**Statistik in R**

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

**Themen**

Übungen mit R

1. Allgemeines zur 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

Für Materialien gehe zu

www.ipds.uni-kiel.de/cm/statistikR

Warum Statistik?

(a) Datenreduktion auf einige relevante Kennwerte: Prozente, Mittelwert, Standard-abweichung, 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 (Zugehörigkeit ausschließlich)
  b) Quantitatives Merkmal: z.B. Körpergröße
- Variable: Merkmalsausprägung werden in Zahlen überführt
  a) diskrete Variable: z.B. Geschlecht
  b) kontinuierliche Variable: z.B. Körpergröße

## 1 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 Null-elements
Ableitbare Interpretationen: Gleichheit, Relationen, Gleichheit und Verschiedenheit von Verhältnissen

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

**Befehle in R**

Unterschied Skalar, Vektor, Matrize
- c
- seq
- rep

Strings
- paste
- substring

Indizierung in R
```
ii=studies$geschl=="w"
studies$groesse[ii]
```

Häufigkeiten
- hist

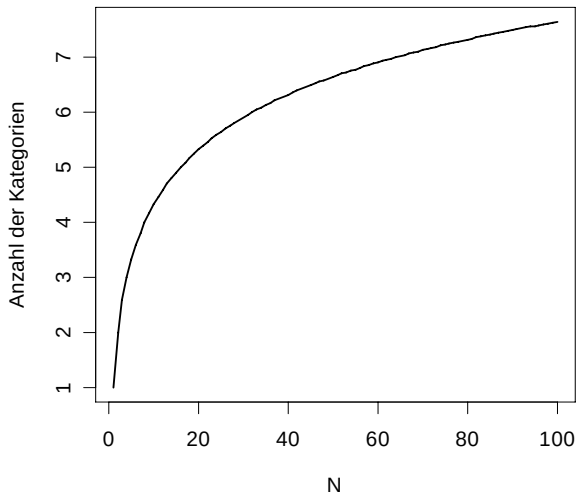
z.B. nn=hist(studies$groesse)

**Aufgabe 1:**
a) Zähle in einer Tabelle pro vorkommendem Alter die Anzahl der Studenten und berechne so die absolute bzw. relative Häufigkeit. (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:**
a) Lade die Datei *formants* in R. Es handelt sich hierbei um akustische Messungen zum vokalischen Mittelpunkt in drei verschiedenen Lautstärken (L=loud, N=normal, S=soft) gesprochen. Verwende den Befehl load
b) Stelle die Vokaldauern (vdur) graphisch dar.
c) Stelle die Vokaldauern für die einzelnen Lautstärken graphisch mit dem Befehl 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?

## 2 Häufigkeitsverteilung

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



# 3 Maße der zentralen Tendenz

## *3.1 Modus (engl. mode)*

**Def.**: Der Modalwert ist derjenige Werte 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?

## *3.2 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:

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

Für gerades N:

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

Für gruppierte Daten:

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

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?

### *3.3 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.

Lösung in R?

Gewichtete arithmetische Mittel siehe Leonhart

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:**

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.

**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.

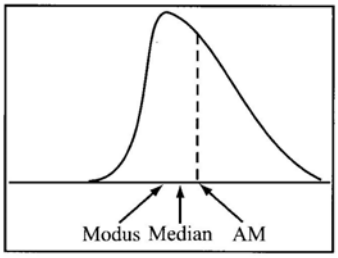



**Abbildung 3.1:** Linkssteile Verteilung

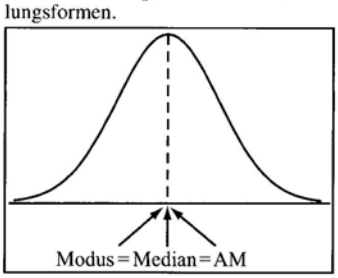



**Abbildung 3.2:** Symmetrische Verteilung

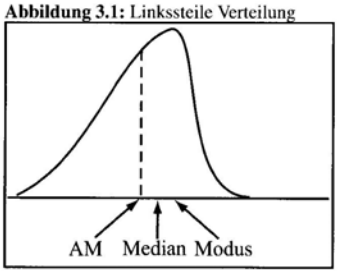



**Abbildung 3.3:** Rechtssteile Verteilung

# 4 Maße der Dispersion

## *4.1 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.

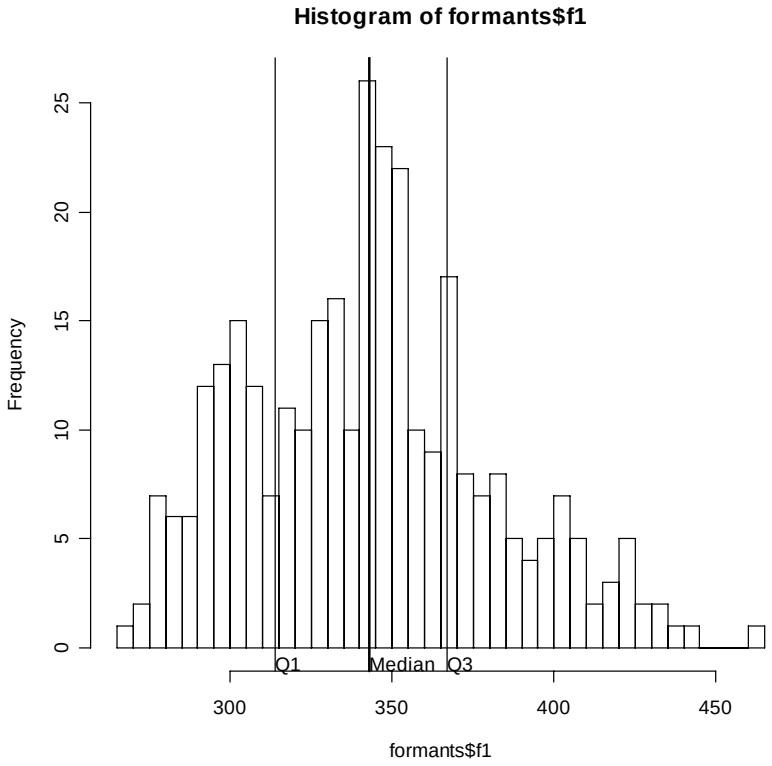
## *4.2 Quartile, Interquartilabstand (interquartile range)*

**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




### *4.3 AD-Streuung (average deviation)*

**Def.**: Durchschnitt der absoluten Abweichungen aller Messwerte vom Mittelwert

Wird kaum verwendet, da kleine Abweichungen vom Mittelwert einen ähnlichen Einfluss haben können als einige große.

### *4.4 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

### *4.5 Standardabweichung(standard deviation)*

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

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

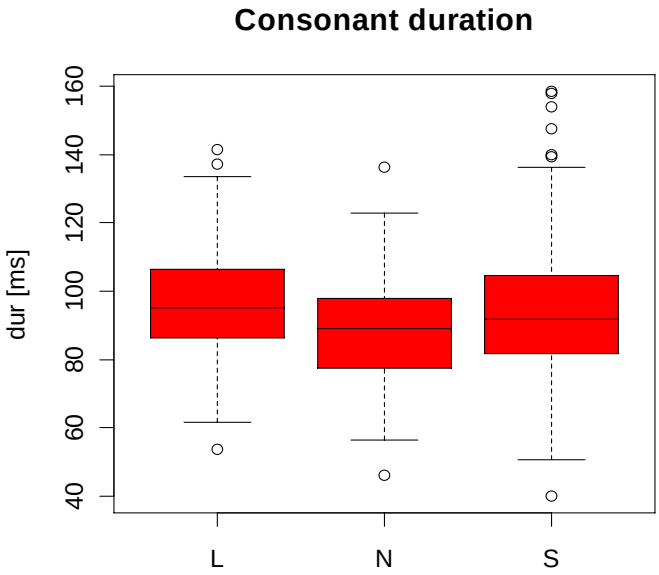
**Berechnung der Standardabweichung in der Stichprobe:**

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

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



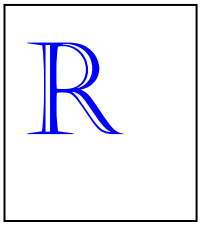
### *4.6 Variabilitätskoeffizient*

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:**
Lade die Datei formants.Rdata. Wir wollen 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.