

# Cheatsheet WS – PVA 1

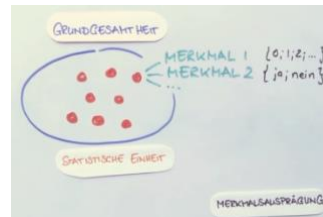
## GRUNDLAGEN

### Typologie

- **Deskriptiv:** Übersicht durch beschreibende Grafiken bestehender Daten
- **Induktiv** (Interferenzstatistik): Beobachtungen verallgemeinern, Fehler feststellen und kontrollieren
- **Explorativ:** Neues erschliessen, mit Hilfe einer Kombination aus bestehenden Daten und Algorithmen, Kombination von Verfahren aus deskriptiver und induktiver Statistik

### Statistische Grundbegriffe

- **Statistische Einheit:** was wird gemessen? Auch «Merkmalsträger» →  $e_i$ . Müssen gleichem Typ entsprechen (Äpfel, Personen, Unternehmen)
- **Grundgesamtheit:** alle gemessenen Einheiten, müssen sinnvoll festgelegt werden (sachlich, räumlich, zeitlich).  
→  $G = \{e_1, e_2, e_3, \dots, e_n\}$   
Weitere Begriffe: Teilgesamtheit, Vollerhebung, Teilerhebung
- **Merkmale:** Farben, Gewichte, Werte, Daten →  $\alpha^{(v)}$
- **Merkmalsausprägung:** Werte des Merkmals  
Wenn es sich um reelle Zahlen handelt:  
Statistische Variable:  $X^{(v)}$   
Ausprägungen:  $x_i^{(v)}$   
 $x^{(1)}$  [0, 50] in EUR/Jahr



**Geordnete Datensätze:** Sortierung mit  $\leq$ . Messwert  $x_{ij}$  der statistischen Einheit  $e_i$  an Stelle  $i$  des geordneten Datensatzes.

Position = Rang  $rg(x_{ij})$

Wenn Rangfolge unklar (hell/dunkel) muss diese explizit definiert werden.

### Skalenniveaus

Skala  $\Omega$  = Zusammenstellung aller möglichen Ausprägungen eines Merkmals als Massstab. Skalenpunkte:  $\omega$

Skalenniveaus	= / ≠	< / >	+ / -	* /	
Nominal	😊				Äpfel, Männlich, Weiblich Es kann nur unterschieden werden ob Elemente gleich oder unterschiedlich sind.
Ordinal	😊	😊			Süss, Sauer, Hell, Dunkel Elemente haben zusätzlich eine Rangordnung.
Intervall	😊	😊	😊		Datum, Temperatur Es gibt berechenbare Abstände. Skalenpunkte müssen also Zahlen sein.
Verhältnis	😊	😊	😊	😊	Grösse, Gewicht, Wert Auch Verhältnisse zwischen Skalenpunkten liefern sinnvolle Ergebnisse. Nullpunkt ist nicht willkürlich festgelegt. Massstabanpassungen sind legitim.

Kardinalskala oder metrische Skala sind Überbegriffe für Intervall- und Verhältnisskalen. → Skalenpunkte sind Zahlen.

### Gruppierte und klassierte Daten

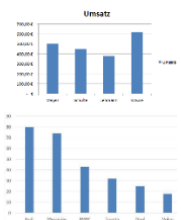
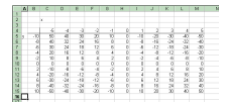
- **Diskret:** Skala oder Merkmal mit endlich vielen Skalenpunkte (keine Zwischenwerte, wie Anzahl Personen)
- **Stetig:** unendlich viele Skalenpunkte, Zwischenwerte existieren (Länge, Zeit)
- **Quasi-stetig:** individuelle Verarbeitung der Datenpunkte ist sinnlos (beispielsweise bei Jahresumsatz, Massband) → Messen in der Praxis

- **Gruppierung:** bei mehrmaligem Auftreten des gleichen Messwertes → Übersichtlichkeit
- **Klassierung:** keine Gruppierung, aber Aufteilung in Teilintervalle → Klassen (müssen nicht gleich gross sein)  
 $K_j = [x_j; x_{j+1})$   
Viele Rechnungen verwenden die Klassenmitte  $x'_j$  für Klasse  $j$  mit stellvertretenden Werten.  
Berechnung = (Klassenuntergrenze + Klassenobergrenze)/2

## UNIVARIANTE ANALYSEN

### Darstellung des kompletten Datensatzes

- **Wertetabelle**  
komplette Liste kann erstellt werden, sofern Messwerte nicht «zu gross» sind
- **Wertediagramm**  
Für metrische Skalen (selten auch Ordinalskalen) möglich
- **Geordnetes Wertediagramm**  
Für metrische Skalen, Min/Max jeweils am Rand



### Eindimensionale Häufigkeitsverteilungen

**Empirische Häufigkeitsfunktion und Stabdiagramm:** Gruppierungseffekt nutzen und Vorkommen von Merkmalen zählen.

**Absolute empirische Häufigkeitsfunktion:** Zuordnung der absoluten Häufigkeiten ( $n_k$ ) zu Messwerten ( $x_k$ ) pro Gruppe ( $k$ ):

$$f_h(x) = \begin{cases} n_k & \text{für } x = x_k \\ 0 & \text{sonst} \end{cases}$$

**Relative empirische Häufigkeitsfunktion:** Häufigkeiten  $n_k$  der Messwerte wird durch Anzahl  $n$  aller Messwerte geteilt (Resultat in Prozentzahlen):

$$f_h^*(x) = \begin{cases} \frac{n_k}{n} = n_k^* & \text{für } x = x_k \\ 0 & \text{sonst} \end{cases}$$

**Stabdiagramm:** Veranschaulichung der Häufigkeitsfunktion. Auf x-Achse werden verschiedene Ausprägungen des Merkmals dargestellt, auf y-Achse gibt man absolute oder relative Häufigkeiten an.

- Also: Wertediagramm: statistische Einheiten & Merkmalsausprägungen  
Stabdiagramm: Merkmalsausprägungen & Häufigkeiten

### Klassenhäufigkeitsfunktion und Histogramm

Bei stetigen Merkmalen (viele verschiedenen Ausprägungen) sind Histogramme besser geeignet.

Faustregel für Klassenbildung (Anzahl):  $m \approx \sqrt{n}$

Bei unterschiedlicher Breite muss Höhe proportional angepasst werden (dreimal so breit = 1/3 Höhe). Normalklassenbreite = häufigste vorkommende Klassenbreite oder grösste Klasse.

$$\text{Dehnungsfaktor } c_j = \frac{\text{Normalklassenbreite}}{\text{Breite der Klasse } K_j}$$

- Jede Klassenhäufigkeit mit jeweiligem Dehnungsfaktor multiplizieren. So erhält man die empirische Häufigkeitsdichtefunktion  
$$f_h(x) = \begin{cases} c_j * n_j & \text{für } x \in [x_j; x_{j+1}) \\ 0 & \text{sonst} \end{cases}$$

Kreisdiagramm

Wenn es nur wenige verschiedene Messwerte gibt (Gruppierungseffekt) hilft das Kreisdiagramm/Tortendiagramm, insbesondere bei nominalen Daten.  
Vorgehen:

- 1. Häufigkeit pro Messwert bestimmen & relative empirische Häufigkeiten bestimmen
- 2. Multiplikation mit 360° mit jeder relativen Häufigkeit (Grösse Kreissektor)
- 3. Messwerte mit geringem Anteil zusammenfassen → «Sonstige» (Faustregel: nicht mehr als sieben Sektoren)
- 4. Sektoren nach Grösse ordnen



Empirische Verteilungsfunktion

Anzahl Beobachtungswerte unter bestimmten Grenzwerten.

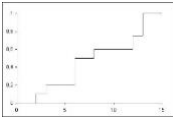
Absolute empirische Verteilungsfunktion (empirische Summenhäufigkeitsfunktion):

F\_h(x) = { 0 für x < x\_1, sum\_{x\_i <= x} f\_h(x\_i) für x in [x\_1, x\_n], 1 für x >= x\_n }

- 1. Alle Messwerte von min. nach max. ordnen
- 2. Für jeden Messwert Häufigkeit seines Vorkommens bestimmen
- 3. Durch Aufsummierung aller Häufigkeiten bis zum jeweiligen Messwert erhält man die absolute empirische Verteilungsfunktion

Anteilswerte (%) können ebenfalls gerechnet werden: absolute Summenhäufigkeiten werden durch Anzahl aller Messwerte geteilt.

Relative empirische Verteilungsfunktion: F\_h^\*(x) = { 0 für x < x\_1, (1/n) \* sum\_{x\_i <= x} f\_h(x\_i) für x in [x\_1, x\_n], 1 für x >= x\_n }



Graphische Darstellung: Treppendiagramm  
Hilfreich wenn nicht nach genauen Werten, sondern nach Bandbreite gesucht werden soll, was bei stetigen oder quasi-stetigen Merkmalen gut funktioniert und übersichtlich wird.

Kontinuität	Skalenniveau	Wertetabelle	Wertediagramm	Häufigkeitstabelle	Stabdiagramm	Klassenhäufigkeitstabelle	Histogramm	Kreisdiagramm	Verteilungstabelle	Treppendiagramm
diskret	Nominal	Wenn n nicht zu gross ist	Piktogramme	😊	😊	😊	😊	😊	😊	😊
	Ordinal			😊	😊	😊	😊	😊	😊	😊
	Metrisch (wenige verschiedene Messwerte)		Wenn n nicht zu gross ist	😊	😊	😊	😊	😊	😊	😊
stetig quasi-stetig	Metrisch			😊	😊	😊	😊	Klassierte Daten	😊	😊

Masszahlen zur Beschreibung statistischer Verteilungen

Zusammenfassung von Datensätzen zu einzelnen Werten → Masszahlen oder Parameter

Mittelwerte (Lagemasse/Lageparameter)

- Modalwert (häufigster Wert)
- Median (Mittelwert 50:50)
- Arithmetisches Mittel (Durchschnitt)
- Geometrisches Mittel (Wachstumsfaktor)
- Harmonisches Mittel (Durchschnitt über Verhältniszahlen)

Modalwert (Modus): x\_Mod

Häufigster Wert im Datensatz: f\_h(x\_Mod) ≥ f\_h(x\_i) für alle i

Also: Häufigkeit des Modalwertes ≥ absolute empirische Häufigkeit des i-ten Messwertes

Prinzipiell auf allen Skalenniveaus zulässig, aber besonders geeignet für Nominalskalen.

Median (Zentralwert): x\_Z oder x\_Med

Liegt genau in der Mitte des sortierten Datensatzes: 50% ≤ x\_Z ≤ 50%

x\_Z = { x\_{[n+1]/2} falls n ungerade ist, (x\_{[n/2]} + x\_{[n/2+1]})/2 falls n gerade ist }

Eckige Klammern → geordneter Datensatz

Bestimmung:

- 1. Datensatz nach Grösse ordnen, aufsteigend
- 2. Feststellen ob gerade oder ungerade Anzahl Werte
- 3. Dann Wert an entsprechender Stelle suchen und bei Bedarf berechnen

Der Median ist unempfindlich gegenüber Ausreissern → robuster Mittelwert

Arithmetisches Mittel (Durchschnitt): x

x = (1/n) \* sum\_{i=1}^n x\_i = (1/n) \* (x\_1 + x\_2 + ... + x\_n)

Gewichtetes arithmetisches Mittel: bei gruppierten Daten (Messwerte werden mit Häufigkeit multipliziert)

x = (1/n) \* sum\_{k=1}^r n\_k \* x\_k = sum\_{k=1}^r n\_k^\* / n \* x\_k

Beispiel: ((Menge1 \* Stückpreis1) + (Menge2 \* Stückpreis2))/Gesamtmenge = Preis/Menge

Wenn Daten nur klassiert vorliegen, kann eine Näherung mit den Klassenmitten gemacht werden.

Geometrisches Mittel (Mittelwert aufeinander aufbauender Veränderungen/Wachstum): x\_g

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 * x_2 * \dots * x_n}$$

- Wird bei **Wachstum** und **Schrumpfung** verwendet, bzw. wenn **Entwicklungsschritte aufeinander aufbauen**.  
(voneinander unabhängige Wachstumswerte werden mit gewichtetem arithmetischem Mittel gemessen)

Bei gruppierten / klassierten (nur als Annäherung) Daten würde man die Klassenmitten gewichten, indem man diese durch ihre Häufigkeit potenziert.

**Harmonisches Mittel** (Durchschnitt über Verhältniszahlen):  $\bar{x}_h$

$$\bar{x}_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Also: Anzahl Messwerte/(Summe der (Kehrwerte aus allen Messwerten))

Beispiele: Durchschnitte von km/h, Erwerbstätige pro Haushalt, etc.

Wenn Gewichte unterschiedlich sind, berechnet man das harmonische Mittel für gruppierte Daten. Liegen Daten klassiert vor, würde man als Näherung mit Klassenmitten rechnen.

## Streuparameter

Nebst Lage ein wichtiger Parameter → Schwankungen

Unterschiedliche Masszahlen:

- Spannweite** (Max – Min)
- Varianz** (häufigste Verwendung)
- Standardabweichung** (Wurzel der Varianz)
- Variationskoeffizient

**Spannweite** (Range): R

Abstand zwischen Minimum und Maximum:  $R = x_{[n]} - x_{[1]}$

- Empfindlich gegenüber Ausreissern, unempfindlich gegenüber Schwankungen

**Varianz** (häufigste Verwendung):  $s_x^2$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} * ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2)$$

Berücksichtigt alle Werte. Von jedem Messwert muss erst arithmetisches Mittel abgezogen werden. Danach jede dieser Differenzen quadrieren, am Ende alle Werte addieren.

→ Differenzen können jeweils positiv oder negativ sein, diese summieren sich alle immer exakt auf 0. Durch Quadrieren werden alle diese Werte positiv, was die 0 «umgeht», aber für grössere «Abweichungen» sorgt.

**Verschiebungssatz** (Umstellen weil praktischer): Summe der quadrierten Messwerte; durch n teilen; Quadrat des arithmetischen Mittels abziehen.

**Standardabweichung** (Ausgleich Varianz):  $s_x$

$$s_x = \sqrt{s_x^2}$$

- Korrektur des Grössenverhältnisses und der Einheit des Ergebnisses (EUR gibt's nicht im Quadrat)

**Variationskoeffizient** (Risiko):  $v_x$

$$v_x = \frac{\sqrt{s_x^2}}{\bar{x}} = \frac{s_x}{\bar{x}}$$

Durch Division durch  $\bar{x}$  bekommen wir eine Art prozentuale Streuung, was bei Risikoanalysen hilfreich sein kann.

→ laut Moodle bei intervallskalierten Daten nicht zulässig

## Quantile und Boxplots

Quantile sind eher Lageparameter, sobald mehrere berechnet werden wird aber Streuung gut visualisiert.

Ein p-Quantil ( $x_{Q[p]}$ ) ist der Wert für den gilt, dass p\*100% der Werte des Datensatzes kleiner oder gleich und gleichzeitig (1-p)\*100% der Werte des Datensatzes grösser oder gleich sind:

$$x_{Q[p]} = \frac{1}{2} (x_{[n \cdot p]} + x_{[n \cdot p + 1]}) \text{ wenn } n \cdot p \in \mathbb{N}; x_{Q[p]} = x_{[n \cdot p]} \text{ wenn } n \cdot p \text{ nicht } \in \mathbb{N}$$

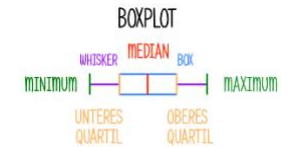
$$x_{Q[p]} = \begin{cases} \frac{1}{2} (x_{[n \cdot p]} + x_{[n \cdot p + 1]}) & \text{wenn } n \cdot p \in \mathbb{N} \\ x_{[n \cdot p]} & \text{wenn } n \cdot p \text{ nicht } \in \mathbb{N} \end{cases}$$

→ Gauss'sche Klammer (oben Ecken, unten nicht) heisst: aufrunden

Also: wenn n\*p keine natürliche Zahl ist, soll aufgerundet werden.

Vorgehen:

- Aufsteigend sortieren und prüfen ob n\*p eine natürliche Zahl ist
  - Dann entweder Werte an entsprechenden Stellen suchen und arithmetisches Mittel nehmen, oder aufrunden.
- Median = 50%-Quantil



Die Formel ist nicht ganz eindeutig. Quantile sind auch eng verwandt mit Verteilungsfunktion.

**Boxplots:** grafische Darstellung von Median, Quartilen und Minimum/Maximum

Mittelwerte	Modalwert	Median	Arithmetisches Mittel	Geometrisches Mittel	Harmonisches Mittel
Haupts. Anwendung	Bei nominal skalierten Merkmalen (einfach häufigster Wert)	Bei ordinal skalierten Daten, aber auch bei metrischen Daten (z.B. wenn Randwerte bzw. Ausreisser keinen Einfluss haben sollen)	Bei metrisch skalierten Datensätzen, wenn alle Werte gleichermassen einfließen sollen	Für Durchschnittsbildung über aufeinander aufbauenden Wachstumsraten	Für Durchschnittsbildung über Verhältniszahlen, wenn die gegebene Bezugseinheit bzw. gegebene Gewichte sich auf den Zähler beziehen
Symbol	$\bar{x}_{Mod}$	$\bar{x}_Z$	$\bar{x}$	$\bar{x}_g$	$\bar{x}_h$
Formel	$f_h(\bar{x}_{Mod}) \geq f_h(x_i)$	$\begin{cases} x_{[\frac{n+1}{2}]} & n \text{ ungerade} \\ \frac{1}{2} (x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}) & n \text{ gerade} \end{cases}$	$\frac{1}{n} \sum_{i=1}^n x_i$	$\sqrt[n]{\prod_{i=1}^n x_i}$	$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$
Nominal	☹️	☹️	☹️	☹️	☹️
Ordinal	☺️	☺️	☹️	☹️	☹️
Intervall-skaliert	Bei vielen Mehrfachnennungen	☺️	☺️	☹️	☹️
Verhältnis-skaliert		☺️	☺️	☺️	Solange kein x = 0

Streuemasse	Spannweite	Varianz	Standard-abweichung	Variations-koeffizient	Quantil(-e)
Symbol	$R$	$s_x^2$	$s_x$	$v_x$	$x_{Q[p]}$
Formel	$x_{[n]} - x_{[1]}$	$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	$s_x = \sqrt{s_x^2}$	$\frac{\sqrt{s_x^2}}{\bar{x}} = \frac{s_x}{\bar{x}}$	Wert, der die aufsteigend geordnete Beobachtungsreihe in p : 1-p teilt
Nominal	☹	☹	☹	☹	☹
Ordinal	☹	☹	☹	☹	☹
Intervall-skaliert	☺	☺	☺	☺	☺
Verhältnis-skaliert	☺	☺	☺	☺	☺

## Grundprinzipien der Kombinatorik

«Wie viele Möglichkeiten gibt es, Dinge auszuwählen und anzuordnen?»

Zwei grundlegende Dinge:

- Darf das selbe Element mehrmals gewählt werden? (→ mit Zurücklegen)
- Ist die Reihenfolge der Auswahl wichtig? (→ mit Beachtung der Reihenfolge)

Daraus ergeben sich vier Ziehvorschriften.

**Fundamentalprinzip:** Kriterium1 \* Kriterium2 = Möglichkeiten

- Zwei Donut-Teige, drei Glasuren = 2\*3 Varianten

$$T = n_1 * n_2 * \dots * n_k = \prod_{i=1}^k n_i$$

**Wichtig:** alle Schritte müssen voneinander unabhängig sein.

**Permutationen (P):**

$$P = n * (n - 1) * (n - 2) * \dots * 2 * 1 = n!$$

**Bestimmung der Anzahl von Auswahlmöglichkeiten**

n: Anzahl Wahlmöglichkeiten k: Anzahl Schritte		Reihenfolge	
		Mit	Ohne
Zurücklegen	Mit	$n * n * \dots * n = n^k$ Fundamentalprinzip	$\frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k}$ $= \binom{k+n-1}{n-1}$ Beispiel Suppe & Zutaten.
	Ohne	$n * (n-1) * (n-2) * \dots * (n-(k-1)) = \frac{n!}{(n-k)!}$ Permutationen durch Permutationen der nicht verwendeten Elemente (es bleiben ev. leere Positionen)	$\frac{n!}{k!(n-k)!} = \binom{n}{k}$ Reihenfolge eliminieren, indem durch Zahl der Möglichkeiten, k anzuordnen (k!) dividiert wird → ergibt Binomialkoeffizient

**0! = 1;**  $\binom{n}{k} \rightarrow \binom{8}{3} = \frac{8*7*6}{3!}$  Zähler: Fakultät bis k Stellen; **WICHTIG:** bei kombinierten Möglichkeiten multiplizieren!

## Zufallsexperimente

Definition als Vorgang:

- Er resultiert in einem von mehreren möglichen bekannten Ergebnissen (Ampel ist rot oder grün)
- Ergebnis ist nicht mit Sicherheit vorhersagbar
- Lässt sich unter den gleichen Randbedingungen wiederholen

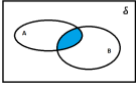
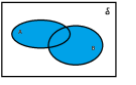
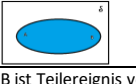
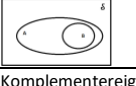
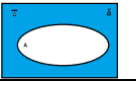

**Ergebnisräume** (Menge aller Ergebnisse):  $\delta$

**Ereignisse** (Sachverhalt): A

→ gilt als eingetreten, wenn das Ergebnis eines Zufallsexperiments ein Element der Teilmenge A ist.

- Elementarereignis:** Ergebnis, welches aus einem Ergebnis besteht
- Sicheres Ereignis:** umfasst alle Ergebnisse ( $A = \delta$ )
- Unmögliches Ereignis:** enthält kein Element aus  $\delta$

Rein mathematisch sind Ereignisse und Ergebnisräume Mengen:

Bezeichnung & Symbolik	Erläuterung	Beispiel
Ereignisschnitt $A \cap B$ 	$A \cap B$ besteht aus den Ergebnissen, die sowohl in A als auch in B vorkommen. Beobachtet man ein Ergebnis, welches zur Schnittmenge von A und B gehört, so ist sowohl A als auch B eingetreten.	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde höchstens zweimal. B: beim dreimaligen Besuch eines online-Shops kauft ein Kunde mindestens einmal. $A \cap B$ : Beim dreimaligen Besuch eines online-Shops kauft ein Kunde ein- oder zweimal. Gilt $A \cap B = \emptyset$ , sind A und B disjunkt (Überschneidungsfrei)
Vereinigung $A \cup B$ 	$A \cup B$ besteht aus allen Ergebnissen, die entweder zu A oder zu B oder zu beiden gehören.	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde höchstens zweimal. B: beim dreimaligen Besuch eines online-Shops kauft ein Kunde mindestens einmal. $A \cup B$ : Beim dreimaligen Besuch eines online-Shops kauft ein Kunde kein, ein-, zwei oder dreimal.
Gleichheit zweier Ereignisse $A = B$ 	A und B setzen sich aus exakt denselben Ergebnissen zusammen	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde exakt einmal. B: beim dreimaligen Besuch eines online-Shops kauft ein Kunde exakt zweimal nichts.
B ist Teilereignis von A $B \subset A$ 	Jedes Ergebnis aus B gehört auch zu A. Tritt B ein, ist somit auch automatisch A eingetreten. Man sagt auch, B impliziert A	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde mindestens einmal. N: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde exakt einmal.
Komplementereignis $\bar{A}$ 	$\bar{A}$ besteht aus allen Ergebnissen von $\delta$ , die nicht zu A gehören; A und $\bar{A}$ ergänzen sich zu $\delta$ .	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde mindestens einmal. $\bar{A}$ : Beim dreimaligen Besuch eines online-Shops kauft ein Kunde bei keinem Besuch.
Differenz $A \setminus B = A \cap \bar{B}$ 	Das Differenzereignis tritt ein, wenn A, aber nicht B eintritt.	A: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde höchstens zweimal. B: Beim dreimaligen Besuch eines online-Shops kauft ein Kunde mindestens einmal. $A \setminus B$ : Beim dreimaligen Besuch eines online-Shops kauft ein Kunde kein einziges Mal.

**Wahrscheinlichkeiten (P):**

- Der **klassische (Laplacesche) Wahrscheinlichkeitsbegriff**: Bsp. Würfel
- Der statistische (frequentische, empirische) Wahrscheinlichkeitsbegriff
- Der subjektivistische Wahrscheinlichkeitsbegriff

**Klassischer Wahrscheinlichkeitsbegriff**

Wurde im Zusammenhang mit der Analyse von Glücksspielen entwickelt (Pierre-Simon Laplace).

Berechnung: indem Anzahl aller Ergebnisse von A durch Anzahl aller möglicher Ergebnisse des Ergebnisraumes dividiert werden:

$$P(A) = \frac{\text{Anzahl der Ergebnisse in } A}{\text{Anzahl der Ergebnisse in } \Omega}$$

Annahme: alle Ereignisse sind gleich wahrscheinlich. Jedes erhält die Wahrscheinlichkeit  $1/n$ .

**Statistischer Wahrscheinlichkeitsbegriff**

Berechnung: indem Häufigkeit des Eintritts in der Vergangenheit ( $n(A)$ ) durch Häufigkeit der Zufallsexperimente ( $n$ ) dividiert werden:

$$P(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n}$$

Diese Definition ist eher eine Behauptung, scheint aber plusminus zu funktionieren.

**Subjektivistischer Wahrscheinlichkeitsbegriff**

Wenn Elementarereignisse nicht gleich wahrscheinlich sind und es zu wenige Beobachtungen gibt, nimmt man das «Bauchgefühl». Dies ist aus statistischer Sicht problematisch, weil Objektivität und Wiederholbarkeit nicht gegeben sind.

**Axiome von Kolmogorov**

Welche Eigenschaften müssen erfüllt werden, damit man widerspruchsfrei mit ihnen rechnen kann?

Drei Regeln:

- $P(A) \geq 0$ : es gibt pro Ereignis immer eine konkrete nichtnegative Wahrscheinlichkeitszahl  $P(A)$
- $P(\Omega) = 1$ : Ein sicheres Ereignis hat die Wahrscheinlichkeit 1
- $P(A \cup B) = P(A) + P(B)$  für  $A \cap B = \emptyset$ : Wahrscheinlichkeit der Vereinigung zweier ausschliessender Ereignisse ist = Summe der einzelnen Wahrscheinlichkeiten

Daraus kann man weitere Regeln ableiten:

Regel	Beispiel
$P(\bar{A}) = 1 - P(A)$	Männlicher Mitarbeiter gewinnt die Reise: $a - P(\text{Gewinner} = \text{weiblich}) = 0.5$
$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = P(\Omega) = 1$	Männlicher oder weiblicher Mitarbeiter gewinnt die Reise = 1
$P(\emptyset) = 0$	Niemand gewinnt die Reise = 0
$P(A \setminus B) = P(A) - P(A \cap B)$	A: Eine Mitarbeiterin gewinnt die Reise. B: Frederike gewinnt die Reise. $P(A \setminus B) = \frac{12}{24} - \frac{4}{24} = \frac{1}{3}$
$B \subset A \Rightarrow P(B) \leq P(A)$	A: Eine Mitarbeiterin gewinnt die Reise. B: Frederike gewinnt die Reise.

	$B \subset A; P(B) = \frac{1}{6} < \frac{1}{2}$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	A: Eine Mitarbeiterin gewinnt die Reise. B: Jemand der bereits dreimal gewonnen hat, gewinnt. $P(A \cup B) = \frac{12}{24} + \frac{9}{24} - \frac{6}{24} = \frac{5}{8}$
$P(A \cap B) \geq 1 - [P(\bar{A}) + P(\bar{B})]$	A: Ein/e Mitarbeiter/in, der/die mindestens bereits zweimal gewonnen hat, gewinnt. B: Hagen gewinnt die Reise. $P(A \cap B) = \frac{6}{24} \geq 1 - \left[ \frac{1}{24} + \frac{18}{24} \right] = \frac{5}{24}$

Beispiele:

12 Sockenpaare, 4 davon mit Loch. Wahrscheinlichkeit, 1 oder 2 Paare mit Loch zu erwischen:  $\binom{8}{3} * \binom{4}{1} + \binom{8}{2} * \binom{4}{2}$

## Cheatsheet WS – PVA 2

### BEDINGTE WS UND STOCHASTISCHE UNABHÄNGIGKEIT

#### Bedingte Wahrscheinlichkeiten: $P(A|B)$

Wahrscheinlichkeit von A unter der Bedingung B. Es geht um folgende Fragen:

- Wie verhält sich die Wahrscheinlichkeit für ein Ereignis A? (z.B. «Ich bekomme meinen Traumjob.»)
- ... unter der Bedingung, dass ein anderes Ereignis B eintritt (z.B. «Ich habe einen Hochschulabschluss.»)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ für } P(B) > 0$$

Die Formel wird wahrscheinlicher, wenn man mit absoluten Werten statt Wahrscheinlichkeiten arbeitet. Vierfeldertafel:

	A	$\bar{A}$	
B	0.3 ( $A \cap B$ )	0.1 ( $\bar{A} \cap B$ )	0.4
$\bar{B}$	0.2 ( $A \cap \bar{B}$ )	0.4 ( $\bar{A} \cap \bar{B}$ )	0.6
	0.5	0.5	1

Wie hoch ist die Wahrscheinlichkeit, dass A eintritt, wenn B gilt?  $\rightarrow 0.3/0.4 = 0.75 = 75\%$

Wie hoch ist die Wahrscheinlichkeit, dass B gilt, wenn ich A habe?  $\rightarrow 0.3/0.5 = 0.6 = 60\%$

Wenn 75% von B für A gilt, dann ist Wahrscheinlichkeit, dass B gilt wenn ich A habe  $\rightarrow (0.75 * 0.4) / 0.5 = 0.6 = 60\%$  (Bayes)

**Stochastische Unabhängigkeit:** wenn Eintritt des Ereignisses B keinen Einfluss auf die Wahrscheinlichkeit des Ereignisses A hat:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A)$$

Nur bei stochastisch unabhängigen Ereignissen lässt sich die Wahrscheinlichkeit für **Ereignisschnitte** berechnen (wegen  $P(A|B) = P(A)$ ):

$$P(A \cap B) = P(A) * P(B)$$

Dieser sogenannte **Multiplikationssatz** für stochastisch unabhängige Ereignisse ist Voraussetzung für viele Wahrscheinlichkeitsmodelle, und lässt sich wie folgt verallgemeinern:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) * P(A_2) * \dots * P(A_n)$$

**Satz von Bayes:** stellt eine Beziehung her von bedingten Wahrscheinlichkeiten und dem was wir berechnen wollen.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$P(A) = P(B \cap A) + P(\bar{B} \cap A) = P(B) * P(A|B) + P(\bar{B}) * P(A|\bar{B})$$

$$A \cap B \neq A * B \quad \rightarrow \text{bei Abhängigkeit!}$$

$$(\bar{B}|A) = \frac{B}{A} \quad \rightarrow \text{Nicht B, wenn nicht A wahr ist: nicht B geteilt durch nicht A}$$

**Beispiel:** Geräte A, B und 2\*C; C\*C = Ausfall x.

Alle laufen:  $A * B * (1-x) = y$ ; Alle laufen 7 Tage:  $y^7$

### ZUFALLSVARIABLEN

Symbol: **X** (oder Y, Z, ...). Wahrscheinlichkeitskonzept wird so erweitert, dass Ereignisse mit möglichen Werten in Zusammenhang gebracht werden. Allgemein gibt es zwei Typen von Zufallsvariablen:

- Diskrete Zufallsvariablen (1,2,,3,4,...N)
- Stetige Zufallsvariablen (R)

#### DISKRETE ZUFALLSVARIABLEN

**Wahrscheinlichkeitsfunktion:** bei diskreten Zufallsvariablen gibt es für jede Ausprägung  $x_i$  eine Wahrscheinlichkeit. Die Zuordnung der Wahrscheinlichkeit heisst Wahrscheinlichkeitsfunktion (bzw. Massenfunktion, Wahrscheinlichkeitsmassenfunktion)

$$f(x) = P(X = x) = \begin{cases} p_i & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases}$$

$P(X=x) \rightarrow$  Wahrscheinlichkeit, dass Variable X den Wert x annimmt;  $p_i \rightarrow$  Symbol für Einzelwahrscheinlichkeit. Bei mehreren Werten Verknüpfung mit Multiplikation. Visualisierung mit Stabdiagramm.

**Verteilungsfunktion:** Frage, mit welcher Wahrscheinlichkeit ein bestimmter Wert nicht überschritten wird. Symbol: **F(x)**.

- Die Verteilungsfunktion hat immer einen Wert zwischen 0 und 1 (inkl.)
- F(x) ist eine monoton steigende Funktion (nicht zwingend streng)
- Grenzwert strebt immer gegen minus unendlich nach links, plus unendlich nach rechts

Berechnung: Aufaddieren von Werten, die kleiner oder gleich dem gesuchten Wert sind:  $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$

#### STETIGE ZUFALLSVARIABLEN

**Dichtefunktion:** Stetige Funktion nötig, Symbol **f(x)**. Auch Dichte oder Wahrscheinlichkeitsdichtefunktion genannt. Wenn gesamte Wahrscheinlichkeitsmasse von 1 auf unendlich viele Werte verteilt wird, streben einzelne Werte gegen 0, weswegen Berechnung über Teilintervalle, indem man Integral über der Dichtefunktion bildet.

$$P(a < X < b) = \int_a^b f(x) dx \text{ für alle } a, b \text{ mit } a \geq b$$

$P(a < X < b) \rightarrow$  Wahrscheinlichkeit, dass Zufallsvariable einen Wert zwischen a und b annimmt; um Wert zu erhalten muss Integral von a bis b über Dichtefunktion gebildet werden.

f(x) darf selbst keine negative Werte annehmen und das Integral über dem ganzen Definitionsbereich muss = 1 sein.

**Verteilungsfunktion:** Frage nach Wahrscheinlichkeit, dass ein bestimmter Wert nicht überschritten wird. Symbol: **F(x)**.

- Die Verteilungsfunktion hat immer einen Wert zwischen 0 und 1 (inkl.)
- F(x) ist eine monoton steigende Funktion (nicht zwingend streng)
- Grenzwert strebt immer gegen minus unendlich nach links, plus unendlich nach rechts

Berechnung:  $F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$

Die Verteilungsfunktion wird auch für Wahrscheinlichkeitsberechnung beliebiger Teilintervalle zu berechnen.  $\rightarrow$  F(x) an Ober- und Untergrenze bestimmen und den zweiten vom ersten Wert abziehen. Sinnvoll bei schwierig zu integrierenden Dichtefunktionen.

	Diskret	Stetig
Beispiel	X: Zahl unzustellbarer Sendungen pro Monat N → Stück/Monat	X: Dauer zwischen Bestellung und Zustellung einer Sendung R+ → Stunden
f(x)	Wahrscheinlichkeitsfunktion $f(x) = P(X = x) = \begin{cases} p_i & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases}$	Dichtefunktion $P(a < X < b) = \int_a^b f(x) dx$
F(x)	Verteilungsfunktion $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$ Verteilungsfunktion gibt an mit welcher Wahrscheinlichkeit die Zufallsvariable höchstens einen bestimmten Wert x annimmt.	Verteilungsfunktion $F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$

## ERWARTUNGSWERT VON ZUFALLSVARIABLEN

**Erwartungswert** (Symbol **E**) gibt das «Mittel» des Wertes an, im Unterschied zum arithmetischen Mittel geht es aber um theoretisch mögliche Werte.

Bei diskreten Zufallsvariablen:  $EX = \sum_i x_i * f(x_i) \rightarrow$  Wert mit Wahrscheinlichkeit multiplizieren, dann alles summieren.

Bei stetigen Zufallsvariablen mit Dichte f(x):  $EX = \int_{-\infty}^{\infty} x * f(x) dx \rightarrow$  Dichte mit Wert multiplizieren, dann integrieren.

Bei praktischen Fragestellungen ist meist keine Dichtefunktion da. Bei vorliegenden Messwerten entweder eine Dichtefunktion schätzen und Formel für stetige Zufallsvariablen anwenden, oder Klassen bilden und relative Klassenhäufigkeiten als Wahrscheinlichkeiten interpretieren (analog zum Mittelwert bei klassiert vorliegenden Daten).

## VARIANZ VON ZUFALLSVARIABLEN

**Varianz** (Symbol **var**) als Mass für **Streuung** der Realisationsmöglichkeiten einer Zufallsvariable um den Erwartungswert. Wichtig bei Vorhersagen. Eng verwandt mit empirischer Varianz bei gruppiert/klassiert vorliegenden Daten. Unterschied: Mass für Streuung theoretisch möglicher Werte.

$$varX = E(X - EX)^2 = EX^2 - (EX)^2$$

$(X-EX)^2 \rightarrow$  von jedem möglichen Wert X muss Erwartungswert abgezogen werden, Resultat quadrieren  
 $E(X-EX)^2 \rightarrow$  Erwartungswert bilden; also quadrierte Differenz mit zugehöriger Wahrscheinlichkeit gewichten und Summe bilden (wenn diskret) oder Integrieren (wenn stetig).

Für diskrete Zufallsvariablen:  $varX = \sum_i (x_i - EX)^2 * f(x_i)$

Für stetige Zufallsvariablen:  $varX = \int_{-\infty}^{\infty} (x_i - EX)^2 * f(x) dx$

## RECHNEN MIT ZUFALLSVARIABLEN

**Transformation** von Zufallsvariablen:

Grundlegende Operationen: feste Werte hinzuaddieren/subtrahieren oder multiplizieren, oder andere Zufallsvariablen addieren/subtrahieren, so ergeben sich neue Zufallsvariablen. Der Erwartungswert hat folgende Eigenschaften:

$$E(a + bX) = a + b * EX; \quad E(X^{(1)} + X^{(2)}) = EX^{(1)} + EX^{(2)}$$

Standardisierung:  $X^* = \frac{X-EX}{\sqrt{varX}}$

X\* ist eine neue Zufallsvariable  $\rightarrow$  'Standardisierte Zufallsvariable'; X wird eigener Erfahrungswert EX abgezogen, dann durch Wurzel der Varianz varX von X geteilt.

Der **Erwartungswert** einer standardisierten Zufallsvariable ist stets 0, ihre **Varianz** ist stets 1 und sie ist immer einheitslos.

## THEORETISCHE WAHRSCHEINLICHKEITSMODELLE

Allgemeine Formeln zur Berechnung von Wahrscheinlichkeiten.

### GLEICHVERTEILUNG

**Gleichverteilung** (Symbol **X<sub>a</sub>**): behauptet, dass jede Realisationsmöglichkeit dieselbe Wahrscheinlichkeit besitzt (gilt für diskret und stetig)  $\rightarrow$  also bei 'gleichverteilten' Variablen (wie Würfeln).

**Diskrete Gleichverteilung**:  $f_G(x|r) = \begin{cases} \frac{1}{r} & \text{für } x \text{ Element von } \{x_1; x_2; \dots; x_r\} \\ 0 & \text{sonst} \end{cases}$

r  $\rightarrow$  Realisationen, z.B. 6 bei einem Würfel;  $\frac{1}{r}$  für x Element von  $\{x_1; x_2; \dots; x_r\} \rightarrow$  für jedes x gleiche Wahrscheinlichkeit von  $\frac{1}{r}$

**Stetige Gleichverteilung**:  $f_G(x|a; b) = \begin{cases} \frac{1}{b-a} & \text{für } x \text{ Element von } [a; b] \\ 0 & \text{sonst} \end{cases}$

Auch Rechteckverteilung genannt. [a;b]  $\rightarrow$  Intervall von a nach b.

Davon die Verteilungsfunktion:  $F_G(x|a; b) = \begin{cases} 0 & \text{für } x < a \\ \frac{x-a}{b-a} & \text{für } x \in [a; b] \\ 1 & \text{für } x > b \end{cases}$

Allgemeine Formel für Erwartungswert:  $\frac{b+a}{2}$  Varianz:  $varX_G = \frac{1}{12} (b-a)^2$

### BERNOULLI-MODELLE

Beispiel Sonnentage zählen, mit 1 oder 0 als Werte. Dann aufaddieren und Prozentwert berechnen.

**Bernoulli-Variablen** (Symbol: **X<sub>B</sub>**): Zufallsvariablen mit festen Werten im Zufallsexperiment. Allgemein: Wert=1 für Wahrscheinlichkeit Theta  $\Theta$ .

**Wahrscheinlichkeitsfunktion**:  $f_B(x|\Theta) = \begin{cases} 1-\Theta & \text{für } x = 0 \\ \Theta & \text{für } x = 1 \\ 0 & \text{sonst} \end{cases}$

**Verteilungsfunktion**:  $F_B(x|\Theta) = \begin{cases} 0 & \text{für } x < 0 \\ 1-\Theta & \text{für } 0 \leq x < 1 \\ 1 & \text{für } x \geq 1 \end{cases}$

**Erwartungswert**:  $EX_B = \Theta$  **Varianz**:  $varX_B = \Theta * (1-\Theta)$

Allgemein ist das Bernoulli-Modell anwendbar, wenn Zufallsexperimente

- n-mal wiederholt werden können
- Bei jeder Wiederholung nur zwei Realisierungen mit Werten 1 oder 0 möglich sind
- Die Eintrittswahrscheinlichkeit immer unverändert bleibt

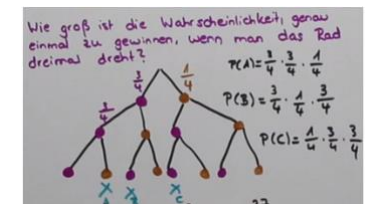
### BINOMIALVERTEILUNG

Man kann eine Zufallsvariable bilden, die einen über n Schritte laufenden Zählvorgang beschreibt, wenn man n Bernoulli-Variablen addiert. Bedingung: Wahrscheinlichkeit  $\Theta$  für Eintritt der Realisationsmöglichkeiten bleibt immer gleich. Anzahl Erfolge in einer Serie von gleichartigen und unabhängigen Versuchen, mit genau zwei möglichen Ergebnissen (0 oder 1).

**Binominalvariable**:  $X_B = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_B$

Für die **Anzahl der Möglichkeiten**:  $\binom{n}{k} = \frac{n!}{k!(n-k)!} \rightarrow n = \text{Anzahl Wahlmöglichkeiten/Ereignisse}; k = \text{Anzahl Ziehungen}$

Daraus erhält man die allgemeine Wahrscheinlichkeitsfunktion:



Anzahl:  $\binom{n}{k} \rightarrow \binom{3}{1}$   
 Beispiel: 20 Meetings mit 95% Erfolg:  
 $0.95^{20} = 0.358 = 35.8\%$  für alle Meetings



**Binomialverteilung:**  $f_b(x|\theta; n) = \begin{cases} \binom{n}{x} * \theta^x * (1 - \theta)^{n-x} & \text{für } x = 0, 1, 2, \dots, n \\ 0 & \text{sonst} \end{cases} \rightarrow \text{für einen Fall}$

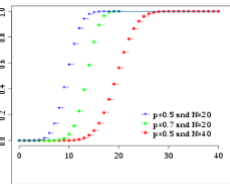
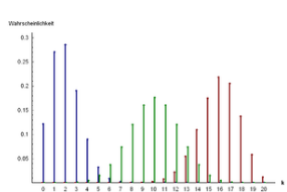
$n \rightarrow$  Anzahl Schritte;  $x \rightarrow$  insgesamt x-maliges Auftreten  
Beispiel: Berechnung für Wahrscheinlichkeit für 3 pünktliche Fahrzeuge.

Die **Verteilfunktion** berechnet Wahrscheinlichkeiten für ein höchstens x-maliges Auftreten des interessierenden Ergebnisses bei n Schritten:  
 $\rightarrow$  Fälle a bis ...

$$F_b(x|\theta; n) = \begin{cases} 0 & \text{für } x < 0 \\ \sum_{k=0}^x \binom{n}{k} * \theta^k * (1 - \theta)^{n-k} & \text{für } x = 0, 1, 2, \dots, n \\ 1 & \text{für } x > n \end{cases}$$

**Erwartungswert:**  $EX_b = n * \theta$

**Varianz:**  $varX_b = n * \theta * (1 - \theta)$



$\rightarrow$  Grosse Zahlen dürfen durch Normalverteilung berechnet werden!

## POISSON-VERTEILUNG

Mit der Bino mialverteilung berechnet man Wahrscheinlichkeiten dafür, dass x-viele Elemente aus einer Stichprobe vom Umfang n eine bestimmte Eigenschaft tragen. Mit Hilfe der **Poisson-Verteilung**  $f_p(x|\lambda)$  kann man dafür Wahrscheinlichkeiten berechnen. Die Zugehörige **Zufallsvariable**  $X_p$  ist  $=0, 1, 2, 3, \dots$  und steht für die Häufigkeiten mit der das interessierende Ereignis in der Reihe beobachtet wird.

Anzahl Ereignisse, die bei konstanter mittlerer Rate unabhängig voneinander in einem festen Zeitintervall oder räumlichen Gebiet eintreten.

Sofern das Auftreten des Ereignisses **zufällig** und **unabhängig** von anderen ist, funktioniert alles ähnlich wie ein Spezialfall der Binomialverteilung:

- Man weiss aus Erfahrung dass ein Ereignis durchschnittlich 4-mal ( $\lambda$ ) pro Monat auftritt (z.B. Vertragsabschluss)
- Wenn man Monat in 8 Teilschritte teilt, tritt Ereignis im Durchschnitt 0.5-mal auf
- Interpretiert man Anzahl der Teilschritte als n Schritte, dann ergibt sich daraus sofort die Wahrscheinlichkeit  $\theta$  für das Auftauchen des interessierenden Ereignisses in jedem Schritt, weil  $n * \theta$  immer dem durchschnittlich erwarteten Wert für die gesamte Bezugsgrösse entsprechen muss
- Bei der Poisson-Verteilung ist dieser durchschnittlich erwartete Wert das Symbol  $\lambda$  und ist der **Erfahrungswert**:  $EX_p = n * \theta = \lambda \rightarrow \theta = \frac{\lambda}{n} \rightarrow x$ : wie oft tritt Ereignis **höchstens** auf?
- Damit man tatsächlich die Wahrscheinlichkeitsfunktion findet, muss man Bezugsgrösse (hier Monat) nicht in 8, sondern in unendlich viele Abschnitte zerlegen

**Poisson-Verteilung:**  $f_p(x|\lambda) = \begin{cases} \frac{\lambda^x}{x!} * e^{-\lambda} & \text{für } x \text{ element von } N \text{ inkl. } 0 \\ 0 & \text{sonst} \end{cases}$

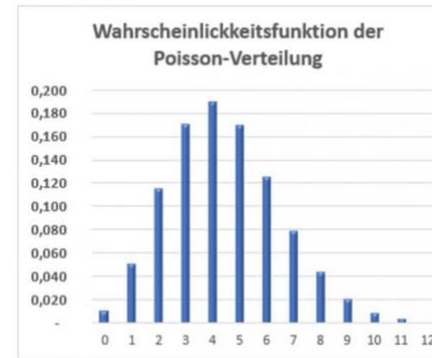
$\lambda \rightarrow$  Häufigkeit, mit der man Ereignisse durchschnittlich in Bezugsgrösse beobachtet;  $e \rightarrow$  Eulersche Zahl (2.71828...)

**Verteilungsfunktion:**  $F_p(x|\lambda) = \begin{cases} 0 & \text{für } x < 0 \\ e^{-\lambda} * \sum_{k=0}^x \frac{\lambda^k}{k!} & \text{für } x \text{ element von } N \text{ mit } 0 \\ 1 & \text{für } x \rightarrow \infty \end{cases}$

**Erwartungswert:**  $EX_p = \lambda$  **Varianz:**  $varX_p = \lambda$  **Parameter  $\lambda$  konvergiert gegen Unendlich;  $0! = 1$**

Die Poisson-Verteilung ist eine Grenzverteilung zur Binomialverteilung für  $n \rightarrow \infty$ , liefert aber auch im Endlichen gute Näherungen für Binomialverteilung wenn n gross und  $\theta$  klein ist. Nützlich, weil weniger Rechenaufwand.

Für Näherungen prüft man ob  $n \geq 100$  und  $\infty \leq 0.1 \rightarrow$  dann kann man lambda bestimmen aus  $n * \theta = \lambda$ .



## ZUSAMMENFASSUNG ZÄHLVERTEILUNGEN

Name	Anwendung	Bezugsgrösse ist	
		Diskret	«Stetig»
		Häufigkeit, mit der das interessierende Merkmal auftritt	Nur Anteilswert $\theta$ ist bekannt: Binomialverteilung
			Poisson-Verteilung
Binomial-verteilung	Vorgang läuft von Schritt zu Schritt unabhängig ab, die Wahrscheinlichkeit für das Auftreten des interessierenden Ereignisses beträgt in jedem Schritt $\theta$ . Gesucht sind Wahrscheinlichkeiten für eine bestimmte Anzahl x von Beobachtungen des interessierenden Merkmals bei n Schritten.	$f_b(x \theta; n) = \binom{n}{x} * \theta^x * (1 - \theta)^{n-x}$ $F_b(x \theta; n) = \sum_{k=0}^x \binom{n}{k} * \theta^k * (1 - \theta)^{n-k}$ $EX_b = n * \theta$ $varX_b = n * \theta * (1 - \theta)$	Eine Maschine produziert Ausschuss mit Wahrscheinlichkeit $\theta = 0.02$ je Stück. Wie gross ist die Wahrscheinlichkeit, dass in einer Produktion von $n=1'000$ Stück mehr als 50 ( $x \geq 50$ ) Ausschussstücke enthalten sind?
Poisson-Verteilung	Ein interessierendes Ereignis tritt erfahrungsgemäss innerhalb einer zeitlichen, räumlichen oder anderweitig stetigen Spanne im Schnitt mit der Häufigkeit $\lambda$ auf. Gesucht ist die Wahrscheinlichkeit für ein x-maliges Auftreten dieses Ereignisses innerhalb dieser Spanne. Grenzverteilung zur Binomialverteilung bei grossen n ( $n \geq 100$ ) und kleinen $\theta$ ( $\theta \leq 0.1$ ).	$f_p(x \lambda) = \frac{\lambda^x}{x!} * e^{-\lambda}$ $F_p(x \lambda) = e^{-\lambda} * \sum_{k=0}^x \frac{\lambda^k}{k!}$ $EX_p = \lambda$ $varX_p = \lambda$	An einer Tankstelle tanken pro Nacht im Durchschnitt $\lambda = 7$ Fahrzeuge. Wie gross ist die Wahrscheinlichkeit, dass in einer Nacht gar kein Fahrzeug tanken wird ( $x = 0$ )?



## STANDARDNORMALVERTEILUNG

Wird für viele empirische Verteilungen verwendet (z.B. Abweichungen von der Norm, Umsätze, Messfehler, usw.). Ausserdem: viele theoretische Verteilungen konvergieren gegen die Normalverteilung und lassen sich so gut annähern.

Wir beginnen mit dem Spezialfall der **Standardnormalverteilung**, die Definitionsmenge ist ganz  $\mathbb{R}$ , es ist eine stetige Verteilung.

Ausgangspunkt für die Suche nach einer standardnormalverteilten Zufallsvariable ist der Wunsch nach einer Dichte,

- Die symmetrisch um Null liegt und
- Deren Werte immer kleiner werden, je weiter man sich von der Null entfernt;
- Die Abnahme der relativen Änderung der Dichte soll dabei proportional sein zur Abweichung selbst.

### Standardnormalverteilung:

$$f_N(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Statt  $x$  wird hier  $z$  für Realisierungsmöglichkeiten verwendet, wegen herausragender Stellung dieser Funktion (Zufallsvariable =  $Z$ ) Die zugehörige Verteilungsfunktion beantwortet die Frage «Mit welcher Wahrscheinlichkeit nimmt  $Z$  höchstens den Wert  $z$  an?» und ist das Integral der Dichte (hier wird  $z$  durch  $u$  ersetzt).

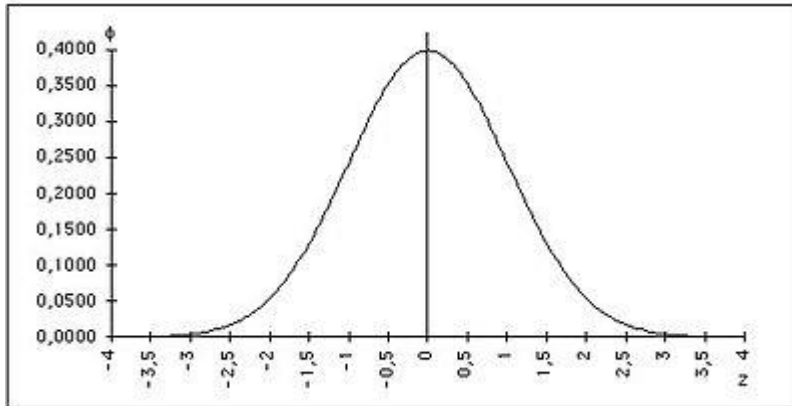
### Verteilungsfunktion:

$$F_N(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}u^2} du$$

- $\frac{1}{\sqrt{2\pi}}$  → Normierungsfaktor, damit Fläche im Gesamten den Definitionsbereich 1 ergibt.
- Die Dichte hat bei  $z = 0$  ein Maximum.
- Bei  $z = -1$  und  $z = 0$  besitzt die Dichte Wendepunkte
- Nach links konvergiert nach links und rechts gegen 0
- Die Dichte wird oft als Glockenkurve bezeichnet
- Kann nicht analytisch gelöst werden

**Erwartungswert:**  $EZ = 0$

**Varianz:**  $\text{var}Z = 1$



## ALLGEMEINE NORMALVERTEILUNG

Die allgemeine Normalverteilung kann einen beliebigen Erwartungswert und eine beliebige Varianz haben. Die Dichte entspricht jeweils immer noch der Glockenform. Sie ist über  $\mathbb{R}$  definiert. Die Funktion selbst besitzt keine Wahrscheinlichkeitsbedeutung weil es sich um stetige Verteilung handelt.

**Zentraler Grenzwertsatz:** wenn Stichprobe gross genug ist → normalverteilt!

### Normalverteilung:

$$f_N(x|\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\mu$  → wie weit die Mitte der Glockenkurve nach rechts oder links aus 0 verschoben ist

$\sigma$  → Standardabweichung der Normalverteilung, gibt an wie stark die Glockenform gestreckt oder gestaucht wurde (wenn gross, flach; wenn klein, spitz)

### Verteilungsfunktion:

$$F_N(x|\mu;\sigma) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

Auch diese Verteilungsfunktion kann nicht analytisch gelöst werden.

- Verläuft symmetrisch zu  $\mu$
- Wendepunkte liegen bei  $x = \mu - \sigma$  und  $x = \mu + \sigma$

**Erwartungswert:**  $EX_N = \mu$

**Varianz:**  $\text{var}X_N = \sigma^2$

**Standardabweichung:**  $\sqrt{n} * \sigma$

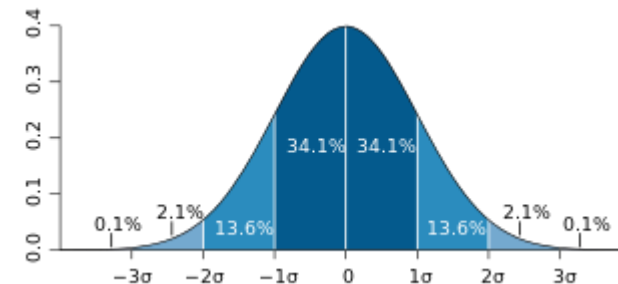
**Störung:**  $E(x_i) - \mu = \sigma$

Es gibt unendlich viele Normalverteilungen, aber alle Normalverteilungen lassen sich durch Standardisierung auf die Standardnormalverteilung zurückführen:

### Transformation zur Normalverteilung:

$$F_N\left(\frac{x-\mu}{\sigma} \mid \mu; \sigma\right) = F_N(z)$$

**Normalisieren:**  $\frac{x-\mu}{\sigma}$  oder  $\frac{\sum \text{Störungen}}{\sqrt{n} * \sigma}$



→ Wahrscheinlichkeit ist immer Fläche (Dichte, p-Wert) → z-Wert als vertikale Begrenzung

### Beispiel:

- Mittel: 450'000 →  $\mu$
- Standardabweichung: 170'000 →  $\sigma$
- Miete (Grenze): 110'000 →  $x$

$$z = \frac{110'000 - 450'000}{170'000} = -2$$

wenn negativ =  $|z| = 1 - |z|$

→ Symmetrie immer beachten!

**Beispiel Wahrscheinlichkeit:** 3 Geräte B fallen zu je 3% aus, mind. 2 müssen laufen.

Alle drei funktionieren:  $0.97^3$ ; zwei funktionieren:  $3 \text{ (Möglichkeiten)} * 0.97^2 * 0.03$   
Wahrscheinlichkeit dass mind. Zwei laufen: beide Fälle summieren!

## Cheatsheet WS – PVA 3

### GRENZWERTSÄTZE

Grenzwertsätze zeigen auf, dass viele Verteilungen gegen Normalverteilung konvergieren, das die Bestimmung von Wahrscheinlichkeiten oft erleichtern. Frage: «wie gut kann man Anteilswert in Grundgesamtheit durch Stichprobe schätzen?»

#### GESETZ DER GROSSEN ZAHLEN

Verwendet arithmetisches Mittel. Gesetz der grossen Zahlen: je grösser die Stichprobe, umso genauer das Ergebnis.

**Ausgangspunkt:**  $S_n = \sum_{i=1}^n X_i$

Der Summand kann eine *Bernoulli-Variable* oder *beliebige Zufallsvariable* sein. Alle Variablen  $X_i$  sollen aber den gleichen Erwartungswert  $EX_i = \mu$  und dieselbe Varianz  $\text{var}X_i = \sigma^2$  haben. Beispiel: sehr viele Elemente, oder Elemente 'mit Zurücklegen'.

**Durchschnitt** wird errechnet:  $\bar{X}_n = \frac{1}{n} * (x_1 + x_2 + \dots + x_n)$

**Erwartungswert** von  $\bar{X}_n$  wird errechnet:  $E\bar{X}_n = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n EX_i = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} * n * \mu = \mu$

Also: Durchschnittswert aller Schätzungen entspricht exakt dem Erwartungswert der einzelnen Zufallsvariable.

Für stochastisch unabhängige Schritte für  $\bar{X}_n$

gilt für **Varianz:**  $\text{var}\bar{X}_n = \frac{\sigma^2}{n}$  und **Standardabweichung:**  $\sqrt{\text{var}\bar{X}_n} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

→ Bei immer grösser werdenden  $n$  läuft die Varianz gegen 0. Daraus lässt sich das schwache Gesetz der grossen Zahlen ableiten, aus welchem sich folgende Aussage ergibt: «Die Wahrscheinlichkeit, dass der aus einer Stichprobe berechnete Mittelwert vom tatsächlichen Mittelwert der Grundgesamtheit deutlich abweicht, geht für sehr grosse  $n$  gegen 0.»

Stärkere Fassung der Konvergenz als Gesetz der grossen Zahlen ergibt  $P=1$ .

#### BERNOULLIS GESETZ DER GROSSEN ZAHLEN

Spezialfall des allgemeinen Gesetzes – hier geht es um Anteilswerte (3/5, 20%) und nicht um das arithmetische Mittel.

**Herleitung:**  $\bar{X}_n = \frac{1}{n} * (x_1 + x_2 + \dots + x_n)$  → jetzt müssen zwingend Bernoulli-Variablen verwendet werden!

$x_i$  ist also entweder 0 oder 1, Summen von 0-n. Durch Division durch  $n$  erhält man Anteilswert in Prozent.

Wahrscheinlichkeitsfunktion mit Binomialverteilung. Beispiel:

- Anteilswert der Grundgesamtheit  $\theta = 0.64$  (=64%), Stichprobe mit Umfang  $n$
- Wenn 64% Statistik mögen und 20 Befragungen stattfinden, berechnet sich die Wahrscheinlichkeit, dass 14 mit 'Ja' antworten:  $\binom{20}{14} * 0.64^{14} * 0.36^6 \approx 0.163$
- Anteilswert der Stichprobe =  $\frac{\text{Anzahl Eigenschaftsträger}}{\text{Stichprobenumfang}}$

**Anteilvariable  $P_b$**  erhält man, wenn Binomialvariable  $X_b$  durch Stichprobenumfang  $n$  dividiert wird. Wahrscheinlichkeitsfunktion dazu ist Binomialverteilung:

$$f_b\left(\frac{x}{n} | \theta; n\right) = \begin{cases} \binom{n}{x} * \theta^x * (1-\theta)^{n-x} & \text{für } x = 0, 1, 2, \dots, n \\ 0 & \text{sonst} \end{cases}$$

**Erwartungswert:**  $EP_b = E\left(\frac{X_b}{n}\right) = \frac{EX_b}{n} = \frac{n * \theta}{n} = \theta$  **Varianz:**  $\text{var}P_b = \text{var}\left(\frac{X_b}{n}\right) = \frac{\text{var}X_b}{n^2} = \frac{n * \theta * (1-\theta)}{n^2} = \frac{\theta * (1-\theta)}{n}$

Bei steigendem  $n$  wird Streuung um Anteilswert immer geringer. Also: «Je grösser die Stichprobe, umso grösser die Wahrscheinlichkeit, dass ein aus der Stichprobe geschätzter Anteilswert dem Anteilswert in der Grundgesamtheit entspricht.»

#### HAUPTSATZ DER MATHEMATISCHEN STATISTIK

Wenn mit Zufallsstichproben ganze empirische **Verteilungsfunktionen** geschätzt werden, wird das Ergebnis bei grösseren Stichproben auch jeweils genauer. Dies ist der

**Hauptsatz der mathematischen Statistik:**  $P\left(\lim_{n \rightarrow \infty} P_n(x) = F_h^*(x)\right) = 1$

→ Je grösser die Stichprobe, umso grösser die Wahrscheinlichkeit, dass die auf Stichprobenbasis geschätzte Verteilungsfunktion  $P_n(x)$  der eigentlichen empirischen Verteilungsfunktion  $F_h^*(x)$  entspricht.

#### ZENTRALER GRENZWERTSATZ

Wenn mit geschätztem Mittelwert aus Stichprobe gearbeitet wird, handelt es sich hier um eine Zufallsvariable. Zentraler Grenzwertsatz beantwortet folgende Frage: «Wie würde die Wahrscheinlichkeitsfunktion einer Mittelwertschätzung über eine Zufallsstichprobe aus einer beliebigen Grundgesamtheit aussehen?»

Bedingung:  $n$ -viele unabhängige Zufallsvariablen  $X_i$ , alle mit demselben Erwartungswert  $EX_i = \mu$  und selbe Varianz  $\text{var}X_i = \sigma^2$ . Daraus wird Mittelwertvariable  $E\bar{X}_n$  berechnet:

- Mittelwertvariable hat Erwartungswert  $E\bar{X}_n = \mu$  und Varianz  $\text{var}\bar{X}_n = \frac{\sigma^2}{n}$
- Ihre Standardisierte ist  $Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$
- Zentraler Grenzwertsatz als Beweis, dass Verteilungsfunktion dieser Standardisierten mit wachsendem  $n$  gegen Standardnormalverteilung konvergiert:  $\lim_{n \rightarrow \infty} P\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z\right) = F_n(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}u^2} du$

#### GRENZWERTSATZ VON MOIVRE UND LAPLACE

Verteilungsfunktion einer binomial verteilten Zufallsvariable konvertiert gegen Normalverteilung für steigendes  $n$ .

→ Sonderfall des zentralen Grenzwertsatzes; spart einiges an Rechenaufwand. **Für Binomialverteilung.**

- Faustregel:** wenn  $n * \theta * (1 - \theta) \geq 9$  → dann Normalverteilung zulässig (je näher  $\theta$  an 0.5 liegt und je grösser  $n$ , umso besser)
  - Parameter der Normalverteilung** ( $\mu$  und  $\sigma$ ) berechnen:  $\mu = n * \theta$  und  $\sigma = \sqrt{n * \theta * (1 - \theta)}$
  - Stetigkeitskorrektur** für bessere Näherungsergebnisse: (macht Bereich immer ein bisschen grösser)
    - Binomialverteilung kann nur diskrete Werte annehmen
    - Normalverteilung ist stetig und liefert über Integral
    - Unterschied zwischen Diskretheit und Stetigkeit ausgleichen: bei Näherung wird jedem Wert  $x$  der Binomialverteilung immer stetiger Bereich von 0.5 vor/nach dem  $x$ -Wert zugerechnet
- Ergebnis: 1-p-Wert (bei Moodle-Beispielen, Konventionalstrafe)

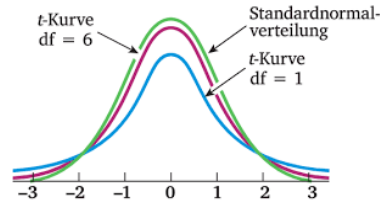
## T-VERTEILUNG

Ergibt nur Sinn beim **Schätzen** und **Testen**. Gehört zu «theoretischen Wahrscheinlichkeitsmodellen». Meist mit Tabellen.

Berechnung, wie stark ein aus einer Zufallsstichprobe geschätzter Mittelwert vom Mittelwert der Grundgesamtheit abweicht. Normalverteilte Zufallsvariable  $X_N$  wird durch Wurzel einer  $\chi^2$ -verteilten Zufallsvariable dividiert. Daraus entsteht neue,

t-verteilte Zufallsvariable  $X_t$ : 
$$X_t = \frac{X_N}{\sqrt{\frac{\chi^2}{n}}}$$

- Parameter der t-Verteilung ist n für «Freiheitsgrade» (Stichprobenumfang – Zahl aus Probe geschätzter Streuungsparameter)
- Je höher n ist, umso ähnlicher der Standardnormalverteilung



## SCHÄTZEN VON PARAMETERN

«Wie ermittelt man mit Zufallsstichproben Schätzwerte für eine Grundgesamtheit?»

Grundsätzlich wird geprüft ob aus Stichprobe gewonnene Aussage inwiefern gültig ist.

### PUNKTSCHÄTZUNGEN

Wenn man als Schätzung nur einen einzelnen Wert angibt. Zwei Größen werden besonders oft verwendet:

**Mittelwerte** (Durchschnitt) und **Anteilswerte** (Prozent).

Die **Schätzfunktion** ist jeweils die entsprechende Rechenvorschrift. Je nach dem gibt es unterschiedliche Alternativen, wobei es explizit eine Rolle spielt, dass mit einer begrenzten Stichprobe gerechnet wird.

Parameter sind:

- $\mu$  für arithmetisches Mittel der Grundgesamtheit
- $\theta$  für Anteilswert
- $\sigma$  für Standardabweichung und  $\sigma^2$  für Varianz
- $M$  für die Träger einer Eigenschaft,  $N$  für Anzahl aller Elemente

Ergebniswerte sind:

- $\bar{x}$  für arithmetisches Mittel
- $p$  für Anteilswert
- $s$  für Standardabweichung und  $s^2$  für Varianz (oder  $s^*$  oder  $s^{*2}$ )
- $m$  für die Träger einer Eigenschaft,  $n$  für Umfang der Stichprobe

Gesuchter Wert	Grundgesamtheit	Stichprobe
Mittelwert	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Anteilswert	$\theta = \frac{M}{N}$	$p = \frac{m}{n}$
Varianz	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_a)^2$	$s^{*2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Standardabweichung	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_a)^2}$	$s^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

### Erwartungstreue von Schätzfunktionen

Erwartungstreue heisst, dass der Erwartungswert aller möglichen Schätzwerte zu einem Parameter der Grundgesamtheit; dem Wert des Parameters in der Grundgesamtheit entspricht. Zieht man eine Zufallsstichprobe, ist

- $\bar{x}$  die Realisation einer Zufallsvariable  $\bar{X}$  mit  $E\bar{X} = \mu$
- $p$  die Realisation einer Zufallsvariable mit  $P$  und  $EP = \theta$

Also:  $\bar{x}$  und  $p$  sind erwartungstreu. Im Durchschnitt würde mit der Formel der Grundgesamtheit im Durchschnitt die Varianz zu niedrig geschätzt, was mit der erwartungstreuen Schätzfunktion korrigiert wird.

Grundsätzlich: je grösser n, umso geringer die Unterschiede.

Erwartungstreue von  $s^{*2}$  ohne Zurücklegen:  $s^{*2} = \frac{N-1}{N} * \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  (als zusätzliche Korrektur)

### INTERVALLSCHÄTZUNGEN

Vernünftigerweise Intervall angeben, da Punkte nie genaue Schätzungen sind. Das Intervall selbst heisst **Konfidenzintervall**.

Das Konfidenzintervall umfasst einen Bereich, in dem logisch begründbar der gesuchte Wert der Grundgesamtheit mit hoher Wahrscheinlichkeit liegt. Wesentliche Punkte:

- Erwartungswert  $E\bar{X}$  aller möglichen Schätzungen  $\bar{x}$  liegt genau bei  $\mu$
- Standardabweichung der möglichen Schätzungen (Symbol:  $\sqrt{\text{var}\bar{X}}$ ) um diesen Erwartungswert steht dabei mit der Varianz aller Messwerte der Grundgesamtheit (Symbol:  $\sigma^2$ ) in einem funktionalen Zusammenhang, nämlich:  
$$\sqrt{\text{var}\bar{X}} = \sqrt{\frac{\sigma^2}{n}} \rightarrow \text{Standardfehler} \quad \rightarrow \text{ACHTUNG: muss auf erwartungstreuem Sigma gemacht werden}$$
  
**Schätzung Standardabweichung  $\rightarrow$  dividieren durch Wurzel n für erwartungstreu Standardabweichung**
- Mögliche Schätzungen für  $\bar{x}$  sind **normalverteilt** um  $\mu$
- Also kann man Mithilfe der Standardnormalverteilung berechnen, in welchem Bereich um  $\mu$  Schätzungen  $\bar{x}$  sich mit grosser Wahrscheinlichkeit bewegen werden (Unter- und Obergrenze)
- Gewünschte Wahrscheinlichkeit kann vorgegeben werden und wird **Konfidenzniveau** genannt (Symbol:  $(1 - \alpha)$ )

Also: «Wenn der gesuchte Mittelwert  $\mu$  sowie die Varianz  $\sigma^2$  der Werte der Grundgesamtheit bekannt sind, kann angegeben werden, in welchem Bereich eine Schätzung  $\bar{x}$  für  $\mu$  mit beliebig vorgebarer Wahrscheinlichkeit  $(1 - \alpha)$  liegen wird.»

**Reale Problemstellung:** wir haben die Werte der Stichprobe, kennt aber  $\mu$  und  $\sigma^2$  nicht.

Lösung: wir verschieben das Intervall. Statt  $\mu$  ist nun  $\bar{x}$  die Mitte.

Wenn  $\sigma^2$  nicht bekannt ist, kann es durch  $s^{*2}$  angenähert werden. Wenn  $\theta$  nicht bekannt ist, kann es durch  $p$  angenähert werden.

Folgende **Faustregeln** müssen erfüllt sein:  $n \geq 30$  und  $\frac{N-n}{N-1} \geq 0.9$

Vorgehen zur Bestimmung von Konfidenzintervallen:

- Prüfen der Faustregeln:  $n \geq 30$  und  $\frac{N-n}{N-1} \geq 0.9$
- Ziehen und Vermessen der Stichprobe
- Ermittlung von Punktschätzungen  
 $\rightarrow$  Mittelwert:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  und  $s^{*2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$   
 $\rightarrow$  Anteilswert:  $p = \frac{m}{n}$
- Bestimmung der Varianzen der Schätzfunktionen  
 $\rightarrow$  Mittelwert:  $\text{var}\bar{X} \approx \frac{s^{*2}}{n}$   
 $\rightarrow$  Anteilswert:  $\text{var}P \approx \frac{p*(1-p)}{n}$
- Festlegung des gewünschten Vertrauensniveaus und Ermittlung des zugehörigen z-Wertes der **Standardnormalverteilung**:  
 $\cdot (1 - \alpha) = 0.9 \rightarrow \alpha = 0.1 \rightarrow \frac{\alpha}{2} = 0.05 \rightarrow 1 - 0.05 = 0.95 \rightarrow z\text{-Wert} = 1.65$

$$-(1 - \alpha) = 0.95 \rightarrow z = 1.96$$

$$-(1 - \alpha) = 0.99 \rightarrow z = 2.58$$

- Bestimmung des **Konfidenzintervalls**:

$$\rightarrow \text{Mittelwert: } \mu \in [-z * \sqrt{\text{var}\bar{X}} + \bar{x}; z * \sqrt{\text{var}\bar{X}} + \bar{x}] \text{ zur Konfidenz } (1 - \alpha)$$

$$\rightarrow \text{Anteilswert: } \theta \in [-z * \sqrt{\text{var}P} + p; z * \sqrt{\text{var}P} + p] \text{ zur Konfidenz } (1 - \alpha)$$

#### Intervallschätzungen bei Verletzung der Faustregeln

→ **Regel 1:**  $n \geq 30$

Grund: wenn beim Standardisieren für die Standardabweichung mit der Varianz der Stichprobe gearbeitet wird (z.B. wenn  $\sigma$  nicht bekannt ist), erhält man eine t-Verteilung. Dies spielt ab 30 Elementen keine Rolle mehr, weil da die t-Verteilung nahe genug der Standardnormalverteilung ist. Falls darunter, die Tabelle der t-Verteilung verwenden.

→ **Regel 2:**  $\frac{N-n}{N-1} \geq 0.9$

Grund: Normalerweise werden Stichproben ohne Zurücklegen gezogen (Abhängigkeiten entstehen, das Varianz der Schätzfunktionen um Faktor  $\frac{N-n}{N-1}$  verändert («Korrekturfaktor»). Wenn mindestens 0.9, passiert nicht viel mit dem Ergebnis. Ansonsten muss man wie folgt schätzen:

$$\begin{aligned} \text{Mittelwert: } \text{var}\bar{X} &= \frac{\sigma^2}{n} * \frac{N-n}{N-1} & \text{bzw. } \text{var}\bar{X} &= \frac{s^2}{n} * \frac{N-n}{N-1} \\ \text{Anteilswert: } \text{var}P &= \frac{\theta * (1-\theta)}{n} * \frac{N-n}{N-1} & \text{bzw. } \text{var}P &= \frac{p * (1-p)}{n} * \frac{N-n}{N-1} \end{aligned}$$

#### Ermittlung des notwendigen Stichprobenumfangs

Für praktische Zwecke sollte der Stichprobenumfang sinnvoll gesetzt werden. Ableitung der Formel aus Berechnungsvorschrift von Konfidenzintervallen:

- Um Punktschätzung Konfidenzintervall angeben (weil Punktschätzungen abweichen)
- Abstände (Symbol  $\varepsilon$ , je für Untergrenze und Obergrenze) der Intervallgrenzen hängen von z sowie den Standardabweichungen der Schätzfunktion ab:

$$\begin{aligned} \mu &\in [-z * \sqrt{\text{var}\bar{X}} + \bar{x}; z * \sqrt{\text{var}\bar{X}} + \bar{x}] \text{ zur Konfidenz } (1 - \alpha) \\ \theta &\in [-z * \sqrt{\text{var}P} + p; z * \sqrt{\text{var}P} + p] \text{ zur Konfidenz } (1 - \alpha) \end{aligned}$$

- Je kleiner  $\varepsilon$  ist, umso präziser wird die Intervallschätzung
- Je geringer die Konfidenz  $(1 - \alpha)$ , umso näher sind die z-Werte an 0
- Standardabweichungen berechnen sich nach folgenden Formeln:

$$\sqrt{\text{var}\bar{X}_n} = \sqrt{\frac{\sigma^2}{n}} \quad \sqrt{\text{var}P_\theta} = \sqrt{\frac{\theta * (1-\theta)}{n}}$$

- In beiden Fällen gibt es eine Abhängigkeit von n (Stichprobenumfang). Umstellen nach n:

$$\varepsilon = z * \sqrt{\frac{\sigma^2}{n}} \rightarrow \varepsilon^2 = z^2 * \frac{\sigma^2}{n} \rightarrow n = z^2 * \frac{\sigma^2}{\varepsilon^2}$$

$$\varepsilon = z * \sqrt{\frac{\theta * (1-\theta)}{n}} \rightarrow \varepsilon^2 = z^2 * \frac{\theta * (1-\theta)}{n} \rightarrow n = z^2 * \frac{\theta * (1-\theta)}{\varepsilon^2}$$

- Nun muss man
  - das gewünschte Vertrauen  $(1 - \alpha)$  festlegen
  - akzeptable Abstände  $\varepsilon$  der Konfidenzintervallgrenzen der Punktschätzung festlegen (bzw.  $2 * \varepsilon$ )
- Irgendwie muss man noch Abschätzen wie gross  $\sigma^2$  bzw.  $\theta$  ist:
  - üblicherweise hat man Werte aus Probeerhebungen/alten Daten
  - bei Anteilswertschätzungen kann man «worst case»  $\theta = 0.5$  nehmen
- Moodie Shortcut: Standardabweichung anpassen (z.B. durch x (z.B. 3))** -> n muss dann  $x^2$  Mal grösser sein  
**Beispiel: Konfidenzintervall unter n=16 dritteln**  $\rightarrow 3 = \sqrt{9} \rightarrow 9 * 16 = 144$

#### Zufallsauswahl

Diese Berechnungen können logisch fundiert nur auf Basis einer Zufallsauswahl vorgenommen werden. Zu beachten ist, dass Stichproben in der Realität oft nicht zufällig sind (Passantenbefragungen, Serie 0 von Produkten).

## BEISPIELE

### PUNKTSCHÄTZUNGEN UND NORMALVERTEILUNG

Der Besitzer einer Ladenkette möchte seine Marketingstrategie optimieren. Er will daher seine Kundschaft besser kennenlernen. Unter anderem möchte er etwas über das Alter seiner Kundinnen und Kunden wissen.

Er befragt daher 100 zufällig ausgewählte Kundinnen und Kunden nach ihrem Alter. Dabei ergibt sich ein durchschnittliches Alter von 34,1 Jahren bei einer Standardabweichung von 7,3 Jahren. Bei der Berechnung der Varianz (Quadrat der Standardabweichung) wurde durch 100 geteilt. Der Besitzer geht davon aus, dass das Alter normalverteilt ist.

$$\text{Varianz} = \frac{7.3 * \sqrt{100}}{\sqrt{99}} = 7.34 \quad z\text{-Wert} = 1.28 \text{ (bei gesuchten 80\%, je 10\% pro Rand } \Rightarrow 0.9, \text{ nachgeschlagen: } 0.8997)$$

$$\text{Bereich von 80\% Kunden: } (34.1 + 1.28 * 7.34 = 24.8) \text{ bis } (34.1 - 1.28 * 7.34 = 43.5)$$

Der Besitzerin eines Modegeschäftes möchte Ihr Sortiment künftig stark auf die Altersgruppe 50+ ausrichten. Sie hat den Eindruck, schon jetzt einen erheblichen Anteil von Kundinnen und Kunden über 50 Jahren zu haben, möchte sich aber darüber mehr Klarheit verschaffen.

Sie befragt daher 80 zufällig ausgewählte Kundinnen und Kunden (Stichprobe) zu ihrem Alter. Dabei ergibt sich ein durchschnittliches Alter von 54,5 Jahren bei einer Standardabweichung von 8,1 Jahren. Bei der Berechnung der Varianz (Quadrat der Standardabweichung) wurde durch 80 geteilt. Die Besitzerin geht davon aus, dass das Alter normalverteilt ist.

$$\text{Varianz} = \frac{8.1 * \sqrt{80}}{\sqrt{79}} = 8.15 \quad \text{Standardisieren} = \frac{50 - 54.5}{8.15} = -0.55 \quad z + 0.55 \rightarrow p = 0.7088 \rightarrow 70.88\% \text{ liegen über 50 Jahren}$$

$$\text{Kundenalter der tiefsten 10\%: } p = 0.9 \rightarrow z = 1.28 \quad x = 54.5 - 1.28 * 8.15 = 44.1 \text{ Jahre}$$

$$\text{Altersbereich 80\%: } (54.5 - 1.25 * 8.15 = 44.07) \text{ bis } (54.5 + 1.25 * 8.15 = 64.93)$$

### NORMALVERTEILUNG ZUR APPROXIMATION DER BINOMIALVERTEILUNG

In einer Fabrik wird ein Massenartikel hergestellt. Man weiss aus langer Erfahrung, dass es bei der Herstellung 4% Ausschuss gibt.

Es geht eine Bestellung über 10'000 Stück ein. Gemäss Vertragsbedingungen muss die Fabrik eine Strafe zahlen, wenn es in der Lieferung mehr als 4,7% Ausschuss hat. Andererseits erhält sie eine Prämie, wenn es weniger als 3,8% Ausschuss sind.

Die Direktion der Fabrik möchte die vorhandenen Chancen und Risiken abschätzen.

Dazu betrachtet sie die Zufallsvariable X = Anzahl defekter Stücke in der Lieferung.

Wahrscheinlichkeit Strafe zu zahlen:

$$\text{Erwartungswert} = 10000 * 0.04 = 400; \text{Varianz} = 10000 * 0.04 * (1 - 0.04) = 384; \text{Standardabweichung} = \sqrt{384} = 19.60$$

$$\text{Strafe ab 471 Stück, Stetigkeitskorrektur} = 470.5; \text{Standardisierung } z = \frac{470.5 - 400}{19.60} = 3.60; \quad p = 1 = 100\%; \quad \text{Wahrscheinlichkeit} = 0\%$$

$$\text{Wahrscheinlichkeit Prämie: Standardisierung } z = \frac{379.5 - 400}{19.60} = -1.05 = 1.05 \rightarrow p = 0.8531 = 85.31\% \rightarrow 100\% - 85.31\% = 14.69\%$$

In einer Fabrik werden Geräte hergestellt. Man weiss aus langer Erfahrung, dass es bei der Herstellung 5% defekte Geräte gibt.

Die Fabrik führt einen Auftrag über 5000 Geräte aus. Nach der Fertigstellung erfolgt eine Inspektion durch den Auftraggeber. Dabei wird eine Stichprobe von n = 300 Geräten gezogen und geprüft. Gemäss den Vertragsbedingungen dürfen höchstens 18 defekte Geräte dabei sein, sonst wird eine Konventionalstrafe fällig.

Die Direktion der Fabrik möchte das Risiko, die Strafe bezahlen zu müssen, abschätzen.

Dazu betrachtet sie die Zufallsvariable X = Anzahl defekter Stücke in der Stichprobe.

$$\text{Check ob Approximation durch Normalverteilung erlaubt ist: } \sqrt{np(1-p)} = \sqrt{300 * 0.05 * 0.95} = 3.775 > 3 \rightarrow \text{Ja}$$

Wahrscheinlichkeit Strafe zahlen zu müssen:

Erwartungswert =  $300 * 0.05 = 15$ ; Varianz =  $300 * 0.05 * (1-0.05) = 14.25$ ; Standardabweichung:  $\sqrt{14.25} = 3.775$

Standardisierung  $z = \frac{318.5-15}{3.775} = 0.93 \rightarrow p\text{-Wert} = 0.8238 = 82.38\%$  (keine Strafe)  $\rightarrow 100\% - 82.38\% = 17.62\%$  Strafe

Erhöhung der Strafe ab 25 defekten Geräten:

Stetigkeitskorrektur: 24.5; Standardisierung  $z = \frac{24.5-15}{3.775} = 2.25 \rightarrow p\text{-Wert} = 0.9941 = 99.41\% \rightarrow 100\% - 99.41\% = 0.59\%$

## KONFIDENZINTERVALLE

Die Geschäftsführerin eines Damenbekleidungsgeschäftes möchte sich schnell Informationen über den Durchschnittsbetrag, den die Kundinnen pro Einkauf ausgeben, verschaffen.

Sie zieht daher eine Stichprobe von 16 Kundinnen. Daraus ermittelt sie erwartungstreue Schätzungen für den Durchschnittsbetrag (CHF 1230) und die Standardabweichung (CHF 544) der ausgegebenen Beträge. Sie geht davon aus, dass die Beträge in der Grundgesamtheit (alle Kundinnen) normalverteilt sind.

Die Geschäftsführerin ist sich bewusst, dass die Punktschätzung des Durchschnittsbetrages unsicher ist.

Sie möchte daher das 95% - Konfidenzintervall für den Durchschnittsbetrag bestimmen, d.h. den Bereich, der mit 95% Wahrscheinlichkeit den Durchschnittsbetrag aller Kundinnen enthält.

Standardabweichung des Durchschnittswertes:  $\frac{544}{\sqrt{16}} = 136$ ; t-Verteilung weil  $n = 16 < 30$ ; 15 Freiheitsgrade;  $t = 2.1314$

Konfidenzintervall 95% =  $(1230 - 2.1314 * 136 = 940)$  bis  $(1230 + 2.1314 * 136 = 1520)$

Gesuchte Stichprobengröße wenn Intervall auf 1/3 verringert =  $3^2 * 16 = 144$

Der Geschäftsführer eines Herrenbekleidungsgeschäftes möchte sich schnell Informationen über den Durchschnittsbetrag, den die Kunden pro Einkauf ausgeben, verschaffen. Er zieht daher eine Stichprobe von 50 Kunden. Daraus ermittelt er erwartungstreue Schätzungen für den Durchschnittsbetrag (CHF 1170) und die Standardabweichung (CHF 489) der ausgegebenen Beträge. Er geht davon aus, dass die Beträge in der Grundgesamtheit (alle Kunden) normalverteilt sind.

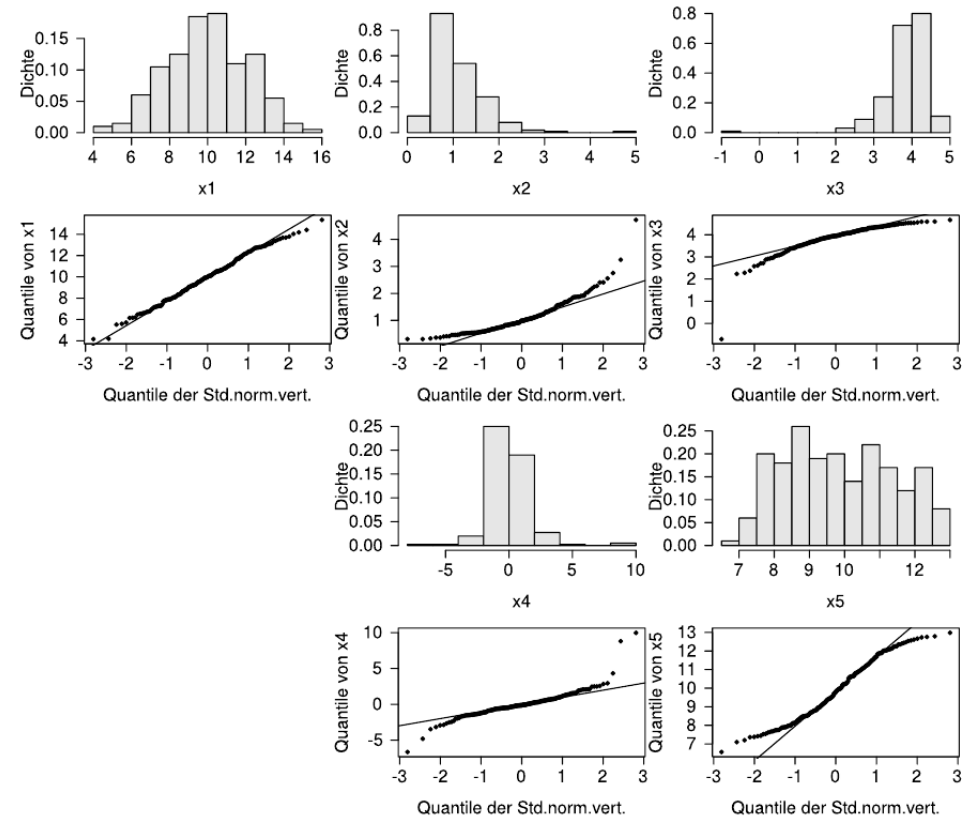
Der Geschäftsführer ist sich bewusst, dass die Punktschätzung des Durchschnittsbetrages unsicher ist.

Er möchte daher das 98% - Konfidenzintervall für den Durchschnittsbetrag bestimmen, d.h. den Bereich, der mit 98% Wahrscheinlichkeit den Durchschnittsbetrag aller Kunden enthält.

Standardabweichung =  $\frac{489}{\sqrt{50}} = 69.16$

Konfidenzintervall (mit je 1% unter/über Mitte)  $\rightarrow p = 0.99 = 2.33 z$  ergibt  $(1170 - 2.33 * 69.16 = 1009)$  bis  $(1170 + 2.33 * 69.16 = 1331)$

Stichprobengröße für 2/3 der Intervallgröße: neun viertel mal grösser =  $\frac{\sqrt{9}}{\sqrt{4}} = \frac{3}{2} \rightarrow n = \frac{9}{4} * 50 = 113$



# Cheatsheet WS – PVA 4

## BIVARIANTE UND MULTIVARIANTE ANALYSEN

### DARSTELLUNG MULTIVARIANTER DATENSÄTZE

#### Streudiagramm

Eignet sich sehr gut für einen ersten Überblick, basiert auf Koordinatensystem. Mindestens ab Intervallskalenniveau anwendbar.

- x-Achse: Merkmal 1

- y-Achse: Merkmal 2

#### Kontingenz- oder Korrelationstabelle

Verwendet zur Darstellung von gemeinsamen Häufigkeiten von zwei Merkmalen. Für beliebige Skalenniveaus einsetzbar.

- «Gemeinsame Häufigkeiten»: jede Kombination von Merkmalsausprägungen ist zu zählen.

- Zahl der Zeilen und Spalten muss überschaubar sein, kann durch Gruppierung oder Klassierung erreicht werden.

- Kopfspalte: Merkmal 1

- Kopfzeile: Merkmal 2

- Im Inneren: relative empirische Häufigkeiten, aussen: Summierung (=Randhäufigkeiten)

- Unten rechts: Wert n, Zahl der statistischen Einheiten

- Rand: Randhäufigkeiten

→ auch relative Häufigkeiten können eingetragen werden (z.B. jede Zeile = 100%); jeweils angeben, was genau dargestellt wird.

	Schwarz	Silber	Pink
Modell A	30	10	40
Modell B	9	9	2

	Schwarz	Silber	Pink	
Modell A	0.3	0.1	0.4	<b>0.8</b>
Modell B	0.09	0.09	0.02	<b>0.2</b>
	<b>0.39</b>	<b>0.19</b>	<b>0.42</b>	<b>1</b>

## TESTTHEORIE

### VORBEMERKUNGEN

Hier geht es um Verfahren, die Haltbarkeit von Behauptungen (Hypothesen) mathematisch fundiert prüfen (anhand Stichproben). Dabei gibt es eine Grundstruktur:

- Beschreibung des zu prüfenden **Sachverhalts**
- Formulierung der **Nullhypothese** und Alternativhypothesen
- Berechnung einer **Teststatistik**
- Auswahl der zugehörigen **Testverteilung**      **Signifikanzniveau**
- Festlegung des **Testniveaus** (aus diesem folgen über die Testverteilung Ablehnungs- und Annahmebereich)
- Füllen eines statistischen und fachlichen **Testurteils**

Überblick über verwendete Testverfahren:

- **Binomialtest**: wenn Hypothesen zu **Anteilswerten** geprüft werden (z.B. «Weniger als 50% der Kinogänger mögen salziges Popcorn.»)
- **t-Test** und **Varianzanalyse**: Hypothesen werden zu **Mittelwerten** geprüft.  
**t-Test** → bei **einer oder zwei Gesamtheiten** (z.B. «Die durchschnittliche Zahlungsbereitschaft für Popcorn liegt unter 3.-/100g.» oder «Die durchschnittliche Zahlungsbereitschaft für Popcorn ist in der Stadt 0.5.-/100g höher als auf dem Land.»)  
**Varianzanalyse** → bei **drei oder mehr Gesamtheiten** (z.B. «Die Zahlungsbereitschaft für Popcorn unterscheidet sich bei verschiedenen Filmgenres – Komödie, Liebesfilm, Thriller – deutlich voneinander.»)
- **X<sup>2</sup>-Test**: wenn Hypothesen zu **ganzen Verteilungen** geprüft werden. (z.B. «Die Zahlungsbereitschaft für Popcorn ist normalverteilt.»).  
Zugehörig: Fragen zum **Zusammenhang von Merkmalen** (z.B. «Die Intensität des Popcornduftes beeinflusst die Zahlungsbereitschaft.»)

### BINOMIALTEST AUF KONKRETE ANTEILSWERTE

Geeignet für Anteilswerte (z.B. «40% der Geschichtsstudenten mögen Stalin.»), einsetzbar für beliebige Skalenniveaus.

**Sachverhalt**: Beispielsweise die Frage, ob Verpackungsänderung eine erhebliche Abwanderung von Kunden verursachen würde.

**Nullhypothese (H<sub>0</sub>)** und Alternativhypothesen: müssen konkret formuliert werden, so dass mathematische Berechnungen möglich sind. Zu jeder Hypothese gibt es eine oder viele Alternativhypothesen («Arbeitsypothesen»). Diese können:

- entweder konkret formuliert sein (z.B. «H<sub>1</sub> ist 10% höher als H<sub>0</sub>.»)

- als Negierung der Nullhypothese vorliegen

Alternativhypothesen spielen bei Berechnungen aber vorerst keine Rolle.

Beispiel für «mindestens 30% wandern ab»:  $H_0: \theta \geq 0.3$

**Teststatistik**: Berechneter Wert, der zum Ergebnis des Tests führt. Kann aus komplizierten Formeln berechnet werden, in anderen Fällen reicht ein Zusammenzählen.

**Testverteilung**: Nullhypothese bezieht sich auf Gesamtheit aller Kunden, die Tests verwenden als Berechnungsbasis eine Stichprobe. Beispiel-Annahmen:

- wenn es insgesamt sehr viele Kunden gibt

- wenn 40 Testkunden zufällig ausgewählt werden

- wenn wir annehmen, dass H<sub>0</sub> stimmt

erhalten wir folgende Berechnung: 
$$f_b(x|0.3; 40) = \begin{cases} \binom{40}{x} * 0.3^x * (0.7)^{40-x} & \text{für } x = 0, 1, 2, \dots, 40 \\ 0 & \text{sonst} \end{cases}$$

→ **Formel für Binomialverteilung; summieren bis 1-alpha überschritten wird**

Dann: summieren, bis Wahrscheinlichkeit bei 0.95 (je nach alpha) bzw. die Zielwahrscheinlichkeit +- alpha beträgt.

Generell: die Testverteilung ist die Wahrscheinlichkeitsverteilung, die angibt welche Teststatistik man aus der Stichprobe mit welcher Wahrscheinlichkeit erhalten würde, wenn H<sub>0</sub> tatsächlich wahr wäre.

**Testniveau**: Ergebnisse in der Nähe des Erwartungswertes (θ \* n) sind am wahrscheinlichsten, aber kein Testergebnis ist völlig ausgeschlossen. Einige Ergebnisse sind aber sehr unwahrscheinlich. Um eine Grenze zu ziehen soll entsprechend das Testniveau α festgelegt werden.

**Fehlerklassen**: Wenn H<sub>0</sub> stimmt, aber das Ergebnis verworfen wird, haben wir einen α-Fehler (Fehler I. Art). Umgekehrt wäre es ein Beta-Fehler. Je kleiner α ist, umso höher steigt die Wahrscheinlichkeit für einen Beta-Fehler, die richtige Balance muss gesucht werden.

Die Aufteilung des Testniveaus wird in einem **Verwerfungsbereich** (oder Ablehnungsbereich), Symbol  $V$ , und einen

**Annahmebereich**,  $\bar{V}$ , eingeteilt. Hier muss man darauf achten, ob der Verwerfungsbereich **ein- oder zweiseitig** gelten soll.

Der Letzte Wert, der gerade noch zum Annahmebereich gehört, ist der **kritische Wert**.

**Testurteil**: Liegt die Teststatistik im Annahmebereich, behält man H<sub>0</sub> bei, ansonsten wird die Nullhypothese verworfen. Es können zwei Arten von Urteil gefällt werden:

- **Statistisches Urteil**: nüchterne Feststellung der Ergebnisse der Zahlen

- **Fachliches Urteil**: Einordnung in sachlichen Zusammenhang – erhält meist Interpretation der Bedeutung des Ergebnisses und ist an sich nicht mehr der Statistik zugehörig.

**Kritischer Wert**: der letzte Wert, der gerade noch zum Annahmebereich gehört.

Mit dem **p-Wert** kann berechnet werden, mit welcher Wahrscheinlichkeit man bei Gültigkeit von H<sub>0</sub> den Wert der Teststatistik erreicht oder überschritten hätte. Wenn der p-Wert kleiner als das Testniveau ist, wird H<sub>0</sub> verworfen.

**Binomialtests können einseitig (auf eine Seite ausschlagend, wie mehr als 25%) oder zweiseitig (=50% oder mehr oder weniger) sein. Zweiseitig: theta auf beide Seiten aufteilen! → immer ein Wert mehr nehmen («aufrunden»); nächster Wert ist safe.**

### BINOMIALTEST BEI GLEICHHEITSHYPOTHESEN

Als Nullhypothese braucht man immer eine konkret genug formulierte Angabe um damit rechnen zu können. In der Realität hat man aber statt exakten Anteilen Annahmen der Form «**A ist besser als B**».

Weil solche Aussagen nicht konkret genug sind, **formuliert man entsprechend um: «A ist genauso wie B»**.  
→ Also: die Nullhypothese muss nicht das gewünschte Ergebnis widerspiegeln, sondern hauptsächlich berechenbar sein.

## FESTLEGUNG DES TESTNIVEAUS

**Generelle Handlungsempfehlungen, da es keine absolut gültigen Regeln gibt:**

- Sich an übliche Praxis halten, typische Niveaus sind: 0.1%, 1%, 5%, 10%
- Bei zweiseitigen Tests sollte Testniveau möglichst symmetrisch aufgeteilt werden
- Nötige Abweichungen von diesen Konventionen sollen entsprechend begründet werden

**Abwägung der Konsequenzen von alpha- und beta-Fehlern:**

Der alpha-Fehler wird jeweils vom Testniveau vorgegeben, woraus Annahme- und Verwerfungsbereich resultieren. Wenn man eine Alternativhypothese als wahr unterstellt, lässt sich hieraus auch der beta-Fehler konkret berechnen. Der beta-Fehler wird in der Praxis aber oft nicht explizit berechnet.

Grundsätzlich gilt: je kleiner der alpha-Fehler, umso grösser der beta-Fehler.

Wenn der Test als Entscheidungsgrundlage verwendet wird, sollten die Konsequenzen beider Fehlertypen im Vorfeld geklärt werden. Wenn kein befriedigender Ausgleich gefunden werden kann, sollte der Stichprobenumfang erhöht werden.

**Berücksichtigung des Stichprobenumfangs:**

- **Wenn Stichprobenumfang klein ist**, können sehr kleine Testniveaus (0.1%, 1%) nicht sinnvoll realisiert werden, oder man erhält einen sehr extremen Verwerfungsbereich. Also: Testniveau auf 5% oder 10% setzen oder Stichprobenumfang erhöhen.
- **Wenn Stichprobenumfang gross ist**, sollte das Testniveau klein gehalten sein. Ansonsten wären die Ergebnisse übertrieben genau.

## T-TEST: VORAUSSETZUNGEN

Der t-Test ist ein Test zum Vergleich von Mittelwerten und ist eng verwandt mit dem Schätzen von Konfidenzintervallen. Konzept: für einen stichprobenbasierten Mittelwert wird getestet, ob er im Intervall der Nullhypothese liegt. Grundsätzlich gibt es drei verschiedene Testsituationen:

- **Einstichproben-t-Test: der Mittelwert einer Gesamtheit** wird getestet (z.B. «Hält sich das Aroma von Kartoffelchips im durchschnitt so lange wie vom Hersteller versprochen?»)
- **Zweistichproben-t-Test: die Mittelwerte aus zwei verschiedenen, unabhängigen Gesamtheiten** werden getestet, um welchen Betrag diese voneinander abweichen (z.B. «Unterscheidet sich die durchschnittliche Aromahaltbarkeit von Kartoffelchips bei Hersteller A von der Aromahaltbarkeit bei Hersteller B?»)
- **Paardifferenzentest: die Mittelwerte aus zwei verbundenen Gesamtheiten** werden getestet, um welchen Betrag diese voneinander abweichen (z.B. «Ist die durchschnittliche Aromahaltbarkeit bei Chips mit Paprika-, Barbecue- als auch Sour Cream-Geschmack durch Verwendung von Kartoffeln aus biologischem Anbau höher als bei herkömmlichen Kartoffeln mit denselben Geschmacksrichtungen?»)

Damit man den t-Test durchführen kann müssen folgende Voraussetzungen erfüllt sein:

- Mindestens Intervallskala
- Stichprobe muss zufällig und stochastisch unabhängig gezogen werden
- Das zu testende Merkmal soll annähernd normalverteilt sein, oder die Stichprobe muss ausreichend gross sein ( $n \geq 30$ )
- Im Zweistichproben-Fall gelten die Regeln für beide Mengen als Summe; die Varianzen beider Mengen sollten ungefähr gleich sein

## EINSTICHPROBEN T-TEST

**Sachverhalt:** Einsatz möglich, wenn es um Annahmen zu durchschnittlichen Merkmalsausprägung geht (z.B. «Das verfügbare Einkommen eines Haushaltes liegt im durchschnitt bei 1'500.-/Monat; «Beim Produkt A kann man pro Woche durchschnittlich 450 Einheiten absetzen.»).

**Nullhypothese:** Formulierung: der Mittelwert des Merkmals soll einem behaupteten Mittelwert entsprechen:  $H_0: \mu = \mu_0$   
Ähnliche mögliche Formulierungen: «Ergebnis soll mindestens/höchstens dem behaupteten Mittelwert entsprechen».

**Teststatistik:** Aus Grundgesamtheit wird eine Zufallsstichprobe  $n$  gezogen, berechnet werden **Stichprobenmittelwert** und **Stichprobenvarianz**:  
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad s^{*2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Die Teststatistik wird wie folgt berechnet: 
$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^{*2}}{n}}}$$
 → Teststatistik grösser als kritischer Wert ist, muss

**Nullhypothese verworfen werden!**

**Testverteilung:** Die Idee der Teststatistik ist es, den Stichprobenmittelwert  $\bar{x}$  zu standardisieren. Wenn die Nullhypothese stimmt, müsste dann die zur Teststatistik gehörende Verteilung **normalverteilt** sein. Wenn die Varianz ebenfalls aus der Stichprobe berechnet wird, muss eine **t-Verteilung** mit  $(n-1)$  Freiheitsgraden angewendet werden.

**Testniveau:** Gemäss üblicher Praxis:  $V = (-\infty; -x_c); \bar{V} = [-x_c; \infty)$  (einseitig oder zweiseitig möglich)

**Vorgehen:**

- Standardabweichung berechnen (erwartungstreu + zurück auf Grundgesamtheit)
- Teststatistik berechnen
- Kritischer Wert (z-Wert) aus Standardnormal- oder t-Test-Tabelle ablesen (aufgrund alpha; ein- bzw. zweiseitig)
- Wenn kritischer Wert < Teststatistik → verwerfen!

## ZWEISTICHPROBEN T-TEST

**Sachverhalt:** Einsatz, wenn es um die Frage geht, ob sich die durchschnittliche Ausprägung eines Merkmals bei zwei verschiedenen Gesamtheiten um einen bestimmten Betrag unterscheidet (z.B. «Das verfügbare Einkommen eines Haushaltes in der Region A liegt im Durchschnitt um 100.-/Monat über dem verfügbaren Einkommen in der Region B.»). Dazu gehört, dass der Unterschied auch Null betragen kann; höchstens/mindestens sind ebenfalls zulässig.

**Nullhypothese:** Mittelwert zum Merkmal der ersten Gesamtheit ( $\mu_1$ ) weicht vom Mittelwert zum Merkmal der zweiten Gesamtheit ( $\mu_2$ ) um Betrag  $\omega_0$  ab →  $H_0: \mu_1 - \mu_2 = \omega_0$

**Teststatistik:** Stichproben ziehen, dann Stichprobenmittelwerte und Stichprobenvarianzen nach bekanntem Schema berechnen.

$$t = \frac{\bar{x}^{(1)} - \bar{x}^{(2)} - \omega_0}{\sqrt{\frac{(n_1-1)s^{(1)*2} + (n_2-1)s^{(2)*2}}{n_1+n_2-2}}} * \sqrt{\frac{n_1 * n_2}{n_1 + n_2}} \quad \text{Achtung! } s(i)^{*2} \text{ ist Stichprobenvarianz!}$$

**Testverteilung:** Die Idee ist es, die Differenz der Stichprobenmittelwerte zu standardisieren. Wenn Nullhypothese stimmt und Mittelwerte der Gesamtheiten entsprechend abweichen, haben wir eine t-Verteilung (**Freiheitsgrad:  $n_1 + n_2 - 2$** ).

**Testniveau:** wie bekannt.

## PAARDIFFERENZENTEST

**Sachverhalt:** Auch hier geht es um Mittelwertunterschiede zwischen zwei Gesamtheiten, allerdings sind hier in irgendeiner Form Paare zwischen den Elementen beider Gesamtheiten. Typisch sind vorher-nachher-Untersuchungen oder über andere Eigenschaften (z.B. «Studierende haben nach Zeitmanagementtraining durchschnittlich vier Stunden mehr Freizeit pro Woche als vorher.»; «Sowohl blaue, als auch rote oder schwarze Lacke können ihren Glanz durch eine Wachsversiegelung länger bewahren.»). Solche Fälle werden «verbundene Gesamtheiten» genannt, wenn man hier den Zwei-Stichproben-t-Test anwenden kann dies zu nichtplausiblen Ergebnissen führen, da dieser auf unabhängigen Ereignissen aufbaut (Paarzusammenhänge werden 'verloren').

**Nullhypothese:** Für Paare werden die Unterschiede ihrer Messwerte betrachtet.  $H_0: \mu_d = \omega_0$

**Teststatistik:** Arithmetisches Mittel der Paardifferenzen sowie Stichprobenstandardabweichung werden bestimmt:



$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n x_i^{(2)} - x_i^{(1)} \quad s^{*2} = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^{(2)} - x_i^{(1)} - \bar{d})^2$$

$$t = \frac{\bar{d} - \omega_0}{\sqrt{\frac{s^{*2}}{n}}}$$

**Testverteilung:** Die Idee ist es, den Mittelwert der Paardifferenzen zu standardisieren. Die Verteilung ist eine t-Verteilung mit (n-1) Freiheitsgraden.

**Testniveau:** wie bekannt.

## X<sup>2</sup>-TEST: ALLGEMEINE VORGEHENSWEISE

Der X<sup>2</sup>-Test («Chi-Test») unterscheidet sich vom Binomial- und t-Test, indem er jeweils ganze Verteilungen und nicht nur zusammengefasste Daten betrachtet. Es gibt in prinzipiell für drei Situationen:

- **Anpassungstest:** untersucht wird, ob eine empirische Verteilung einer theoretisch erwarteten Verteilung entspricht (z.B. «Sind die vergebenen Punkte in der Statistiklausur normalverteilt?»)
- **Homogenitätstest:** untersucht, ob Verteilungen eines Merkmals in verschiedenen Gesamtheiten gleich sind (z.B. «Kann die Verteilung der erreichten Punkte in der Statistiklausur in den Jahrgängen A, B und C als vergleichbar angesehen werden?»)
- **Unabhängigkeitstest:** untersucht, ob zwei Merkmale voneinander unabhängig sind (z.B. «Ist die in der Statistiklausur erreichte Punktezah unabhängig von der Regelmässigkeit, mit der die Statistikvorlesung besucht wurde?»)

Die Merkmale können beliebig skaliert sein.

Allgemeine Vorgehensweise:

- **Sachverhalt** und **Nullhypothese:** hier unterscheiden sich die unterschiedlichen Testarten, wird spezifisch erklärt.
- **Teststatistik:** Die Teststatistik erfordert immer die Bestimmung von n<sub>o</sub> (o = observed) und n<sub>e</sub> (e = expected).
  - bei diskreten Merkmalen kann einfach das Auftreten gezählt werden
  - bei stetigen Merkmalen muss man Klassen bilden und abwägen:
    - für Klassenzahl und Klassengrenzen können folgende Faustregeln verwendet werden:
 
$$m \approx \sqrt{n} \text{ für } n \leq 100 \text{ für Klassenzahl, dann möglichst gleich breite Klassen verwenden.}$$
    - je mehr Klassen, umso höher der Unterschied zwischen empirischer und theoretischer Verteilung
    - Bestimmung der erwarteten Häufigkeiten sind je nach Testart unterschiedlich. Allgemein gilt, dass alle erwarteten Häufigkeiten mindestens den Wert 5 haben müssen.
- **Berechnung Teststatistik:** 
$$X_{emp}^2 = \sum_{j=1}^m \frac{(n_{oj} - n_{ej})^2}{n_{ej}}$$
 hier handelt es sich um eine Normierung, die die prozentuale Abweichung summiert. Je höher das Ergebnis, umso höher ist die Wahrscheinlichkeit, dass die Nullhypothese stimmt. Wenn über dem kritischen Wert, wird Ergebnis verworfen.
- **Testverteilung:** es wird angenommen, dass der Datensatz einer Zufallsstichprobe aus einer Gesamtheit gehört, in der das zu untersuchende Merkmal der theoretischen Verteilung entspricht. Die verwendete Test-Verteilung ist die X<sup>2</sup>-Verteilung. Ein Parameter: Zahl der Freiheitsgrade → Anzahl der quadrierten Normalvariablen, die addiert wurden.
- **Testniveau:** nach üblichem Vorgehen. Abweichungen nach unten oder oben wirken sich immer gegen die Nullhypothese aus. Das Testniveau wird aber immer nur rechts abgetragen.
- **Auswertung:** wenn Teststatistik grösser ist als kritischer Wert, wird Nullhypothese verworfen

## X<sup>2</sup>-TEST: ANPASSUNGSTEST

**Sachverhalt:** wird eingesetzt, wenn geprüft werden soll ob sich ein Merkmal entsprechend eines theoretischen Modells verteilt. Prinzipiell kann es um jede theoretische Verteilung gehen, meist geht es aber um die Gleichverteilung (z.B. «Ziehen die drei

nebeneinanderliegenden Restaurants ungefähr gleichviele Gäste an?») oder eine Normalverteilung (z.B. «Können die Tagesumsätze als annähernd normalverteilt unterstellt werden?»).

**Nullhypothese:** wird immer so formuliert, dass die empirische (beobachtete) Verteilung der theoretischen (erwarteten) Verteilung entspricht:  $H_0: F_{empirisch}(x) = F_{theoretisch}(x)$

**Teststatistik:**

- Erst Gruppen/Klassen berechnen
- Dann jeweils erwartete Häufigkeiten entsprechend der theoretischen Verteilung bestimmen, wobei man in der Summe auf Gesamtstichprobenumfang n kommen muss. Beispiel: n Messwerte bei 4 Ausprägungen → 50 / 4 = 12.5 für jede erwartete Häufigkeit
- Teststatistik entsprechend berechnen

**Testverteilung:** X<sup>2</sup>-Verteilung, Freiheitsgrade wie folgt bestimmen:

- Theoretische Häufigkeiten müssen summiert = n sein, also können nur m-1-viele als freie Zufallsvariablen gelten (der letzte Wert ergibt sich automatisch aus den anderen Zahlen)
- Für jeden Parameter der theoretischen Verteilung, der aus der Stichprobe berechnet werden muss, reduziert sich die Zahl der Freiheitsgrade jeweils um 1
- Formel für Freiheitsgrade: (m-1) – (Anzahl der aus dem Datensatz berechneten Parameter)

**Testniveau:** gemäss allgemeiner Vorgehensweise

## X<sup>2</sup>-TEST: UNABHÄNGIGKEITSTEST

**Sachverhalt:** wird eingesetzt, wenn geprüft werden soll ob zwei Merkmale voneinander unabhängig gelten können (z.B. «Hängen Wasserhärtegrad und der Verschleiss einer Waschmaschine zusammen?» oder «Ist die Zufriedenheit eines Menschen mit seinen Lebensumständen (un-)abhängig von seinem Einkommen?»).

**Nullhypothese:** wird so formuliert, dass es keinen Zusammenhang zwischen den beiden untersuchten Merkmalen gibt. Prinzip der Stochastik: «Wenn X von Y unabhängig wäre, dann wäre die gemeinsame Verteilung F(x,y) gleich dem Produkt der einzelnen Verteilungen:  $H_0: F(x,y) = F(x) * F(y)$

**Teststatistik:** am besten beide Merkmale mit ihren Ausprägungen in einer Matrix (Kontingenztafel) darstellen. Bei Gruppen oder Klassen nach allgemeiner Vorgehensweise vorgehen.

- Für die erwarteten Häufigkeiten gilt:
  - Wenn beide Merkmale unabhängig sind, muss sich gemeinsame Verteilung als Produkt der Randverteilung berechnen lassen.
  - rechnerisch wie Homogenitätstest: in jeder Zelle kann man erwartete Häufigkeit berechnen nach Formel: 
$$\frac{\text{Spaltensumme} \cdot \text{Zeilensumme}}{\text{Anzahl aller Messwerte}} \text{ pro Zelle}$$
 ergibt die theoretische Häufigkeit bei Unabhängigkeit
- Teststatistik wird dann gemäss allgemeiner Vorgehensweise berechnet
- Wenn Ergebnis kleiner als kritischer Wert, kann Nullhypothese beibehalten werden. Es wird von Unabhängigkeit ausgegangen.

**Testverteilung:** nach allgemeinem Vorgehen. Freiheitsgrade nach folgendem Schema bestimmen:

- Von den r Ausprägungen des Merkmals X können (r-1) freie Zufallsvariablen gelten
- Das selbe gilt für q Ausprägungen des Merkmals Y
- Zahl der Freiheitsgrade berechnet sich also: (m-1)(q-1)

**Testniveau:** allgemeine Vorgehensweise

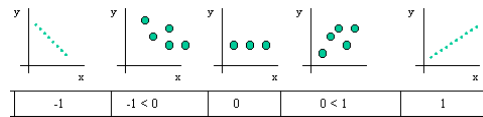
# Cheatsheet WS – PVA 5

## BIVARIANTE UND MULTIVARIANTE ANALYSEN

### KORRELATIONSANALYSE

Es geht darum Zusammenhänge aufzudecken – keine Kausalität!

Hilfreich, wenn man die Ursache für eine bestimmte beobachtbare Entwicklung finden will (Kausalität).



#### Metrischer Korrelationskoeffizient nach Bravais-Pearson

Der metrische Korrelationskoeffizient ( $r_{xy}$ ) drückt aus wie deutlich Messwerte auf einer Gerade liegen und ob wir eine positive oder negative Steigung haben.

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} * \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x * s_y}$$

- Zähler: «Kovarianz» - Mass für die gemeinsame Streuung der Variablen X und Y, alleine wenig aussagekräftig. Korrelation ist die standardisierte Kovarianz (gleich zu lesen).
- Nenner: «individuelle Varianzen» - bezieht sich auf Messwertreihen, durch Wurzel erhält man individuelle Standardabweichungen.
- Berechnung ab Intervallskalenniveau zulässig.
- Achtung: möglicherweise gibt es weitere, direktere Variablen, die wir aber nicht beobachtet haben.
- Korrelationskoeffizient ist für lineare Zusammenhänge konstruiert.
- Ausreisser sollten vor der Anwendung beseitigt werden.

#### Rangkorrelationskoeffizient nach Spearman-Pearson

Die Rangkorrelation ( $r_s$ ) für fortlaufende Ränge (1,2,3,...) wird verwendet, wenn mindestens ein Merkmal ordinal skaliert ist.

$$r_s = 1 - \frac{6 * \sum_{i=1}^n (rg(x_i) - rg(y_i))^2}{n * (n^2 - 1)}$$

- Bei gleichen Werten (die mehrmals gemessen werden, Gruppierungseffekt) müssen diese auch die gleichen Ränge erhalten («gebundene Ränge»). Diese erhalten die mittlere Rangzahl aller durch sie belegten Positionen.
- Ist gegenüber Ausreissern weniger empfindlich.
- Bei gebundenen Rängen führt Formel zu Ungenauigkeiten, Formel kann aber trotzdem verwendet werden. Wenn Optimierung gewünscht ist, kann Bravais-Pearson-Formel verwendet werden.

### REGRESSIONSRECHNUNG

Hiermit können Funktionen bestimmt werden, die sich an Punktwolken möglichst gut anpassen. So können diskret erhobene Werte in stetige Graphen umgewandelt werden.

#### Generelle Vorgehensweise bei der Regressionsrechnung

Gültig wenn alle Variablen mindestens intervallskaliert sind.

- Festlegung der zu berücksichtigenden Variablen (eine abhängige, eine oder mehrere unabhängige)
- Wahl des Regressionsansatzes (linear, logarithmisch, exponentiell, Potenzfunktion, etc.)
- Linearisierung des Regressionsansatzes (falls nötig)
- Aufstellen der Fehlerfunktion  $E = \sum_{i=1}^n (y_i - f(x_i))^2$
- Bildung der ersten partiellen Ableitungen von E nach allen Koeffizienten der Regressionsfunktion und Null setzen und
- Lösen des resultierenden Gleichungssystems.

#### Lineare Einfachregression

Der einfachste Regressionsansatz (=gesuchte Funktion) verwendet die Variablen X und Y und ist linear (=lineare Einfachregression). Der Buchstabe a ist der Schnittpunkt mit der y-Achse, b ist der Steigungsfaktor.

$$y = f(x) = a + b * x$$

#### Aufstellen der Fehlerfunktion (Methode der kleinsten Fehlerquadrate)

Die abweichenden Punkte werden als Fehler ( $e_i$ ) gewertet. Der Gesamtfehler E wird über folgende Fehlerfunktion bestimmt:

$$E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

#### Auswertung

Gefundene Minima entsprechen tatsächlichem globalen Minimum (hinreichende Bedingung für Extremwerte).

Für lineare Einfachregression können folgende allgemeingültige Formeln zur Berechnung von a und b verwendet werden:

$$b = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) * (y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$a = \frac{(\sum_{i=1}^n y_i) - b * (\sum_{i=1}^n x_i)}{n} = \bar{y} - b\bar{x}$$

Bestimmtheitsmass – Mass für die Güte der Anpassung: Korrelationskoeffizient im Quadrat. ( $R^2$ )

ist zwischen 0 und 1; je höher, umso linearer. Ergebnis in Prozent der Varianz (Erklärbarkeit durch linearen Zusammenhang), bezieht sich auf y-Variable. (1. Güteprüfung; 2. Güteprüfung im Nachhinein: passt Modell?)

Allgemein: x ist jeweils die unabhängige Variable.

Interpolieren: Schätzen von Zwischenwerten. Extrapolieren: schätzen über Regressionslinie hinaus.

Bei ordinalskalierten Grössen: Rangkorrelation vornehmen (ist Bester in x auch bester in y?)

## Cheatsheet WS – R

---

```
q()           # quit
library(Rcmdr) # start R commander
library()     # show installed packages; install.packages('xlsx')
?mean        # help for mean()
#            # comment
```

### Navigation

```
load("C:/R/R-3.6.2/tests/arima.rda") # load file
data("mtcars")                       # load data
View(mtcars)                         # view data
help(mtcars)                         # get data info
summary(mtcars)                     # get summary of table
summary(daten$groesse)              # get summary of variable
```

### Datei laden

```
# from RData
load("C:/Users/Lena
Papailiou/Desktop/FFHS/FS20/WSWP/datenAufgSatz01A/datensatz1.RData")
View(datensatz1)

# from excel
library(readxl)
test <- read_excel("C:/Users/Lena
Papailiou/Desktop/FFHS/FS20/WSWP/test.xlsx")
View(test)

# from csv
test <- read.csv("C:/Users/Lena Papailiou/Desktop/FFHS/FS20/WSWP/test.csv",
header=FALSE, sep=";")
View(test)
```

### Datei speichern

```
# as RData
save(test, file ='C:/Users/Lena
Papailiou/Desktop/FFHS/FS20/WSWP/test2.RData')

# as csv
write.csv(mtcarsNew, "C:/Users/Lena
Papailiou/Desktop/FFHS/FS20/WSWP/mtcarsNew", row.names = FALSE)

# as Excel
library(xlsx)
write.xlsx(as.data.frame(result$r), file = "results.xlsx")

# export image: export > save as image
```

### Parameter Handling

```
par()           # get list of parameters

par.default <- par(      # par = parameters; gets current parameters
  no.readonly=TRUE      # readonly-parameter will not be included
)                        # par.default saves standard params

par()           # set new parameters
```

```
mar=c(4.5,4.5,1,1),    # margin of graphic, from bottom clockwise
las=1,                 # las=label style, orientation of labels
                        # 0 = parallel to axis, 1 = always horizontal,
                        # 2 = vertical to axis, 3 = always vertical
cex.axis=1.75,         # (cex = character expansion ratio) scaling
                        # factor for scala labels
cex.lab=2               # (lab = label) scaling factor for labels
)

par(par.default)        # reset parameters to default
```

---

## EINFACHE VARIABLEN- UND TABELLENMANIPULATION

### Variablen erstellen und manipulieren

```
# create id
daten$id <- seq_len(nrow(daten))

daten$bmi <- daten$gewicht / (daten$groesse/100)^2

# create variable and fill column with categories
# right=FALSE: outside of interval
daten$bmiKat <- cut(daten$bmi, c(0, 18.5, 25, 50), labels =
c("untergewichtig", "normalgewichtig", "uebergewichtig"), right = FALSE)

# alter variables (values obtained by calculations do not depend on source)
daten$groesse[100] <- 176
daten$groesse[daten$id == 100] <- 176
daten$groesse <- daten$groesse/100

# rename variable
colnames(daten)[2] <- "gr"
colnames(daten)[colnames(daten) == "groesse"] <- "gr"

# delete variable
daten$falschVar <- NULL # alternative: NA
```

### Tabellen erstellen und manipulieren

```
# create matrix
m <- matrix(c(2.1, 1.8, 1.9, 2.4))

mtcarsT <- table(mtcars$cyl)
mtcarsT <- as.data.frame(mtcarsT)
mtcarsT$relFreq <- (mtcarsT$Freq/sum(mtcarsT$Freq))

# count rows
nrow(daten)

# sorting, nested example
datenNeu <- daten[order(daten$alter, -daten$groesse), ] # minus for desc

# rearrange variables
datenNeu <- daten[, c("id", "gewicht", "groesse", "bmi")]

lg3$Januar <- with(lg3, Januar/Total*100)
lg4 <- stack(lg3[,c("Januar", "Februar")]) # Spalten untereinander stacken
```

## Faktoren

```
# factors: mark variables as non-numeric (i.e. 6 for grades, not the number)
daten$geschlecht <- factor(daten$geschlecht, levels = c(0, 1), labels =
c("m", "f"))
is.factor(daten$geschlecht)
factor(daten$geschlecht)
verbrauchElektroauto <- within(verbrauchElektroauto, {
  typ <- as.factor(typ)
})
```

```
temperaturLang$zeit <- factor(
  temperaturLang$zeit,
  labels=c("Morgen", "Mittag", "Abend")
)
```

```
datensatz3$alter40 <- cut(datensatz3$alter, c(0, 40, 100), labels =
c('below40', '40plus'), right = FALSE)
is.factor(datensatz3$alter40)
```

## EINFACHE OPERATIONEN AUF DATEN

### Bedingungen

```
daten$braunInd <- ifelse(
  haarfarbe=="braun",      # condition to test
  1,                        # if true
  2,                        # else
)
```

### Funktionen anwenden (apply)

```
# 1=per row, 2=per col (second argument)
daten$pulsMittel <- apply(daten[, 2:8], 1, mean)
```

```
# tabular apply - per category
tapply(variable, factor, FUN = function, <arguments to function>)
tapply(daten$groesse, daten$geschlecht, FUN = mean, na.rm = TRUE)
```

### Teile von Daten isolieren

```
# get value [row, col]; leave col or row to get full line; 2:8 --> range
1,2,3,4,5,6,7,8
daten[1,2]
```

```
# use subset of data; condition: variables need to be after each other
daten$pulsMittel <- apply(subset(daten, select=pulsMontag:pulsSonntag), 1,
mean)
```

```
datenReduziert <- subset(
  daten,
  subset = (geschlecht == "weiblich") & (alter >= 50) & (alter != 60)),
  select = c(alter, groesse, gewicht) # Variablen auswählen
)
```

### Tabelle transponieren

```
colnames(temperatur)[colnames(temperatur) == 'tMo'] <- 'temperatur1'
colnames(temperatur)[colnames(temperatur) == 'tMi'] <- 'temperatur2'
colnames(temperatur)[colnames(temperatur) == 'tAb'] <- 'temperatur3'
temperaturLang <- reshape(
  temperatur,
  # data record
```

```
varying=c("t1", "t2", "t3"),      # old variables
timevar="zeit",                  # indicator variable
sep="",                          # separator between "t" and value
direction="long"                 # format-to-be
)
```

```
lg2 <- t(lg)
```

### Relative Häufigkeit nach Faktor

```
table(daten$haarf Farbe)
haarf Farbe.table <- table(daten$haarf Farbe)      # save in global variable
haarf Farbe.table      # show new table
prop.table(haarf Farbe.table)                      # show relative frequency
prop.table(haarf Farbe.table) * 100                # get relative frequency in %
prop.table(geschlHaar.table, 1) * 100              # get row % (1=row, 2=col)
```

```
> table(datensatz2$haarf Farbe)
```

```
      blond      braun      rot schwarz
         2          4          1          2
```

```
> prop.table(table(datensatz2$haarf Farbe))
```

```
      blond      braun      rot schwarz
0.2222222 0.4444444 0.1111111 0.2222222
```

## FUNKTIONEN

### Standard-Funktionen

```
x^(5/2)      # exponent
sqrt(x)      # square root
exp(x)       # natural exponent
log(x)       # natural ln
log10(x)     # log base 10
log2(x)      # log base 2
sin(x)       # sin in rad
cos(x)       # cos in rad
tan(x)       # tan in rad - same for asin(x), acos(x), atan(x)
round(x, n)  # round, for n < 0 for digits before comma
choose(n, k) # binomial coefficient
factorial(n) # factorial n!
```

```
mean(daten$groesse, na.rm = TRUE) # Arithmetisches Mittel
var(daten$groesse, na.rm = TRUE)  # Varianz (durch n-1)
sd(daten$groesse, na.rm = TRUE)   # Standardabweichung
median(daten$groesse, na.rm = TRUE) # Median
quantile(daten$groesse, probs = c(0.1, 0.9), na.rm = TRUE) # Quantile
```

### Zufallszahlen generieren

```
set.seed(<startzahl>) # same numbers can be generated again
sample(x=1:6, size=100, replace=TRUE) # random numbers 1,2,3,...,6
rbinom(n=100, size=10, p=0.2)          # random binomial numbers n = 10 and
p = 0.2
rhyper(nn=100, m=20, n=40, k=5)        # random hypergeometric numbers N =
20+40 = 60, M = 20 and n = 5
rpois(n=100, lambda=2)                 # random poisson numbers λ=2
runif(n=100, min=5, max=10)            # random static numbers [5, 10]
rnorm(n=100, mean=170, sd = 8)         # random normed numbers μ=170 and
standard derivant sd = 8
rexp(n=100, rate=2)                    # random exponential numbers λ=2
```

## DIAGRAMME (UNIVARIANT)

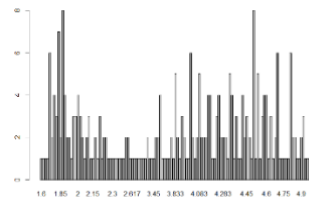
### Balkendiagramm

```
barplot(faithful$eruptions)
```

### Balkendiagramm nach absoluter Häufigkeit

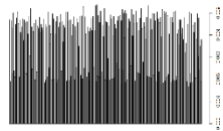
```
# first convert variable to factor
# bar diagram: first prepare variable with table(), then display with
barplot()
```

```
barplot(                                     # creation of bar diagram
  table(                                     # create data table
    daten$haarfارbe
  ),
  xlab="Haarfارbe",
  ylab="Absolute Häufigkeit"
)
```



### Balkendiagramm nach relativer Häufigkeit

```
# create table with relative frequency
barplot(
  prop.table(                               # extracts proportions from table
    table(
      daten$haarfارbe
    )
  ),
  xlab="Haarfارbe",
  ylab="Relative Häufigkeit"
)
```

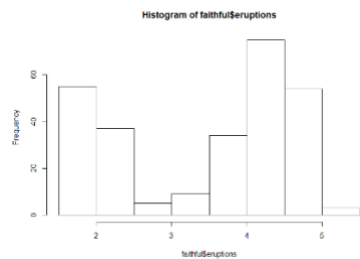


## DIAGRAMME (UNIVARIANT)

### Histogramm

```
# R will automaticall create classes of same width and will display absolute
frequency of each class
```

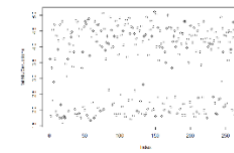
```
hist(
  daten$groesse
)
```



```
hist(
  daten$groesse,
  breaks=c(150, 160, 165, 170, 175, 180, 190),      # class borders
  freq=FALSE,
  # (freq = frequency) display density instead of frequency
  xlab="Grösse [cm]",
  ylab="Dichte",
  main=""      # (main = main title) do not display title
)
```

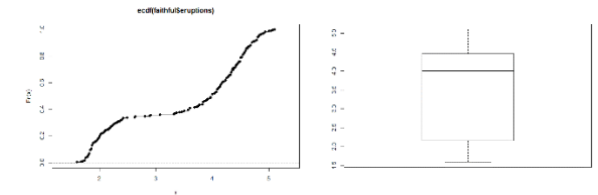
### Streudiagramm

```
plot(faithful$eruptions)
```



### Verteilungsfunktion

```
plot(
  ecdf(daten$bmi),
  pch=NA,
  main="",
  xlab="BMI [kg/m^2]",
  ylab="F"
)
# draws distribution function with ecdf()
# ecdf = empirical cumulative
# distribution function
# (pch = plotting character, NA = not
# available) no dots in diagram
```



### Boxplot

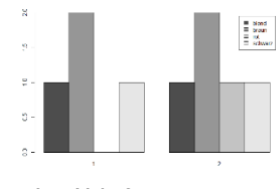
```
boxplot(
  daten$bmi,
  ylab="BMI [kg/m^2]"
)
```

## DIAGRAMME (BIVARIANT)

### Zwei diskrete Variablen

```
# usually grouped bar diagrams used, factorize first; create table, then
plot
```

```
barplot(
  table(
    daten$haarfارbe,
    daten$geschlecht
  ),
  beside=TRUE,
  legend.text=TRUE,
  xlab="Geschlecht",
  ylab="Absolute Häufigkeit"
)
# creates bar plot
# creates cross table
# displayed variable
# grouping variable
# display bars beside, not above
# add legend for displayed variable
```

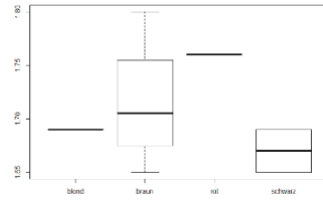


```
# same, with relative frequency
barplot(
  prop.table(
    table(
      daten$haarfارbe,
      daten$geschlecht
    )
  ),
  beside=TRUE,
  legend.text=TRUE,
)
```

### Stetige + Diskrete Variable

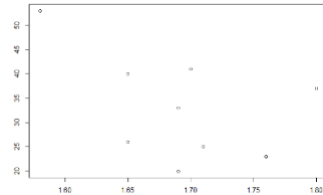
```
# group by discrete variable
# tilde: y ~ x; y is displayed depending on x
```

```
boxplot(
  daten$bmi ~ daten$geschlecht,
  xlab="Geschlecht",
  ylab="BMI [kg/m^2]"
)
```



### Stetige + Stetige Variable

```
plot(
  daten$groesse,
  daten$bmi,
  xlab="Körpergröße [cm]",
  ylab="BMI [kg/m^2]"
)
```



### Stetige + Stetige + Stetige Variable

```
pairs(
  daten[c("groesse", "gewicht", "bmi")],
  labels=c("Größe [cm]", "Gewicht [kg]", "BMI [kg/m^2]")
)
```

### Gruppierung innerhalb Streudiagramm

```
# plot multiple continuous variables, grouped by discrete variable
# pch=1 is circle; pch=19 is circle disk
```

```
plot(
  daten$groesse,
  daten$gewicht,
  pch=c(1,19)[daten$geschlecht],
  xlab="Größe [cm]",
  ylab="Gewicht [kg]"
)
```

### WAHRSCHEINLICHKEITEN UND QUANTILE

# prob/density	# distribution	# quantile	# random
dhyper(x, m, n, k)	phyper(q, m, n, k)	qhyper(p, m, n, k)	rhyper(nn, m, n, k)
dbinom(x, size, prob)	pbinom(q, size, prob)	qbinom(p, size, prob)	rbinom(n, size, prob)
dpois(x, lambda)	ppois(q, lambda)	qpois(p, lambda)	rpois(n, lambda)
dunif(x, min=0, max=1)	punif(q, min=0, max=0)	qunif(p, min=0, max=0)	runif(n, min=0, max=0)
dnorm(x, mean=0, sd=1)	pnorm(q, mean=0, sd=1)	qnorm(p, mean=0, sd=1)	rnorm(n, mean=0, sd=1)
dexp(x, rate=1)	pexp(q, rate=1)	qexp(p, rate=1)	rexp(n, rate=1)
dt(x, df)	pt(q, df)	qt(p, df)	rt(n, df)
dchisq(x, df)	pchisq(q, df)	qchisq(p, df)	rchisq(n, df)
df(x, df1, df2)	pf(q, df1, df2)	qf(p, df1, df2)	rf(n, df1, df2)

# hyper geom.	
# binomial	
# poisson	
# uniform distribution/stetige gleichverteilung	
# normal distribution	
# exponential	
# t distribution	
# chisquare distribution	
# f distribution	

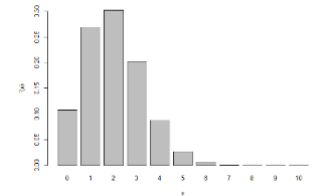
# X	als Zufallsvariable
# f(x) = P(X=x)	als Wahrscheinlichkeitsverteilung (diskret)
# F(x) = P(X <= x)	als Dichtefunktion (diskret)
# f = f(x)	als Dichtefunktion (stetig)
# F(x) = P(X <= x)	als Verteilungsfunktion (stetig)

```
# d() = density; p() = probability; q() = quantile; r() = random
# f = f(x) as probability or density function
# F = F(x) as distribution function
```

### Balkendiagramm aus Zufallszahlen (diskret)

```
xVect <- 0:10 # vector of numbers (1, 2, ..., 10)
fVect <- dbinom(xVect, size=10, prob=0.2) # f(x) for x for 1:10
```

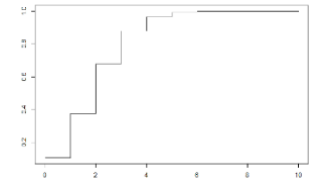
```
barplot(
  fVect,
  names.arg=xVect, # scaling x-axis
  xlab="x",
  ylab="f(x)",
  main=""
)
```



### Treppendiagramm aus Zufallszahlen (diskret)

```
xVect <- 0:10 # vector of numbers (1, 2, ..., 10)
FVect <- pbinom(xVect, size=10, prob=0.2) # F(x) for x for 1:10
```

```
plot(
  xVect,
  FVect,
  type="s", # stair diagram
  xlab="x",
  ylab="F(x)",
  main=""
)
```

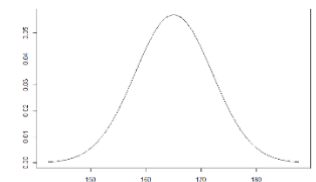


### Wahrscheinlichkeitsfunktion aus Zufallszahlen (stetig)

```
xVect <- seq(from=142, to=188, by=0.1)
fVect <- dnorm(xVect, mean=165, sd=7)
```

```
plot(
  xVect,
  fVect,
  type="l", # line diagram
  xlab="x",
  ylab="f(x)",
  main=""
)
```

```
# vector (142.0-188.0)
# f(x) for x f
```



### Dichtefunktion aus Zufallszahlen (stetig)

```
xVect <- seq(from=142, to=188, by=0.1)
FVect <- pnorm(xVect, mean=165, sd=7)
```

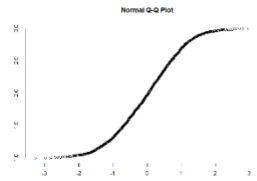
```
plot(
  xVect,
  FVect,
  type="l", # line diagram
  xlab="x",
  ylab="F(x)",
  main=""
)
```

```
# vector (142.0- 188.0)
# F(x) for x
```

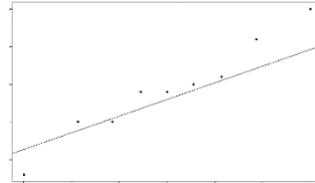


## Q-Q-Plots

```
qqnorm(zufallsdaten$unimean32, pch = 1, frame = FALSE)
```



```
qqnorm(
  daten$groesse,
  main = "",
  xlab = "Quantile der Std.norm.vert.",
  ylab = "Quantile der Körpergrösse",
  pch = 20
)
```



```
# will insert line in above diagram
```

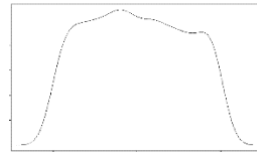
```
qqline(
  daten$groesse
)
```

```
# multiple diagrams in one
```

```
par(mfrow = c(2, 3))      # frame with two rows and 3 colums
plot(...)                 # partial diagram top left
plot(...)                 # partial diagram top center
plot(...)                 # partial diagram top right
frame()                   # empty diagram (space)
plot(...)                 # partial diagram bottom center
plot(...)                 # partial diagram bottom right
par(mfrow = c(1, 1))      # from here, single diagrams again
```

```
# bell curve
```

```
plot(density(zufallsdaten$univar1))
```



## KONFIDENZINTERVALLE

### One Sample t-test

#### Konfidenzintervall (arithmetisch)

```
t.test(
  daten$groesse,
  conf.level=.95
)
data: daten$groesse
t = 79.338, df = 8, p-value = 7.102e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.643037 1.741408
sample estimates:
mean of x
 1.692222
```

#### Konfidenzintervall (Verhältnis)

```
daten$brilleInd <- ifelse(
  daten$sehhilfe=="Brille",
  1, # if true
  2 # else
)

daten$brilleInd <- factor( # define "brilleInd" as factor
  daten$brilleInd,
  levels=1:2,
  labels=c("Brille", "keine Brille")
)
```

```
table( # check with contingency table
  daten$sehhilfe,
  daten$brilleInd
)
```

	Brille	keine Brille
1	5	0
2	0	5

```
# hypothesis test with prop.test(), works with generated table and absolute
frequencies
# then generates confidence interval; confidence level at [...]
# R will use a specific function which will not equal the one from our book
brilleTab <- table(daten$brilleInd) # create table
prop.test(
  brilleTab,
  conf.level=0.95
)
```

### 1-sample proportions test without continuity correction

```
data: brilleTab, null probability 0.5
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2365931 0.7634069
sample estimates:
p
0.5
```

```
remove(brilleTab)
```

```
# cleanup
```

## T-Test

### T-Test

```
t.test(z2, conf.level=0.95)
```

One sample t-test test name

```
data: daten$groesse
t = 79.338, df = 8, p-value = 7.102e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.643037 1.741408
sample estimates:
mean of x
 1.692222
```

Teststatistik  
Freiheitsgrade (muss keine natürliche Zahl sein)  
p-Wert: ist p kleiner oder gleich alpha, wird  
Nullhypothese verworfen  
Konfidenzintervall

## HYPOTHESENTESTS

### Hypothesentest (Binomial)/Anteilstest

```
# first, indicator variable is created
daten$braunInd <- ifelse(
  daten$haarfärbefarbe == "braun",
  1, # if true
  2 # else
)
daten$braunInd <- factor( # define "braunInd" as factor
```



```

    daten$braunInd,
    labels = c("braun", "nicht braun"))
table(      # cross-check
    daten$braunInd,      braun      blond braun rot schwarz
    daten$haarfarbe      nicht braun  2      0      1      2
)

```

```

# then we create a table to get rel freq and make binomial test
braunTab <- table(      # create table for rel freq
    daten$braunInd
)

```

```

binom.test(      # binom test
    braunTab,
    p = 0.6,      # percentage to test on
    alternative = "two.sided",
    conf.level = 0.95
)

```

#### Exact binomial test

```

data: braunTab
number of successes = 4, number of trials = 9, p-value = 0.4984
alternative hypothesis: true probability of success is not equal to 0.6
95 percent confidence interval:
 0.1369957 0.7879915
sample estimates:
probability of success
 0.4444444

```

#### Chiquadrat-Test

```

# make the actual test
braunTab <- table( # Tabelle erzeugen (wird von R nicht angezeigt)
    daten$braunInd
)

```

```

prop.test(      # percentage test with normal derivation
    braunTab,
    p = 0.6,
    alternative = "two.sided",
    conf.level = 0.95
)

```

#### 1-sample proportions test with continuity correction

```

data: braunTab, null probability 0.6
x-squared = 93.126, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.4189204 0.4814685
sample estimates:
p
 0.45

```

#### T-Test (einseitig, eine Stichprobe)

```

t.test(
    daten$groesse,
    alternative='greater', # right sided test
    mu=170,               # mu value
    conf.level=.9         # confidence level
)

```

#### One sample t-test

```

data: daten$groesse
t = -7890.9, df = 8, p-value = 1
alternative hypothesis: true mean is greater than 170
90 percent confidence interval:
 1.662429      Inf
sample estimates:
mean of x
 1.692222

```

#### T-Test (zweiseitig, zwei Stichproben)

```

t.test(
    groesse~geschlecht, # groesse depends on geschlecht
    alternative="two.sided",
    conf.level=.95,
    data=daten
)

```

#### Welch Two Sample t-test

```

data: groesse by geschlecht
t = -2.3236, df = 6.3387, p-value = 0.05689
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.164186464 0.003186464
sample estimates:
mean in group 1 mean in group 2
 1.6475      1.7280

```

#### T-Test (zweiseitig, zwei abhängige Stichproben)

```

t.test(
    daten$blutdruckVor,
    daten$blutdruckNach,
    alternative='two.sided',
    conf.level=.95,
    paired=TRUE      # paired test
)

```

#### Paired t-test

```

data: daten$groesse and daten$alter
t = -8.8086, df = 8, p-value = 2.17e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -39.64402 -23.19375
sample estimates:
mean of the differences
 -31.41889

```

#### Chiquadrat-Anpassungstest

```

chisq.test(
    haarTab,
    p=c(0.4,0.6)      # must add up to 1
)

```

#### Chi-squared test for given probabilities

```

data: braunTab
X-squared = 10.417, df = 1, p-value = 0.001249
chisq.test(x, p=p)      # other example

```

### Chi-Quadrat-Abhängigkeitstest nach Pearson

```
geschlHaarTab <- table(      # create cross table
  daten$geschlecht,         # row variable
  daten$haarfarbe           # col variable
)
chisq.test(
  geschlHaarTab
)

Pearson's Chi-squared test
```

```
data: geschlHaarTab
X-squared = 34.695, df = 30, p-value = 0.2539
```

### Test auf Normalverteilung -> bei grossem p-Wert haben wir Normalverteilung

```
normalityTest(~V1, test="pearson.test", data=Dataset)
```

### KORRELATIONEN & REGRESSIONSTESTS

#### Übersicht

```
# get overview
summary(daten)

# print scatter plot
plot(
  daten$groesse,           # x-axis
  daten$gewicht,          # y-axis
  xlab="Körpergrösse [cm]",
  ylab="Gewicht [kg]"
)

# get correlation coefficient after pearson
cor(
  daten$groesse,
  daten$gewicht
)
```

#### Korrelationstest nach Pearson

```
# get regression test preparation after pearson's product-moment
cor.test(
  daten$groesse,
  daten$gewicht
)

Pearson's product-moment correlation

data: datensatz3$alter and datensatz3$bmi
t = -0.47714, df = 8, p-value = 0.646
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7205080  0.5174806
sample estimates:
      cor
-0.1663437
```

### Lineare Regression

```
# ----- linear regression -----
# do linear regression test
gewichtVsGroesse <- lm(      # lm = linear model
  gewicht~groesse,          # show weight depending on height
  data=daten
)
summary(gewichtVsGroesse)    # get results of lm
```

```
call:
lm(formula = gewicht ~ alter, data = daten)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-5.2305 -4.5029 -1.4171  0.6762 12.7657
```

```
Coefficients:
(Intercept) 77.3125 a      6.8553 11.278 9.63e-06
alter      -0.2724 b      0.2086 -1.306 0.233
```

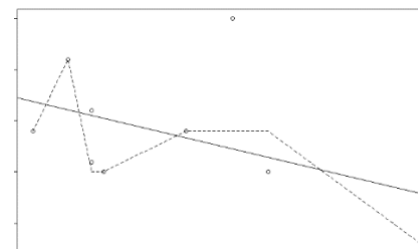
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.209 on 7 degrees of freedom
Multiple R-squared:  0.1959, Adjusted R-squared:  0.08103
F-statistic: 1.705 on 1 and 7 DF, p-value: 0.2329
```

Regressionsgerade =  $f(x) = a + b \cdot x$  ( $x = \text{alter}$ )  $b = \text{Steigung}$   
 $R^2 = \text{so viele Prozent (19.59)}$   
der Variation ist durch Fehler für Güteprüfung:  
linearen Zusammenhang Fehler müssen  
erklärbar Normalverteilt sein und  
Bei vielen unabhängigen gesamthaft etwa  
Variablen dieselbe Varianz haben  
ausgekräftiger (Symmetrie beachten)  
kleine p-Werte sprechen gegen Nullhypothese  
bei einer unabhängigen  
Variable (alter) ist p-Wert =  
Regressionskoeffizient =  
Korrelationskoeffizient

```
# coefficients: our regression coefficients a and b (column estimates)
# multiple r-squared: in percent/100, shows if variation is correlated with
depending variable (height)
# small p-value (last col) points against 0 hypothesis
```

```
# plot result with smooth curve
plot(      # print scatter plot
  daten$groesse,
  daten$gewicht,
  xlab="Körpergrösse [cm]",
  ylab="Gewicht [kg]"
)
abline(gewichtVsGroesse) # draw regression line
lines(      # draw smooth curve
  lowess(      # calculate smoothe curve
    daten$groesse,
    daten$gewicht,
    f=0.5      # amount of accounted points (f = fraction)
  ),
  lty="dashed" # lty = line type
)
```



Linie: Regressionsgerade  
Gestrichelt: Glättungskurve

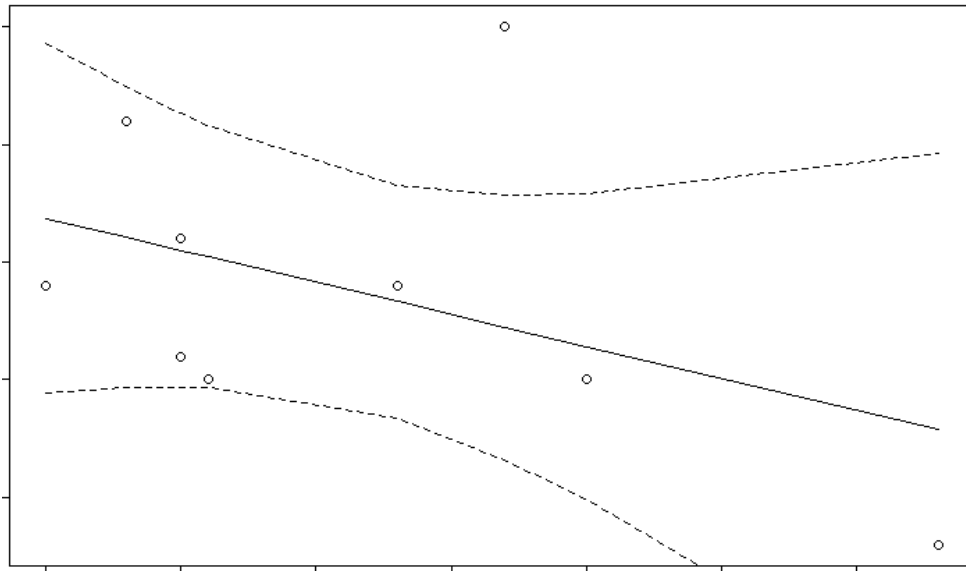
### Regressions-Koeffizienten-Test

```
# create temp data, sorted by height
tempDaten <- regressionsdaten[order(regressionsdaten$groesse), ]

predictMatrix <- predict(          # generate predicted values
  gewichtVsGroesse,               # ... for this model
  interval="confidence",          # calculate conf. interval
  tempDaten                      # predict height by weight in tempDaten
)

plot(                             # print scatter diagram, sorted by height
  tempDaten$groesse,
  tempDaten$gewicht,
  xlab="Körpergröße [cm]",
  ylab="Gewicht [kg]"
)

#predicted values and conf. interval (matlines = matrix lines)
matlines(
  tempDaten$groesse,              # independant variable x
  predictMatrix,                 # depending y values
  lty=c("solid", "dashed", "dashed"), # line types
  col="black"                   # line colors
)
```



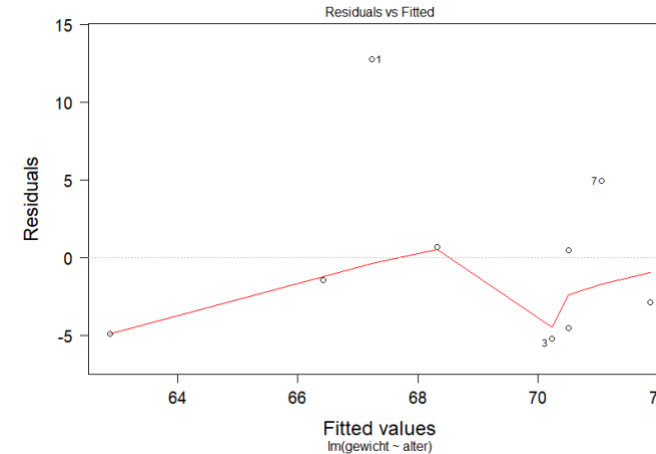
Gestrichelt: Konfidenzintervall mit eingerechnet. Regressionsgerade kann sich innerhalb dieses Bereichs befinden.

### Güte-Prüfung des Modells

```
# check residuals (difference between real and predicted value)
# residuals are normally distributed if test succeeded (and take whole
bandwidth and have similar variance)
```

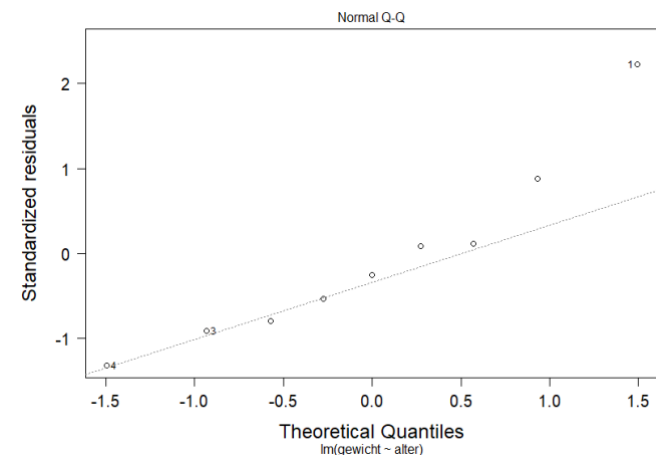
```
# r will make further tests and save them to model. do following to test:
origParam <- par(                # save old parameter and set new
  mar=c(6,4.5,2,1),             # margin around graph
  las=1,                         # orientation of scala labeling
  cex.axis=1.25,                # scaling factor for scaling label
  cex.lab=1.5                    # scaling factor for axis label
)
```

```
plot(gewichtVsGroesse, which=1)  # diagnostic plot 1
```



Residuen vs.  
Vorhergesagte Werte

```
readline(prompt="Weiter mit <Enter>") # pause script
plot(gewichtVsGroesse, which=2)      # diagnostic plot 2
```



QQ-Plot gegenüber  
Normalverteilung mit  
standardisierten  
Residuen.

```
par(origParam)                  # reset
old parameter
```

```
# residuals vs. fitted: residuals should be symmetrical around 0 (straight
line)
# for further optimizing, variables (like height) could be squared.
```