

semestra

Weitere Files findest du auf www.semestra.ch/files

DIE FILES DÜRFEN NUR FÜR DEN EIGENEN GEBRAUCH BENUTZT WERDEN.
DAS COPYRIGHT LIEGT BEIM JEWEILIGEN AUTOR.

 Einführung in die **Statistik** für Wirtschafts- und Sozialwissenschaften

Vorlesung: Prof. Dr. Lutz Dümbgen, Universität Bern

4 ECTS-P

Teil 1: Beschreibende Statistik

1. Allgemeines / Definitionen	S.02
2. Beschreibung kategorialer Merkmale	S.02
3. Beschreibung numerischer Merkmale	
3.1. Graphische Darstellungsmöglichkeiten numerischer Merkmale	
3.1.1. Die (empirische) Verteilungsfunktion	S.03
3.1.2. Histogramme	S.04
3.2. Quantile und Quartile	S.04
3.3. Lageparameter	S.05
3.4. Skalenparameter	S.06
3.5. Lorenzkurve und Gini-Index	S.07
3.6. Formparameter	S.08
4. Simultane Beschreibung zweier Merkmale	
4.1. Kontingenztafeln und Vierfeldertafeln	S.08
4.2. Box-Whisker-Plots	S.09
4.3. Regression und Korrelation	S.10

Teil 2: Wahrscheinlichkeitsrechnung und statistische Modelle

5. Wahrscheinlichkeitsrechnung	
5.1. Grundlagen / Definitionen	S.13
5.2. Wahrscheinlichkeitsverteilungen	
5.2.1. Diskrete Verteilungen	S.14
5.2.2. LaPlace-Verteilungen	S.14
5.2.3. Rechenregeln für Wahrscheinlichkeiten	S.14
5.2.4. Siebformel	S.15
5.2.5. Bonferroni-Ungleichungen	S.15
5.3. Bedingte Wahrscheinlichkeiten	S.16
5.3.1. Bayessche Formel	S.17
5.4. Stochastische Unabhängigkeit	S.18
5.5. Spezielle Verteilungen	
5.5.1. Hypergeometrische Verteilung	S.19
5.5.2. Binomialverteilung	S.20
5.5.3. Geometrische Verteilung	S.21
5.5.4. Poissonverteilung	S.21

Teil I: Beschreibende Statistik

1. Allgemeines / Definitionen (Skript S. 11-12)

Definitionen:

- Datensatz = **Stichprobenumfang = n**
- Beobachtungen umfassen eine oder mehrere Variablen (Stichprobenwerte, Merkmale)
- **Stichprobenwerte = X_i** (und Y_i bei zwei Variablen), Gesamtheit aller Stichprobenwerte:

$$\sum_{i=1}^n X_i$$

Variablentypen:

- **Numerische Variablen** nehmen eine objektive Bedeutung an; ihre Anzahl möglicher Ausprägungen ist theoretisch unbegrenzt; Beispiele: Alter, Körpergröße, Monatsmiete usw.
- **Kategorielle Variablen** können endlich viele Werte annehmen; ihre Anzahl möglicher Ausprägungen ist somit begrenzt; Beispiele: Geschlecht, Geburtsmonat, Zufallsziffer aus 1-10
- **Ordinal(skaliert)e Variablen** sind *kategorielle Variablen*, deren Kategorien durch Zusammenfassung möglicher Stichprobenwerte künstlich geschaffen wurden; Beispiel: Unterteilung Raucherstatistik in „oft“ = 2 (d.h. z.B. $X_i > 10$ Zigaretten pro Tag), „selten“ = 1 ($0 < X_i < 10$ Zigaretten pro Tag) und „nie“ = 0 ($X_i = 0$ Zigaretten pro Tag)

2. Beschreibung kategorieller Merkmale (Skript S. 12-13)

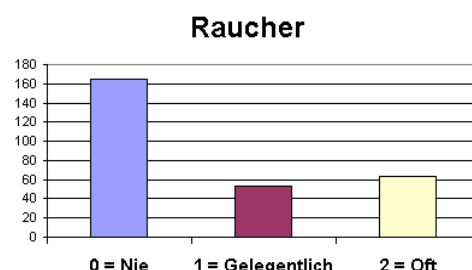
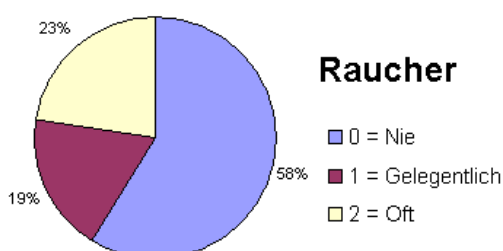
Absolute Häufigkeit (h_j) und relative Häufigkeit (f_j):

- Ausgangslage: ein Stichprobenumfang (n) enthält die Stichprobenwerte X_1, X_2, \dots, X_n , welche verschiedenen Kategorien x_1, x_2, \dots, x_L zugeordnet werden müssen
- Die **absolute Häufigkeit (h_j)** einer bestimmten Kategorie (x_j) ist die Anzahl der Stichprobenwerte, die in diese Kategorie gehören (d.h. Anzahl aller Beobachtungen mit Wert x_j); Die **relative Häufigkeit (f_j)** = absolute Häufigkeit geteilt durch n (Stichprobenumfang)

$$f_j := \frac{h_j}{n}$$

Graphische Darstellung von kategoriellen Variablen

- Möglichkeit 1: bei **Balkendiagramm** entspricht Höhe der Balken entweder h_j oder f_j
- Möglichkeit 2: bei **Kuchendiagramm** ist Fläche pro Kategorie proportional zu h_j bzw., f_j



3. Beschreibung numerischer Merkmale (Skript S. 14-36)

Grundlage: **Sortieren der Stichprobenwerte ($X_i \rightarrow X_{(i)}$)**

- a. da die Reihenfolge der Stichprobenwerte X_i i.d.R. keine Rolle spielt, kann man diese – oft nach deren Grösse – sortieren. So erhält man die **Ordnungsstatistiken $X_{(i)}$** . Fortan gilt:

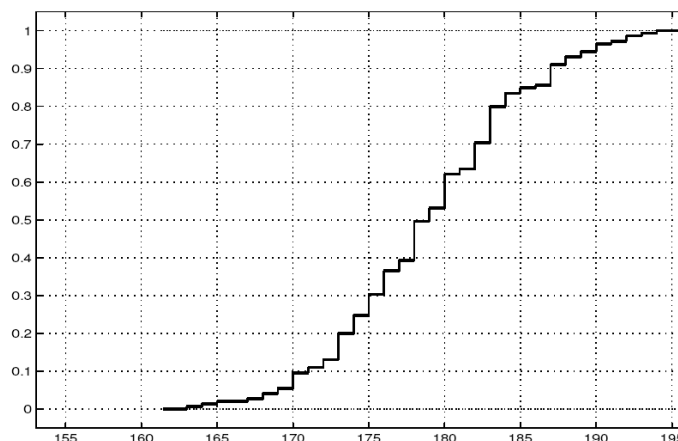
$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

3.1. Graphische Darstellungsmöglichkeiten numerischer Merkmale

3.1.1. Die **(empirische) Verteilungsfunktion** (Skript S. 14-18)

Definition

- b. Die empirische Verteilungsfunktion (**F**) ist eine **monoton wachsende Treppenfunktion** auf der man ablesen kann, wie die X-Werte in der Stichprobe verteilt sind
- c. F **basiert auf** den geordneten Stichprobenwerten (= **Ordnungsstatistiken**)
- d. Die Abszisse listet mögliche Stichprobenwerte auf; Die Ordinate zeigt den bei einem gewissen Wert erreichten relativen Anteil am Stichprobenumfang (n) auf
- e. **F steigt** dabei **sprunghaft** immer dann an, wenn ein gewisser möglicher Wert tatsächlich unter den Beobachtungen vorhanden ist. Der **kleinstmögliche Anstieg** beträgt **1/n**, kann aber selbstredend grösser sein, wenn mehrere Beobachtungen denselben Wert ausweisen (3 Beobachtungen = 3/n)
- f. **Resultat: der empirischen Verteilungsfunktion kann man entnehmen, wie gross der Anteil der Beobachtungen ist, die mindestens einen gewissen (frei bestimmbar) Wert aufweisen**



Tipps und Tricks

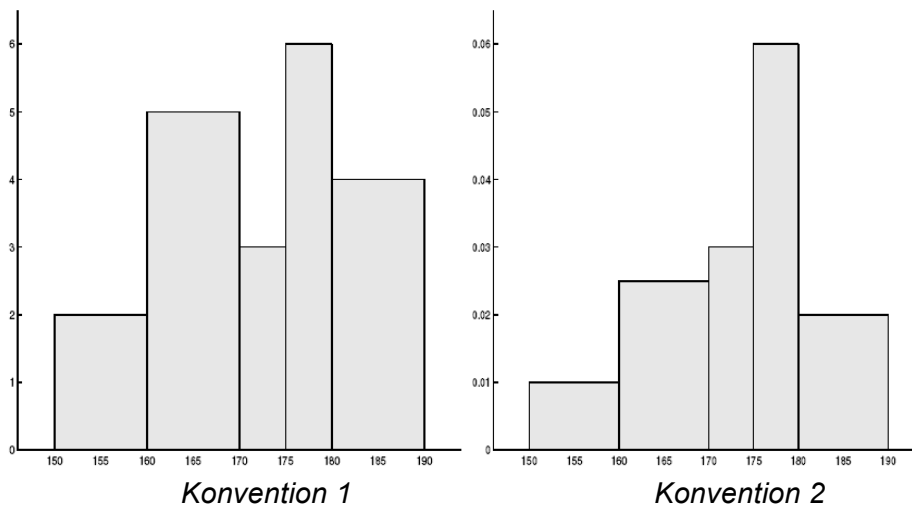
- g. will man **ablesen**, wie hoch der erreichte **Anteil** bei einem bestimmten Wert (= "Schranke" ist) und findet dort einen Anstieg der Funktion vor (in Beispiel z.B. bei 170), so muss der massgebende Anteil beim **oberen Ende des Anstiegs** abgelesen werden (in Beispiel also 0.1 und *nicht* 0.05)

- h. will man **ablesen**, bei welchem **Wert („Schranke“)** ein Anteil erreicht ist (z.B Median = 0.5) und findet dabei eine gerade Linie vor, so ist der **hintere Teil der Linie** gesucht (hier: 179, *nicht* 178)
- i. **Repetition vor Prüfung**: Übungsblatt 1, Aufgabe 4, insbesondere Teilaufgaben (b.4) und (c)

3.1.2. Histogramme (Skript S. 18-20)

Definition

- j. In einem Histogramm werden alle Stichprobenwerte in endlich viele, willkürlich gewählte, nicht überlappende **Intervalle** (I_1, I_2, \dots, I_L) zusammengefasst
- k. Basierend auf den absoluten oder relativen Häufigkeiten zeichnet man für jedes Intervall I_j ein **Rechteck**. Höhe = Häufigkeit, Länge = Länge des Intervalls I_j
- l. 2 verschiedene Darstellungsformen (wobei Konvention 2 klar überlegen, aussagekräftiger!):
 - i. **Konvention 1: Höhe = absolute Häufigkeit** ($= h_j$)
 - ii. **Konvention 2: Höhe = relative Häufigkeit / Länge des Intervalls** ($= F_j / I_j$)
- m. **Kritik an Histogrammen**: Bild hängt sehr stark von Auswahl der Intervalle I_j ab. Einerseits kann die **Wahl der Länge der Intervalle** das Bild stark verzerren, andererseits kann – gar bei Intervallen mit allesamt gleicher Länge – die **Variation des Randpunktes** ein völlig anderes Bild entstehen



Tipps und Tricks

- n. **Repetition vor Prüfung**:
 - i. Übungsblatt 2, Aufgabe 6 → Ermittlung Datensatz aus zwei verschiedenen Histogrammen!
 - ii. (allenfalls: Übungsblatt 2, Aufgabe 5 b): zeichnen von Histogrammen)

3.2. Quantile und Quartile (Skript S. 20-22)

Definition

- o. ein Quantil ist eine **Schranke (Q_β)** mit der Eigenschaft, dass mind. $n\beta$ der X-Werte kleiner oder gleich Q_β sind und mind. $n(1-\beta)$ der X-Werte grösser oder gleich Q_β sind
- p. Spezielle Quantile: **1. Quartil bei $Q_{0.25}$, Median bei $Q_{0.5}$, 3. Quartil bei $Q_{0.75}$**

Berechnungsweise:

- q. wenn $n\beta$ ganze Zahl ist: Q_β hat **Mittelwert** der X mit Ordnungsstatistiken $n\beta$ sowie $n\beta+1$
- r. wenn $n\beta$ **keine ganze Zahl** ist: Q_β hat Wert des X mit **aufgerundeter** Ordnungsstatistik $n\beta$

$$Q_\beta = \frac{X_{(n\beta)} + X_{(n\beta+1)}}{2}$$

Fall 1: $n\beta$ ist ganze Zahl

-5-

$$Q_\beta = X_{(\lceil n\beta \rceil)}$$

Fall 2: $n\beta$ ist keine ganze Zahl

3.3. Lageparameter (Skript S. 22-25)

Definition **Lageparameter**

- s. eine **Zahl**, die **möglichst nah** an allen X-Werten liegt
- t. eine **Zahl**, die einen typischen Wert bzw. die **Größenordnung der X-Werte** angibt

3.3.1. Mittelwert (mean)

- arithmetisches Mittel der Zahlen X_i
- Mittelwert ist **nicht robust** (Definition „nicht robust“: **reagiert empfindlich auf Ausreisser** von X_i)

$$\bar{X} := \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- der Mittelwert ist der **Schwerpunkt aller X-Werte**, da

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

3.3.2. Median

- 50%-Quantil, d.h. Wert des X_i , das genau in der Mitte der Ordnungsstatistiken $X_{(i)}$ liegt
- Median ist **robust** (d.h. reagiert nicht auf einzelne Ausreisser von X_i)

$$\text{Med} := Q_{0.5} = \begin{cases} X_{((n+1)/2)} & \text{falls } n \text{ ungerade} \\ (X_{(n/2)} + X_{(n/2+1)})/2 & \text{falls } n \text{ gerade} \end{cases}$$

Optimalität des Medians

- Berechnung und Überlegung, weshalb Median z.B. bei „Briefkastenproblem“ optimal ist: siehe Skript S. 23 und 24

3.3.3. Getrimmter Mittelwert (trimmed mean)

- modifizierter **Mittelwert**, in dem ein **gewisser Prozentsatz ($= \alpha$)** der **höchstens und tiefsten X-Werte** **weggelassen** wird. (Beispiel unten: $\alpha = 10\%$, d.h. 5% tiefste und 5% höchste Werte fallen weg)
- da die Extremwerte nicht berücksichtigt werden, ist der getrimmte Mittelwert **robust**

$$\bar{X}_\alpha = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} X_{(i)} \quad \text{mit } k := \lfloor n\alpha/2 \rfloor$$

Beispiel: $n = 100$ Beobachtungen, $\alpha = 10\% = 0.1$:
$$\bar{X}_\alpha = \frac{1}{90} \sum_{i=6}^{95} X_{(i)}$$

3.4. Skalenparameter (Skript S. 26-29)

Definition **Skalenparameter**

- u. eine **Zahl**, welche die **Abweichung der X-Werte von ihrem Zentrum** angibt, oder
- v. eine **Zahl**, welche die **Abweichung der X-Werte untereinander** angibt

3.4.1. Spannweite (range)

- Differenz von Maximum und Minimum der Stichprobenwerte
- Spannweite ist **nicht robust**

$$X_{(n)} - X_{(1)}$$

3.4.2. Interquartilabstand (inter quartile range)

- Differenz des 3. und des 1. Quartils
- IQR umfasst somit mind. 50% der X-Werte
- IQR ist **robust**

$$\text{IQR} := Q_{0.75} - Q_{0.25}$$

3.4.3. Standardabweichung (standard deviation)

- **misst die Abweichung der X-Werte vom Mittelwert**
- Standardabweichung ist **nicht robust**
- Die Kenngrösse innerhalb der Quadratwurzel (also gesamte rechte Formel ohne Wurzel) nennt man (Stichproben-) Varianz

$$S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Alternative Formel für Standardabweichung:

- **VORSICHT**: quadrierter Mittelwert darf für Berechnung **nicht gerundet** werden!

$$S = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - \underline{n\bar{X}^2} \right)}$$

3.4.4. Ginis Skalenparameter

- **misst den arithmetischen Mittelwert der Abstände** $|X_i - X_j|$ über alle möglichen Paare von X-Werten
- Ginis Skalenparameter ist **nicht robust** und hat sich in Praxis nicht durchgesetzt

$$G = \frac{2}{n(n-1)} \sum_{i=1}^n (2i - n - 1) X_{(i)}$$

3.4.5. Median der absoluten Abweichung (median absolute deviation)

- **misst die Abweichung der X-Werte vom Median** (also Abwandlung der Standardabweichung)
- Vorgehen: „**doppelter Median**“: zuerst wird Differenz der einzelnen X-Werte zu Median berechnet, in zweitem Schritt nimmt man abermals den Median aus diesen Differenzen; MAD ist **robust**

$$\text{MAD} := \text{Med} \left(|X_1 - \text{Med}|, |X_2 - \text{Med}|, \dots, |X_n - \text{Med}| \right)$$

Hilfreich: wenn **Abstände des Medians zu den beiden anderen Quartilen identisch** ist, **MAD = IQR/2**

Tipps und Tricks zu „Skalenparameter“

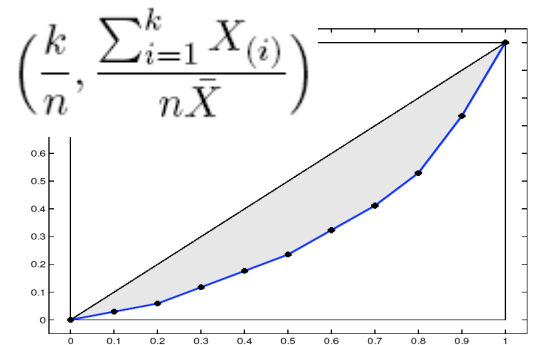
- gute, anschauliche Beispielaufgabe (**Bsp. 2.5**) zu allen Berechnungen auf **S.27/28 im Skript!**
- Vor Prüfung allenfalls repetieren: Übungsblatt 2, Aufgabe 7 (div. Berechnungen in Excel)

3.5. Lorenzkurve und Gini-Index (Skript S.29-32)

3.5.1. Lorenzkurve

- sagt aus, wie Einkommen unter Bevölkerung verteilt ist
- Monoton wachsende, konvexe Kurve
- Ermittlung der einzelnen Kurvenpunkte: siehe Formel
- Bestandteile der Formel
 - $k = 1, 2, \dots, n$; pro k wird ein Punkt ermittelt
 - $n\bar{X}$ (unterhalb Bruchstrich) = Gesamteinkommen
 - Ausdruck über Bruchstrich = Einkommen der k Ärmsten
- Aussage: je weiter Kurve von Winkelhalbierender entfernt ist, desto ungerechter ist die Verteilung

Punkte der Lorenzkurve



3.5.2. Gini-Index

- misst das Ausmass der Ungerechtigkeit der Einkommensverteilung in Lorenzkurve
- Grundgedanke: $GI = 2 * \text{Fläche zwischen Lorenzkurve und 1. Winkelhalbierender}$
- GI somit immer Wert zwischen 0 und 1, wobei gilt: je grösser GI, desto ungleicher die Verteilung

$$GI = \frac{2}{n^2 \bar{X}} \sum_{i=1}^n i \cdot X_{(i)} - \frac{n+1}{n}$$

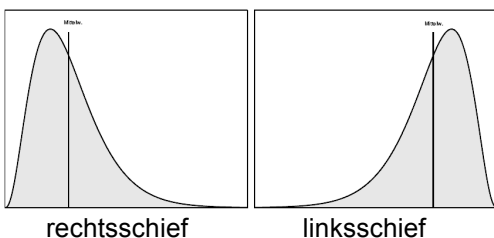
Achtung: Ausdruck $-(n+1)/n$ steht ausserhalb des Summenzeichens und darf erst ganz am Ende subtrahiert werden!

Tipp: Sehr anschauliches, aufschlussreiches Beispiel in Skript (Bsp. 2.6, Seiten 30 und 31)

3.6. Formparameter (Skript S.33-36)

3.6.1. Schiefe (skewness)

- sagt aus, wie die X-Werte um den Schwerpunkt (= Mittelwert) verteilt sind
- rechtsschief wenn Schiefe > 0 (d.h. rechts vom Mittelwert liegen viele X-Werte eher weit entfernt)
- linksschief wenn Schiefe < 0 (d.h. links von Mittelwert liegen viele X-Werte eher weit entfernt)



$$\frac{1}{nS^3} \sum_{i=1}^n (X_i - \bar{X})^3$$

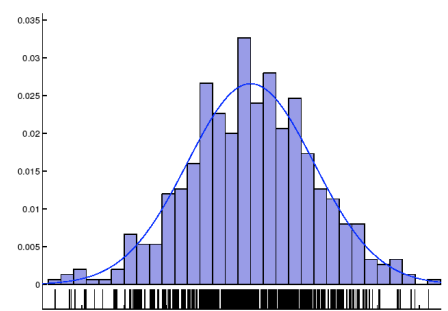
Achtung: S = Standard-Abweichung darf nicht gerundet werden!
Erläuterung: $()^3$ damit Abweichung deutlicher, aber Vorzeichen bleiben

3.6.2. Kurtose (kurtosis)

- sagt aus, wie sehr Verteilung der Werte von Gauscher Glockenkurve abweichen.
- Je tiefer Kurtose-Wert, desto kleiner Abweichung von Glockenkurve

$$\frac{1}{nS^4} \sum_{i=1}^n (X_i - \bar{X})^4 - 3$$

Achtung: Ausdruck -3 steht ausserhalb des Summenzeichens und darf erst ganz am Ende subtrahiert werden!



4. Simultane Beschreibung zweier Merkmale (Skript S. 37-62)

4.1. Kontingenztafeln und Vierfeldertafeln

Grundidee:

Wenn Variable X kategoriell und Variable Y ebenfalls kategoriell ist, kann man deren Verhältnis durch Kontingenztafeln bzw. Vierfeldertafeln auswerten/aufzeigen.

4.1.1. Kontingenztafeln (Skript S. 37-41)

- zwei Variablen, X und Y, werden in derselben Datenbank dargestellt
- Anzahl Zeilen: mögliche Ausprägungen der Variable X (dies sind: x_1, x_2, \dots, x_L)
- Anzahl Spalten: mögliche Ausprägungen der Variable Y (dies sind: y_1, y_2, \dots, y_M)
- Zeilennormierung und Spaltennormierung: siehe Skript S.38

	y_1	y_2	\dots	y_M	
x_1	$N_{1,1}$	$N_{1,2}$	\dots	$N_{1,M}$	$N_{1,+}$
x_2	$N_{2,1}$	$N_{2,2}$	\dots	$N_{2,M}$	$N_{2,+}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_L	$N_{L,1}$	$N_{L,2}$	\dots	$N_{L,M}$	$N_{L,+}$
	$N_{+,1}$	$N_{+,2}$	\dots	$N_{+,M}$	n

4.1.2. Idealisierte Werte

- Aussage: bestände keinerlei Zusammenhang zwischen X und Y, würden die idealisierten Werte erreicht; je mehr die tatsächlichen Werte von den idealisierten Werten abweichen, desto grösser ist der Zusammenhang
- um die idealisierten Werte pro Feld zu erhalten, muss man das Produkt aus ihrer Zeilensumme und ihrer Spaltensumme durch n dividieren (Formel: siehe rechts)
- Zeilensummen = $N_{i,+}$ (siehe roter Kasten in Graphik)
- Spaltensummen = $N_{+,k}$ (siehe grüner Kasten in Graphik)

$$\bar{N}_{j,k} := \frac{N_{j,+} N_{+,k}}{n}$$

Achtung: Zeilensummen und Spaltensummen werden miteinander *multipliziert* (optische Täuschung des „+“ aus $N_{i,+}$)

4.1.3. Chiquadrat-Statistik

- Aussage: Kenngrösse für die Abweichung der tatsächlichen Werte von den idealisierten Werten
- Berechnung: 2 mögliche Formeln (siehe rechts)
- Faustregel zur Einordnung der Chiquadrat-Werte: „echter“ Zusammenhang besteht mit Sicherheit von 95% wenn χ^2 Wert grösser ist als:

$$\chi^2 := \sum_{j=1}^L \sum_{k=1}^M \frac{(N_{j,k} - \bar{N}_{j,k})^2}{\bar{N}_{j,k}}$$

$$\chi^2 := \sum_{j=1}^L \sum_{k=1}^M \frac{N_{j,k}^2}{\bar{N}_{j,k}} - n$$

Achtung: Ausdruck - n steht ausserhalb des Summenzeichens

$$\frac{(L-1)(M-1) + 2\sqrt{2(L-1)(M-1)}}{2}$$

wobei L = Anzahl Zeilen, und M = Anzahl Spalten

Tipp zu Kontingenz- und Vierfeldertafel: Repetition Aufgaben 12 und 13 (Übungsblatt 3)

4.1.4. Vierfeldertafeln

$N_{1,1}$	$N_{1,2}$
$N_{2,1}$	$N_{2,2}$

- sind **spezielle Kontingenztafeln**, die je nur 2 Zeilen und 2 Spalten aufweisen
- mathematisch: $L = M = 2$
- die **Chi-Quadrat-Statistik χ^2** wird deshalb **analog 4.1.3.** (S.8 ZF) berechnet
- wichtigste **Aussagekraft** der Vierfeldertafeln ist der **Chancenquotient**

4.1.5. Chancenquotient (odds ratio)

- wird auch „**Kreuzproduktverhältnis**“ genannt
- kann *nur* bei Vierfeldertafel berechnet werden
- Aussage: siehe Erläuterung **Skript S. 42**
- **Beispiel zu Aussage:** wenn zwei versch. Personengruppen zwei möglichen Merkmalen/Entscheiden usw. gegenübergestellt werden, so sagt „OR“ aus, um welchen Faktor die **Chance** der Gruppe A **größer (OR>0)** bzw. **kleiner (OR<0)** ist, dass das Merkmal bei ihr vorhanden ist, als bei Gruppe B

$$OR := \frac{N_{1,1}N_{2,2}}{N_{1,2}N_{2,1}} = \left\{ \begin{array}{l} \frac{N_{1,1}/N_{1,2}}{N_{2,1}/N_{2,2}} \\ \frac{N_{1,1}/N_{2,1}}{N_{1,2}/N_{2,2}} \end{array} \right.$$

4.1.6. Simpson-Paradoxon

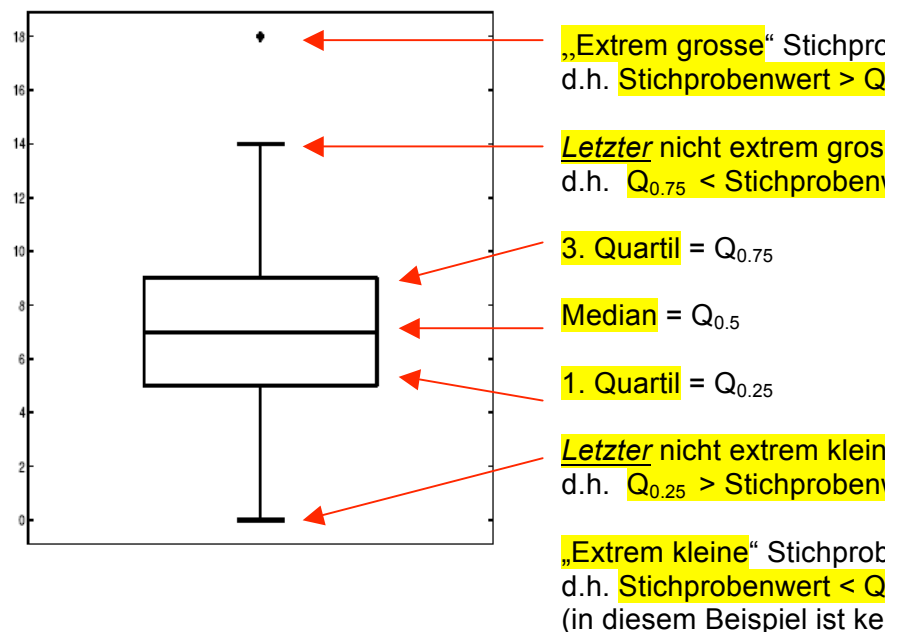
- Hauptaussage: die **Zusammenfassung mehrerer Kontingenztafeln kann zu Fehlschlüssen führen**
- Gutes, erklärendes **Beispiel in Skript auf S. 43**
- Fazit: Zusammenhang sagt nichts über Kausalität aus; **Zusammenhang ungleich Kausalität**

4.2. Box-Whisker-Plots (Skript S. 43-49)

Grundidee:

Wenn **Variable X kategoriell** und **Variable Y numerisch** ist, kann man **pro X ein Box-Whisker-Plot** zeichnen, der **aufzeigt, wie die Stichprobenwerte von Y innerhalb der Quartile verteilt sind**

Anmerkung: **Berechnung Quartile und IQR** siehe S.4 bzw. S.6 dieser Zusammenfassung



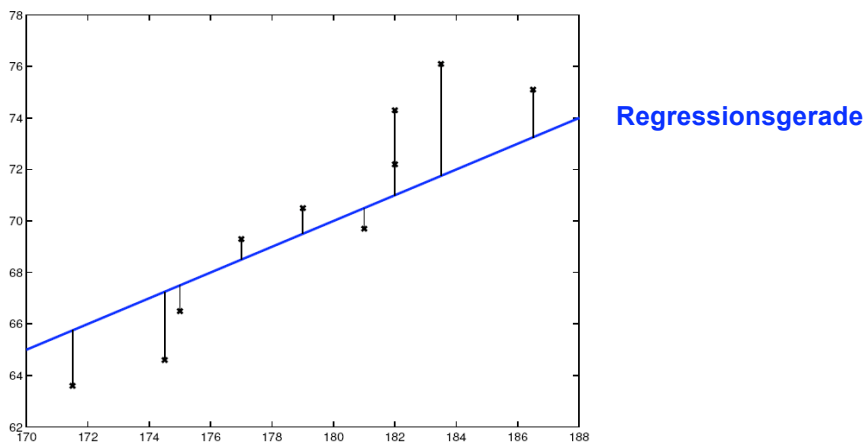
4.3. Regression und Korrelation (Skript S. 49-62)

Grundidee:

Wenn Variable X numerisch und Variable Y ebenfalls numerisch ist, kann man deren Verhältnis durch Streudiagramme darstellen sowie deren Regression und Korrelation berechnen.

4.3.1. Lineare Regression (Skript S. 50-55)

- Frage: inwiefern besteht ein linearer Zusammenhang zwischen X und Y?
- gesucht ist in diesem Zusammenhang eine lineare Funktion, deren Abweichung von den von den tatsächlichen Punkten (X_i, Y_i) möglichst gering ist; man nennt dies **Regressionsgerade**



- die Regressionsgerade verläuft stets durch den Schwerpunkt (\bar{X}, \bar{Y}) der Beobachtungen (X_i, Y_i)
- zur Berechnung der Regressionsgerade ist ein mehrstufiger Vorgang nötig: {

Regressionsgerade =

$$y = \hat{a} + \hat{b}x = \bar{Y} + \hat{b}(x - \bar{X})$$

wobei

$$\hat{a} := \bar{Y} - \hat{b}\bar{X}$$

und

$$\hat{b} := \frac{QS_{XY}}{QS_{XX}}$$

sowie:

$$QS_{XX} := \sum_{i=1}^n (X_i - \bar{X})^2,$$

$$QS_{XX} = \sum_{i=1}^n X_i^2 - n\bar{X}^2,$$

$$QS_{YY} := \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$QS_{YY} = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2,$$

$$QS_{XY} := \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$QS_{XY} = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}$$

Achtung: Ausdrücke stehen ausserhalb des Summenzeichens; nicht runden!

4.3.2. Bestimmtheitsmass und Korrelation (Skript S. 55-59)

- sind Ausdruck davon, wie stark der lineare Zusammenhang von X und Y ist
- **Bestimmtheitsmass R^2**
 Minimalwert = 0 , d.h. kein linearer Zusammenhang
 Maximalwert = 1 , d.h. alle Stichprobenwerte liegen auf Regressionsgerade
- **Korrelationskoeffizient** (nach Bravais-Pearson): r_{XY}
 Minimalwert = -1 , d.h. alle SP-Werte liegen auf Regressionsgerade mit negativer Steigung
 Wert = 0 , d.h. kein linearer Zusammenhang
 Maximalwert = 1 , d.h. alle SP-Werte liegen auf Regressionsgerade mit positiver Steigung

Bestimmtheitsmass:

$$R^2 = \frac{QS_{XY}^2}{QS_{XX} QS_{YY}}$$

Ermittlung QS-Werte: siehe S.10 dieser Zusammenfassung (4.5.1. Lineare Regression)

Korrelationskoeffizient:

$$r_{XY} := \frac{QS_{XY}}{\sqrt{QS_{XX} QS_{YY}}}$$

Zusammenhang zwischen Bestimmtheitsmass und Korrelationskoeffizient:

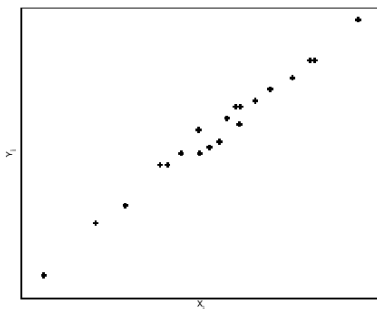
$$R^2 = r_{XY}^2$$

Erlaubte Transformationen bei Korrelationskoeffizient: nur *lineare* Transformationen möglich!

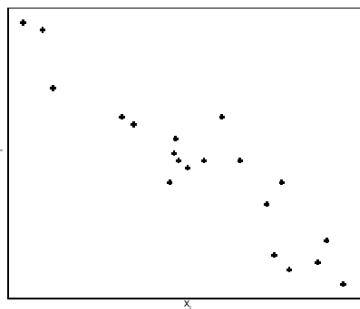
- vertauschen der Rollen von X und Y
- Addition einer Konstante zu allen X- oder Y-Werten
- Multiplikation aller X- oder Y-Werte mit einer Konstante

4.3.3. Graphische Darstellung: Streudiagramm (scatter plot)

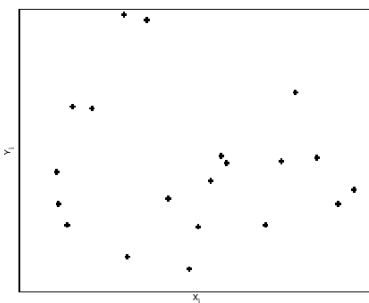
Vorgehen: jedes Datenpaar X_i, Y_i wird als Punkt in einem zweidimensionalen Graphen eingezeichnet:



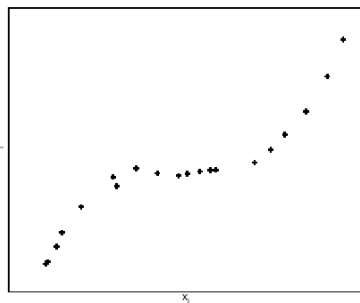
linearer Zusammenhang
stark *positive* Korrelation



linearer Zusammenhang
negative Korrelation



kein Zusammenhang



nicht-linearer Zusammenhang

4.3.4. Rangkorrelation nach Spearman (Skript S. 59-62)

Unterschiede zu Bestimmtheitsmass und Korrelationskoeffizient (nach Bravais-Pearson):

- **alle streng monoton wachsenden Transformationen** der X- und Y-Werte **sind erlaubt**, z.B. also auch Exponentialfunktionen, Logarithmusfunktionen, Quadratwurzel usw.
- Rangkorrelation ist nicht empfindlich gegenüber Ausreißern (d.h. ist **robust!**)
- **Minimum = -1**; alle SP-Werte liegen auf dem Graphen einer *streng monoton fallenden Funktion*
- **Maximum = 1**; alle SP-Werte liegen auf dem Graphen einer *streng monoton steigenden Funktion*

Vorgehen Rangzuordnung:

- **jeder Beobachtung von X, Y wird ein Rang zugeordnet** (kleinster Wert = tiefster Rang)
 - wenn manche Beobachtungen der X, Y-Werte identisch sind, d.h. wenn die Werte *nicht* paarweise verschieden sind, arbeitet man mit **mittleren Rängen**.
- Bsp. Für mittlere Ränge (rechts): Rang 4, 5 und 6 hätten alle denselben Wert (10); der mittlere Rang ist deshalb:
 $(4+5+6)/3 = 5$; Rang 1 und 2 = 0, d.h. beide $(1+2)/2 = 1.5$

<i>i</i>	1	2	3	4	5	6	7
<i>X_i</i>	4	5	0	1	10	13	12
<i>RX_i</i>	3	4	1	2	5	7	6

Zuteilung der Ränge

<i>i</i>	1	2	3	4	5	6	7
<i>X_i</i>	4	10	0	0	10	13	10
<i>RX_i</i>	3	5	1.5	1.5	5	7	5

Sonderfall: „Mittlere Ränge“

Allgemeine Formel (wird verwendet, wenn sowohl X-Werte wie auch Y-Werte **nicht paarweise verschieden** sind; d.h. wenn sowohl bei X wie auch bei Y „mittlere Ränge“ bestehen)

$$r_{Sp} = \frac{\sum_{i=1}^n RX_i RY_i - n(n+1)^2/4}{\sqrt{\left(\sum_{i=1}^n RX_i^2 - n(n+1)^2/4 \right) \left(\sum_{i=1}^n RY_i^2 - n(n+1)^2/4 \right)}}$$

ausserhalb Summe!

Vorgehen wenn X-Werte paarweise verschieden, Y-Werte nicht paarweise verschieden (oder umgekehrt): siehe **Beispiel 2.21** auf Skript S.61

Vereinfachte Formel wenn sowohl X-Werte wie auch Y-Werte **paar** (d.h. keine mittleren Ränge; keiner der X-Werte ist gleich wie ein an

$$\sqrt{n-1} |r_{Sp}| > 2$$

$$r_{Sp} = \frac{\sum_{i=1}^n RX_i RY_i - n(n+1)^2/4}{n(n^2-1)/12}$$

<i>i</i>	1	2	3	4	5	6	7
<i>X_i</i>	4	5	0	1	10	13	12
<i>Y_i</i>	2.1	1.5	1.2	1.3	2.7	4.0	3.5

Faustregel: mit 95%-Sicherheit „echter Zusammenhang“ wenn:

Tipps und Trick zu Regression und Korrelation
NOCH UEBUNGEN EINTRAGEN!

Teil II: Wahrscheinlichkeitsrechnung und statistische Modelle

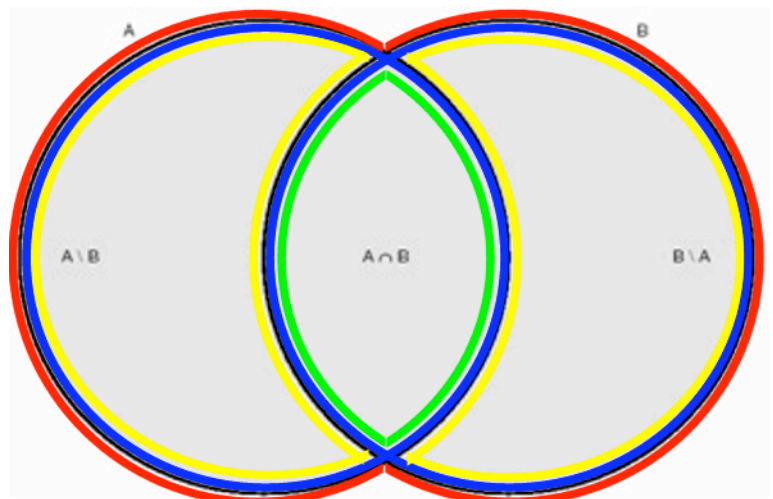
5. Wahrscheinlichkeitsrechnung

5.1. Grundlagen, Definitionen (Skript S. 65 – 68)

- **Grundraum Ω** = Menge aller möglichen Resultate, die das Zufallsexperiment liefern kann
- **Elementarereignis ω** = ein einzelnes Element aus Ω
- **Ereignis** (z.B. A, B usw.) = klar definierte Teilmenge aus Ω

Verknüpfung von Ereignissen:

- **$A + B$**
d.h. beide Ereignisse einzeln,
somit wird $A \cap B$ doppelt gezählt
- **$A \cap B$**
Schnittmenge aus A und B
- **$A \cup B$**
Ganzer grauschraffierter Bereich;
d.h. $A \cap B$ nur einfach gezählt
- **$A \setminus B$ bzw. $B \setminus A$**
 $A \setminus B$ = Ereignis A ohne $A \cap B$;
Schnittmenge wird abgezogen



Zuordnen von Wahrscheinlichkeiten

- **$P(A)$ ist die Wahrscheinlichkeit, dass Ereignis A eintritt (Minimum = 0, Maximum = 1)**
- zwei verschiedene Deutungsweisen:
 - o **$P(A)$ als Wetteinsatz (subjektivistische Deutung)** = subjektives Mass dafür, wie sicher man sich ist, dass Ereignis A eintritt. Anmerkung: Wette ist fair, wenn $E/G = P(A)$, wobei E = Einsatz, G = Gewinn; falls $E/G < P(A)$ so hat der Spieler einen Vorteil, fall $E/G > P(A)$ hat Spielanbieter einen Vorteil
 - o **$P(A)$ als Grenzwert (frequentistische Deutung)** = Wahrscheinlichkeit die sich einstellt, wenn man ein (Zufalls-)Experiment unendlich oft *unabhängig* voneinander durchführt.
- **Zusammenhang zwischen den beiden Deutungen: siehe Skript S.67** (unten)

5.2. Wahrscheinlichkeitsverteilungen (Skript S. 68 – 76)

5.2.1. Diskrete Verteilungen

- jedem Elementarereignis ω wird Wahrscheinlichkeit zugeordnet
- die Summe der Wahrscheinlichkeiten aller ω ist 1
- die Wahrscheinlichkeit eines Ereignisses ist die Summe der Wahrscheinlichkeiten aller ω , die im Ereignis A enthalten sind
- die Wahrscheinlichkeiten der einzelnen ω können verschieden sein (Beispiel „gezinkter Würfel“, siehe unten):

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

$$P(A) := \sum_{\omega \in A} p(\omega)$$

Beispiel: „gezinkter Würfel“

ω	1	2	3	4	5	6
$p(\omega)$	0.1	0.15	0.15	0.3	0.2	0.1

5.2.2. Laplace-Verteilung

- Spezialfall der diskreten Verteilung. Voraussetzung: jedes Elementarereignis ω tritt mit genau derselben Wahrscheinlichkeit ein; somit gilt: $p(\omega) = 1 / \#\Omega$
- Man nennt die Laplace-Verteilung auch „uniformelle Verteilung“. Sie beschreibt die „rein zufällige“ Auswahl eines Elementes von Ω

$$P(A) := \frac{\#A}{\#\Omega} \quad \left(\frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}} \right)$$

- die Anzahl günstiger Fälle (d.h. die Anzahl Fälle, die das Ereignis A beinhaltet) kann mit den Mitteln der Kombinatorik ermittelt werden
- **Tipps und Trick** zu Kombinatorik:
 - o Manchmal ist es kombinatorisch einfacher und deshalb sinnvoll, die Wahrscheinlichkeit des Komplementärereignisses A^c zu bestimmen. Dabei nützt man aus dass $P(A) = 1 - P(A^c)$
 - o Gute Beispiele zu Kombinatorik:
 - Beispiel 3.2. Skript S.69, v.a. Ermittlung der Ereignisse C und D
 - Aufgaben 28 und 29 (v.a. Teilaufgabe c), Übungsblatt 7
 - evt. Aufgabe 25, Übungsblatt 6 (leichte Aufgabe, aber gut für Grundverständnis)

5.2.3. Rechenregeln für Wahrscheinlichkeiten

- für zwei beliebige Ereignisse: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- bei disjunkten Ereignissen (= Ereignisse haben keine Schnittmengen): $P(A \cup B) = P(A) + P(B)$
- Komplementärereignis: $P(A) = 1 - P(A^c)$
- bei $A \subset B$ („A c B“ bedeutet, dass A vollständig in B enthalten ist): $P(A) \leq P(B)$
- **Tipp:** eine gute Anwendungsübung für die oberste Regel ist Aufgabe 27 a) – Übungsblatt 7

5.2.4. Siebformel

- **Idee:** unter „5.2.3. Rechenregeln“ wurde festgehalten, dass $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Nun interessiert die **Verknüpfung von mehr als nur 2 Ereignissen, also $P(A_1 \cup A_2 \dots \cup A_n)$**

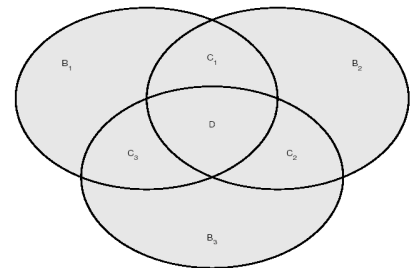
$$\begin{aligned}
 P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_i P(A_i) \\
 &\quad - \sum_{i < j} P(A_i \cap A_j) \\
 &\quad + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) \\
 &\quad \mp \dots \\
 &\quad + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n).
 \end{aligned}$$

- **Bemerkungen zu genereller Vorgehensweise:** **abwechslungsweise werden immer kleinere Schnittmengen subtrahiert bzw. addiert**

Konkretes Vorgehen bei **Verknüpfung von 3 und 4 Ereignissen**

- Anmerkung: aufgrund der Komplexität ist davon auszugehen, dass **in Prüfung 3 bis höchsten 4 Ereignisse miteinander verknüpft werden müssen!** Deshalb sind diese Varianten hier aufgeführt.
- **Ermittlung $P(A_1 \cup A_2 \cup A_3)$** , d.h. von 3 verknüpften Ereignissen:

$$\begin{aligned}
 P(A_1 \cup A_2 \cup A_3) &= \\
 &P(A_1) + P(A_2) + P(A_3) \\
 &- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
 &+ P(A_1 \cap A_2 \cap A_3).
 \end{aligned}$$



Erläuterung in Worten anhand Graphik: alle drei Kreise (A_1, A_2, A_3) werden addiert, dann die Schnittmengen aus je zwei Ereignisse abgezogen. Weil nun die Mitte ($D; A_1 \cap A_2 \cap A_3$) komplett wegfallen würde, wird diese anschliessend wieder addiert

- **Ermittlung $P(A_1 \cup A_2 \cup A_3 \cup A_4)$** , d.h. von 4 verknüpften Ereignissen:
 - o Graphisch nicht darstellbar
 - o Vorgehensweise siehe **Beispiel 3.2. auf Seite 75 Skript**
 - o Achtung: **Binominalkoeffizienten als Vorfaktoren nicht vergessen!**

5.2.5. Bonferroni-Ungleichungen

lässt sich Siebformel nicht anwenden, Schnittmengen nicht ermittelbar, so können dank Bonferroni-Ungleichungen zumindest **Schranken (Größenordnungen) für $P(A_1 \cup A_2 \dots \cup A_n)$** ermittelt werden. Beispiel: siehe **Aufgabe 30 Übungsblatt 7**

1. BF-Ungleichung:	$P(A_1 \cup A_2 \dots \cup A_n) \leq P(A_1) + P(A_2) + \dots + P(A_n)$
2. BF-Ungleichung:	$P(A_1 \cup A_2 \dots \cup A_n) \geq \max[P(A_1) ; P(A_2) \dots P(A_n)]$

5.3. Bedingte Wahrscheinlichkeiten (Skript S. 76 – 80)

Grundidee: es bestehen zwei Ereignisse A und B. Aus irgendeinem Grund weiss man, dass B mit Sicherheit eintreten wird (bzw. eingetreten ist). Die Wahrscheinlichkeit, dass A ebenfalls eintreten wird, nennt man bedingte Wahrscheinlichkeit von A, gegeben B.

Grundlegende Formeln:

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

sowie:

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

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

$$P(B | B) = 1 \quad \text{sowie:} \quad P(B^c | B) = 0$$

Hilfsmittel Vierfeldertafel:

mit Hilfe der Vierfeldertafel kann man die Wahrscheinlichkeiten der Ereignisse P(T), P(K), P(T^c), P(K^c) in Form von „Schnitten“ schreiben; dies erlaubt basierend auf den obigen, grundlegenden Formeln die Umrechnung gegebener bedingter Ereignisse zu anderen bedingten Ereignissen.

	T	T ^c	
K	P(K ∩ T)	P(K ∩ T ^c)	P(K)
K ^c	P(K ^c ∩ T)	P(K ^c ∩ T ^c)	P(K ^c)
	P(T)	P(T ^c)	

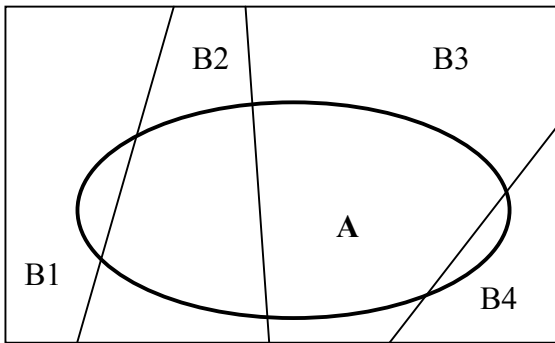
Beispiel für Zweck der Vierfeldertafeln: **„Spam Filter“**, Bsp. 3.6 (Skript S. 78-79):

- Gegeben: P(K) = 0.8 ; P(T | K) = 0.75 ; P(T | K^c) = 0.02 Gesucht: P(K | T)
- Vorgehensweise: 4 Schritte:
 1. Verwendung grundlegende Formel Nr. 1 zwecks Suche P(K | T)
 2. Ersetzen P(T) durch „Schnitte“ aus Vierfeldertafel
 3. Ersetzen „Schnitte“ durch Anwendung grundlegende Formel Nr. 2b
 4. Einsetzen der gegebenen Werte



$$\begin{aligned}
 P(K | T) &= \frac{P(K \cap T)}{P(T)} && 1. \\
 &= \frac{P(K \cap T)}{P(K \cap T) + P(K^c \cap T)} && 2. \\
 &= \frac{P(K)P(T | K)}{P(K)P(T | K) + P(K^c)P(T | K^c)} && 3. \\
 &= \frac{0.8 \cdot 0.75}{0.8 \cdot 0.75 + 0.2 \cdot 0.02} = \frac{0.6}{0.604} \approx 0.9934 && 4.
 \end{aligned}$$

5.3.1. Die Bayessche Formel



Idee: vollständiges definieren eines Ereignisses A mit Hilfe von seinen Schnitten mit anderen Ereignissen (B_j)

Voraussetzung:

$$\Omega = B_1 \cup B_2 \cup \dots \cup B_M$$

d.h. die Summe aller Ereignisse B_j füllt den Grundraum *vollständig* aus

Grundlegende Formeln:

$$P(A) = \sum_{j=1}^M P(A \cap B_j) \rightarrow$$

$$P(A) = \sum_{j=1}^M P(B_j)P(A | B_j)$$

Formalisierung der Idee

formalisierte Idee: ersetzen

bedingte Wahrsch. (siehe S.16)

Umformung der

Schnitt durch

Gute Beispiel: Signalübertragung (Skript S.80)

- B = versendetes Signal (00;01;10;11)
- A = empfangenes Signal (00;01;10;11)
- $P(A | B_j)$ = Wahrscheinlichkeit dass A bei Empfänger ankommt, wenn B_j abgeschickt
- Gesucht: $P(B_j | A)$, d.h. Wahrscheinlichkeit, dass B_j versandt, wenn A empfangen

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{P(A)}$$

Mit **Hilfe $P(A)$** lässt sich $P(B_j | A)$ errechnen

5.4. Stochastische Unabhängigkeit (Skript S. 81 – 85)

5.4.1. Stochastische Unabhängigkeit zweier Ereignisse

stochastische Unabhängigkeit bedeutet, dass Ereignisse völlig unabhängig voneinander eintreten:

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

$$P(A | B) = P(A)$$

WICHTIG: Unabhängigkeit bleibt erhalten, wenn man A durch A^c bzw. B durch B^c ersetzt!

- $P(A \cap B^c) = P(A) P(B^c)$
- $P(A^c \cap B) = P(A^c) P(B)$
- $P(A^c \cap B^c) = P(A^c) P(B^c)$

5.4.2. Stochastische Unabhängigkeit beliebig vieler Ereignisse

$$P(A_{i(1)} \cap A_{i(2)} \cap \dots \cap A_{i(k)}) = P(A_{i(1)})P(A_{i(2)}) \dots P(A_{i(k)})$$

also z.B.

$$P(A \cap B \cap C \cap D) = P(A) P(B) P(C) P(D)$$

Bemerkungen:

- Regel somit intuitiv richtig abgeleitet, analog stochastischer Unabhängigkeit zweier Ereignisse
- Auch hier gilt: Ereignisse können problemlos durch ihre Komplementärereignisse ersetzt werden; z.B: $P(A \cap B^c \cap C^c \cap D) = P(A) P(B^c) P(C^c) P(D)$
- Auch hier gilt $P(A | B) = P(A)$ analog für alle möglichen Ereigniskombinationen
- **WICHTIG:** Paarweise Unabhängigkeit bedeutet nicht zwingend stochastische Unabhängigkeit (paarweise Unabhängigkeit: $P(A_i \cap A_j) = P(A_i) P(A_j)$)
- verschiedene Deutungen (zeitliche etc.): siehe Skript S. 82)

Repetition vor Prüfung: Skript S.83-85 („n-facher Münzwurf, Befragung, „Geburtstagsproblem“)

5.5. Spezielle Verteilungen (Skript S. 89 – 102)

5.5.1. Hypergeometrische Verteilung

Definitionen

N = Total Kugeln in Urne

L = Markierte Kugeln in Urne

n = Anzahl Kugeln, die ohne zurücklegen aus Urne gezogen werden

X = Anzahl markierte Kugeln unter den n gezogenen

P(X = k): Wahrscheinlichkeit, dass sich unter den n gezogenen Kugeln **k** markierte Kugeln befinden

$h_{N,L,n}(k)$ = $P(X = k)$ = Wahrscheinlichkeitsgewichte der hypergeometrischen Verteilung

Hyp(N,L,n) = hypergeometrisch verteilte Zufallsvariable mit Parametern N,L,n.

$$h_{N,L,n}(k) := \binom{L}{k} \binom{N-L}{n-k} / \binom{N}{n}$$

$$P(X = k) = \binom{n}{k} \frac{[L]_k [N-L]_{n-k}}{[N]_n}$$

Umschreibung in Worten:

- n Kugeln ohne zurücklegen aus einer Urne ziehen (d.h. stochastische Unabhängigkeit nicht gegeben!)
- in Urne sind L von insgesamt N Kugeln markiert
- Frage: wie viele der n gezogenen Kugeln sind markiert?

Anmerkungen:

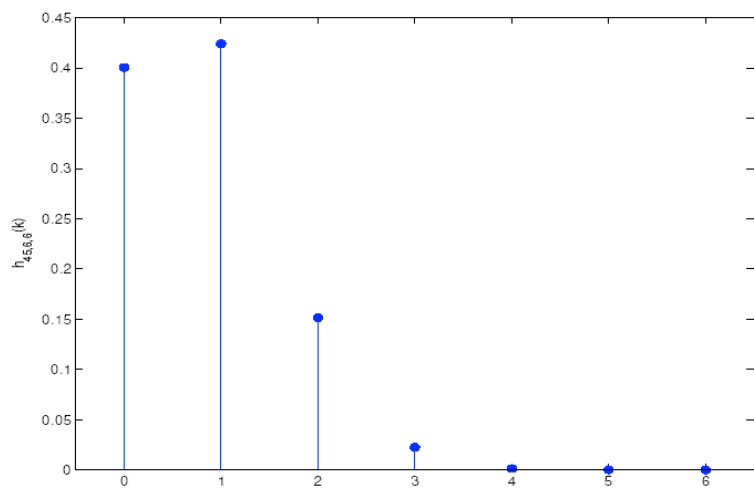
- obige Berechnungsformeln sind gleichwertig
- man darf die Parameter L und n vertauschen!
- klassisches Beispiel: Lottoziehung

$$P(A_i) = \frac{L}{N} \quad A_i := \text{[bei der } i\text{-ten Ziehung eine markierte Kugel]}$$

A_i sind aber nicht stochastisch unabhängig; jede Ziehung verändert Zusammensetzung Urne

Erwartungswert:

$$E(X) = \frac{nL}{N}$$



Beispiel: Wahrscheinlichkeitsgewichte der Lottoziehung

5.5.2. Binominalverteilung

Umschreibung in Worten:

- grundsätzlich gleiche **Situation wie bei hypergeometrischer Verteilung**
- einziger Unterschied: **N ist sehr gross** (im Verhältnis zu n)
- dies führt dazu, dass näherungsweise **stochastische Unabhängigkeit vermutet** wird
- klassisches **Beispiel:**
Befragung der Bevölkerung (N = Anzahl Menschen in Bevölkerung ist im Verhältnis zu n = Anzahl befragte Personen extrem gross; L/N strebt gegen Konstante p)

$$P(A_i) = p \quad A_i := [\text{bei der } i\text{-ten Ziehung eine markierte Kugel}]$$

A_i sind **stochastisch unabhängig** (Annahme, weil N sehr gross ist)

Definitionen

N = Kugeln in Urne (sehr grosse Zahl)

n = Anzahl Kugeln, die aus Urne gezogen werden

p = Wahrscheinlichkeit, dass eine markierte Kugel gezogen wird

X = Anzahl markierte Kugeln unter den n gezogenen

P(X = k): Wahrscheinlichkeit, dass sich unter den n gezogenen Kugeln k markierte Kugeln befinden

$b_{n,p}(k) = P(X = k)$ = Wahrscheinlichkeitsgewichte der Binominalverteilung

Bin(n,p) = binominalverteilte Zufallsvariabel mit Parametern n,p.

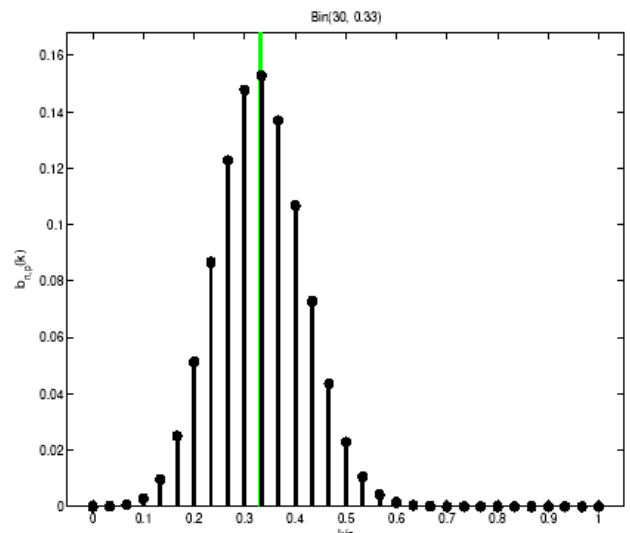
$$b_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}$$

Erwartungswert:

$$E(X) = np$$

Verlauf der Gewichtsfunktion $b_{n,p}(k)$

- Gewichte $b_{n,p}(k)$ sind **maximal, wenn $k/n = p$**
- in Graphik rechts;
 1. auf Abszisse eingetragen: k/n (so sieht man Einfluss n besser)
 2. **grüne Linie = p**; **Gewichte verteilen sich um p herum**
 3. folglich: **je höher p, desto weiter rechts** sind die Wahrscheinlichkeitsgewichte
 4. **je grösser n, desto kleiner die einzelnen Gewichte und desto näher sind sie um p herum verteilt** (in dieser Graphik nicht ersichtlich; siehe Skript S.95)



5.5.3. Geometrische Verteilung

Umschreibung in Worten:

- Ausgangslage sind **identische, stochastisch unabhängige Ereignisse**
- **Wartezeit** – wie lange (wie viele Versuche) dauert es, **bis das Ereignis erstmals eintritt?**
- $P(X = k) = g_p(k) =$ geometrische Wahrscheinlichkeitsgewichte
- **Geom(p)** = geometrisch verteilte Zufallsvariable mit Parameter p.

$$g_p(k) := (1 - p)^{k-1} p.$$

Aussage: wie gross ist die Wahrscheinlichkeit, dass das Ereignis erstmals **beim k-ten Versuch eintritt?**

$$P(X > k) = (1 - p)^k$$

Aussage: wie gross ist die Wahrscheinlichkeit, dass das Ereignis **nach k Versuch nicht eingetreten ist?**

„Gedächtnislosigkeit“ der geometrischen Verteilung:

$$P(X = \ell + k | X > \ell) = P(X = k)$$

$$P(X > \ell + k | X > \ell) = P(X > k)$$

Erwartungswert:

$$E(X) = \frac{1}{p}$$

5.5.4. Poissonverteilung

Umschreibung in Worten:

- Ausgangslage sind **seltene, stochastisch unabhängige Ereignisse**
- Grundidee: eine **Binominalverteilung** mit **grossem Parameter n und kleinem Parameter p** kann man **durch die Poissonverteilung approximieren** (graphische Darstellung siehe Skript S.100)
- **Poiss(λ)** = poissonverteilte Zufallsvariable mit Parameter λ .

$$p_\lambda(k) := \exp(-\lambda) \frac{\lambda^k}{k!}$$

wobei **$\exp(-\lambda) = e^{-\lambda}$**

Erwartungswert:

$$E(X) = \lambda$$

Aufschlussreiche Beispiel 3.25 „Telefonauskunft“: $\lambda = 5 =$ Mittlere Zahl von Anfragen während 1h

$$P(X = 0) = P[\text{keine Anfrage}] = \exp(-5) \approx 0.0067,$$

$$P(X = 1) = P[\text{genau eine Anfrage}] = \exp(-5)5 \approx 0.0337,$$

$$P(X > 5) = P[\text{mehr als 5 Anfragen}]$$

$$= 1 - P(X \leq 5) = 1 - \exp(-5) \left(1 + 5 + \frac{5^2}{2!} + \dots + \frac{5^5}{5!} \right) \approx 0.3840$$

$$P(X > 10) = P[\text{mehr als 10 Anfragen}] \approx 0.0137.$$