Statistik in R
Christine Mooshammer

Ziele:
- Theoretische Grundlagen der deskriptiven Statistik und der Prüfstatistik
- Anwendung in der Phonetik
- Berechnungen mit R

Materialien:
Baayen, R.H. *Analyzing Linguistic Data: A practical introduction to Statistics*
http://www.mpi.nl/world/persons/private/baayen/publications/baayenCUPstats.pdf

Dalgaard, Peter (2002). *Introductory Statistics with R*. New York: Springer.

Johnson, Keith (in press). Quantitative Methods in Linguistics. Blackwell.
http://corpus.linguistics.berkeley.edu/~kjohnson/quantitative/

Leonhart, Rainer (2004). *Lehrbuch Statistik: Einstieg und Vertiefung*. Bern: Hans Huber Verlag.

Vasishth, S. *The foundations of statistics: A simulation-based approach*
http://www.ling.uni-potsdam.de/~vasishth/SFLS.html

Für die R Programmiersprache siehe auch:
Harrington, J. *The Phonetic Analysis of Speech Corpora*
Kapitel 2, The basics of R
http://www.phonetik.uni-muenchen.de/~jmh/research/emupapers/pasc.htm

Themen

Übungen mit R

1. Deskriptive Statistik
2. Maße der zentralen Tendenz und der Dispersion
3. Maße der Dispersion
4. Normalverteilung, z-Transformation
5. Prüf- oder Inferenzstatistik. T-test, F-test
6. Korrelation und Regression, lineare Regression
7. Einfaktorielle Varianzanalyse mit festen Effekten, post-hoc tests
8. Mehrfaktorielle Varianzanalyse mit festen Effekten
9. Mehrfaktorielle Varianzanalyse mit Messwiederholungen

Warum Statistik?

(a) Datenreduktion auf einige relevante Kennwerte: Prozente, Mittelwert, Standardabweichung, Varianz etc. (deskriptive Statistik)
(b) Hypothesen testen: F-Test, t-test, Varianzanalysen (Prüfstatistik)
(c) Beziehungen zwischen einzelnen Variablen herstellen: Korrelation und Regression
(d) Vorhersagen und Wahrscheinlichkeiten: stochastische Modellierung

Deskriptiven Statistik

- Datenerhebung: messen bzw. beobachten
- Merkmal und Merkmalsausprägung: Eigenschaft eines Objekts
  a) Qualitatives Merkmal: z.B. Geschlecht
  b) Quantitatives Merkmal: z.B. Körpergröße
- Variable: Merkmalsausprägung werden in Zahlen überführt
  a) diskrete Variable: z.B. Geschlecht
  b) kontinuierliche Variable: u.B. Körpergröße

Skalenniveaus

- Datenerhebung durch Messen
- Art des Skalenniveaus hängt von der Messung ab
- Skalenniveaus in aufsteigender Reihenfolge

1. Nominalskala
Einer Kategorie wird ein **Name** gegeben.
Geschlecht
Bsp. Phonetik?
Eigenschaften: Identität
Ableitbare Interpretation: Gleichheit oder Verschiedenheit

2. Ordinalskala
Zwischen den Werten wird eine *Ordnung* bzw. Reihenfolge erstellt.
Noten
Bsp. Phonetik?
Eigenschaften: Identität, Geordnetheit, Umkehrbarkeit (besser, schlechter)
Ableitbare Interpretationen: Gleichheit, Größer-, Kleiner-Relationen

3. Intervallskala
Werte werden auf einer Skala gemessen, bei der es keinen absoluten Nullpunkt gibt. Zwischen den Werten können Intervalle berechnet werden.
Temperatur in Celsius
Bsp. Phonetik?
Eigenschaften: Identität. Geordnetheit, Umkehrbarkeit, Definiertheit der Abstände
Ableitbare Interpretationen: Gleichheit, Relationen, Gleichheit und Verschiedenheit von Intervallen

4. Verhältnisskala (metrische Skala, Rationalskala)
Die Werte können in ein Verhältnis gesetzt werden, da es einen absoluten Nullpunkt gibt. Aussagen wie doppelt so hoch, lang, schwer sind möglich
Körpergröße
Bsp. Phonetik
Eigenschaften: Identität, Geordnetheit, Definiertheit der Abstände, Existenz eines Nullelements
Ableitbare Interpretationen: Gleichheit, Relationen, Gleichheit und Verschiedenheit von Verhältnissen

Weitere Beispiele: Leonhart S. 25, Aufgabe S. 30

Übung: Tabelle mit Initialen, Alter und Größe der Seminarteilnehmer

Befehle in R

Unterschied Skalar, Vektor, Matrize

- c
- seq
- rep
- cbind
- rbind

Strings

- paste
- substring

Indizierung in R

- ii=geschl=="w"
- groesse[ii]

Häufigkeiten

- hist

z.B. nn=hist(groesse)

- table

Übung: Tabelle mit Initialen, Alter und Größe der Seminarteilnehmer

- data.frame(*studies*)

Dataframes sind sowas wie Matrizen, nur dass sie gemischt Strings und Zahlen enthalten dürfen.

Tabellen bzw. Data Frames

Die ersten 10 Zeilen und 12 Spalten aus *formants*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
	lab	*f1*	*f2*	*f3*	*rms*	*vdur*	*utt*	*cdur*	*clab*	*cons*	*vp*	*CVdur*
1	l.F.s	369	2372	3070	69.8431	115.184	bd:BDLLETNSF02	94.695	l	L	bd	209.879
2	l.F.s	399	2434	3074	69.2922	113.438	bd:BDLLETNSF03	106.031	l	L	bd	219.469
3	l.F.s	356	2361	3114	71.8566	99.830	bd:BDLLETNSF04	86.115	l	L	bd	185.945
4	l.F.s	354	2403	3054	71.0318	126.328	bd:BDLLETNSF05	97.449	l	L	bd	223.777
5	l.F.s	343	2340	3108	72.8582	114.958	bd:BDLLETNSF06	85.515	l	L	bd	200.473
6	l.F.s	375	2382	3122	71.3071	109.069	bd:BDLLETNSF07	99.894	l	L	bd	208.963
7	l.F.s	346	2379	3125	72.6612	108.650	bd:BDLLETNSF08	98.320	l	L	bd	206.970
8	l.F.s	354	2403	3091	70.8650	109.013	bd:BDLLETNSF09	83.923	l	L	bd	192.936
9	l.F.s	342	2381	3108	73.4754	102.398	bd:BDLLETNSF10	81.242	l	L	bd	183.640
10	l.F.s	367	2407	2800	72.6564	145.689	bd:BDLSETNSF02	130.732	z	S	bd	276.421

Zeilennamen:

```
rownames(formants)
```

Spaltennamen:

```
colnames(formants)
```

Indizierungsmöglichkeiten:

1. Direkt: dataframename[row, *col*], ergibt einen Skalar
   z.B. formants[3, *6*] =99.830
2. Spaltennamen: dataframename$*spalte*[row], ergibt einen Skalar
   z.B. formants$*vdur*[3] = 99.830
3. Ganze Spalte: dataframename$*spalte*, ergibt einen Vektor
   z.B. formants$vdur
4. Ganze Zeile: dataframename[row,], ergibt einen Vektor
   z.B. formants[3,]
5. Mehrere Zeilen bzw. Spalten: dataframename[a:b, c:d], ergibt wieder eine Matrix bzw. einen *dataframe*
   z.B. formants[1:4,] Matrix mit den ersten vier Zeilen
6. Auswahl von einzelnen Zeilen bzw. Zellen über logische Operatoren (=subsets)
   z.B. formants[formants$loud=="N",] ergibt alle Zeilen, bei denen *loud* den Wert N annimmt

Aufgabe 1:

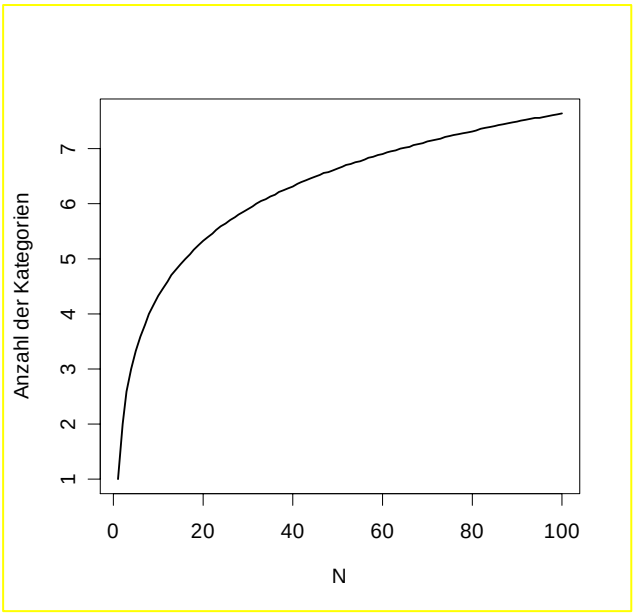
a) Zähle in einer Tabelle pro vorkommendem Alter die Anzahl der Studenten und berechne so die absolute bzw. relative Häufigkeit. (z.B. `plot(alt, freqalt, type="b")`)
b) Welches Alter bzw. welche Größe kommt bei den Seminarteilnehmern am häufigsten vor?
c) Gibt es einen Größen- bzw. Altersunterschied zwischen den anwesenden Männern und Frauen?

Aufgabe 2: mit Musterlösung

a) Lade die Datei *formants* in R
b) Stelle die Vokaldauern (vdur) graphisch dar.
c) Stelle die Vokaldauern für die einzelnen Lautstärken graphisch mit hist in einer Abbildung dar (Tipp: verwende `add=T`)
d) Welche Vokaldauer kommt bei laut am häufigsten vor, welche bei normal und welche bei leise?
e) Haben alle drei Lautstärkestufen die gleiche Anzahl von Items?

Maße der zentralen Tendenz und der Dispersion

1. Häufigkeitsverteilung



- Frage: welche Merkmalsausprägung kommt wie häufig vor?
- Kategorisierung bei diskreten Merkmalsausprägungen: table
- Kategorisierung bei unendlich vielen Merkmalsausprägungen: hist
- bins
- Regel für Kategorisierung:
  Anzahl der Kategorien= 1+3.32*lg(N)
  (immer gerundet)
- Offene Intervalle, wenn Ausreißer vorkommen

2. Maße der zentralen Tendenz

2.1 Modus (engl. mode)

Def.: Der Modalwert ist derjenige Wert einer Verteilung, welcher am häufigsten besetzt ist.

Eigenschaften:

- stabil gegenüber Extremwerten
- kann für alle Skalenniveaus verwendet werden
- Maximum einer Verteilung
- unimodale vs. bimodale vs. multimodale Verteilungen
- wird oft bei nominalskalierten Daten und bei Daten mit asymmetrischer Verteilung verwendet
- Bsp. gehörte Kategorie

Lösung in R?

2.2 Median

Def.: Der Median ist derjenige Wert, der die geordnete Reihe der Messwerte in die oberen und unteren 50 Prozent aufteilt.

Berechnung:

Für ungerades N:

$$Md = x_{\frac{N+1}{2}} \quad (3.4)$$

Für gerades N:

$$Md = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} \quad (3.5)$$

Für gruppierte Daten:

$$Md = \text{untere Grenze } f_k + \frac{\frac{N}{2} - cum\, f_{k-1}}{f_k} \cdot \text{Kategorienbreite} \quad (3.6)$$

mit

x_i : der gemessener Wert der i-ten Versuchsperson in der geordneten Rangreihe

f_k : absolute Häufigkeit in der Kategorie k, in der der Median liegt

$cum\, f_k$: kumulierte absolute Häufigkeit der Kategorie k des Medians

Aus Leonhart (2004), S. 37.

Eigenschaften:

- Anzahl der Messwerte über und unter dem Median ist gleich (entspricht einem Prozentrang von 50)
- mindestens Ordinalskalenniveau
- stabil gegenüber Extremwerten

Lösung in R?

2.3 Arithmetisches Mittel (mean, arithmetic average)

Def.: Das arithmetische Mittel ist die Summe aller Messwerte, geteilt durch deren Anzahl N.

Berechnung:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Eigenschaften:

- Summe der Zentralen Momente ergibt Null.
  Zentrales Moment= (x_i – xbar)
- Summe der quadrierten zentralen Momente ergibt ein Minimum (*sum of squared deviations SS)*
- Bei kleinen Stichproben sehr abhängig von Extremwerten
- Die Daten müssen mindestens intervallskaliert sein.

Gewichtete arithmetische Mittel siehe Leonhart

Vergleich Modus, Median und Mittelwert

Die Form einer Verteilung kann mittels des dritten und vierten Zentralen Moments (siehe 3.4, Seite 52) exakt definiert werden. Doch auch ohne diese detaillierten Maße sind schon Aussagen zur Verteilungsform über einen Vergleich der Maße der zentralen Tendenz (Median, Modalwert und Arithmetisches Mittel) möglich. Man spricht von symmetrischen, linkssteilen (=rechtsschiefen) und rechtssteilen (=linksschiefen) Verteilungsformen.

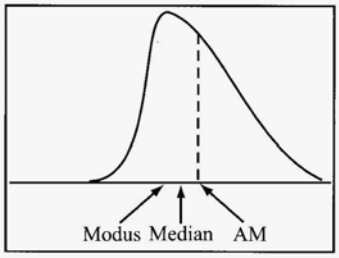


Abbildung 3.1: Linkssteile Verteilung

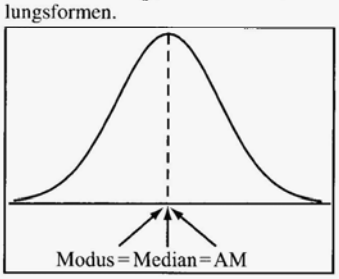


Abbildung 3.2: Symmetrische Verteilung

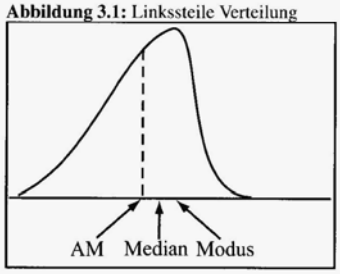


Abbildung 3.3: Rechtssteile Verteilung

Aus Leonhart (2004), S. 42.

R Befehle:

- hist
- which.max
- sort
- nrow
- sum
- cumsum

Abbildungen:

abline (mit Option *col*)	zeichnet eine Gerade in eine Graphik
par(mfcol=c(2,1))	zwei Graphiken nebeneinander
text()	

Aufgabe 3:

a) Lade die Datei segs.txt in R mit load(„segs.txt“) (Laden von Daten im R Format)
b) Berechne die verschiedenen Maße der zentralen Tendenz und zeichne sie in das Histogramm mit Beschriftung
c) Vergleiche die Maße der zentralen Tendenz der Lang- und Kurzvokale miteinander und stelle sie nebeneinander in zwei Abbildungen dar (wiederum mit Berechnung und Beschriftung des Modalwerts, des Medians und des Mittelwerts)

Aufgabe 4 (mit Musterlösung):

a) Lade die Datei *formants.Rdata*
b) Berechne die verschiedenen Maße der zentralen Tendenz und zeichne sie in das Histogramm mit Beschriftung für die Variable *cdur* (Konsonantendauer)
c) Vergleiche die Maße der zentralen Tendenz für die Konsonanten L und S (/l/ aus Lena, Lenor und /z/ aus Sehnen, Senat) miteinander und stelle sie nebeneinander in zwei Abbildungen dar (wiederum mit Berechnung und Beschriftung des Modalwerts, des Medians und des Mittelwerts) oder überlagert in einer Abbildung aber mit unterschiedlichen Farben. Achte dabei auch auf Achsenbeschriftung und Überschriften.

Maße der Dispersion

3.1 Variationsbreite (range):

Def.: Bei kontinuierlichen Daten Differenz zwischen Maximum und Minimum; bei nominalskalierten Daten die Anzahl der Kategorien

Vorteile:
- sehr einfach zu berechnen
- kann für alle Skalenniveaus verwendet werden

Nachteile:
- sehr abhängig von nur 2 Werten
- keine Aussage über die dazwischen liegenden Werte
- kann nicht für theoretische Verteilungen verwendet werden, da z.B. die Normalverteilung für einen Bereich von $\pm\infty$ definiert ist.

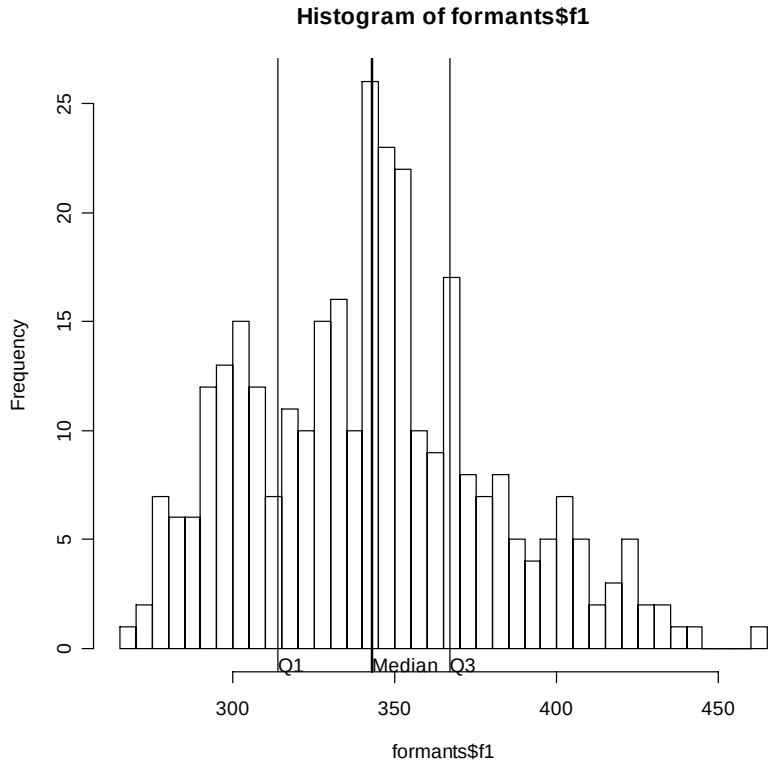
3.2 Quartile, Interquartilabstand (interquartile range)

Def.: Als Quartile werden jene Punkte Q_1, Q_2 und Q_3 bezeichnet, welche eine Verteilung in vier gleich große Abschnitte aufteilen. Das mittlere *Quartil* Q_2 entspricht dem Median, die untere Quartile Q_1 einem Prozentrang von 25 und die obere Quartile Q_3 von 75. Die Differenz von Q_3 und Q_1 wird als Interquartilabstand (IQA) bezeichnet.

Vorteile:
- Kann auch auf ordinalskalierte Daten angewendet werden.
- Der Interquartilabstand bezieht sich nur auf die mittleren 50 % der Daten, weshalb Ausreißer keine Rolle spielen.

Nachteil:
- Die Werte außerhalb werden nicht berücksichtigt.

Vgl. auch Perzentile



3.3 *Varianz (*variance*)*

Definition: Die Varianz wird durch Summierung der quadrierten Abweichungen der einzelnen Messwerte vom Mittelwert und teilen durch die Stichprobengröße, beziehungsweise den Freiheitsgrad, berechnet.

Berechnung der Varianz in der Population:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \qquad (3.37)$$

Bei der Berechnung der Populationsvarianz wird durch N geteilt.

Berechnung der Varianz in der Stichprobe:

$$s_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \qquad (3.38)$$

Durch die Berechnung der Stichprobenvarianz soll die Populationsvarianz geschätzt werden. Für diese Schätzung wird die quadrierten Abweichungen der Messwerte vom Mittelwert am Freiheitsgrad (degree of freedom) relativiert.

- Zentrales Moment zweiter Ordnung
- Quadrieren, da einfache Summe null ergeben würde ➔ unterschiedliche Stichproben können verglichen werden
- Mittelwert aller Abweichungsquardrate
- Unterschied Population (griechische Buchstaben) und Stichprobe (lateinische Buchstaben)

Def.: Freiheitsgrade (**degrees of freedom**): beschreibt die Anzahl der frei wählbaren Werte. Durch die Berechnung eines Kennwerts aus N Messwerten wird ein Messwert „unfrei“.

df=N-1

3.4 *Standardabweichung(*standard deviation*)*

Definition: Die Standardabweichung entspricht der Wurzel aus der Varianz.
Berechnung der Standardabweichung in der Population:

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \qquad (3.55)$$

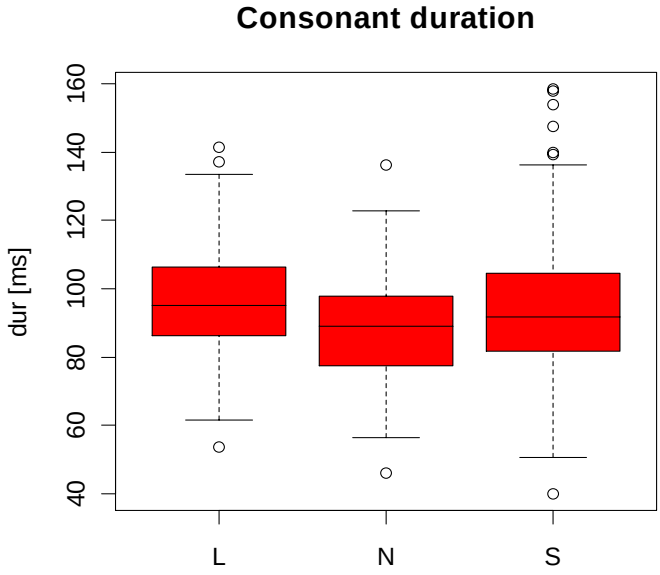
Berechnung der Standardabweichung in der Stichprobe:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \qquad (3.56)$$

Da die Abweichungen für die Varianz quadriert wurden, muss die Wurzel gezogen werden, um wieder die gleiche physikalische Einheit der Messwerte zu erhalten.

Exkurs Boxplot
Darstellungsmethode

- Strich innerhalb der Boxen: Median
- Boxen: Interquartilsabstand
- *Whiskers*: 1.5 * Interquartilsabstand an den äußeren Rändern der Box
- Bedeutung: innerhalb der „whiskers“ liegen 95% der Daten (entspricht 1.96* s_x)
- Ausreißer bzw. *outlier*: Werte außerhalb der whiskers



3.5 Variabilitätskoeffizient

Die Standardabweichung hängt von der Größe des Mittelswert ab, d.h. je größer der Mittelwert umso größer auch die Standardabweichung. Um feststellen zu können, ob zwei Stichproben mit sehr unterschiedlichen Mittelwerten unterschiedlich stark streuen, wird der Variabilitätskoeffizient berechnet.

Def.: Der Variabilitätskoeffizient gibt an, wie viel Prozent des arithmetischen Mittels die Standardabweichung beträgt.

$$s_x*100/xbar$$

R Befehle

summary	
mean	
median	
sd	
quantile	
tapply	tapply(formants$cdur, formants$loud, mean)
as.vector	
boxplot	boxplot(cdur ~ loud, data=formants)

Aufgabe 5 (mit Musterlösung):
Lade die Datei formants.Rdata. Wir wollen nun feststellen, ob die Intensität (berechnet als RMS) ein geeignetes Maß zur Unterscheidung der drei Lautstärken ist.

a) Zeichne Histogramme für die drei Lautstärken. Die relevanten Variablen heißen formants$rms und formants$loud.
b) Erstelle eine Tabelle mit den Medianen, den Mittelwerten, den Quartilen, den Standardabweichungen und den Variabilitätskoeffizienten für die drei Lautstärken einzeln und für die gesamte Verteilung.
c) Stelle die Werte in Boxplots dar.
d) Interpretiere kurz die Daten.

4. Normalverteilung

(Auch Gauß'sche Normalverteilung oder „Glockenverteilung", *normal distribution*)
Der Ausgangspunkt ist, dass Messungen in Experimenten meist zufälligen Variationen unterliegen (Reaktion der Versuchsperson, Messmethode etc.). Ist diese Annahme korrekt, so ergibt eine genügend große Anzahl an Messungen eine symmetrische Verteilung um einen zentralen Wert, der am häufigsten auftritt und durch den Mittelwert widergegeben werden kann.
Johnson (2004, p.14) beschreibt diese mittlere Tendenz als das zugrundeliegende Merkmal, das wir bei Experimenten herausfinden wollen, das aber durch zufällige Fehler „verfälscht" wird. Für die zufälligen Fehler gilt, dass die größeren Abweichungen seltener auftreten, weshalb sich die Verteilung zu den Rändern hin an null annähert.

Definition: Die Gauß'sche Normalverteilung ist durch die folgende Funktionsgleichung bestimmt:

$$f(x) = \frac{1}{\sigma_x \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma_x^2}} \qquad (3.75)$$

Eigenschaften:

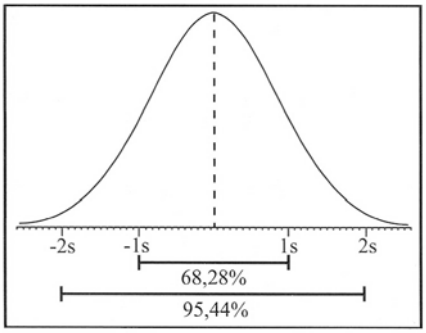
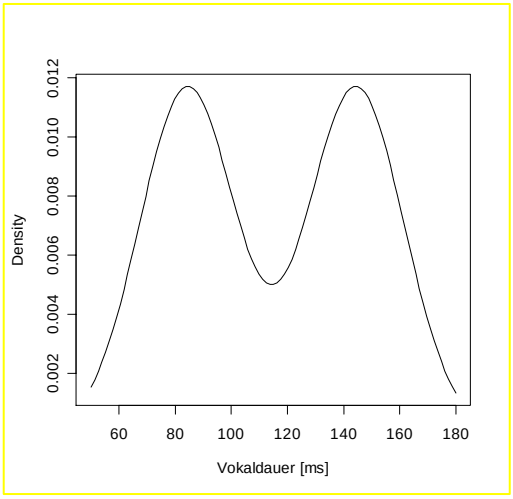


Abbildung 3.8: Normalverteilungskurve

- Datenreduktion: Mit den beiden Kenngrößen μ und σ kann die Wahrscheinlichkeit für das Auftreten einzelner Messwerte vorhergesagt werden.
- Die Fläche unterhalb der Kurve ist immer 1, d.h. Normalverteilungen mit einem Mittelwert, der eine geringe Häufigkeit aufweist, haben eine große Standardabweichung („flach und breit") und umgekehrt („spitz und schmal")
- **Dichte** (*density*): gibt die Wahrscheinlichkeit an, dass ein Maß sehr nah an einem Messwert liegt. Wahrscheinlichkeiten liegen zwischen 0 und 1 mit steigender Wahrscheinlichkeit. Die Wahrscheinlichkeit, dass ein Wert über oder unterhalb einem bestimmten Wert liegt, kann mit dem **Integral** der Normalverteilung berechnet werden.
- Bei normalverteilten Daten liegen 68,28% der Daten innerhalb eines Bereiches von ± 1Standardabweichung und 95,44 % im Bereich von ± 2 SD
- Im statistischen Sinne normale Daten liegen zwischen -1,96 * SD und +1,96*SD. Alle außerhalb dieser 95% Marke liegenden Daten sind Ausreißer.

Abweichungen von der Normalverteilung



1. Mehrere Gipfel (**bimodal** bis multimodal)
bedeutet meist, dass die Quelle der Variation **nicht** zufällig ist, z.B. Vokaldauern, wenn Kurz- und Langvokale in einem Datensatz analysiert werden.

2. **Asymmetrie** (*skewness*)
Vgl. Abbildung S. 6, oben. Die Schiefe wird mit dem zentralen Moment dritter Ordnung berechnet.
a_3=0: Symmetrie
a_3<0: rechtssteil
a_3>0: linkssteil

Definition: Die **Schiefe** einer Verteilung wird über das dritte Zentrales Moment berechnet:

$$a_3 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N \cdot s_x^3} \qquad (3.59)$$

3. „Gipfelichkeit“, Exzess, Breite, **Kurtosis**

Definition: Der **Exzess**, die Breite einer Verteilung, wird über das viertes Zentrale Moment berechnet:

$$a_4 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N \cdot s_x^4} \qquad (3.61)$$

a_4=3: normal
a_4<3: platykurtisch (breit)
a_4>3: leptokurtisch (spitz)

Kurtosis: auf 0 normalisiert

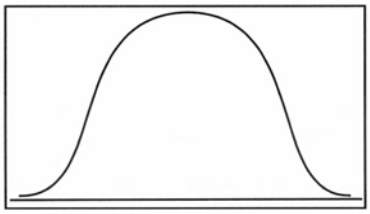


Abbildung 3.6: Eine breitgipflige Verteilung

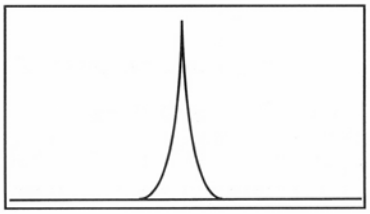
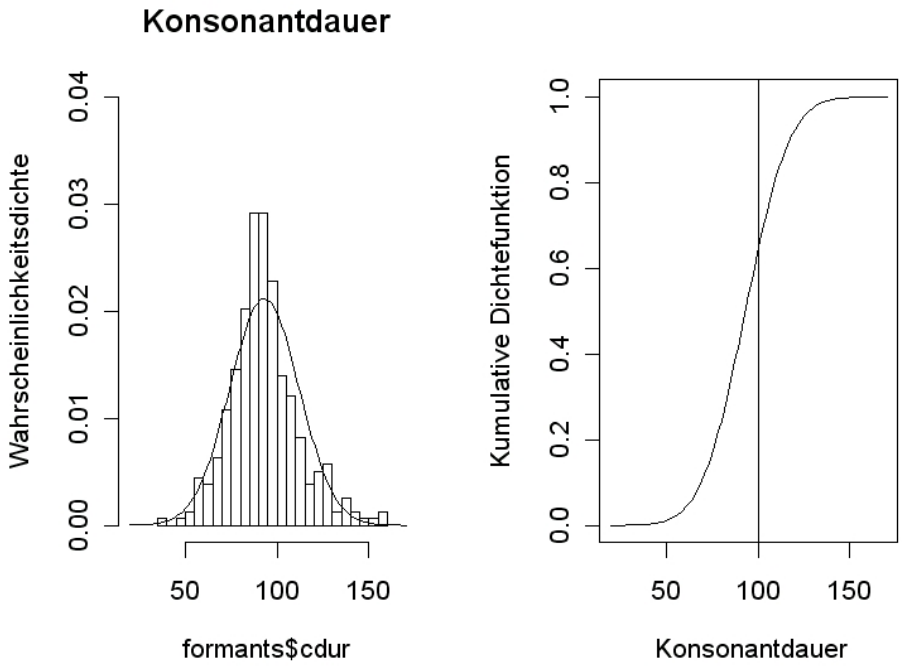


Abbildung 3.7: Eine schmalgipflige Verteilung

R-Befehle

rnorm Erzeugen von zufallsverteilten Daten, def. durch xbar und sd
dnorm Berechnung der Normalverteilung
pnorm(x, *mean*, *sdev*) Dichtefunktion, gibt an, wie groß die Wahrscheinlichkeit ist, dass Werte kleiner als x vorkommen. Für Werte größer als x 1-pnorm(....)
hist(...., freq=F)
plot(function(x) dnorm(x, *mean*, *sdev*), 10, 180, add=T)

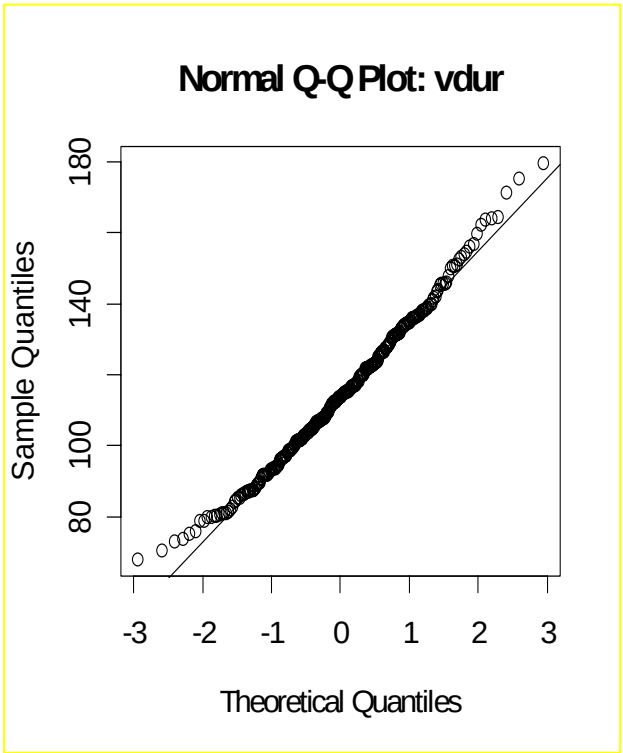
Zum Beispiel:



Einige Befehle zu sogenannten Q-Q Plots, die die Abweichung einer empirischen Verteilung von der Normalverteilung darstellen:

qqplot
qqnorm erzeugt die Kringel
qqline erzeugt die Linie, auf denen die Punkte liegen müssten, wenn es sich um normalverteilte Daten handelt.

Hinweis: weitere Erklärungen vgl. Johnson Kap. 1.



Aufgabe 6 (mit Musterlösung):
a) Teste, ob die Variable *rms* aus dem Datensatz *formants* normalverteilt ist.

- Erzeuge hierfür „künstliche“ normalverteilte Daten mit Mittelwert und Standardabweichung von *rms*.
- Stelle Histogramm und Normalverteilung zusammen in einer Graphik dar.
- Berechne das dritte und vierte Moment. Sind die Daten symmetrisch und „normal gipfelig“?
- Stelle einen Q-Q Plot her. Sind die Daten normal verteilt?

b) Wie hoch ist die Wahrscheinlichkeit, dass

- Vokale vorkommen, die länger als 140 ms sind

- mit einer RMS von weniger als 60 dB produziert wurden

Standardnormalverteilung und die z-Transformation

Die Standardnormalverteilung hat einen Mittelwert von 0 und eine Standardabweichung von 1. Die Überführung geschieht durch die z-Transformation in die sogenannten *z scores*.

$$z_i=(x_i\text{-xbar})/s_x$$

Eigenschaften der z-Verteilung:

- Die Fläche ist wiederum 1 bzw. 100%.
- Berechnung des Prozentrangs: wie viel Prozent der Daten liegen unterhalb oder sind gleich einem bestimmten Wert? Z.B. wie viel Prozent der Individuen wiegen unter 85 kg?

Wichtige Anwendung in der Phonetik: **Sprechernormalisierung**

Problem: Formanten sind nicht nur von der Vokalqualität sondern auch von sprecherspezifischen Merkmalen des Ansatzrohres abhängig.

Lösung:

1. z-Transformation mit sprecherspezifischen Mittelwerten und Standardabweichungen = **Lobanov-Transformation**

$$F_{n.norm}=(F_n-F_{n.mean})/F_{n.sd}$$

$F_{n.norm}$ wird für jeden einzelnen Sprecher berechnet.

n entspricht jeweils dem n-ten Formanten (F1, F2 etc.)

Die Daten können zurück transformiert werden, indem die Daten mit der Gesamtstandardabweichung multipliziert und zum Gesamtmittelwert werden.

2. Daten werden auf den maximalen *Range* der einzelnen Sprecher normalisiert = **Gerstman-Transformation**

$$F_{n.norm}=(F_n-F_{n.min})/(F_{n.max}-F_{n.min})$$

Hinweis: Schlag nach bei Harrington & Cassidy (1999) S. 76-78

Aufgabe 7 (bitte alle vorbereiten):
Vergleiche die Formanträume der Vokale des Deutschen in betonten und unbetonten Silben.
Schritte:

a) Lade hierfür die Datei *speakernorm.Rdata*. Die geladene Tabelle heißt gervow. Sie besteht aus Messungen der ersten beiden Formanten für alle 15 Vollvokale des Deutschen in betonten und unbetonten Silben. Dabei bedeutet ein „+“ im Vokallabel *gespannt* und ein „-„ *ungespannt*.

b) Verwende für einen ersten graphischen Überblick den Befehl eplot.
Um die betonten und unbetonten Vokale übereinander drucken zu können, sollte zweimal nacheinander der eplot Befehl ausgeführt werden (beide Male colour=F angeben, dazwischen par(lwd=2, new=F), damit betonte und unbetonte in eine Abbildung gelangen). Außerdem muss die Achsenlänge (xlim, ylim) angegeben werden.

c) Berechne sprechernormalisierte Formanten nach Lobanov und Gerstman und vergleiche die Ergebnisse. Da hierzu die Mittelwerte für die einzelnen Sprecher berechnet werden müssen etc., empfiehlt es sich eine Schleife zu verwenden.

```
for (zähler in anf : ende) {
        befehl1
        befehl2
        etc.
}
```

Hinweise:

- Spätestens bei der Verwendung von Schleifen sollte man die Befehlsfolgen in einen Dateneditor (z.B. *nedit*) schreiben und anschließend speichern.
- Die Anzahl der Versuchspersonen lässt sich über den Befehl levels herausfinden.
- Die beiden transformierten Formantwerte sollten wieder in eine Matrix geschrieben werden. Diese sollte man möglichst vorher schon in der richtigen Größe definieren mit dem Befehl matrix.

Prüf- oder Inferenzstatistik

Hypothesen über die gesamte Population anhand von Stichprobenkennwerten. Dadurch möchte man z.B. feststellen, ob 2 Stichproben aus einer oder aus verschiedenen Populationen stammen.

Schätzung von Populationskennwerten

- Stichprobenkennwerte als Schätzmaße für den Populationsmittelwert
- Punktschätzung: nur ein Wert wird angegeben
- Intervallschätzung: ein Konfidenz- bzw. Vertrauensintervall wird angegeben

Ausgangspunkt: Mittelwert der Gesamtpopulation ist unbekannt. Entnimmt man der Gesamtpopulation gleichgroße Stichproben mit dem Umfang N, so sind die Mittelwerte dieser Stichproben $xbar_i$ wiederum normalverteilt. Je größer der Umfang der Stichproben, umso besser schätzt der Mittelwert der Stichprobenmittelwerte den wahren Populations-mittelwert (**Zentraler Grenzwertsatz**).
Das **Konfidenzintervall** bestimmt die Grenzen, in denen der Populationsmittelwert mit 95% Wahrscheinlichkeit liegt.
Die Berechnung des Konfidenzintervalls hängt vom *Standardfehler des Mittelwerts* ab.

$$s_{xbar}=s_x/sqrt(N)$$

d.h. der Standardfehler wird kleiner, je größer die Stichprobe ist.

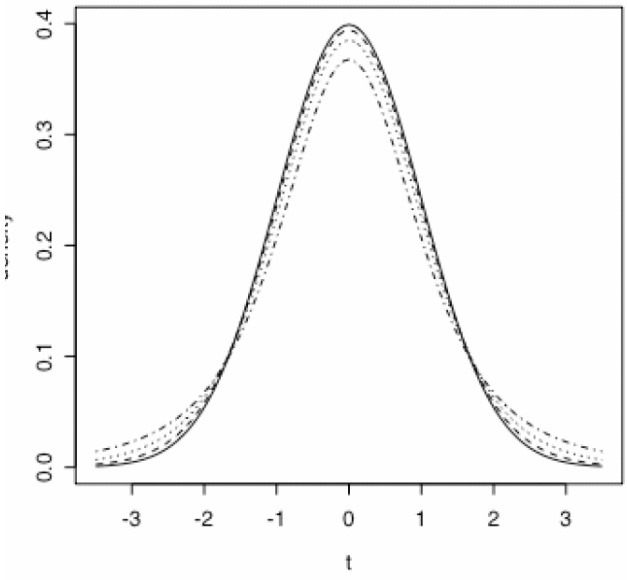
Für Stichproben mit N<30 bzw. wenn der Populationsmittelwert nicht bekannt ist, wird zur Berechnung der Wahrscheinlichkeiten nicht die Normalverteilung sondern die t-Verteilung verwendet.
Konfidenzintervall:

$$xbar-t_{95\%,df}*s_{xbar} \leq xbar \leq xbar +t_{95\%,df}*s_{xbar}$$

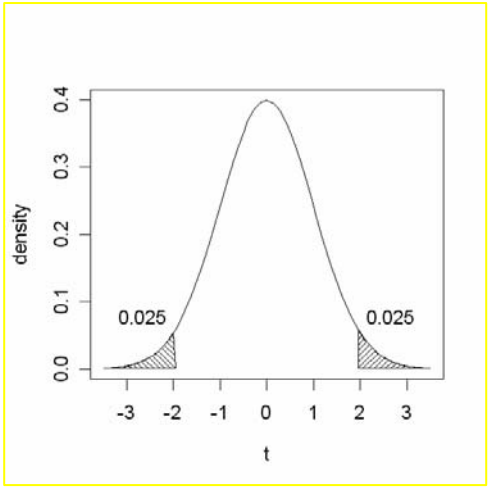
Bei einer zufälligen Stichprobe beträgt also die Wahrscheinlichkeit, dass der Mittelwert zwischen diesen beiden Grenzen liegt 95%.

t-Verteilung:

- $t=(xbar- \mu_0)/SE$
- trägt der Tatsache Rechnung, dass die Mittelwerte von zufällig entnommenen Stichproben erst bei einem großen N normalverteilt sind.
- Für ein großes N entspricht $t_{95\%}$ 1.96, d.h. mit einer 95% Sicherheit weicht der Mittelwert der Stichprobe nicht stärker als 2 (exakt 1.96) Standard-fehler vom Populationsmittelwert ab.
- Der t-Wert hängt von den Freiheits-graden ab. Freiheitsgrade (df) = N-1
- Je kleiner die Anzahl der Messwerte, desto größer wird der t-Wert;
  z.B. df=3: t=4.54; df=10: t=2.23; df=∞: t=1.96
- Die t-Verteilung wird deshalb auch als *konservativer* als die Normalverteilung bezeichnet, da bei kleinem N kleinere Standardabweichungen nötig sind.



Normalverteilung (durchgezogene Linie) im Vergleich zur t-Verteilung mit n= 3 (strich-punkt), n=7 (gepunktet) und n=21 (gestrichelt)



Beispiel:

- xbar=50, sd=5, N=25
- SE=5/sqrt(25)
- $t_{95\%,24}$=2.063 (aus Tabelle, oder mit R qt(0.025, 24))
- Konfidenzintervall: $47.93 \leq 50 \leq 52.06$

➔ Liegt der Mittelwert einer weiteren Stichprobe oberhalb oder unterhalb dieser Grenzen, so beträgt die Wahrscheinlichkeit, dass die neue Stichprobe aus der gleichen Population stammt, weniger als 5%.

Die Einheit der t Werte ist Standardfehlern vom Mittelwert.
Bei bekanntem t Wert kann die Wahrscheinlichkeit berechnet werden, ob sich eine Stichprobe von einem angenommenen Wert unterscheidet.

Je kleiner das Konfidenzintervall ist, desto exakter ist unsere Schätzung des Populationsmittelwertes.
Da das Konfidenzintervall vom Standardfehler abhängt, wird unsere Schätzung umso genauer, je größer unsere Stichprobe und je kleiner die Standardabweichung ist.

Hypothesen

Nullhypothese: es existiert kein Unterschied zwischen zwei Mittelwerten (z.B. Stichprobenmittelwert und Populationsmittelwert, oder Mittelwert und einem angenommenen Mittelwert, oder zwischen 2 Stichprobenmittelwerten) H_0: $\mu_1=\mu_2$

Alternativhypothese: Mittelwerte unterscheiden sich.
ungerichtete Alternativhypothese: es gibt einen Unterschied H_1: $\mu_1 \neq \mu_2$
gerichtete Alternativhypothese gibt eine Richtung an (< oder >) H_1: $\mu_1 < \mu_2$

R Befehle:

dnorm()	Normalverteilung
dt()	Dichtefunktion der t-Verteilung
pt()	kumulative t-Verteilung
qt()	t-Wert für eine bestimmte Wahrscheinlichkeit
t.test()	

Aufgabe 8 (mit Musterlösung):

1. Bestimme Konfidenzintervalle für die Variablen
   - F1 bei normalem Sprechen,
   - F1 bei lautem Sprechen,
   - rms bei leisem Sprechen von Versuchsperson DP
2. Aus der Literatur wissen wir, dass für den Vokal /e/ im Deutschen mit einem Wert für F1 von ca. 320 Hz produziert wird.
   - Formuliere die entsprechenden Hypothesen
   - Untersuche getrennt für lautes, normales und leisen Sprechen, ob die Nullhypothese angenommen oder abgelehnt werden muss. Verwende hierfür t.test(...., mu=320) und ein 95% Konfidenzintervall

Die Datenbasis bildet die Tabelle formants

Kurze Wiederholung:

Wozu brauchen wir eine Normalverteilung?

_____

_____

Wie unterscheidet sich die t-Verteilung von der Standardnormalverteilung?

_____

_____

α-Niveau

- Die Nullhypothese wird verworfen, wenn der empirisch ermittelte Kennwert außerhalb des Konfidenzintervalls liegt.
- Abhängig von $t_{\alpha,df}$
- Je kleiner α ist, desto größer muss der Mittelwertunterschied sein, um *signifikant* zu sein.
- Signifikant ≈ statisch relevant
- α-Niveau legt die Wahrscheinlichkeit fest, mit der die Nullhypothese abgelehnt wurde.
- Irrtumswahrscheinlichkeit bzw. Restrisiko für eine Fehlentscheidung gegen eine gültige Nullhypothese
- Umgangssprachlich ausgedrückt: wir haben blöderweise eine Stichprobe gezogen, die an den seitlichen Rändern der theoretischen Verteilungskurve aller Stichprobenmittelwerte liegt.
- α-Fehler, Fehler erster Art, *Type I error*

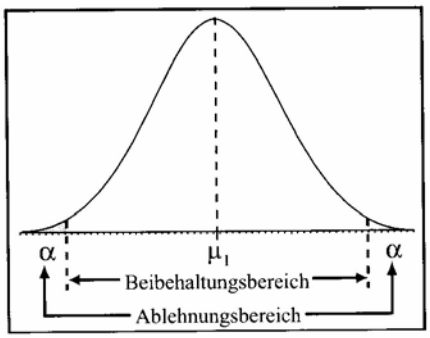


Abbildung 7.5: Das α-Niveau bei zweiseitiger Testung

Prüfung der Signifikanz hängt vom α-Niveau ab:

α=0.1	marginal signifikant	.
α=0.05	signifikant	*
α=0.01	hoch signifikant	**
α=0.001	höchst signifikant	***

Sind die Konsequenzen einer fälschlichen Ablehnung der Nullhypothese sehr gravierend, so setzt man das α-Niveau auf einen kleineren Wert (1% oder 1 Promille).

Testen von Hypothesen: zwei Mittelwerte, x_1 und x_2, sollen miteinander verglichen werden. Wir wollen feststellen, ob sie aus der gleichen Population stammen (= ____-Hypothese) oder aus verschiedenen (=____ - Hypothese). Bei einem α-Niveau von 5 % ist die Wahrscheinlichkeit, dass x_1 und x_2, wenn sie außerhalb des Beibehaltungsbereichs liegen, trotzdem aus der gleichen Population stammen, gleich 5%.

Bei einem **beidseitigen Test** entsprechen die beiden Ränder jeweils α/2. Der Beibehaltungsbereich ist 1-α.

FRAGE: Wie groß sind die α-Bereiche bei einem beidseitigen Test mit einem Beibehaltungsbereich von

- 95%
- 99%
- 99.9%

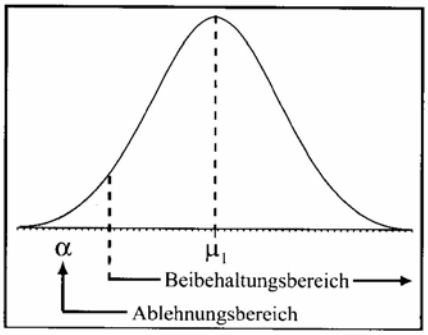


Abbildung 7.4: Das α-Niveau bei einseitiger Testung

Bei einem **einseitigen Test** wissen wir aus der Literatur, dass einer der beiden Mittelwerte größer (kleiner) sein sollte als der andere, d.h. wir nehmen eine Richtung an.
Vorteil: der t-Test wird schon bei einem geringeren Mittelwertsunterschied signifikant.

β-Fehler

= Beibehaltung der falschen Nullhypothese bei gültiger Alternativhypothese
= Fehler zweiter Art, Type II error

Fehler 1. Art: Ablehnung einer gültigen Nullhypothese
Fehler 2. Art: Beibehaltung der falschen Nullhypothese

		Realität	
		H0 ist wahr	H0 ist falsch
Entscheidung	akzeptiere H0	korrekt (es brennt nicht, kein Alarm)	Fehler 2. Art (es brennt, aber kein Alarm)
	lehne H0 ab	Fehler 1. Art (es brennt nicht, aber Alarm)	korrekt (es brennt und Alarm)

1-β: **Teststärke** (*test power*) ist die Wahrscheinlichkeit, dass ein in der Population vorhandener Unterschied bei statistischer Testung aufgedeckt wird.

β-Fehler ist abhängig von

- α-Niveau: je höher das vorher festgelegte α-Niveau, desto kleiner wird die Wahrscheinlichkeit für einen β-Fehler
- Einseitige vs. zweiseitige Testung: höhere Wahrscheinlichkeit für einen Fehler 2. Art bei zweiseitiger Testung
- Streuung des Merkmals: je einheitlicher sich die Stichprobenteilnehmer bezüglich eines Merkmals verhalten, umso geringer die Streuung. Je kleiner die Streuung umso kleiner ist auch der Standardfehler. Je kleiner der Standardfehler umso eher erhält man ein signifikantes Ergebnis.
- Stichprobenumfang: je größer die Stichprobe, umso kleiner der Standardfehler
- Mittelwertsunterschied: je größer der Unterschied zwischen zwei Stichproben (oder Faktorstufen) umso eher ein signifikantes Ergebnis
- β ist kleiner für abhängige als für unabhängige Stichproben
- Skalenniveau: je höher das Skalenniveau, desto kleiner β

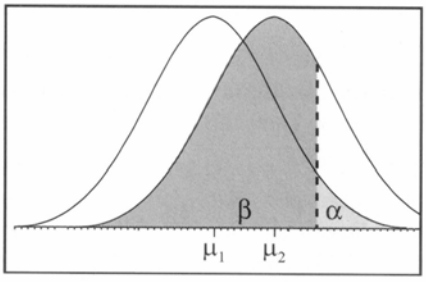


Abbildung 7.8: α- und β-Fehler, Teil 3

R Befehle

Formelschreibweise in R: *AbhängigeVar ~ Faktor*
(entspricht sprachlich: Variable wird durch Faktor beschrieben bzw. hängt von dem Faktor ab)

t.test mit paired=T (t Test für abhängige Stichproben)

power.t.test bei gegebener Teststärke, Signifikanzniveau, Mittelwertsunterschied und Standardabweichung kann so die Anzahl der Versuchspersonen bestimmt werden

> Exkurs: Teststärke
> hängt ab von der Effektgröße $d=(x_1 - x_2)/s_x$
> Daumenregel: $d \geq 0.2$ ➔ kleiner Effekt, $d \geq 0.5$ ➔ mittlerer Effekt, $d \geq 0.8$ ➔ großer Effekt
> Effekt entspricht power

barplot Balkendiagramme

Aufgabe 9 (mit Musterlösung):

1. Nenne Beispiele für gerichtete und ungerichtete Alternativhypothesen in der Phonetik
2. Die Größe des Konfidenzintervalls hängt von zwei Faktoren ab. Nenne diese und beschreibe, auf welche Weise sie den Beibehaltungsbereich beeinflussen.
3. Teste anhand des Dataframes GERVOW die folgenden Hypothesen. Formuliere dabei auch die Null- und die Alternativhypothese. Verwende ein α von 0.05. Stelle die Ergebnisse jeweils als Boxplots und als Balkendiagramme dar
   - F1 des gespannten /i/ unterscheidet sich für verschiedene Betonungsstufen.
   - F1 des betonten gespannten /u/ ist niedriger als F1 für das unbetonte gespannte /u/.
   - F1 unterscheidet sich nicht für die Sprecher RW und CG
4. Gegeben sei ein Signifikanzniveau von 0.01 und eine Standardabweichung von 50. Wie verändert sich die erforderliche Stichprobengröße für
   - kleine, mittlere und große Effekte bei einem Mittelwertsunterschied von 10
   - Mittelwertsunterschiede von 5, 20 und 30 für einen großen Effekt

Überblick über verschiedene Testverfahren

Tabelle 7.6: Übersicht zu den inferenzstatistischen Prüfverfahren der zentralen Tendenz

Anzahl der Stichproben	Skalenniveau: Nominalskala	Ordinalskala	Intervallskala
eine Stichprobe			
	Binomial-Test (9.1)		z-Test (8.2)
			t-Test für eine Stichprobe (8.3)
zwei Stichproben			
unabhängig	Vierfelder-χ^2-Test (9.2)	Mediantest (9.5)	t-Test für homogene Varianzen (8.6)
		U-Test von Mann-Whitney (9.6)	t-Test für heterogene Varianzen (8.7)
abhängig	McNemar-Test (9.3)	Vorzeichen-Test (9.7)	t-Test für abhängige Stichproben (8.4)
		Vorzeichenrangtest von Wilcoxon (9.8)	
> zwei Stichproben			
unabhängig	kxm-Felder-χ^2-Test (9.2)	H-Test von Kruskal & Wallis (9.9)	Varianzanalyse (14)
abhängig	Cochran-Test (9.4)	Friedman-Test (9.10)	Varianzanalyse für Meßwiederholung (17.1)

Anmerkung: Eine Spalte für das Verhältnisskalenniveau ist nicht dargestellt, da die vorgestellten Verfahren höchsten Intervallskalenniveau voraussetzen.

parametrische Verfahren:
- setzen voraus, dass die Daten einer theoretischen Verteilung folgen (meist Normalverteilung)
- Daten müssen intervallskaliert sein.
- vgl. Tabelle letzte Spalte

non-parametrische Verfahren:

- keine Voraussetzungen über die Verteilung ➔ verteilungsfreie Testverfahren
- Daten müssen nicht intervallskaliert sein

Weitere Kriterien für die Auswahl eines Testverfahrens:
- Stichprobengröße
- Stichprobenanzahl
- Varianzhomogenität
- abhängige oder unabhängige Stichproben

Schritt 1: wie sind die Daten skaliert
➔ intervallskaliert
Schritt 2: handelt es sich um abhängige Stichproben
(bei Messwiederholungen)
Beispiele: _____
JA➔ t-Test für abhängige Stichproben

Beispiele in R mit t.test.
Beispieldatensatz: Laryng.Rdata

Definition: Beim t-Test für abhängige Stichproben wird nur mit den Differenzen der Messwertpaare gerechnet. Es gilt:

$$t_{N-1} = \frac{\bar{x}_D}{\frac{s_D}{\sqrt{N}}} \quad (8.17)$$

mit

$$x_{D_i} = x_{1i} - x_{2i} \quad (8.18)$$

$$\bar{x}_D = \bar{x}_1 - \bar{x}_2 \quad (8.19)$$

$$s_{\bar{x}_D} = \frac{s_D}{\sqrt{N}} \quad (8.20)$$

$$s_D = \sqrt{\frac{\sum_{i=1}^{N}(x_{D_i} - \bar{x}_D)^2}{N-1}} \quad (8.21)$$

Hier bezeichnet N die Anzahl der Messungen in der Stichprobe, also die Anzahl der Messwertpaare.

NEIN: t-Test für unabhängige Stichproben:

Beispiele: _____

Schritt 3: Sind die Varianzen der beiden Stichproben homogen? TEST

(nur wichtig bei unabhängigen Stichproben)
Varianzhomogenität = die Varianzen zweier Stichproben ist gleich
Test: F-Test nach Fisher

$$F = \sigma_1^2 / \sigma_2^2$$
$$df_1 = n_1 - 1$$
$$df_2 = n_2 - 2$$

H0: Varianzen sind identisch = Varianzhomogenität
Bei Ablehnung: heterogene Varianzen
Je größer der F-Wert, desto wahrscheinlicher (in Abhängigkeit von den Freiheitsgraden), dass die Varianzen sich unterscheiden.

a) Varianzen sind homogen:

Test für unabhängige Stichproben mit homogenen Varianzen

Definition: Der t-Test für unabhängige Stichproben mit homogenen Varianzen ist definiert über

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2} \cdot \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (8.37)$$

mit

$$df = n_1 + n_2 - 2. \quad (8.38)$$

b) Varianzen sind heterogen:

Test für unabhängige Stichproben mit heterogenen Varianzen

Definition: Der t-Test für unabhängige Stichproben mit heterogenen Varianzen ist definiert über

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}} \quad (8.49)$$

mit

$$df = \frac{(n_1-1)\cdot(n_2-1)}{(n_2-1)\cdot c^2 + (n_1-1)\cdot(1-c)^2} \quad (8.50)$$

und

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \quad (8.51)$$

d.h. die Freiheitsgrade werden korrigiert. Somit muss der t-Wert höher sein, um signifikant zu werden.

Schritt 1: wie sind die Daten skaliert

➔ nominalskaliert, d.h. es handelt sich um Häufigkeiten

Beispiele: _____

χ^2-Test

Die Nullhypothese beim χ^2-Test ist, dass die beobachteten Häufigkeiten den erwarteten entsprechen.

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

R Befehle

var.test	F-Test nach Fisher für zwei Stichproben
bartlett.test	Varianzhomogenität für mehr als zwei Stichproben
shapiro.test	Test auf Normalverteilung (umstritten)
t.test	var.equal=T ➔ homogene Varianz var.equal=F ➔ heterogene Varianz
pairwise.t.test	t-Test für mehr als zwei Stichproben
chisq.test	
read.table	Einlesen von Textdateien
factor	Umwandeln einer Variablen in einen Faktor (am besten nachdem Untergruppen ausgewählt wurden)

Aufgabe 10 (mit Musterlösung):

Die Daten aus der Datei **cherokee.dat** stammen von einer Erhebung zu Plosiven, gesprochen von den Cherokee-Indianern. Die Aspirationsdauern wurden über mehrere Jahrzehnte hinweg aufgenommen.

Zeige, ob sich die VOT für die stimmlosen Plosive zwischen 1971 und 2001 verändert haben. Stelle die entsprechenden Hypothesen auf. Verwende für die Darstellung der Ergebnisse folgende Befehle und begründe Deine Antwort mit statistischen Tests:

tapply
barplot
boxplot
var
t.test
und ggfs. var.test

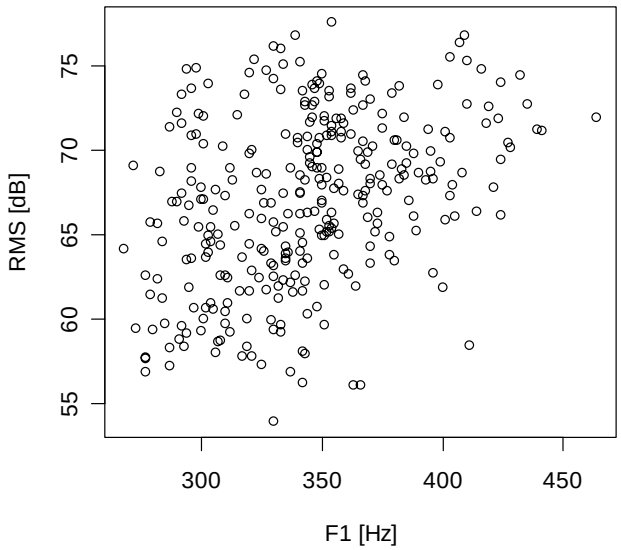
Zusätzliche Fragen:

- Sind die Daten normalverteilt?
- Welchen Test können wir verwenden, wenn wir zusätzlich auch die Messungen der stimmlosen Plosive aus dem Jahre 1961 verwenden möchten?

Korrelation und Regression

Zusammenhang zwischen mehreren kontinuierlichen Variablen
Beispiele:

1. Körpergröße – Körpergewicht,
2. Zungenhöhe und Gaumenkontakt,
3. Frequenzwerte von F1 und f0



1. Kovarianz

= Maß für den linearen Zusammenhang zwischen zwei Variablen X und Y *(covariance)*
cov_{xy}=

$$\frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Interpretation:

- Zähler: Summe der Abweichungen vom Mittelwert für Variable X multipliziert mit den Abweichungen für Y.
- Fall 1: Beide Messwerte weichen stark positiv oder stark negativ von ihren Mittelwerten ab ➔ Produkt positiv
- Fall 2: Ein Messwert weicht negativ und der andere positiv von den Mittelwerten ab ➔ Produkt ist negativ
- Verhalten sich nun alle Differenzen einheitlich, d.h. sind sie für einen vorgegebenen Zusammenhang alle positiv oder negativ, so wird der Betrag der Summe größer als bei ständig wechselnden Vorzeichen.
- Nenner: Division durch die Anzahl der Messwerte ist notwendig, da die Summe automatisch größer wird, je mehr Messwerte in die Berechnung eingehen. Für Stichproben wird durch (n – 1) geteilt.

2. Korrelation

Die Kovarianz ist stark vom Maßstab der Daten abhängig ➔ Standardisierung durch Division durch die Standardabweichungen auf einen Wert zwischen -1 und 1

$$\frac{\sum_{i=0}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n} = \frac{\sum_{i=0}^{n}(z_x)(z_y)}{n} = r_{xy}$$

(wiederum geteilt durch n-1 für Stichproben)
= **Pearsons Produkt-Moment Korrelation**
Voraussetzungen für Produkt-Moment Korrelation:

1. Beide Variablen müssen intervallskaliert sein.
2. Beide Variablen müssen normalverteilt sein.
3. Der Zusammenhang zwischen beiden Variablen sollte linear sein.
4. Homoskedastizität: für alle Individuen *i*, die ein gleiches x_i haben, müssen die entsprechenden y_i normalverteilt sein.

Andere Korrelationskoeffizienten (vgl. Tabelle 11.1 in Leonhart):

1) **Spearmans Rangkorrelation**: Alle Beobachtungen werden durch ihren Rang ersetzt.
   Verwendung: wenn keine Normalverteilung vorliegt, bei kleinem N,
   Voraussetzung: die Daten müssen mindestens ordinalskaliert sein.
   Nachteil: sehr anfällig für Ausreißer
2) **Kendalls τ**: die übereinstimmenden und nicht-übereinstimmenden Paare werden gezählt. Übereinstimmendes Paar = Differenz der x Koordinate hat gleiches Vorzeichen wie die Differenz der y Koordinate.
   Verwendung: bei ordinalskalierten Daten mit vielen Ausreißern
   Nachteil: Da dabei alle Ränge miteinander vergleichen werden, ist der Rechenaufwand sehr hoch.

Interpretation

r=-1	negative Winkelhalbierende, perfekter Zusammenhang, kommt in der Realität nicht vor
-1<r<0	**negative Korrelation** negativ gerichtete Punktwolke, „je größer x umso kleiner y“ z.B. Intelligenzquotient und Lösungszeit für eine Aufgabe, f0 und F1 bei Vokalen
r=0	ca. kreisförmige Punktwolke ➔ es existiert kein Zusammenhang zwischen den Variablen
0<r<1	**positive Korrelation** positiv gerichtete Punktwolke, „je größer y umso größer x“ z.B. Körpergröße und Körpergewicht, Körpergröße und f0
r=1	positive Winkelhalbierende, kommt in der Realität nicht vor.

JE NÄHER DER KORRELATIONSKOEFFIZIENT BEI 1 ODER -1 LIEGT UND JE SCHMALER DIE PUNKTWOLKE, DESTO DEUTLICHER ODER AUSGEPRÄGTER IST DER ZUSAMMENHANG ZWISCHEN ZWEI VARIABLEN

Vorsicht vor Fehlinterpretationen: auch aus einer hohen Korrelation darf kein kausaler Zusammenhang geschlossen werden!!!

Signifikanzprüfung

Nullhypothese: Der Stichprobenkorrelationskoeffizient unterscheidet sich nicht signifikant von 0 (=Populationskorrelationskoeffizient).
Alternativhypothese: $r \neq 0$
Zur Überprüfung wird wiederum ein t-Test verwendet.

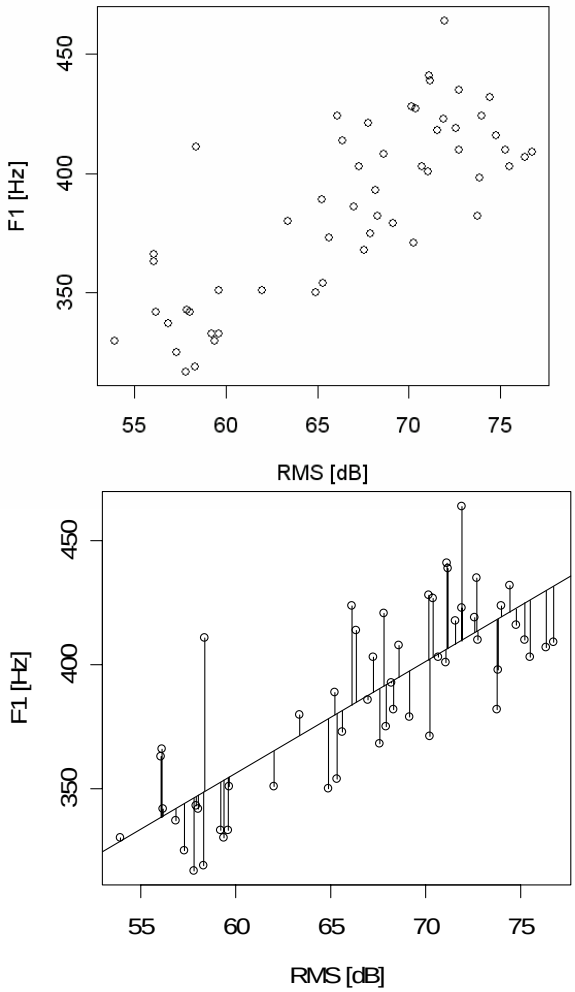
Vorsicht: Bei großen Stichproben werden auch niedrige Korrelationskoeffizienten leicht signifikant

R

Beispiele mit den Variablen f1, f2, rms und f0 aus der Tabelle *formants* (nur aus den lexikalisch und satzbetonten Daten, d.h. substring(lab,3,5)==“F.s“
Welche Variable korreliert stark, welche schwach? Welche signifikant?

3. Lineare Regression

- Vorhersage von Werten einer Variablen (=Kriteriumsvariablen y) durch eine Prädiktorvariablen (x)
- Zusammenhang wird nicht beschrieben (wie bei der Korrelation) sondern modelliert.
- Dabei wird ein linearer Zusammenhang angenommen, d.h. die Punktwolke soll durch eine **Regressionsgerade** modelliert werden.
  $\hat{y}_i = b_{y,x}\, x_i + a_{y,x}$,
  wobei $b_{y,x}$ die Steigung der Regressionsgeraden (Regressionskoeffizient β, *slope*), $a_{y,x}$ das *Intercept* (additive Konstante)
- Der Fehler der Vorhersage sollte dabei minimiert werden = Methode der kleinsten Quadrate (*method of least squares*), d.h.
  $\Sigma(y_i-\hat{y}_i)^2$=minimal
  Die Regressionsgerade ist also diejenige Gerade, bei der die Summe der quadrierten Vorhersagefehler minimal ist.



R

Ausgabe R von summary(lm(f1 ~ rms)):

```
Call:
lm(formula = f1 ~ rms)

Residuals:
    Min      1Q  Median      3Q     Max
-36.239 -19.542  -3.628  13.113  62.028

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.488     32.536   2.627   0.0113 *
rms            4.512      0.486   9.284 1.27e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.03 on 52 degrees of freedom
Multiple R-Squared: 0.6237,     Adjusted R-squared: 0.6165
F-statistic: 86.19 on 1 and 52 DF,  p-value: 1.268e-12
```

`Residuals`: entsprechen den senkrechten Linien in der Abbildung (je größer umso schlechter die Anpassung)
`Coefficients`:
`Intercept`: Schnittpunkt mit y-Achse, ist signifikant unterschiedlich von 0
`rms`: Steigung der Geraden, ist signifikant unterschiedlich von 0
`Residual standard error`: Standardfehler der Residuen, Maß für die Güte der Anpassung

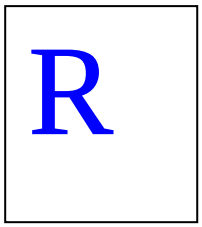
`Multiple R-Squared`: erklärte Varianz (*variance accounted for*) η^2 (Eta square). Beschrieben wird mit diesem Wert der Anteil der Varianz, der durch den Regressionszusammenhang erklärt wird.

R Befehle

plot(x,y) Scatterplots, d.h. einzelne Datenpunkte werden zweidimensional dargestellt. Marker können mit pch geändert werden (siehe help(points)

```
plot(x,y, pch=as.numeric(factor))
legend(55,450, legend=c("bd", "dp", "ga", "nu", "sb", "sz"), pch=1:6)
```

cov(x,y)
cor(x,y)
cor.test(x,y)
segment()
abline()
lm()
summary(lm())



Aufgabe 11 (mit Musterlösung 1):

Es soll untersucht werden, ob der Zusammenhang zwischen der RMS und den Frequenzen der ersten beiden Formanten signifikant ist. Als Datengrundlage dient der Dataframe *formants*

- Berechne die Korrelationskoeffizienten und teste, ob sie sich signifikant von 0 unterscheiden.
- Können die Frequenzen von F1 und F2 durch den RMS-Wert vorhergesagt werden? Berechne zur Beantwortung dieser Frage Lineare Modelle und stelle sie grafisch dar.

Die Berechnungen sollen für alle Sprecher und einzeln für die Sprecher NU und SZ durchgeführt werden.

Aufgabe 12 (mit Musterlösung 2):

Die Datei vowel_epg.txt enthält artikulographische, elektropalatographische und akustische Daten zu Vokalen, gesprochen von 3 männlichen Sprechern der Deutschen.
Lies die Daten mit dem Befehl *variablenname*=read.table(*dateiname*, header=T) ein.
Untersuche für jeden Sprecher, ob es einen Zusammenhang gibt zwischen

- F1 und JAWY (=Unterkieferhöhe)
- F2 und BACKX (=horizontal Zungenrückenposition, höher für weiter hinten)
- F0 und JAWY
- BACKY (Zungenrückenhöhe) und APPC (Kontaktfläche am Gaumen in Prozent)

Erstelle für die Ergebnisse (Korrelationskoeffizienten, Slope, Intercept, erklärte Varianz sowie deren Signifikanzen) eine Tabelle
Wie lassen sich die signifikanten Ergebnisse phonetisch interpretieren?

Einfaktorielle Varianzanalyse mit festen Effekten

1. WARUM?

Varianzanalysen werden durchgeführt, um Hypothesen zu testen, d.h. ob sich ein oder mehrere Faktoren signifikant auf einen Messwert auswirken.
Bisher: _____
Problem I: Faktor hat mehr als 2 Stufen
Bisher: _____
Folgeproblem: Inflation des α Fehlers
= je mehr Paarvergleiche umso höher wird die Wahrscheinlichkeit einen α Fehler zu begehen und zwar exponential mit der Anzahl der Vergleich m

$$p(\text{Fehler}) = 1 - (1- \alpha)^m$$

z.B. für Faktor Konsonant aus der Klausur:
Anzahl der Paarvergleich: 15 R Befehl choose(6, 2)
$p= 1-(1-0.05)^{15}= 0.54$
d.h. die Wahrscheinlich bei 15 Paarvergleichen einen α Fehler zu begehen ist 54%

Lösung 1: **Bonferroni**-Korrektur = das α-Niveau für jeden Einzeltest wird soweit herabgesetzt, dass das Gesamtniveau nur noch 0.05 beträgt (α/Anzahl der Tests).
Lösung 2: Mehrstufige Faktoren können getestet werden ohne Korrektur mittels der Varianzanalyse (=**An**alysis **o**f **Va**riance = **ANOVA**)
Problem II: mehrere Faktoren könnten sich auf eine abhängige Variable auswirken (z.B. Geschlecht und Akzent auf Grundfrequenz).

2. VORAUSSETZUNGEN

1) Mindestens Intervallskalenniveau und Normalverteilung innerhalb der Stichprobe bei der abhängigen Variablen
2) Mindestens 20 Elemente pro Stichprobe (Gruppe, Zelle)
3) Ähnlich stark besetzte Gruppen (gleiches N)
4) Varianzhomogenität der abhängigen Variablen zwischen den einzelnen Stichproben (s. Bartlett-Test)

3. BERECHNUNG

Beispiel Kieferdaten während des Konsonanten von Sprecher AW mit k=9 Messungen pro Faktorstufe, p=6 Faktorstufen und insgesamt N = 9*6 = 54 Messungen .

Spalten		**i Faktorstufen**					
Zeilen	JC AW	S	$	T	D	N	L
j Messwerte	1	1.259	1.318	1.665	1.081	0.283	-1.133
	2	1.339	1.206	1.735	0.804	-0.351	-1.533
	3	1.291	0.909	1.569	0.374	-0.498	-1.846
	4	1.174	1.040	1.342	0.298	-0.066	-1.287
	5	1.178	1.004	1.127	0.274	-0.840	-1.284
	6	1.128	1.052	1.381	0.261	-0.335	-1.730
	7	1.298	1.129	1.469	0.963	0.154	-1.206
	8	1.325	0.827	1.495	0.404	-0.444	-1.900 → Messwert x_{ij}
	9	1.260	1.041	1.530	0.428	-1.500	-1.375

= Vergleich der Varianzen der einzelnen Faktorstufen mit der Gesamtvarianz. Ist die Varianz der einzelnen Faktorstufen wesentlich größer als die zufällige Gesamtvarianz in den Daten, dann hat der Faktor einen signifikanten Einfluss.

Die Gesamtvarianz lässt sich demnach zerlegen in folgende Quadratsummen (**SS**=**S**um of **S**quares)

$$SS_{total} = SS_{treatment} + SS_{error}$$

$SS_{treatment}$= Varianz, die sich aus den Faktorstufen ergibt (auch $SS_{between}$)
SS_{error} = Varianz, die sich aus mehreren Messungen ergeben (auch SS_{within})

Wichtig: Zusammenhang Quadratsummen und Varianz: SS=var*df

Schritte:

1. Berechnung Faktorstufenvarianzen SS $_{treatment}$

Summe der Abweichungsquadrate innerhalb der Faktorstufen = $SS_{treatment}$
(fettgedrucktes x bedeutet im Folgenden Mittelwert, i bezieht sich auf die Faktorstufen und j auf die Messungen).

	S	$	T	D	N	L	**x**..
$\mathbf{x}_{i.}$	1.25	1.06	1.48	0.54	-0.40	-1.48	0.41
$\mathbf{x}_{i.}$-**x**..	0.84	0.65	1.07	0.13	-0.81	-1.89	
$(\mathbf{x}_{i.}$-**x**..$)^2$	0.71	0.42	1.14	0.02	0.65	3.6	Sum 6.5

$$SS_{treatment} = r\sum(\bar{x}_{i.} - \bar{x}_{..})^2 \quad \text{Sum of squares - treatment}$$

$$MS_{treatment} = \frac{SS_{treatment}}{df_{treatment}} \quad \text{Mean Square (variance) - treatment}$$

$SS_{treatment}$ = Sum * Anz.Messwerte pro Faktorstufe= 6.5*9 = 58.5
$MS_{treatment} = SS_{treatment}/df_{treatment}$ = 58.5 / 5
(r=Anzahl der Messwerte pro Faktorstufe = 9)

2. Berechnung Fehlervarianz SS $_{error}$

= Varianz, die durch die Abweichungen vom Faktormittelwert bei z.B. mehrfachen Wiederholungen entstehen („weil die Versuchsperson nicht exakt immer das Gleiche gemacht hat“).

$$SS_{error} = \sum(x_{ij} - \bar{x}_{i.})^2$$

SS_{error} = sum(tapply(pos_awJC, pos_awcons, var))*8 = 4.25
$MS_{error} = SS_{error}/df$ = 4.25/48

3. Berechnung Gesamtvarianz SS $_{total}$

$$SS_{tot} = \sum(x_{ij} - \bar{x}_{..})^2$$

SS_{tot} = var(pos_aw$JC) * (9*6-1) = 62.79

4. Berechnung F Wert

Zugrundeliegende Modellgleichung

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Jeder Messwert x_{ij} setzt sich zusammen aus dem Gesamtmittelwert, dem Einfluss des Faktors τ_i und zufälliger Variation ε_j, die nicht auf den Faktor zurückzuführen ist.
Ob der Faktor nun wichtiger ist als die Fehlervarianz, läßt sich durch den F-Wert schätzen

$$F = MS_{treatment}/MS_{error}$$

Berechnung der Freiheitsgrade

$df_{treatment}$ = Faktorstufen -1 p-1
df_{error} = Gesamt – Faktorstufen N – p

Nullhypothese:
H0: $\tau_s = \tau_S = \tau_t = \tau_d = \tau_l = \tau_n = 0$

4. INTERPRETATION

Ergebnis aus R mit

```
anova(lm(pos_aw$JC ~ pos_aw$cons))
```

```
Analysis of Variance Table

Response: pos_aw$JC
            Df Sum Sq Mean Sq F value    Pr(>F)
pos_aw$cons  5 58.543  11.709  132.18 < 2.2e-16 ***
Residuals   48  4.252   0.089
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

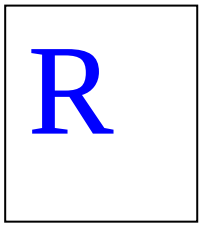
- `pos_aw$cons` entspricht Werten für *treatment*
- `Residuals` entspricht den Werten für *error*
- SS sind in Spalte `Sum Sq` und MS in Spalte `Mean Sq`
- Da $MS_{treatment}$ wesentlich größer ist als MS_{error} , wird der F-Wert ziemlich groß (132.18).
- Ein großer F-Wert ist meistens auch signifikant (siehe Tabellen in Statistikbüchern).
- Freiheitsgrade:
  - $df_{treatment}$ = Anzahl der Faktorstufen – 1 = 6-1=**5**
  - df_{error} = Gesamtanzahl – Anzahl der Faktorstufen = 54 – 6 = **48**
- Ergebnis (wie es in wissenschaftlichen Artikeln, Magisterarbeiten und Statistik-klausuren berichtet werden sollte): der Konsonant hat einen signifikanten Einfluss **(F(5,48)=132.18, p<0.001)** auf die Kieferposition während des Konsonanten.

R Befehle

pairwise.t.test(...., p.adj="bonf")

lm() linear model (da Varianzanalyse ein Spezialfall der Regressionsanalyse ist)

anova(lm())



Aufgabe 13 (Musterlösung zu zweit)

Die Varianzanalyse wird in Johnson (2007), vgl. Kapitel 4, S. 104 - 113 http://corpus.linguistics.berkeley.edu/~kjohnson/quantitative/ anhand eines Beispiels zur Reaktionszeitmessung bei Priming Experimenten erläutert. Dieses Beispiel soll mit Erläuterungen und den entsprechenden Analysen in R in einem Kurzreferat vorgeführt werden.

Aufgabe 14 (Musterlösung)

In der 1. Aufgabe der Statistikhausaufgabe vom 6. Dezember sollten die Kieferpositionen und -bewegungen für verschiedene Konsonanten für die einzelnen Sprecher miteinander verglichen werden. Vergleiche die verwendeten Tests mit den Ergebnissen der Varianzanalysen für jeden Sprecher einzeln.

Wiederholung ANOVA

1. Warum heißt dieses Verfahren Varianzanalyse?
2. Erläutere die folgenden Modellgleichungen
   a) $x_{ij}= \mu+\tau_i$
   b) $x_{ij}= \mu+\tau_i+\varepsilon_{ij}$
3. Was bedeutet: $SS_{total} = SS_{treatment} + SS_{error}$
4. Was bedeutet folgende Tabelle

```
Analysis of Variance Table

Response: form_ga$vdur
              Df Sum Sq Mean Sq F value    Pr(>F)
form_ga$prom   3 9191.7  3063.9  85.689 3.214e-14 ***
Residuals     28 1001.2    35.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Wie sollten die Ergebnisse dokumentiert werden?

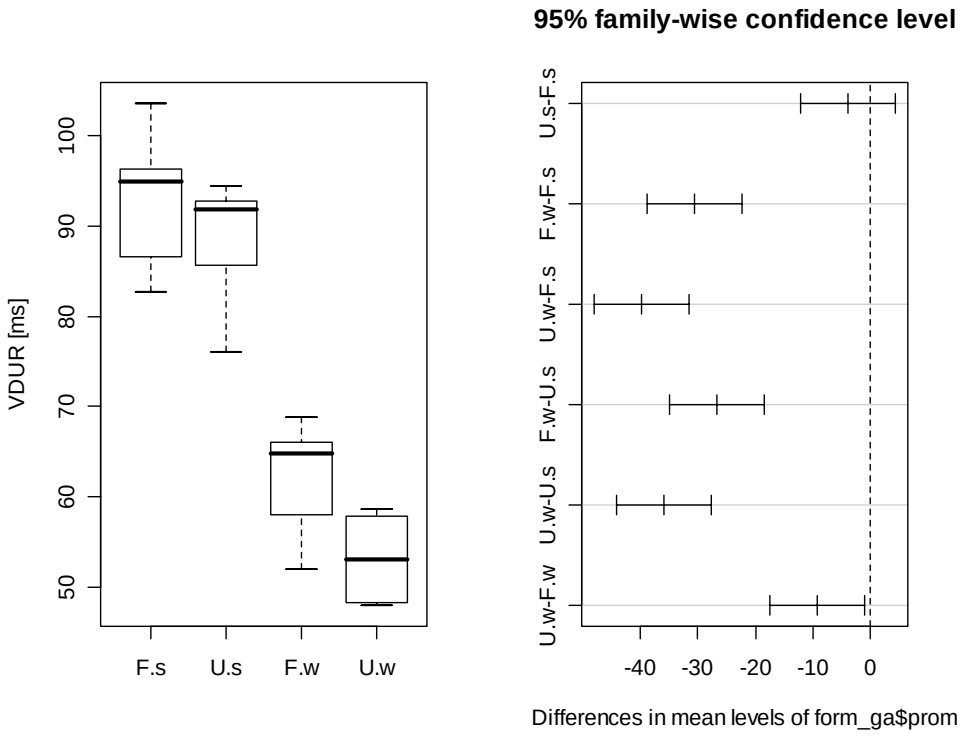
5. POST HOC TESTS

- Ziel: Welche Mittelwerte unterscheiden sich signifikant bei mehrstufigen Faktoren?
- **Nachdem** mittels einer ANOVA ein signifikanter Effekt festgestellt wurde, können so genannte *Post hoc Tests* durchgeführt werden.
- Es wird keine Unabhängigkeit der Stichproben gefordert.
- Automatische Anpassung des α-Niveaus

Tests:

- Sehr gebräuchlich: *Scheffé* Test (sehr konservativ)
- Pairwise.t.test mit Bonferroni Anpassung
- Auch in R implementiert *Tukey HSD* („honestly significant difference“)

Beispiel:



```
TukeyHSD(aov(form_ga$vdur ~ form_ga$prom))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = form_ga$vdur ~ form_ga$prom)

$"form_ga$prom"
            diff       lwr        upr
U.s-F.s  -3.9245 -12.08765   4.238651
F.w-F.s -30.5125 -38.67565 -22.349349
U.w-F.s -39.7125 -47.87565 -31.549349
F.w-U.s -26.5880 -34.75115 -18.424849
U.w-U.s -35.7880 -43.95115 -27.624849
U.w-F.w  -9.2000 -17.36315  -1.036849
```

```
par(mfcol=c(1,2))
boxplot(form_ga$vdur ~  form_ga$prom, ylab="VDUR [ms]")
plot(TukeyHSD(aov(form_ga$vdur ~ form_ga$prom)))
```

Mehrfaktorielle Varianzanalyse mit festen Effekten

- Ziel: Testen, ob mehrere Faktoren einen signifikanten Einfluss auf eine Variable haben, z.B. Geschlecht und Akzent auf f0
- Zweifaktorielles Design:

Stress	strong	weak
Accent		
Focus	F.s	F.w
Unfocussed	U.s	U.w

Modellgleichung:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_k$$

Haupteffekte:

Stress: α_1=strong
α_2=weak

Accent: β_1=focus
β_2=unfocussed

Stress		strong	weak	xbar
Accent				
Focus		F.s	F.w	
	Messung 1	103.62	66.250	77
	Messung 2	96.72	52.001	
	Messung 3	96.00	68.850	
	Messung 4	82.78	65.679	
	.k.			
Unfocussed		U.s	U.w	
	Messung 1	92.100	48.03	71
	Messung 2	94.406	49.31	
	Messung 3	86.880	57.03	
	Messung 4	91.510	58.57	
	.k.			
				μ
xbar		91	58	74

```
anova(lm(vdur ~ accent+stress, data=form_ga))
```

```
Analysis of Variance Table

Response: vdur
          Df Sum Sq Mean Sq  F value    Pr(>F)
accent     1  344.5   344.5   9.4534  0.004561 **
stress     1 8791.5  8791.5 241.2428 1.362e-15 ***
Residuals 29 1056.8    36.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- höchst signifikanter Effekt von Wortakzent ($F(1,29)= 241.24$, $p<0.001$) auf die Vokaldauer,
- hoch signifikanter Effekt von Satzakzent ($F(1,29)= 9.45$, $p<0.01$)) auf die Vokaldauer
- **beide Haupteffekte wirken sich signifikant auf die Vokaldauer aus.**

Interaktionen:

$$x_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_k$$

Interaktionen treten auf, wenn die Unterschiede zwischen den Faktorstufen eines Faktors nicht für alle Faktorstufen des zweiten Faktors gleich sind.

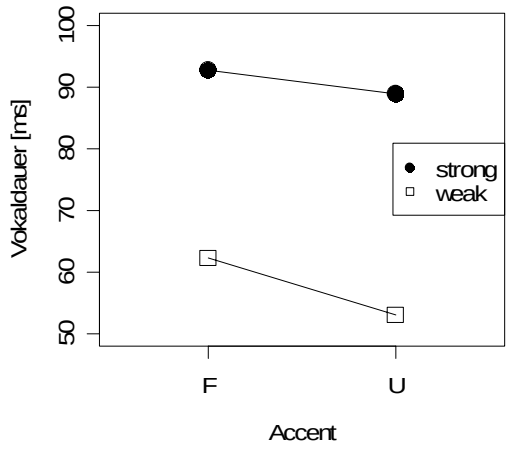
```
anova(lm(vdur ~ accent+stress+accent:stress, data=form_ga))
anova(lm(vdur ~ accent*stress, data=form_ga))          (Kurzform)
```

```
Analysis of Variance Table

Response: vdur
              Df Sum Sq Mean Sq  F value    Pr(>F)
accent         1  344.5   344.5   9.6348  0.004336 **
stress         1 8791.5  8791.5 245.8739 2.142e-15 ***
accent:stress  1   55.7    55.7   1.5567  0.222483
Residuals     28 1001.2    35.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zum Ausprobieren

```
interaction.plot(form_ga$accent, form_ga$stress, form_ga$vdur)
```



Berechnung der Freiheitsgrade:

- Haupteffekte: $p - 1=1$, $q-1=1$
- Interaktion: $(p-1)*(q-1)=1$
- Fehler: $pq(n-1) =2*2*7=28$

- Deakzentuierung hat immer eine reduzierende Wirkung auf die Vokaldauer, unabhängig vom Wortkazent.
- Wortakzentuierte Vokale (strong) sind immer länger als wortunbetonte Vokale, unabhängig vom Satzakzent.

➔ Linien verlaufen ungefähr parallel
➔ Keine Interaktionen

Signifikante Interaktionen

```
anova(lm(cdur ~ accent+stress+accent:stress, data=form_ga))
```

```
Analysis of Variance Table

Response: cdur
              Df Sum Sq Mean Sq F value    Pr(>F)
accent         1   62.7    62.7  1.0554 0.3130594
stress         1 5678.0  5678.0 95.5573 1.586e-10 ***
accent:stress  1 1017.4  1017.4 17.1231 0.0002899 ***
Residuals     28 1663.7    59.4
---
Signif.  codes:    0  '***'
0.001  '**'  0.01  '*'  0.05
'.' 0.1 ' ' 1
```

- Deakzentuierung hat eine reduzierende Wirkung auf die Konsonantdauer für Konsonanten in starken Silben und den umgekehrten Effekt in schwachen Silben, unabhängig vom Wortkazent.

➔ Signifikante Interaktion

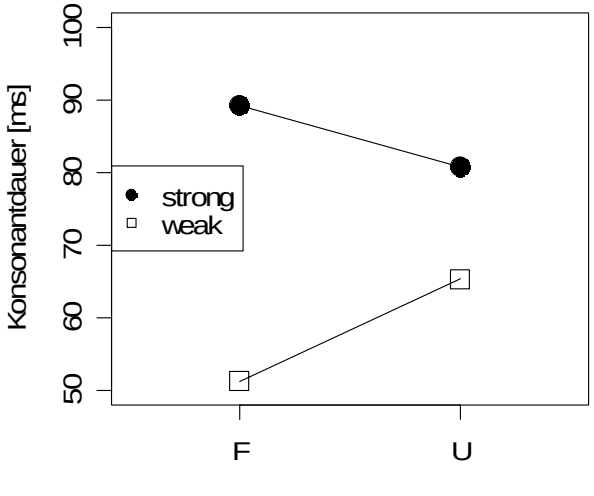
- Wortakzentuierte Konsonanten (strong) sind immer länger als wortunbetonte Konsonanten, unabhängig vom Satzakzent.

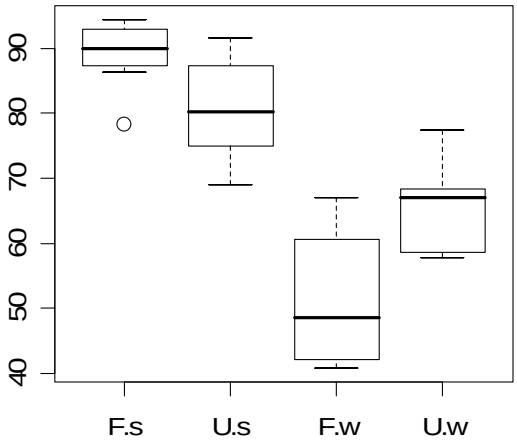
➔ Signifikanter Haupteffekt stress

- Durch die signifiante Interaktion streut die Konsonantdauer so stark, dass der Akzent keinen signifikanten Haupteffekt hat

Kennzeichen der Interaktion

➔ Linien verlaufen nicht parallel





Aufgabe 15 (mit Musterlösung):
Untersuche anhand der Datenbasis *JawPositions.Rdata* für die Sprecher AW, KH und UR mittels zweifaktorieller Varianzanalysen, ob die Kieferposition während des Konsonanten (JC) vom Artikulationsmodus (cons) und von der Lautstärke (loudness) beeinflusst wird. Nimm hierzu auch Abbildungen und Post hoc Tests zuhilfe.

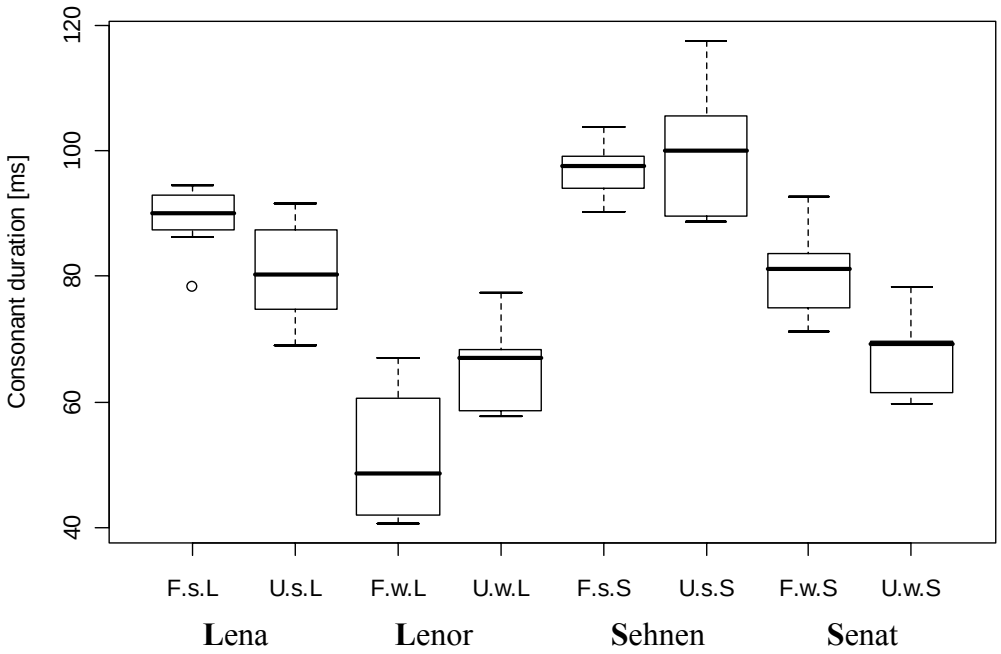
Aufgabe 16:
Interaktionen bei mehrfaktoriellen Varianzanalysen werden in Johnson (2007), vgl. Kapitel 4, S. 113 - 117
http://corpus.linguistics.berkeley.edu/~kjohnson/quantitative/
anhand eines Beispiels zur Reaktionszeitmessung bei Priming Experimenten erläutert. Dieses Beispiel soll mit Erläuterungen und den entsprechenden Analysen in R in einem Kurzreferat vorgeführt werden.

Zusammenfassung Vorgehensweise:

1. Prüfen, ob die Voraussetzungen für eine Varianzanalyse gegeben sind
2. Berechnen einer vollständigen Varianzanalyse

```
anova(lm(cdur ~ accent*stress*cons, data=formls))
```

```
Response: cdur
                  Df  Sum Sq Mean Sq  F value    Pr(>F)
accent             1    23.6    23.6   0.4242   0.51753
stress             1 10392.4 10392.4 187.0678 < 2.2e-16 ***
cons               1  3361.5  3361.5  60.5090 1.789e-10 ***
accent:stress      1    46.6    46.6   0.8381   0.36387
accent:cons        1   257.7   257.7   4.6390   0.03557 *
stress:cons        1    21.4    21.4   0.3844   0.53779
accent:stress:cons 1  1465.8  1465.8  26.3856 3.663e-06 ***
Residuals         56  3111.0    55.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Weitere Vorgehensweise je nach Ergebnis:

I. Mehrstufiger Haupteffekt ist signifikant ➔ Post hoc Test

II. Interaktion ist signifikant ➔ 2 Möglichkeiten

a) „Aufdröseln“, d.h. ANOVAs getrennt für die Faktorstufen eines Faktors, z.B. **cons**

```
anova(lm(cdur ~ accent*stress, subset=cons=="L"))
```

```
Response: cdur
             Df Sum Sq Mean Sq F value    Pr(>F)
accent        1   62.7    62.7  1.0554 0.3130594
stress        1 5678.0  5678.0 95.5573 1.586e-10 ***
accent:stress 1 1017.4  1017.4 17.1231 0.0002899 ***
Residuals    28 1663.7    59.4
```

```
anova(lm(cdur ~ accent*stress, subset=cons=="S"))
```

```
Response: cdur
             Df Sum Sq Mean Sq F value    Pr(>F)
accent        1  218.6   218.6  4.2285  0.049179 *
stress        1 4735.8  4735.8 91.6210 2.506e-10 ***
accent:stress 1  495.0   495.0  9.5756  0.004441 **
Residuals    28 1447.3    51.7
```

b) Erstellen eines mehrstufigen Faktors, z.B. promcons aus stress, accent und cons

```
promcons=paste(accent, stress, cons, sep=".")
pairwise.t.test(cdur, promcons, p.adj=”bonf”)
```

➔ Nachteil: sehr viele t-Tests, d.h. kaum noch Signifikanzen

Mehrfaktorielle Varianzanalyse mit Messwiederholungen

- ANOVA bisher: nur für eine Versuchsperson, unterschiedliche Bedingungen (Faktorstufen), mehrere Wiederholungen pro Bedingung
- Ziel: Generalisierung für die gesamte Population anhand einer Stichprobe
- Annahme bei **ANOVA mit festen Faktoren**: die Stichproben sind unabhängig voneinander. Das heißt, für jede Messung (auch Faktorstufe) wurde ein neuer Sprecher verwendet
- Messwiederholungen bedeutet, dass ein Sprecher unter verschiedenen Bedingungen (z.B. beim Sprechen in verschiedenen Lautstärken) gemessen wurde = ***within subject factor***.
  Analog zu _____
- Unterschiedliches Verhalten der einzelnen Versuchspersonen wird dabei als Zufallseffekt angesehen.
- Vergleich von Sprechergruppen (z.B. Stotterer vs. Nicht-Stotterer, bzw. Deutsche vs. Koreaner) = ***between subjects factor***
- Strukturgleichung: $y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}$ mit
  α_j fester Effekt der Faktorstufe j,
  π_i Zufallseffekt der Versuchsperson i
  ε_{ij} Fehler

Datenbasis bei einfaktoriellem Design

	Stress	Strong	Weak
Sprecher		F0	F0
BD		s	w
	Messung 1	126.507	106.072
	Messung 2	132.109	106.348
	Messung 3	147.076	112.089
	Messung 4	125.079	107.023
		…	…
	xbar	**137**	**108**
BP		s	w
	Messung 1	125.687	111.952
	Messung 2	113.411	103.334
	Messung 3	128.160	105.855
	Messung 4	119.850	107.437
		…	…
	xbar	**129**	**102**

...
Weitere Sprecher
…

- Errechnen der **Zellenmittelwerte**, da sonst jede Messung wie eine neue Versuchsperson behandelt wird
- Zellenmittelwert = Mittelwerte für jeden Sprecher und jede Bedingung über die Wiederholungen

Zellenmittelwerte in R:

```
lab=paste(formNF$vp, substring(formNF$lab,5,5),"")
f0=as.vector(tapply(formNF$f0, lab, mean))
labnew=names(tapply(formNF$f0, lab, mean))
```

Erzeugen einer Matrize mit den Faktoren Sprecher und Stress sowie der Variablen f0:

_____

_____

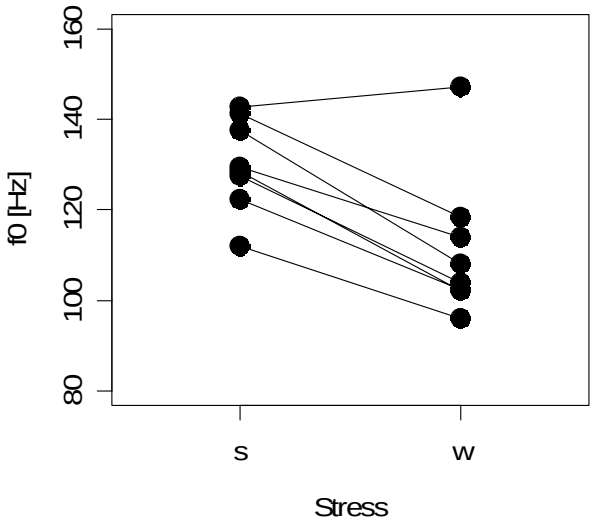
_____

Erstellen des Modells in R

Falsch:

```
summary(aov(f0~stress), data=mat))
```

```
             Df   Sum Sq  Mean Sq  F
value  Pr(>F)
stress          1  1397.28  1397.28
7.6925 0.01494 *
Residuals   14 2542.99  181.64
---
Signif. codes:  0 '***' 0.001 '**'
0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Warum falsch????

Richtig:

```
summary(aov(f0~stress                +
   Error(sp/stress), data=mat))
```

```
Error: sp
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  7 2158.38  308.34

Error: sp:stress
          Df  Sum Sq Mean Sq F value   Pr(>F)
stress     1 1397.28 1397.28  25.430 0.001492 **
Residuals  7  384.62   54.95
```

`Error(sp/stress)` bedeutet, dass Änderungen in der Grundfrequenz aufgrund des Faktors *stress* immer innerhalb des Subjekts betrachtet werden sollte.

Aufgabe 17 (mit Musterlösung) Mehrfaktorielle ANOVA
Untersuche anhand der Datenbasis *formls.Rdata* für einen Sprecher wie die Vokaldauer (*vdur*) durch die Faktoren *accent*, *stress* und *cons* beeinflusst werden. Welche Folgeanalysen zur Ermittlung der einzelnen Faktoreinflüsse müssen durchgeführt werden, wenn die experimentelle Fragestellung sich nur auf die Prosodie bezieht? Wie können die Ergebnisse interpretiert werden? Nimm hierzu auch Abbildungen und Post hoc Tests zuhilfe.

Aufgabe 18 (mit Musterlösung) ANOVA mit Messwiederholungen
Untersuche anhand der Datenbasis *formants.Rdata* für die Teilmenge der akzentuierten (*accent*=F) und wortbetonten Daten (*stress* =s), ob die Lautstärke einen signifikanten Einfluss auf die Variable *f0* hat. Erzeuge hierfür eine Matrix mit den Zellenmittelwerten. Wie können die Ergebnisse interpretiert werden? Nimm hierzu auch Abbildungen und Post hoc Tests zuhilfe.

Mehrfaktorielles Design

```
summary(aov( JC ~ cons * loudness + Error(subj/(cons*loudness)), data=pos))
```

```
Error: subj
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  4 1.62608 0.40652

Error: subj:cons
          Df Sum Sq Mean Sq F value    Pr(>F)
cons       5 44.477   8.895  26.904 3.103e-08 ***
Residuals 20  6.613   0.331
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subj:loudness
          Df  Sum Sq Mean Sq F value  Pr(>F)
loudness   1 1.04635 1.04635  7.7388 0.04972 *
Residuals  4 0.54084 0.13521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subj:cons:loudness
              Df  Sum Sq Mean Sq F value  Pr(>F)
cons:loudness  5 1.16398 0.23280  3.6258 0.01697 *
Residuals     20 1.28412 0.06421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```