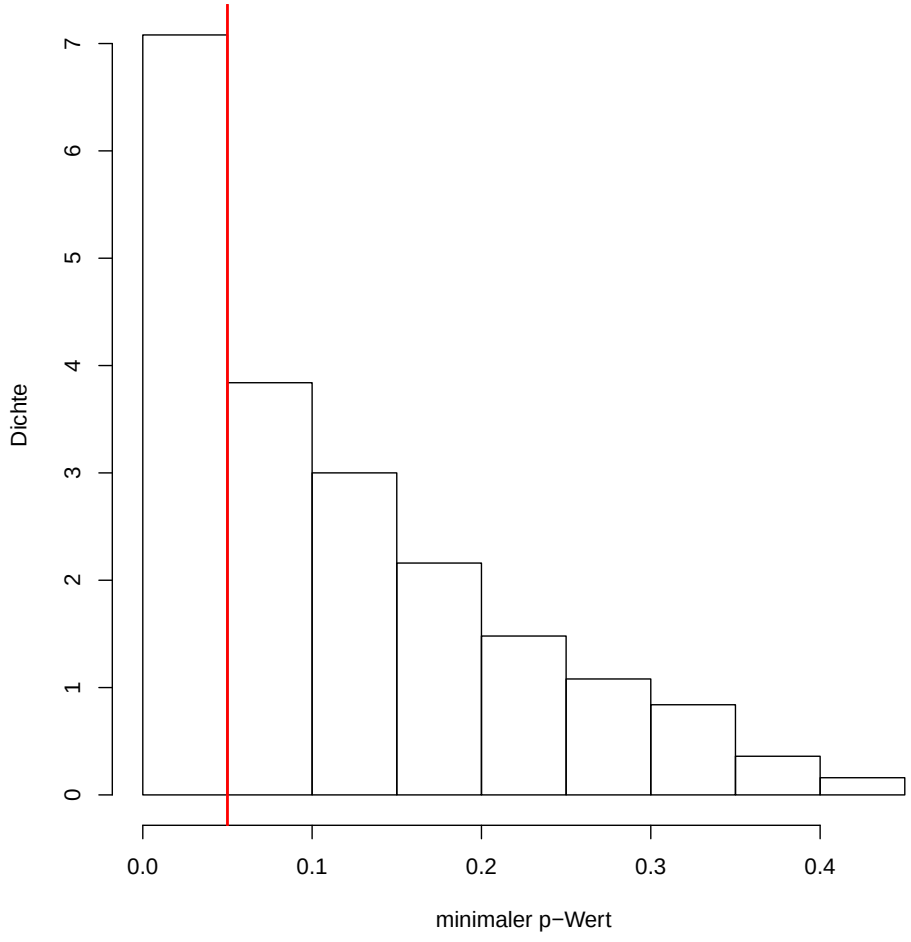
Angenommen, an jedem Tag $i = 50, 51, \ldots, 100$ geht der Experimentator die $i - 50 + 1$ möglichen Messreihen

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

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.