Tests für Mittelwerte

Marcus Hudec

Mittelwertsvergleich bei 2 gebundenen Stichproben

- Liegen 2 Beobachtungen an n Objekten vor, spricht man von einer gebundenen Stichprobe
- Typische Struktur bei "stimulus-response" Versuchen

	Obj.1	Obj.2	...	Obj.n
Beobachtung-1	x_1	x_2	...	x_n
Beobachtung-2	y_1	y_2	...	y_n
Differenz	d_1	d_2	...	d_n

Mittelwertsvergleich bei gebundenen Stichproben

- Im Falle einer gebundenen Stichprobe kann die Fragestellung durch Differenzbildung der einzelnen Beobachtungen (Übergang auf $d_i=x_i-y_i$) auf den Einstichprobenfall reduziert werden.
- Diese Vorgangsweise ist auch effizienter als die Anwendung des Zweistichprobentests für Mittelwerte
- Wenn bei gegebenen Daten eine paarweise Differenzbildung sinnvoll möglich ist, ist dies die adäquate Vorgangsweise
- Versuchsplanung: ==> Anstreben von gebundenen Stichproben

Beispiel

- 2 Düngemittel A und B werden auf 8 Versuchsfeldern unter konstanten Bedingungen getestet

Feld	1	2	3	4	5	6	7	8	Mittel
A	8,2	8,1	7,5	8,2	8,5	8,4	7,8	8,0	8,09
B	8,1	7,3	7,2	7,8	6,9	8,2	7,2	7,1	7,48
D	0,1	0,8	0,3	0,4	1,6	0,2	0,6	0,9	0,61

- Frage: besteht ein signifikanter Unterschied (α=0,05) in bezug auf den Ernteertrag pro Hektar?

Beispiel

$$H_0 : \mu_A = \mu_B \equiv H_0 : \mu_D = 0$$

$$\bar{x}_A = 8{,}0875 \; \bar{x}_B = 7{,}475 \; \bar{d} = 0{,}6125$$

$$s_d = 0{,}4883$$

$$s_{\bar{d}} = s_d / \sqrt{n} = 0{,}4883 \; / \sqrt{8} = 0{,}1726$$

$$t_{7,0,975} = 2{,}3646$$

$$t = \frac{0{,}6125}{0{,}1726} = 3{,}5482$$

Lösungsansatz: 1-Stichproben-test angewendet auf die Differenzen pro Versuchsfeld

$$p - value = 0{,}0093$$

Signifikantes Ergebnis

H0 wird abgelehnt

t-Test 2 verbundene Stichproben mit R

R Console

```
> A<-c(8.2, 8.1, 7.5, 8.2, 8.5, 8.4, 7.8, 8.0)
> B<-c(8.1, 7.3, 7.2, 7.8, 6.9, 8.2, 7.2, 7.1)
> t.test(A,B,paired=T)

        Paired t-test

data:  A and B
t = 3.5482, df = 7, p-value = 0.009367
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2043087 1.0206913
sample estimates:
mean of the differences
                 0.6125

>
```

```
> # t-Test verbundene Stichprobe
> # ============================================================================
>
> Ertrag <- c(8.2, 8.1, 7.5, 8.2, 8.5, 8.4, 7.8, 8.0,
+             8.1, 7.3, 7.2, 7.8, 6.9, 8.2, 7.2, 7.1)
> Mittel <- c(rep("A", 8), rep("B", 8))
>
> t.test(Ertrag ~ Mittel, paired=T)

        Paired t-test

data:  Ertrag by Mittel
t = 3.5482, df = 7, p-value = 0.009367
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2043087 1.0206913
sample estimates:
mean of the differences
                 0.6125

> t.test(Ertrag ~ Mittel, paired=F)

        Welch Two Sample t-test

data:  Ertrag by Mittel
t = 2.9563, df = 12.122, p-value = 0.01189
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1615923 1.0634077
sample estimates:
mean in group A mean in group B
         8.0875          7.4750
```

Mittelwertsvergleich bei 2 unabhängigen Stichproben

- <u>Fall1: Varianzen seien bekannt</u>
- Wir betrachten 2 unabhängige Stichproben von 2 Grundgesamtheiten
- Parameter der Grundgesamtheiten:

$$\mu_1, \mu_2 \quad \sigma_1^2, \sigma_2^2 \quad N_1, N_2$$

- Parameter der Stichproben:

$$\bar{x}_1, \bar{x}_2 \quad s_1^2, s_2^2 \quad n_1, n_2$$

Fall -1 Bekannte Varianzen

Modellannahmen

- Die beiden Stichproben sind unabhängig
- Entweder stammen die Stichproben aus normalverteilten Grundgesamtheiten, oder die Stichprobenumfänge n_1, n_2 sind so groß, dass mit dem zentralen Grenzwertsatz die Normalverteilung der Mittelwerte gerechtfertigt werden kann
- Die Grundgesamtheiten N_1, N_2 sind so groß, dass der Korrekturfaktor für endliche Grundgesamtheiten vernachlässigt werden kann.
- Die Varianzen der Grundgesamtheit sind bekannt

Teststatistik im Fall 1

- Unter den obigen Annahmen ist

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

- Und unter H_0: $\mu_1 = \mu_2$

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Teststatistik im Fall 1

- Im Spezialfall konstanter (homogener Varianzen) vereinfacht sich der Ausdruck für die Teststatistik wie folgt:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}$$

Fall-2: Varianzen seien unbekannt

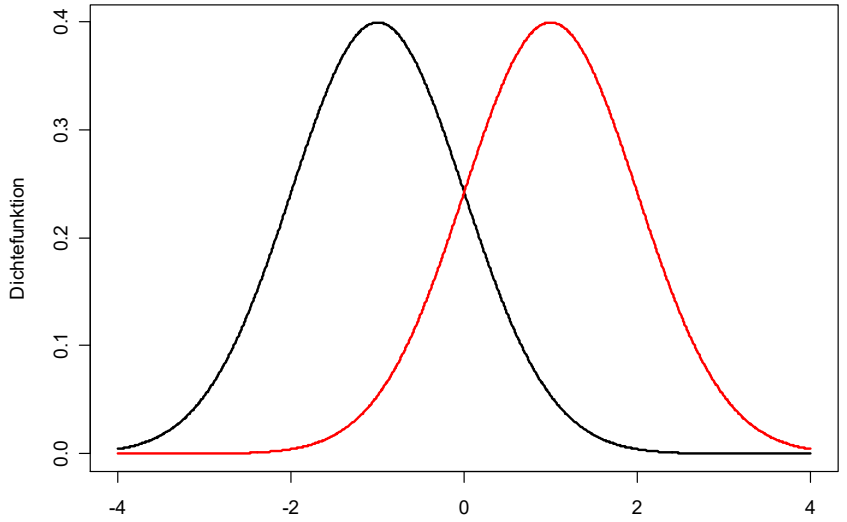
- 2a) Annahme homogener Varianzen
- Parameter der Grundgesamtheiten:

$$\mu_1, \mu_2 \qquad \sigma_1^2 = \sigma_2^2 = \sigma^2 \qquad N_1, N_2$$

- Parameter der Stichproben:

$$\bar{x}_1, \bar{x}_2 \qquad s_1^2, s_2^2 \qquad n_1, n_2$$

Modellvorstellung in Bezug auf die Alternativen



Modellannahmen Fall2a

- Die beiden Stichproben sind unabhängig
- Entweder stammen die Stichproben aus normalverteilten Grundgesamtheiten, oder die Stichprobenumfänge n_1, n_2 sind so groß, dass mit dem zentralen Grenzwertsatz die Normalverteilung der Mittelwerte gerechtfertigt werden kann
- Die Grundgesamtheiten N_1, N_2 sind so groß, dass der Korrekturfaktor für endliche Grundgesamt-heiten vernachlässigt werden kann.
- Die Varianzen der Grundgesamtheit sind unbekannt aber **in beiden Gruppen gleich**

Teststatistik im Fall 2a

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}$$

Wir greifen auf die Formel von Folie 10 zurück, müssen jetzt aber die Varianzen schätzen

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \quad \text{mit} \quad \hat{\sigma} = s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

s..."pooled variance estimate„
Gewogener Mittelwert aus den beiden Varianzschätzungen der beiden Stichproben

Die Teststatistik t ist t-verteilt mit n_1+n_2-2 Freiheitsgraden

Beispiel zu Fall 2a

- Die durchschnittliche Intelligenz zweier Personengruppen soll verglichen werden
- Annahmen:
  - IQ ist in jeder Gruppe normalverteilt und die Varianz ist in beiden Gruppen gleich groß

$$n_1 = 12 \quad \bar{x}_1 = 130 \quad s_1 = 2{,}2$$

$$n_2 = 10 \quad \bar{x}_2 = 127 \quad s_2 = 1{,}8$$

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

$$t_{20,0,995} = \pm 2{,}845$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{11 \cdot 2{,}2^2 + 9 \cdot 1{,}8^2}{20} = 4{,}12$$

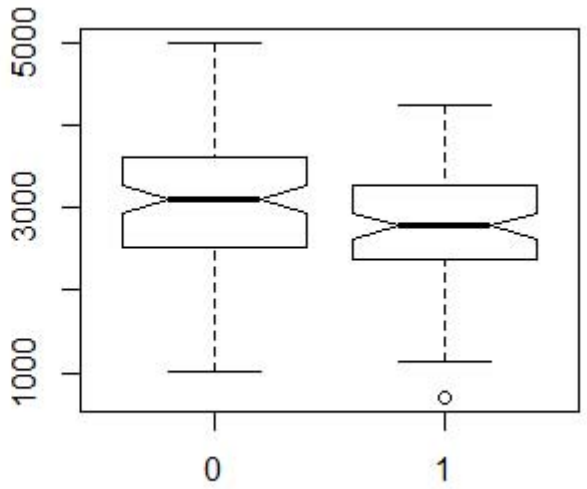
$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}} = \frac{130 - 127}{\sqrt{4{,}12}\sqrt{\frac{22}{120}}} = 3{,}45$$

Beispiel mit R

R Console

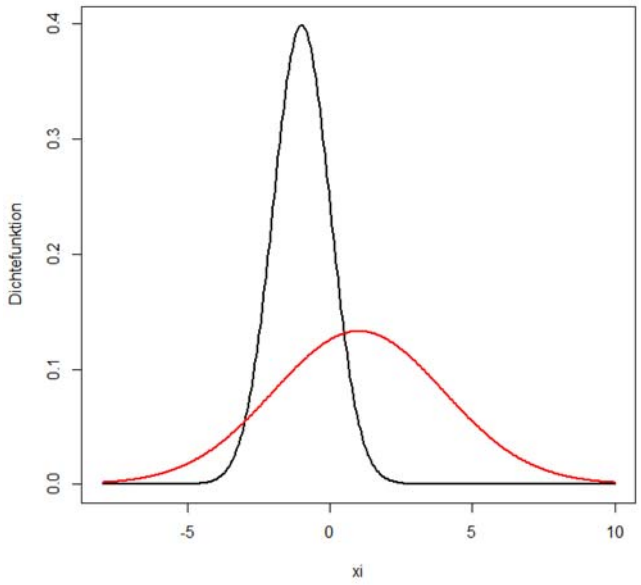
```
> boxplot(BWT ~ SMOKE, notch=T)
> t.test(BWT ~ SMOKE, var.equal=T)

        Two Sample t-test

data:  BWT by SMOKE
t = 2.6336, df = 187, p-value = 0.009156
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  70.69274 492.73382
sample estimates:
mean in group 0 mean in group 1
       3054.957        2773.243

>
```

R Graphics: Device 2 (ACTIVE)



Mittelwertsvergleich bei 2 unabhängigen Stichproben

- Fall2: Varianzen seien unbekannt
- 2b) Varianzen sind verschieden
- 2b1) große Stichproben
  - Einsetzen der Stichprobenschätzer für die unbekannten Varianzen ist bei großen Stichproben unproblematisch
- 2b2) kleine Stichproben
  - Dem Einsetzen der Stichprobenschätzer für die unbekannten Varianzen muss bei kleinen Stichproben Rechnung getragen werden ➔ t-Verteilung



Teststatistik im Fall 2b1

- Fall2: Varianzen seien unbekannt
- 2b) Varianzen sind verschieden
- 2b1) große Stichproben
- Entweder stammen die Stichproben aus normalverteilten Grundgesamtheiten, oder wir können mit dem zentralen Grenzwertsatz die Normalverteilung der Mittelwerte rechtfertigen
- Unter H_0 (siehe Folie 10):

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

Beispiel zu Fall 2b1)

- 2 Übungsgruppen von Studenten

$$n_1 = 40 \quad \bar{x}_1 = 74 \quad s_1 = 8$$

$$n_2 = 50 \quad \bar{x}_2 = 78 \quad s_2 = 7$$

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

$$z_{0,975} = \pm 1{,}96$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{74 - 78}{\sqrt{\frac{8^2}{40} + \frac{7^2}{50}}} = -2{,}49$$

$$\text{Entscheidung: } H_1$$

$$p - \text{value} = 0{,}01277$$

Teststatistik im Fall 2b2

- Fall2: Varianzen seien unbekannt
- 2b) Varianzen sind verschieden
- 2b2) kleine Stichproben
- Die Stichproben stammen aus normalverteilten Grundgesamtheiten
- Fisher-Behrens-Problem
- Approximation nach Welch: Bei Gültigkeit von H_0:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df} \quad \text{wobei} \quad df = \frac{\left(1 + \frac{s_1^2 n_2}{n_1 s_2^2}\right)^2}{\left(\frac{s_1^2 n_2}{n_1 s_2^2}\right)^2 / (n_1 - 1) + \frac{1}{n_2 - 1}}$$

Beispiel zu Fall 2b2)

- 2 Gruppen von Autofahrern

$$n_1 = 15 \quad \bar{x}_1 = 53 km/h \quad s_1 = 22,8$$

$$n_2 = 20 \quad \bar{x}_2 = 41 km/h \quad s_2 = 21,5$$

$$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{53 - 41}{\sqrt{\frac{22,8^2}{15} + \frac{21,5^2}{20}}} = 1,56$$

$$df = \frac{(1+1,5)^2}{1,5^2/14 + 1/19} = 32,02 ==> 32$$

$$t_{32,0,95} = 1,6939$$

$$Entscheidung : H_0$$

Bei der Approximation nach Welch werden die Freiheitsgrade der t-Verteilung verändert!

Lösung mit R

- Variant 2b1

```
> n.1 <- sum(SMOKE==0)
> n.2 <- sum(SMOKE==1)
> mean.0 <- mean(BWT[SMOKE==0])
> mean.1 <- mean(BWT[SMOKE==1])
> var.0 <- var(BWT[SMOKE==0])
> var.1 <- var(BWT[SMOKE==1])
> B_2_1 <- (mean.0-mean.1)/sqrt(var.0/sum(SMOKE==0)+var.1/sum(SMOKE==1))
> B_2_1
[1] 2.709457
> 2*(1-pnorm(B_2_1))
[1] 0.006739335
> 
```

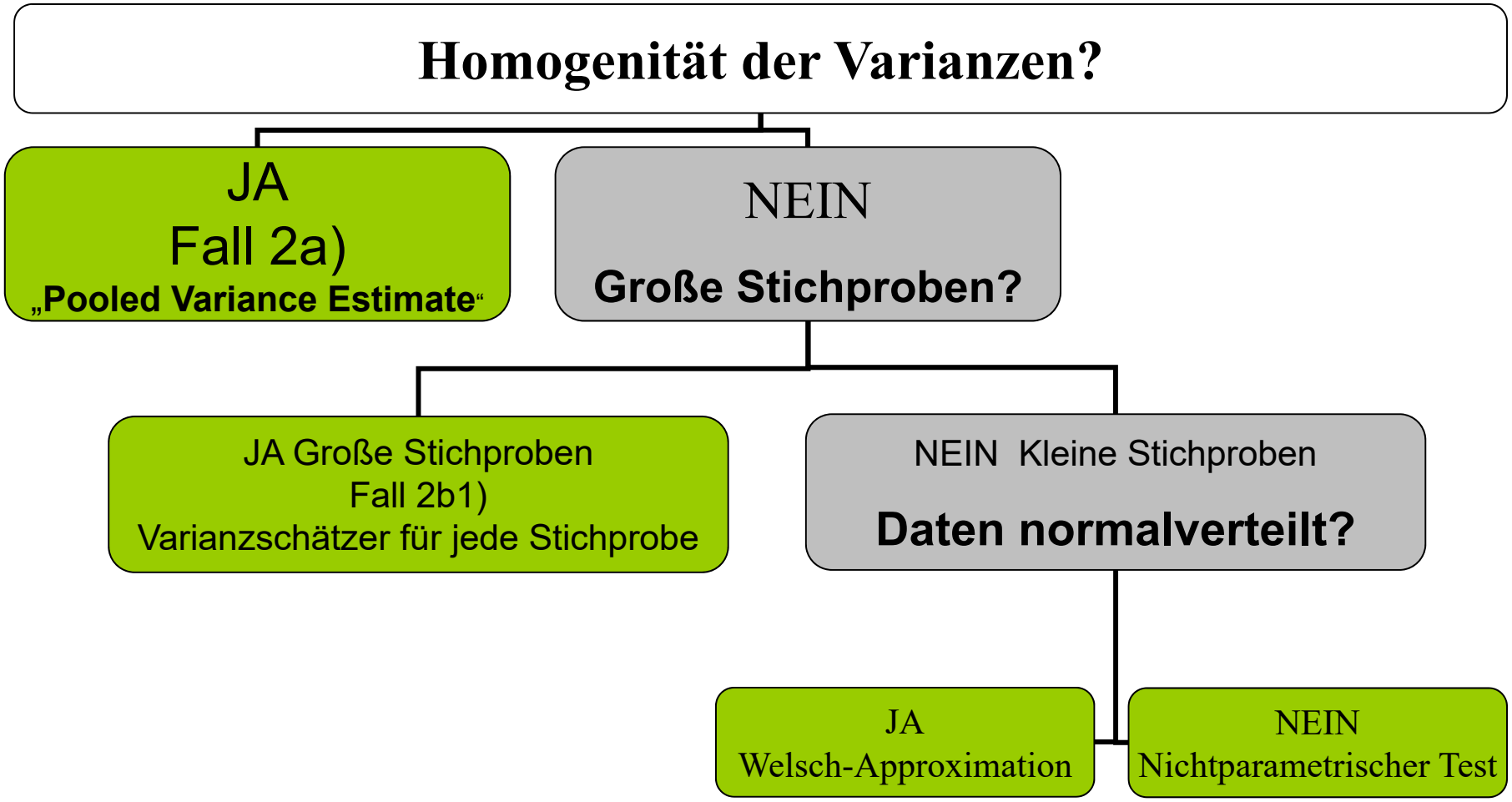
- Variante 2b2

$$df = \frac{\left(1+\frac{s_1^2 n_2}{n_1 s_2^2}\right)^2}{\left(\frac{s_1^2 n_2}{n_1 s_2^2}\right)^2 / (n_1 - 1) + \frac{1}{n_2 - 1}}$$

```
> 
> n.0
[1] 115
> n.1
[1] 74
> n.0 + n.1 - 2
[1] 187
> df.z <- (1+var.0*n.1/(n.0*var.1))^2
> df.n <- (var.0*n.1/(n.0*var.1))^2/(n.0-1)+1/(n.1-1)
> df.z/df.n
[1] 170.0013
> 
```

```
> t.test(BWT ~ SMOKE)

        Welch Two Sample t-test

data:  BWT by SMOKE
t = 2.7095, df = 170.001, p-value = 0.00743
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  76.46677 486.95979
sample estimates:
mean in group 0 mean in group 1
       3054.957        2773.243
```

2-Stichprobenfall mit unbekannter Varianz



Test auf Gleichheit der Varianzen

- Natürlich existieren auch Testverfahren, mit dem man testen kann, ob die Varianzen von 2 Gruppen gleich oder verschieden sind.
- In manchen Büchern wird empfohlen, die Wahl der Methode in Abhängigkeit von diesem Test zu treffen.
- Wenngleich dies in der Praxis häufig angewendet wird, ist dies im streng konfirmatorischen Sinn kein valides Vorgehen, da ja auch der Test auf Gleichheit der Varianzen fehlerbehaftet ist (α-, β-Fehler)
- Die Wahl der Methode (des konkreten Testverfahrens) ist streng konfirmatorisch unbedingt vor Ansicht der Daten zu treffen, wenn man das α-Niveau einhalten will.
- Der Test auf Gleichheit kann als diagnostisches Hilfsmittel verstanden werden, ob das ex ante gewählte Modell passt.

Vergleich von 2 Varianzen mit F-Test

- Unter der Annahme, dass die Daten aus 2 normalverteilten Grundgesamtheiten stammen, existiert ein einfacher Test auf Gleichheit der beiden Varianzen.
- Dabei bildet man den Quotienten der beiden Stichprobenvarianzen, welcher bei Gültigkeit der H_0 F-verteilt ist mit Freiheitsgraden n_1-1 und n_2-1
- Beachte, dass beim einseitigen Test jene Varianz im Zähler des Bruchs steht, von der wir zeigen wollen, dass sie größer ist.
- Beim zweiseitigen Test steht immer die größere Varianz im Zähler des Bruchs.

Vergleich von 2 Varianzen

	einseitiger Test		zweiseitiger Test
Hypothese	$H_0 : \sigma_1^2 \geq \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
Testgröße (F-Verteilung)	$F = \frac{s_2^2}{s_1^2}$	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{\text{größere Varianz}}{\text{kleinere Varianz}}$
Freiheitsgrade	$df_1 = n_1 - 1$	$df_2 = n_2 - 1$	
Rückweisung	H_0 ablehnen, falls $F > F_\alpha$		H_0 ablehnen, falls $F > F_{\alpha/2}$

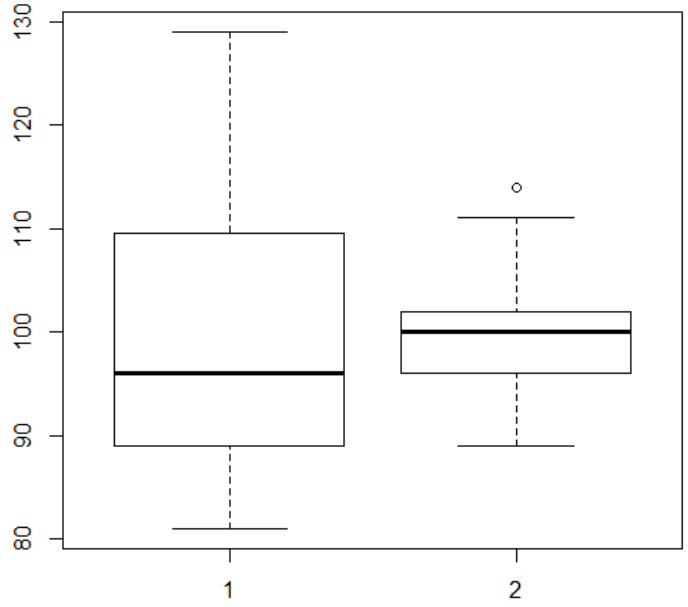
R Console

```
> var.0
[1] 566119.3
> var.1
[1] 435699.2
> var.test(BWT ~ SMOKE)

        F test to compare two variances

data:  BWT by SMOKE
F = 1.2993, num df = 114, denom df = 73, p-value = 0.229
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8469514 1.9550579
sample estimates:
ratio of variances
          1.299335
```

Beispiel zum Vergleich von 2-Varianzen

- Wir haben Beobachtungen aus 2 Gruppen

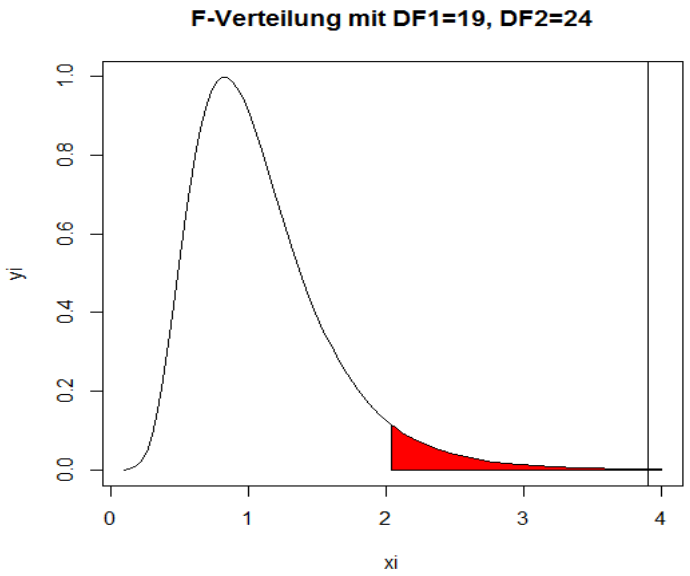
```
> x1
 [1]  93 109  88  81  90  96  90 111 100 113  96  82 129  99  90  98  84 110 110  81
> x2
 [1] 103  98  94 111 101 103  89  92 100 100  91  98  96 100 114  98  90 101 101 110 102  96 108 100  92
```



Der Boxplot zeigt deutlich, dass in Gruppe 1 die Streuung wesentlich größer zu sein scheint.

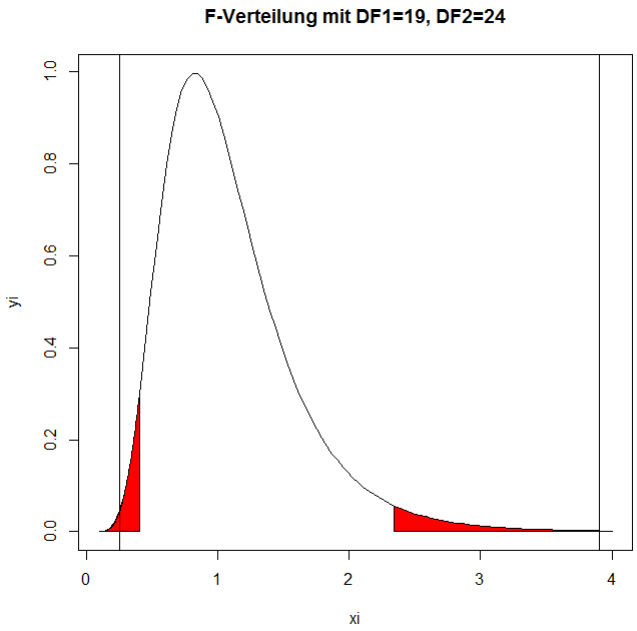
Einseitiger Signifikanztest

- Ho: $\sigma_1^2 \leq \sigma_2^2$
- H_A: $\sigma_1^2 > \sigma_2^2$
- Alpha = 0,05
- Kritischer F-Wert: df = 19, df = 24 ➔ 2,04
- s_1^2 = 164,16 s_2^2 = 42,09
- Varianz-Quotient: 3,90
- Testwert > kritischer Wert
  ➔ signifikantes Ergebnis
- p-value = 0,00103



Zweiseitiger Signifikanztest

- Ho: $\sigma_1^2 = \sigma_2^2$
- H_A: $\sigma_1^2 \neq \sigma_2^2$
- Alpha = 0,05
- Kritischer F-Wert: df = 19, df = 24 ➔ 2,3452
- s_1^2 = 164,16 s_2^2 = 42,09
- Varianzquotient: 3,90
- Testwert > kritischer Wert
  ➔ signifikantes Ergebnis
- p-value = 0.0029

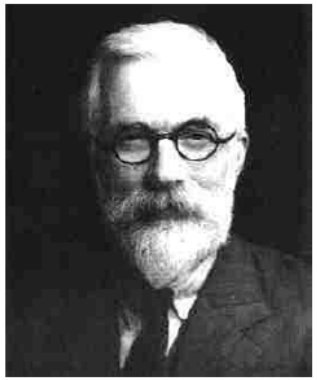


Mittelwertsvergleich bei mehr als 2 Stichproben

- Im Fall von 2 Gruppen kennen wir nunmehr verschiedene Varianten des t-Tests für den Vergleich der Mittelwerte
- Liegen k > 2 Gruppen von Beobachtungen vor und man möchte testen, ob sie sich die Gruppen signifikant in Bezug auf den Mittelwert unterscheiden, ist es nicht adäquat einfach mehrere paarweise Vergleiche durchzuführen.
- Als Standardmethode wird bei Annahme der Normalverteilung und konstanter Varianz in den Gruppen die **einfache Varianzanalyse** verwendet.

Einfache Varianzanalyse

- Methode geht auf Sir Ronald Fisher zurück

- Ein Faktor teilt eine Grundgesamtheit in k-Gruppen ein, wobei k>2 sei (ONEWAY ~ einfache Varianzanalyse, d.h. nur ein Faktor wird berücksichtigt, Ein-Weg-Varianzanalyse)
- Allgemeiner: ANOVA … Analysis of Variance

Beispiel

Fog-Index: Lesbarkeit von Texten

0,4*(durchschnittl. Anzahl Wörter pro Satz+Prozentsatz der Wörter mit mehr als 3 Silben)

Newsweek	10,21	9,66	7,67	5,12	4,88	3,12
Scientific American	15,75	11,55	11,16	9,92	9,23	8,2
Sports Illustrated	9,17	8,44	6,1	5,78	5,58	5,36

Aus jedem Journal der 3 Journale wurden 6 zufällige Artikel ausgewählt und jeweils der Fog-Index ausgewertet.

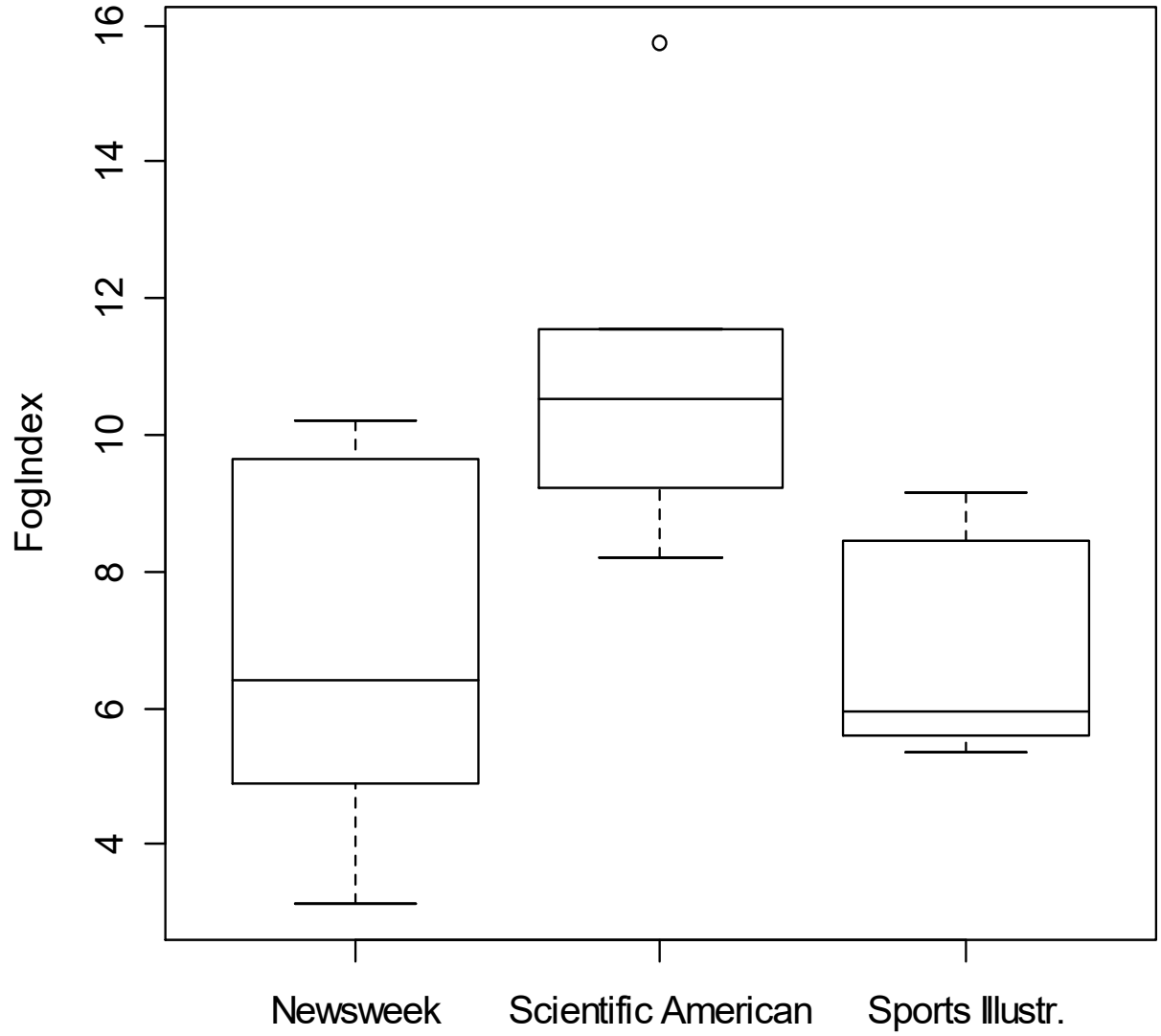
Fragestellung:

- Sind die beobachteten Mittelwerts-Differenzen zufällig oder Ausdruck eines systematischen Effekts?
- Bevor wir eine inferenzstatistische Analyse durchführen, ist eine deskriptive Darstellung der empirischen Daten sinnvoll

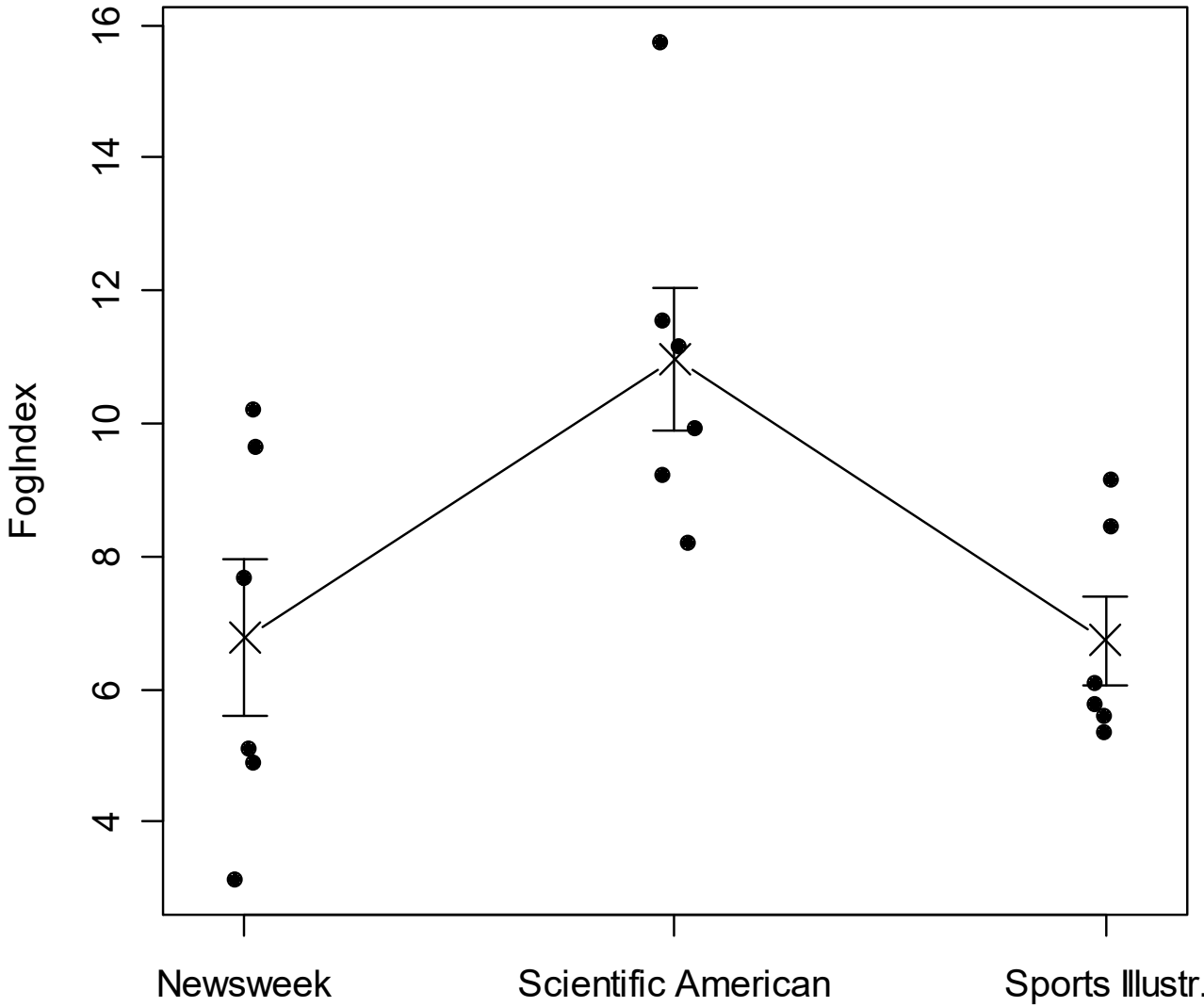
	Mittelwert	Varianz	Standardabw.
Newsweek	6,78	8,12	2,85
Scientific American	10,97	7,00	2,65
Sports Illustrated	6,74	2,68	1,64

Gesamtmittel **8,16**

Deskriptiver Vergleich mit Boxplots



Alternative Visualisierung



Statistisches Modell

- Faktor A mit Ausprägungen $a_1, \ldots, a_k$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \ldots, k \quad j = 1, \ldots, n \quad \sum_i \alpha_i = 0$$

Gesamtmittel — Effekt der Gruppe

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

Die Modellgleichung unterstellt, dass sich ein Messwert y_{ij} konzeptionell wie folgt zusammensetzt:

Gesamtmittelwert
plus
spezifischer Effekt der Gruppe - i
plus
zufällige Streuung der Beobachtung j in Gruppe – i

Im Beispiel:

A Zeitung

a1 Newsweek

a2 Scient. Amer

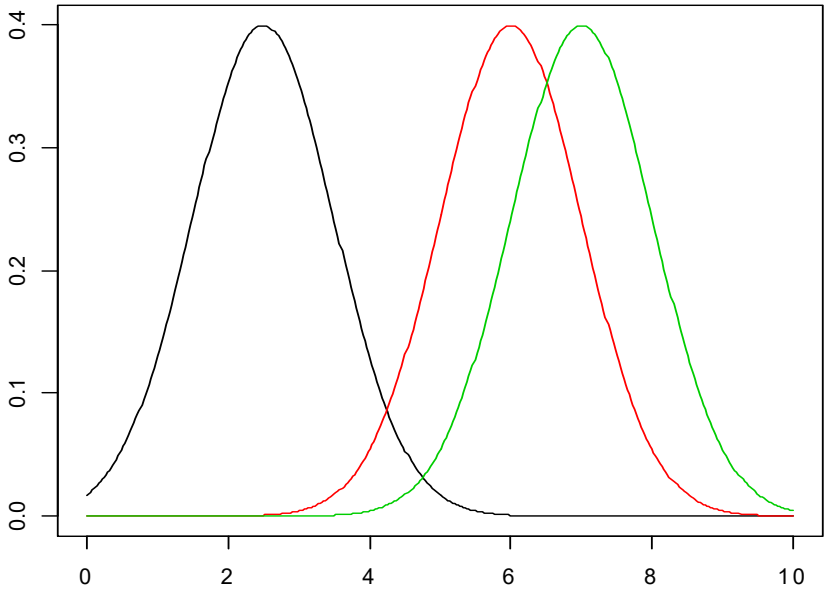
a3 Sports Illustrated

Fehler-Modell

- Bezüglich des Fehlers nehmen wir an, dass die Beobachtungen innerhalb einer Gruppe jeweils mit der gleichen Varianz um den gruppenspezifischen Mittelwert entsprechend einer Normalverteilung streuen

$$E(y_{ij}) = \mu + \alpha_i \quad i = 1,\ldots,k \quad j = 1,\ldots,n \quad \sum_i \alpha_i = 0$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



Modellannahme:
Varianzhomogenität

Interessierende Hypothese

- Wir sind interessiert zu testen:
- H_0: $\alpha_1 = \ldots = \alpha_k = 0$ gegen
- H_1: $\alpha_i \neq 0$ für zumindest ein $i \in \{1, 2, \ldots, k\}$.
- Das entspricht der Nullhypothese, dass die Erwartungswerte aller Gruppen gleich sind:
  H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- Falls H_0 abgelehnt wird, ist der Einfluss des Faktors zum gewählten Niveau statistisch abgesichert. Alternativ können wir auch sagen, dass zumindest eine Gruppe einen signifikant unterschiedlichen Erwartungswert aufweist.

Notation für k Gruppen mit je n Beobachtungen

y_{11}	y_{12}	…	y_{1n}
…			
y_{i1}	y_{i2}	…	y_{in}
…			
y_{k1}	y_{k2}	…	y_{kn}

Mittelwert in Gruppe i

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

$$\bar{y}_{..} = \frac{1}{n \cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$$

Gesamtmittelwert über alle k Gruppen

<u>Für Notation vereinfachende Annahme</u>:
konstante Anzahl von n Beobachtungen in jeder der k Gruppen

Konzept der Varianzanalyse

- Zur Analyse des Unterschiedes in Bezug auf den Mittelwert zwischen den k Gruppen wird die gesamte in der Datenmatrix enthaltene Streuung in zwei Komponenten zerlegt:
  - Variabilität zwischen den Gruppen
  - Variabilität innerhalb der Gruppen
- Der beobachtete Unterschied ist umso bedeutsamer, je geringer der Anteil der Variabilität in den Gruppen und je größer daher der Anteil der Variabilität zwischen den Gruppen ist

Rechenschema der Varianzzerlegung

- Spezialfall: Jede Ausprägung des Faktors wird genau n-mal wiederholt. Insgesamt gibt es dann N=n*k Beobachtungen

Source of Variation	SS	degrees of freedom	Sum of Squares	Mean Squares
Treatment	SSA	k-1	$n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$	MSA=SSA/(k-1)
Error	SSE	k(n-1)	$\sum_{ij}(y_{ij}-\bar{y}_{i.})^2$	MSE=SSE/k(n-1)
Total	SST	kn-1	$\sum_{ij}(y_{ij}-\bar{y}_{..})^2$	

SSTreatment ~ SSBetween ~ SSExplained
SSError ~ SSWithin

Im Beispiel:
n=6, k=3, N=18

Teststatistik

- Falls die Modellvoraussetzungen erfüllt sind, weist die Teststatistik F-Test = *MSA*/*MSE* unter der Nullhypothese eine $F_{k-1,\, nk-k}$ –Verteilung auf
- siehe die Spalte *df* (**d**egrees of **f**reedom) in der ANOVA-Tabelle
- H_0 wird genau dann zu einem Niveau α abgelehnt, falls die F- Teststatistik größer als das entsprechende 1- α Quantil der F-Verteilung mit k-1 und k(n-1) Freiheitsgraden ist

Detaillierte Berechnung (siehe XLS)

Newsweek	10,21	9,66	7,67	5,12	4,88	3,12
Scientific American	15,75	11,55	11,16	9,92	9,23	8,2
Sports Illustrated	9,17	8,44	6,1	5,78	5,58	5,36

	Mittelwert	Varianz	Standardabw.
Newsweek	6,78	8,12	2,85
Scientific American	10,97	7,00	2,65
Sports Illustrated	6,74	2,68	1,64

Gesamtmittel **8,16**

$SS_{TREATMENT}$	**70,93**

SS_{TOTAL}	4,20	2,25	0,24	9,25	10,77	25,41
	57,59	11,48	8,99	3,09	1,14	0,00
	1,02	0,08	4,25	5,67	6,66	7,85
						159,94
$SS_{Residuals}$	11,79	8,31	0,80	2,74	3,60	13,37
	22,86	0,34	0,04	1,10	3,02	7,66
	5,91	2,90	0,41	0,92	1,34	1,90
						89,01

Ergebnis

	Df	Sum of Squares	Mean Squares	F-value	p-value
$SS_{TREATMENT}$	2	70,93	35,46	5,98	0,01234
$SS_{Residuals}$	15	89,01	5,93		
SS_{TOTAL}	17	159,94	9,41		

Der F-Test testet die Hypothese, ob es überhaupt Unterschiede zwischen den Mittelwerten der Gruppen gibt.

Die Teststatistik ist definiert durch den Quotienten $MS_{TREATMENT}/MS_{Residuals}$ und folgt unter H_0 eine F-Verteilung mit k-1 und N-k [k(n-1)] Freiheitsgraden.

Große Werte signalisieren, dass die Variabilität zwischen den Gruppen im Verhältnis zur Variabilität in den Gruppen bedeutsam ist ➔ kleiner p-value ➔ signifikante Differenzen.

Multiple Paarvergleiche

Der F-Test testet die Hypothese, ob es überhaupt Unterschiede zwischen den Gruppen gibt.

Angenommen der F-Test liefert ein signifikantes Resultat, so bedeutet dies, dass nicht alle Gruppenmittelwerte gleich sind.

Aber: Welche der Gruppenmittelwerte unterscheiden sich wirklich (statistisch nachweisbar) voneinander?

Keine adäquate Strategie ist es, einfach alle paarweisen Vergleiche zu rechnen.

Multiple Paarvergleiche

Bei k-Gruppen gibt es $\binom{k}{2}$ paarweise Vergleiche.

k-Gruppen	Paarvergleiche
3	3
4	6
5	10
6	15
7	21
8	28
9	36
10	45

Da sich die einzelnen Fehlerwahrscheinlichkeiten der vielen Paarvergleiche kumulieren, ist der Fehler 1.Art wesentlich größer als die für jeden einzelnen Paarvergleich gewählte Irrtumswahrscheinlichkeit.

Multiple Vergleiche

- Das Problem multipler Vergleiche (multiple comparisons) tritt beim Hypothesentesten immer dann auf, wenn durch das wiederholte Anwenden von Hypothesentests im Rahmen einer empirischen Untersuchung der Fehler 1.Art vergrößert wird.
- Ziel muss es sein, den studienweiten Fehler für die gesamte Untersuchung zu kontrollieren (experimentelle Fehlerrate, experimentwise oder familywise error rate).

Multiple Vergleiche

- Führen wir im Rahmen einer Erhebung m unabhängige Vergleiche durch, und nehmen wir an, dass jede einzelne Nullhypothese zutrifft, so erhöht sich der Fehler 1.Art für die gesamte Untersuchung wie folgt:
- α…gewählte Signifikanzniveau pro Einzeltest
- α^*… Signifikanzniveau dieser Untersuchung

$$\alpha^* = 1 - (1 - \alpha)^m$$

$\alpha = 0{,}05$

m	α^*
1	0,05
2	0,10
3	0,14
4	0,19
5	0,23
10	0,40
20	0,64
50	0,92

Bonferoni Korrektur

Für den Fall, dass wir m nicht unabhängige Vergleiche durchführen, können wir aufgrund der Boole-schen Ungleichung folgende Aussage treffen

$$\alpha^* \leq m \cdot \alpha$$

Wir können diese Aussage dazu verwenden, um mit einem einfachen Trick sicherzustellen, dass die gesamte Fehlerrate unterhalb eines gewünschten Niveaus α^* liegt. Indem wir für jeden der m Einzeltests mit α^*/m testen. Die Fehlerrate der einzelnen Tests wird also so berechnet, dass man die gewünschte experimentelle Fehlerrate durch die Gesamtzahl der durchzuführenden Tests dividiert.
Diese konservative Vorgehensweise ist unter dem Namen Bonferroni-Korrektur bekannt.

Bonferoni Korrektur

Will man also bei k Gruppen nachtesten, welche paarweisen Unterschiede signifikant sind, muss man das Signifikanzniveau adjustieren („multiple comparisons problem“)

Die einfachste (konservative) Methode: Bonferoni-Korrektur

$$\alpha^* = \alpha / \binom{k}{2}$$

Dementsprechend wird jeder Paarvergleich zum Niveau $2\alpha^* /(k^2-k)$ getestet werden, um insgesamt ein Niveau von α^* aufrechtzuhalten.

Bei k=3 bedeutet dies, dass um die experimentelle Fehlerrate von 5% zu erhalten, alle 3 Paarvergleiche
$H_o: \mu_1=\mu_2$, $H_o: \mu_1=\mu_3$ und $H_o: \mu_2=\mu_3$ zu einem Niveau von 1,6667% getestet werden müssen. Dies entspricht einer Multiplikation der rohen p-values mit 3.

Einfache Varianzanalyse mit R

```
> res.aov <- aov(Fog.Index~Zeitschrift, data=daten)
> res.aov
Call:
   aov(formula = Fog.Index ~ Zeitschrift, data = daten)

Terms:
                Zeitschrift Residuals
Sum of Squares     70.92888  89.01250
Deg. of Freedom           2        15

Residual standard error: 2.436015
Estimated effects may be unbalanced
> summary(res.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
Zeitschrift  2  70.93   35.46   5.976 0.0123 *
Residuals   15  89.01    5.93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> pairwise.t.test(Fog.Index, Zeitschrift, p.adj="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  Fog.Index and Zeitschrift

                    Newsweek Scientific American
Scientific American 0.028    -
Sports Illustr.     1.000    0.027

P value adjustment method: bonferroni
```

Bartlett Test

- Prüft, ob *k* Stichproben aus Grundgesamtheiten mit gleichen Varianzen stammen.
- Der Bartlett Test setzt eine Normalverteilung in den *k* Gruppen voraus und reagiert sensitiv auf die Verletzung dieser Voraussetzung.

$$H_0 : \sigma_1^2 = ... = \sigma_k^2 \text{ vs. } H_1 : \exists i, j \text{ mit } \sigma_i^2 \neq \sigma_j^2$$

$$X^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{n_i - 1}) - \frac{1}{N-k}\right)}$$

mit $N = \sum_{i=1}^{k} n_i$ und $S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2$, der gepoolten Varianz.

- Die Teststatistik ist unter H_0 approximativ Chi²-verteilt mit k-1 Freiheitsgraden

Levene Test

- Der **Levene-Test** ist ein Signifikanztest, der es erlaubt auf die Gleichheit der Varianzen (*Homoskedastizität*) von zwei oder mehr Grundgesamtheiten (Gruppen) zu prüfen, ohne dass wir die Annahme einer Normalverteilung benötigen.
- Die Nullhypothese ist, dass alle Gruppenvarianzen gleich sind. Die Alternativhypothese ist, dass mindestens ein Gruppenpaar ungleiche Varianzen besitzt (Heteroskedastizität).
- Der Levene-Test ist eine robuste Alternative zum Bartlett-Test, der Normalverteilung voraussetzt.

Levene Test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

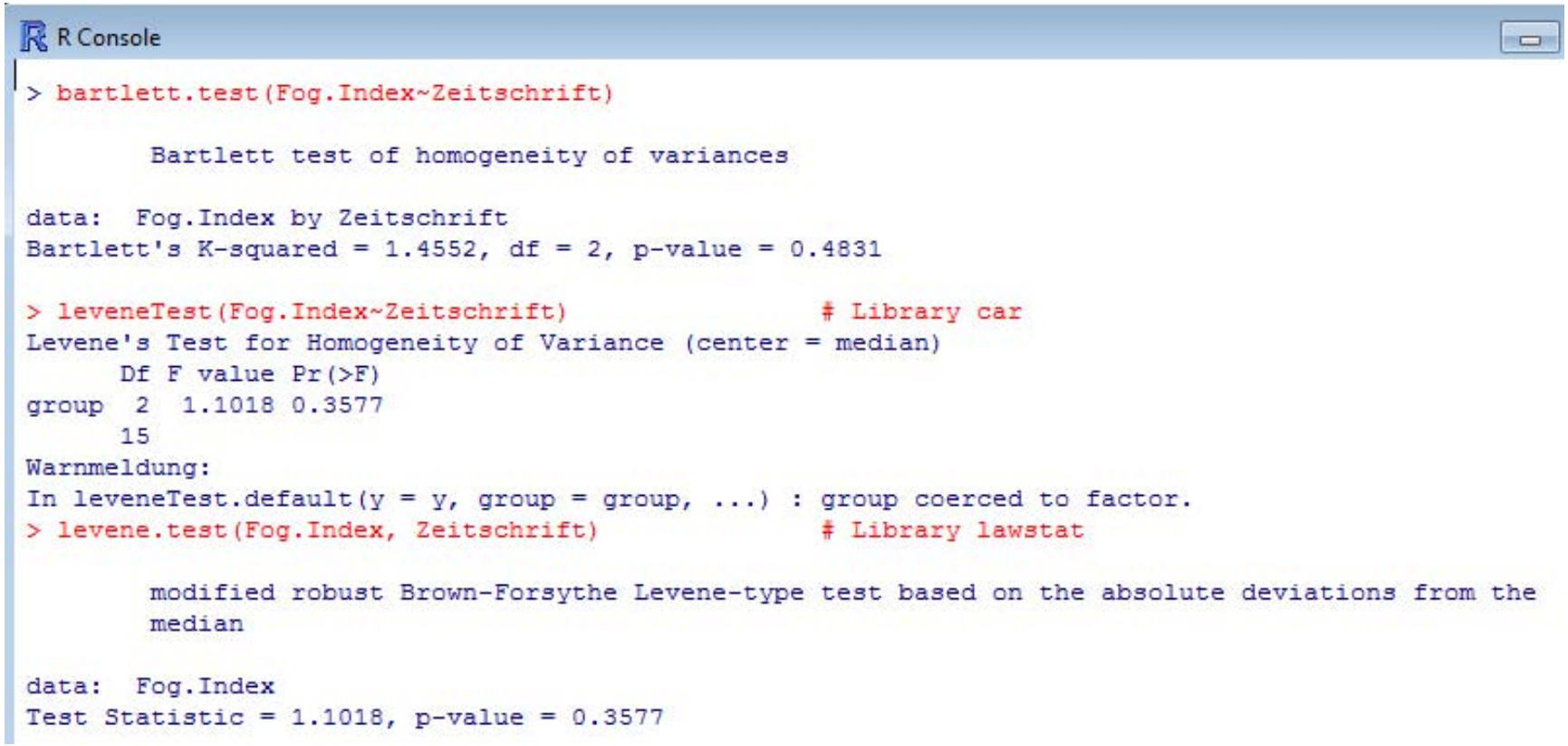
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \quad \text{für mindestens ein Gruppenpaar i,j mit } i \neq j$$

$$Z_{ij} = \left| Y_{ij} - \bar{Y}_i \right| \quad i = 1, \ldots, k$$

$$L = \frac{n\sum_{i=1}^{k} (\bar{Z}_i - \bar{Z})^2 \,/\, (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n} (Z_{ij} - \bar{Z})^2 \,/\, (N-k)} \sim F_{k-1;N-k}$$

$$\bar{Z}_i = \frac{1}{n}\sum_{j=1}^{n} Z_{ij} = \frac{1}{n}\sum_{j=1}^{n} \left| Y_{ij} - \bar{Y}_i \right|$$

- Die Teststatistik ist formal identisch zur Teststatistik der einfachen Varianzanalyse für die Gleichheit von k-Gruppenmittelwerten. Allerdings repräsentiert ein „Gruppenmittelwert Z_i“ eine robuste Schätzung der Gruppenvariabilität.

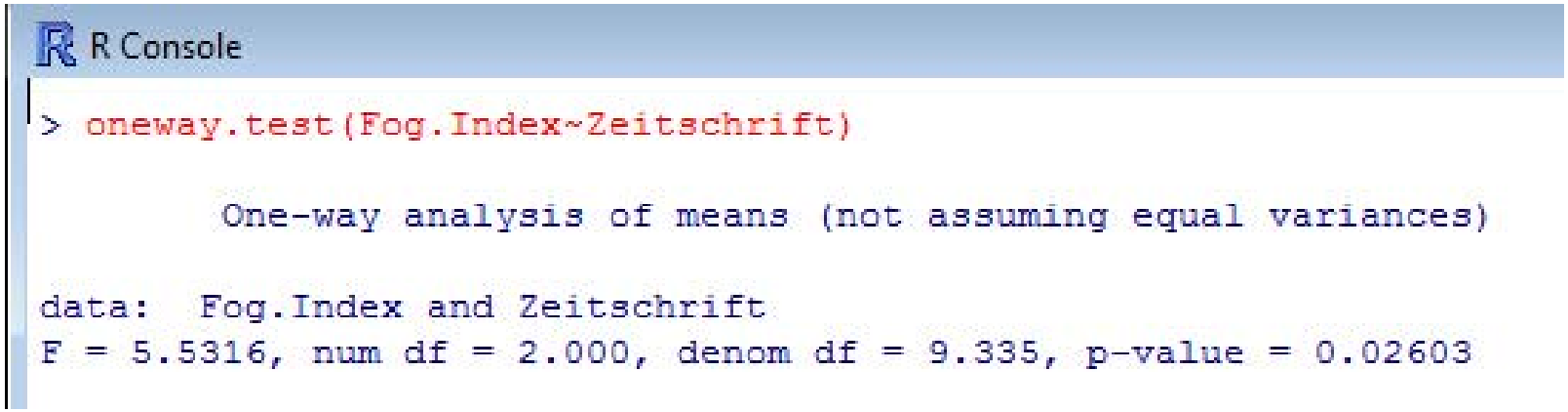

R Console
> bartlett.test(Fog.Index~Zeitschrift)
Bartlett test of homogeneity of variances
data: Fog.Index by Zeitschrift
Bartlett's K-squared = 1.4552, df = 2, p-value = 0.4831
> leveneTest(Fog.Index~Zeitschrift) # Library car
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 2 1.1018 0.3577
15
Warnmeldung:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.
> levene.test(Fog.Index, Zeitschrift) # Library lawstat
modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median
data: Fog.Index
Test Statistic = 1.1018, p-value = 0.3577

Modell-Annahmen

- Normalverteilung: F-Test robust gegenüber Verletzung, sofern Abweichung von Normalverteilung nicht zu stark!
- Annahme gleicher Varianzen in den Gruppen (Homoskedastizität): Wichtige Annahme!
- Wenn Gruppen verschiedene Varianzen haben, ist der Fehler erster Art nicht mehr garantiert.
- Mögliche Auswege:
  Varianzstabilisierende Transformationen
  Verwenden der Welch-Korrektur beim F-Test (analog zum 2-Gruppenfall)

Variante ohne Annahme gleicher Varianzen

Verallgemeinerung der Welsh-Korrektur für den Fall von k>2

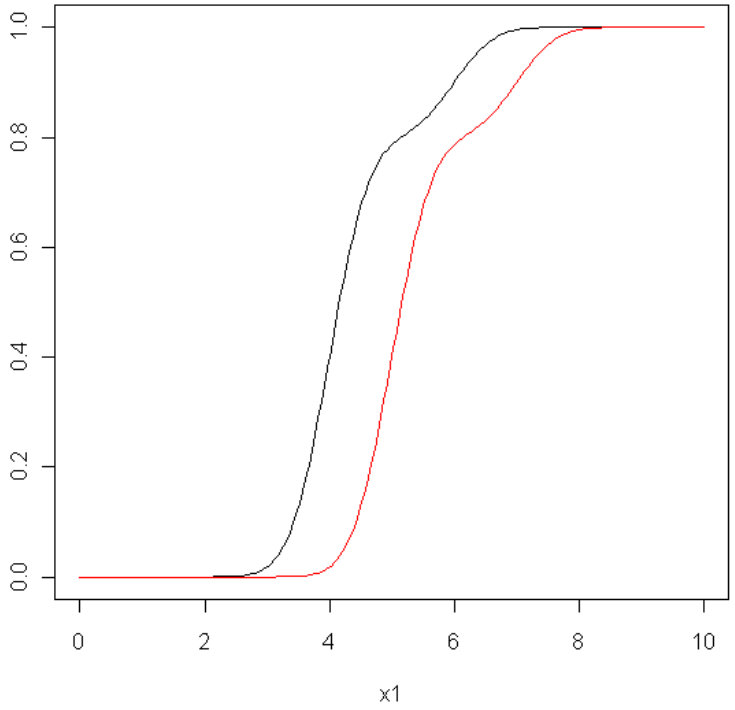


Nichtparametrische Alternative

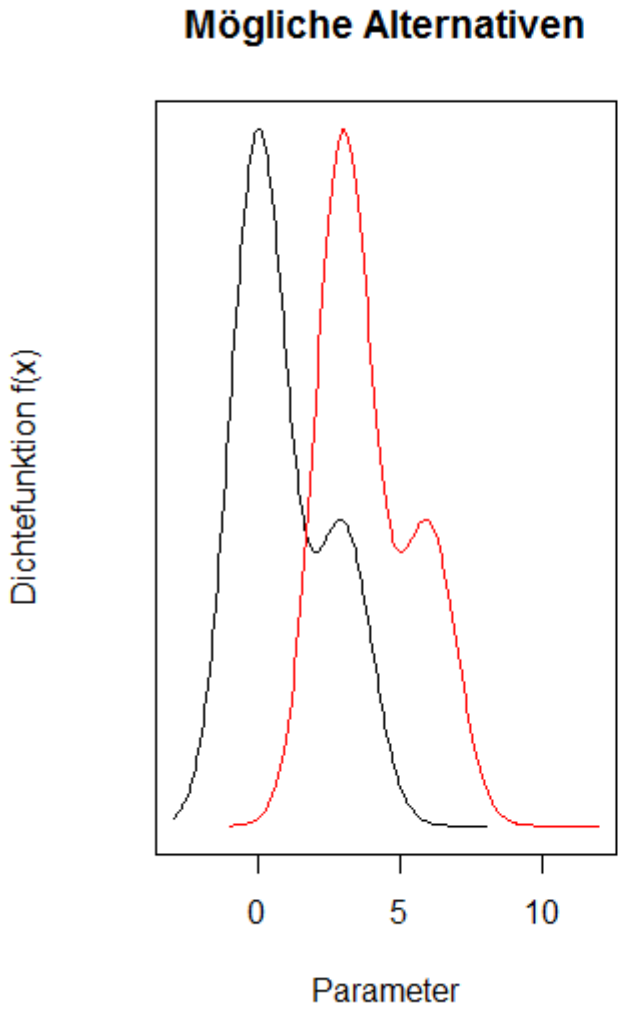
- Häufig gibt es Situationen, bei denen man der Modellvoraussetzung der Normalverteilung für den t-Test (bzw. der einfachen Varianzanalyse) zum Vergleich der Mittelwerte von 2 (bzw. k>2) unabhängigen Stichproben nicht trauen kann.
- Im Fall von 2 Gruppen eignet sich der U-Test nach Mann und Whitney (auch Wilcoxon Test für 2-unabhängige Stichproben genannt) als echte Alternative.

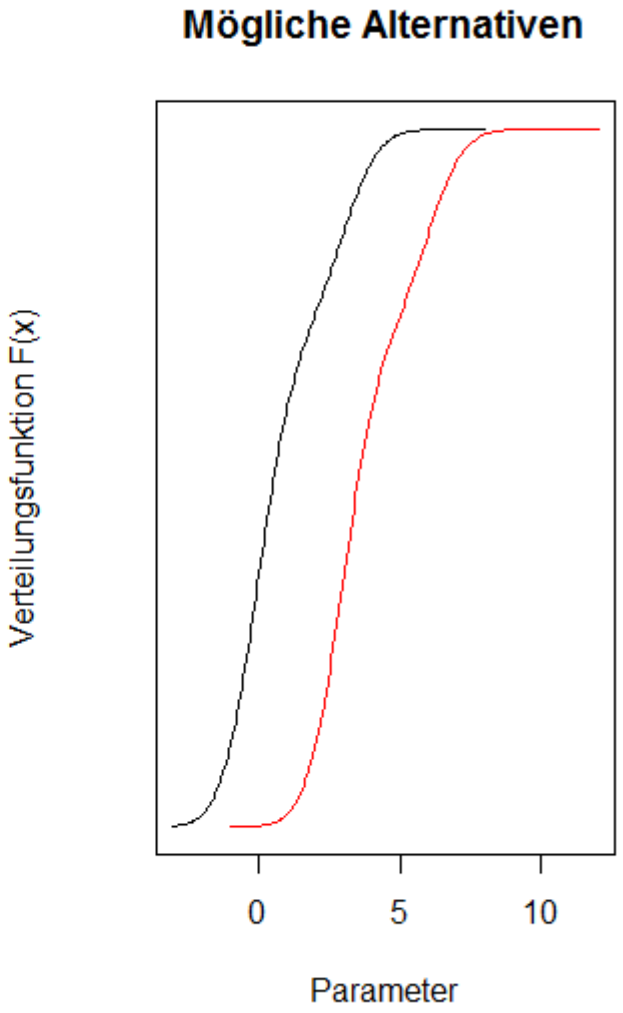
Mann Whitney U-Test

- Der U-Test ist für den Fall nicht-normaler Merkmals-Verteilungen geeignet, um auf die Gleichheit der Mittelwerte zu testen.
- Als einzige Voraussetzung gilt die Bedingung, dass die Form der Verteilungen in den beiden zu vergleichenden Gruppen gleich und stetig sein müssen.

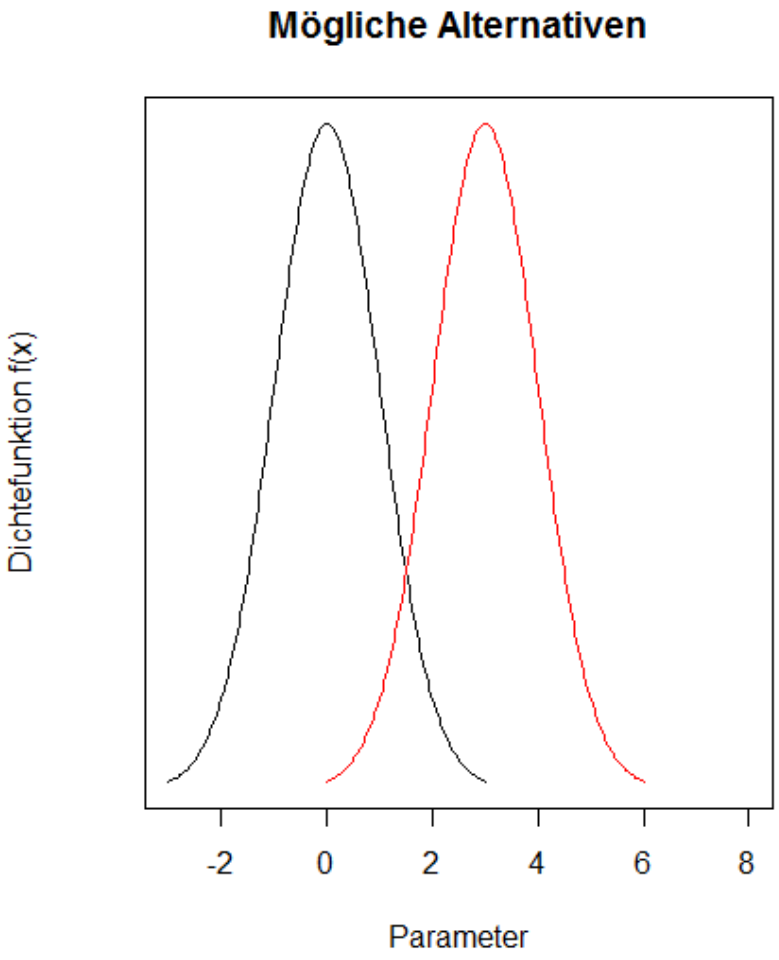


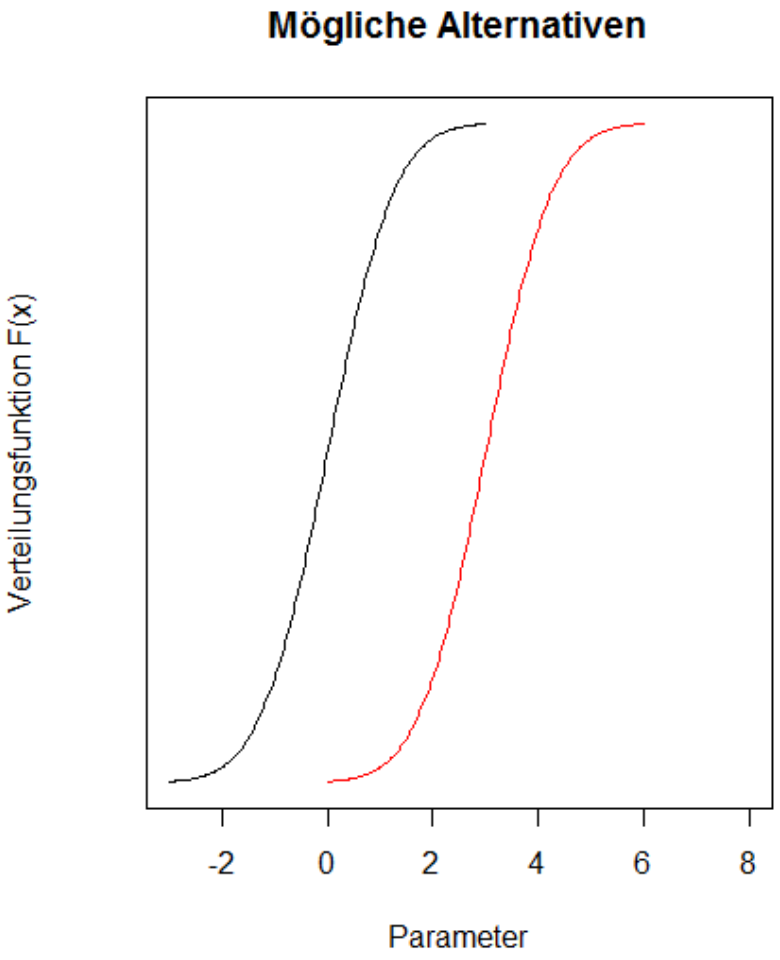
Mann Whitney U-Test (Wilcoxon Test)



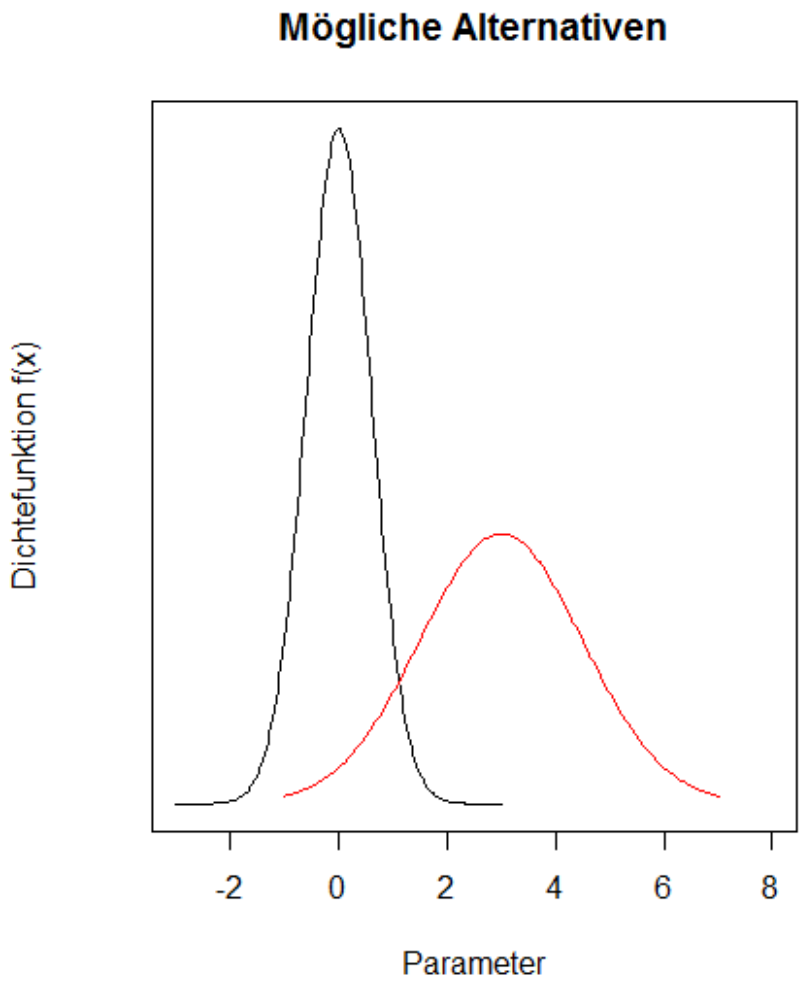


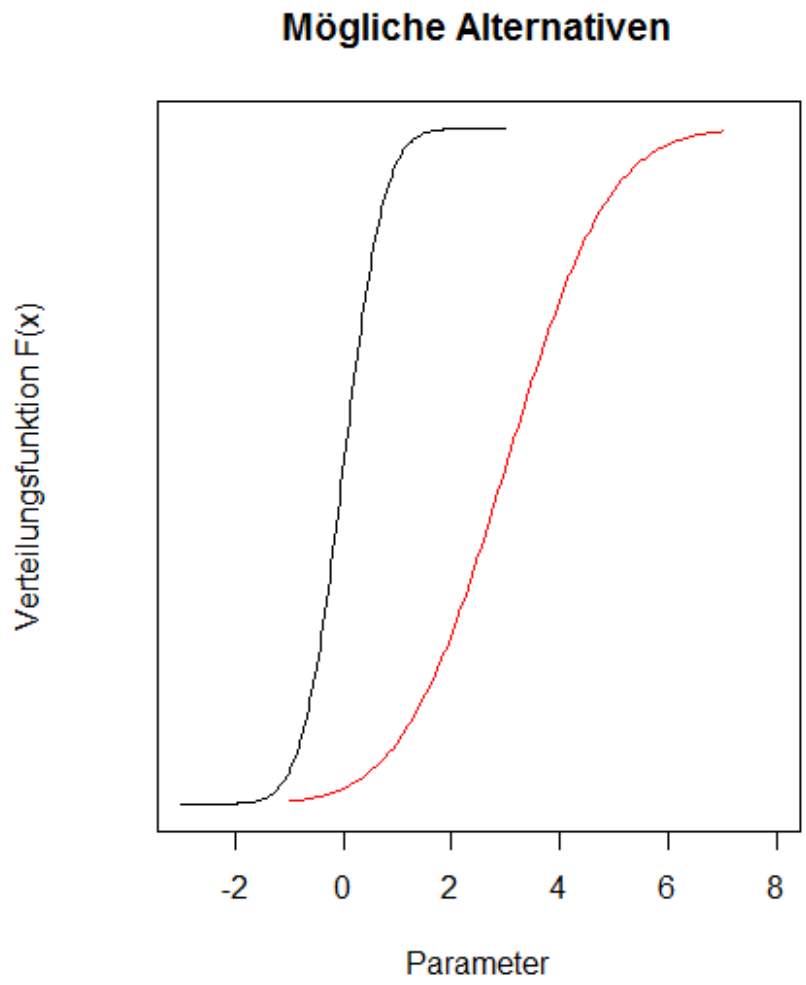
Modell: t-Test mit homogenen Varianzen





Modell: t-Test mit verschiedenen Varianzen





Mann Whitney U-Test

- Der U-Test prüft also folgende Nullhypothese:

Die Wahrscheinlichkeit einer Beobachtung aus den beiden Grundgesamtheiten ist für jede der beiden Grundgesamtheiten gleich (d.h. die Verteilungen sind gleich):

$$H0: F_1(x) = F_2(x)$$

versus der Alternative, dass eine Verschiebung vorliegt

$$H1: F_1(x) = F_2(x+a)$$

Grundidee: Mann Whitney U-Test

- Das Prinzip des U-Tests basiert auf folgender Überlegung: Sortiert man die Messwerte der beiden Stichproben in einer gemeinsamen Liste in aufsteigender Reihenfolge, so werden die Rangsummen für jede der beiden Stichproben sich nur dann stark unterscheiden, wenn sich die beiden Verteilungen unterscheiden (also die eine Gruppe systematisch kleinere Werte aufweist als die andere).
- Zur Berechnung der Prüfgröße U werden die Stichproben also gemeinsam sortiert und jeweils festgehalten, welcher Messwert zu welcher Stichprobe gehört.

Mann Whitney U-Test

- Grundgedanke: Wie verteilen sich die Ränge in einer gemeinsamen Stichprobe?
- Ermittle die Rangsumme für jede Stichprobe und teste, ob diese extreme Werte annimmt.
- Ordne alle Beobachtungen und zähle die Ränge der ersten Gruppe (R_1 ~Rangsumme der Gruppe 1) sowie der zweiten Gruppe (R_2 ~ Rangsumme der Gruppe 1).
- Bestimme die Teststatistik U nach der Formel links unten
- Für deren Verteilung gibt es bei kleinen Stichproben eigene Tabellen; bei großen ➔ Normalverteilung

$$U_1 = n_1 n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$$

$$U = \min(U_1, U_2)$$

$$falls \quad n_1 \geq 8; n_2 \geq 8$$

$$z = \frac{\left| U - \frac{n_1 \cdot n_2}{2} \right|}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}}$$

Beispiel für 2 Gruppen

					RANG
Newsweek	10,21		Newsweek	3,12	**1**
Newsweek	9,66		Newsweek	4,88	**2**
Newsweek	7,67		Newsweek	5,12	**3**
Newsweek	5,12		Newsweek	7,67	**4**
Newsweek	4,88		Scientific American	8,2	**5**
Newsweek	3,12		Scientific American	9,23	**6**
Scientific American	15,75		Newsweek	9,66	**7**
Scientific American	11,55		Scientific American	9,92	**8**
Scientific American	11,16		Newsweek	10,21	**9**
Scientific American	9,92		Scientific American	11,16	**10**
Scientific American	9,23		Scientific American	11,55	**11**
Scientific American	8,2		Scientific American	15,75	**12**
R1	26	4,33	U1	31	
R2	52	8,67	U2	5	
	78				
			U	5	

Wilcoxon Test mit R

R Console

```
> wilcox.test(Fog.Index[1:12]~Zeitschrift[1:12])

        Wilcoxon rank sum test

data:  Fog.Index[1:12] by Zeitschrift[1:12]
W = 5, p-value = 0.04113
alternative hypothesis: true location shift is not equal to 0

>
```

```
> wilcox.test(BWT ~ SMOKE)

        Wilcoxon rank sum test with continuity correction

data:  BWT by SMOKE
W = 5243.5, p-value = 0.007109
alternative hypothesis: true location shift is not equal to 0

>
```

Nichtparametrische Varianzanalyse

- Verallgemeinerung der Idee des Mann Whitney U-Tests auf den Fall von k-Gruppen
- Kruskal Wallis Varianzanalyse