Wahrscheinlichkeitstheorie & Statistik

Alexander Heckmann

28. Mai 2020

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Grundlagen der Statistik

1.1.1 Grundlegende Definitionen

Begriff	Beschreibung
Grundgesamtheit	Menge aller Personen, Einheiten oder Objekte, die in Hinblick auf ein Untersuchungsziel von Relevanz sind
Merkmalträger	Element der Grundgesamtheit
Vollerhebung	Untersuchung aller Merkmalsträger der Grundgesamtheit
Stichprobe	Untersuchung einer Teilmenge der Grundgesamtheit
Repräsentativität	geeignete Auswahl einer Stichprobe als unverzerrtes Abbild der Grundgesamtheit
Merkmal	interessierende Eigenschaft eines Merkmalträgers

1.2 Daten

Symbol	Beschreibung
x	Merkmal

Daten können in verschiedener Ausführung auftreten und lassen sich entsprechend unterteilen, sie können entweder quantitativer Natur sein und als Zahlen vorliegen oder kategorialer Natur sein.

1.2.1 Quantitative Daten

Daten, die in quantitativer Form vorliegen, lassen sich wieder in zwei Kategorien unterteilen, sie können entweder diskreter oder stetiger Art sein.
Zahlen in der Statistik haben dabei die Eigenschaft, dass sie mit groben und zufälligen Fehlern behaftet sind, durch die zwei eigentlich verschiedene Zahlen aufgrund ihrer Streuung trotzdem gleich sein können. Zudem besitzen sie, im Vergleich zur Mathematik, eine konkrete physikalische Bedeutung, die nur in einen Kontext gesetzt Sinn ergibt.

Diskrete Daten

Quantitative Daten heißen diskret, wenn es endlich viele oder abzählbar unendlich viele Ausprägungen gibt.
Endliche viele Ausprägungen gibt es dann, wenn man den Wertebereich im Vorhinein schon einschränken kann.

Beispiel: Anzahl der Fußballspieler bei einem Länderspiel

Abzählbar unendlich heißt, dass eine Bijektion zwischen der Menge der Ausprägungen A und den natürlichen Zahlen $\mathbb{N}$ existiert, d.h., dass alle Elemente $x_1, \ldots, x_n \in A$ durchnummeriert werden können.
Dies ist dann der Fall, wenn man keine wirkliche Einschränkung machen kann, wie viele Ausprägungen es geben kann.

Beispiel 1: Zählung aller Fahrradfahrer an der Fahrradbrücke

Stetige Daten

Ein quantitatives Datum heißt stetig, wenn man es beliebig genau messen kann. Dabei ist gemeint, dass man zwischen zwei nahe beieinanderliegenden Ausprägungen x_1 und x_2 theoretisch immer eine weitere Ausprägung durch genauere Messung einfügen lässt, also $x \in \mathbb{R}$.

Beispiel: Körpergröße in cm

Dabei macht es oft keinen Sinn, diese Daten nach Häufigkeit einer bestimmten Ausprägung zu untersuchen. Bei stetigen Daten untersucht man deswegen beliebig groß, aber sinnhaft gewählte Intervalle $I_1, \ldots, I_n$.

$$(x_0, x_i], (x_i, x_j], (x_j, x_k], (x_k, x_n]$$

1.2.2 Kategoriale Daten

Kategoriale Daten sind Merkmale, die nicht in numerischer Form vorliegen.

Nominale Daten

Ausprägungen nominaler Datenbestände besitzen keine feste Reihenfolge.

Beispiel: Geschlechter

Nominale Merkmale können als Zahlen codiert sein, man kann mit diesen Zahlen aber nicht rechnen.

Bsp.: Zugehörigkeit zu Arbeitsgruppe, von *Tandem 01* bis *Tandem 30*

Ordinale Daten

Im Vergleich zu nominalen Daten können ordinale Datenbestände in einer natürlichen Reihenfolge angeordnet werden.

Beispiel: Noten von *sehr gut* bis *ungenügend*

1.3 Grundlagen von R

R ANWENDUNG IN R: Grundlagen

- `library("package")`: lädt das Paket *package*
- Mal-Zeichen · *immer* notwendig, ansonsten wird Multiplikation nicht erkannt
- `i:n`: Ausgabe aller ganzzahligen Elemente $[i,n]$
- `rep(i, n`: wiederholt Werte und Vektoren n mal
- `seq(i, n, c)`: Ausgabe aller Elemente $[i,n]$ mit Schrittweite c
- `x <- a`: Deklaration einer Variable
- `v <- c(x,y,z)`: Deklaration eines Vektors
- `m <- matrix(v, nrow = a, ncol = b)`: Deklaration einer Matrize mit Inhalt des Vektors `v`, `a` = Zeilen, `b` = Spalten
- `d <- data.frame(col1 = v.1, ..., coln = v.n)`: Deklaration eines DataFrames mit den benannten Spalten `col1`, ..., `coln` mit Inhalt der Vektoren `v.1`, ..., `v.n`
- `View(Object)`: Anzeigen von `Object`
- `table(v)`: Aufzählen aller Ausprägungen mit ihrer Anzahl
- `table(v, w)`: Aufzählen aller Ausprägungen mit ihrer Anzahl in Abhängigkeit einem zweiten Vektor

```
> x <- 9
> v <- c(1:x)
> m <- matrix(v, nrow = 3, ncol = 3)
> w <- c('m','w','m','m','m','m','m','m','w')
> d <- data.frame(number = v, gender = w)
> View(d)
> table(d)
```

1.3.1 Arithmetische Operatoren

Bei arithmetischen Operatoren müssen die Argumente arithmetische Ausdrücke sein, d.h. es müssen einzelne Zahlen bzw. Variablen, Vektoren oder Matrizen sein, das Ergebnis ist dementsprechend wieder eine Zahl, ein Vektor oder eine Matrix.
Da die Matrix-Multiplikation nur definiert ist, wenn die Anzahl der Spalten des linken Argumentes mit der Anzahl der Zeilen des rechten Argumentes übereinstimmen, muss diese Voraussetzung bei der Anwendung dieses Operators erfüllt sein.
Auch wenn Faktoren, d.h. nominale Ausprägungen, als numeric hinterlegt sein können, kann mit ihnen nicht gerechnet werden.

Operator	Funktion
+	Addition
−	Subtraktion
*	Multiplikation
/	Division
^	Exponentiation
%%	Modulo
%/%	ganzzahlige Division
%*%	Matrix-Multiplikation

1.3.2 Primitive Datentypen

Datentyp	Beschreibung
numeric	Zahlen, integer & double
character	Zeichen, Strings
logical	boolesche Werte, `TRUE` == 1, `FALSE` == 0
factor	nominale Ausprägung

R ANWENDUNG IN R: Primitive Datentypen

- `mode(x)`: gibt Datentyp des Elements / der Elemente an
- `as.logical(x)`: Umwandlung von `x` in Wahrheitswert, $x \neq 0 \Leftrightarrow$ `TRUE`, $x = 0 \Leftrightarrow$ `FALSE`
- `as.numeric(x)`: Umwandlung in Zahl
- `as.character(x)`: Umwandlung in Zeichen
- `as.factor`: Umwandlung in Ausprägung, gibt an welche unterschiedlichen nominalen Ausprägungen angenommen werden
- `factor(w)`: Umwandlung aller Elemente des Objekts `w` von Ursprungsdatentyp zu nominaler Ausprägung, Aufzählung aller unterschiedlichen Ausprägungen
- `levels(x)`: gibt alle unterschiedlichen Ausprägungen aus, falls x bereits ein Faktor ist

```
> x <- c(1,0,1)
> mode(x)
> as.logical(x)
> as.character(x)
> x <- as.factor(x)
> levels(x)
```

1.3.3 Komplexe Datentypen

Datentyp	Beschreibung
Vektor	Vektor bestehend aus mehreren Elementen des selben Datentyps
Matrix	Matrize bestehend aus mehreren Elementen des selben Datentyps, mit a Zeilen & b Spalten
DataFrame	Objekt, das Kombination verschiedener Datentypen zu einer Tabelle ermöglicht, mit a Zeilen & b Spalten

Vektoren

R ANWENDUNG IN R: Vektoren

- `v[i]`: Ausgabe des Elements an Stelle i des Vektors, `v[1]` = erstes Element
- `v[-i]`: Ausgabe aller Elemente außer das an Stelle `i`
- `v[i:j]`: Ausgabe aller Elemente an den Stellen `i` bis `j`
- `v[-(i:j)]`: Ausgabe aller Elemente außer denen an den Stellen `i` bis `j`
- `v[c(i,j,k)]`: Ausgabe der Elemente an den Stellen `i`, `j` & `k`
- `v[-c(i,j,k)]`: Ausgabe der Elemente außer denen an den Stellen `i`, `j` & `k`
- `v[v == i]`: Ausgabe aller Elemente, die dem Wert `i` entsprechen
- `v[v < i]`: Ausgabe aller Elemente, die kleiner als `i` sind
- `length(v)`: Länge
- `sort(v)`: aufsteigende Sortierung

```
> v <- c(1,3,2,4,5)
> sort(v)
> v[-3]
> w <- c('m','w','m','m','m','m','m','m','w')
> w <- factor(w)
> length(w)
```

Matrizen

ANWENDUNG IN R: Matrizen

- `m[j,k]`: Ausgabe des Elements in Zeile j und Spalte k
- `m[j,]`: Ausgabe aller Elemente in Zeile j
- `m[,k]`: Ausgabe aller Elemente in Spalte k
- `dim(m)`: Anzahl Zeilen, Anzahl Spalten
- `nrow(m)`: Anzahl Zeilen
- `ncol(m)`: Anzahl Spalten
- `sum(m)`: Summe aller Werte
- `rowSums(m)`: Summe der Werte innerhalb der Zeilen
- `colSums(m)`: Summe der Werte innerhalb der Spalten
- `apply(m, margin, fun)`: wendet Funktion `fun` an, `margin` $= 1 \rightarrow$ Zeile, `margin` $= 2 \rightarrow$ Spalte

```
> m <- matrix(c(1:6), nrow = 3)
> m[2,3]
> nrow(m)
> apply(m,1,min)
> rowSums(m)
```

DataFrames

ANWENDUNG IN R: DataFrames

- `d$col`: Ausgabe aller Elemente von `d` in Spalte `col`
- `d.copy <- subset(d, d$col > a)`: erstellt Untermatrix mit allen Elementen, die dem logischen Ausdruck entsprechen

```
> v <- c(0,1,2,3)
> w <- c('a','b','c','d')
> d <- data.frame(numbers = v, letters = w)
> d.copy <- subset(d, d$numbers >= 2)
```

1.3.4 Lösen von Gleichungen

ANWENDUNG IN R: Lösen von Gleichungen

`uniroot(f, i)`: Analytische Lösung einer Gleichung, *f* ist eine Funktion, der der zu erreichende Wert abgezogen wird, *i* ein Intervall, in dem die Lösung liegen muss

Bsp.: Wahrscheinlichkeit, dass von 100 Autos mit einer Wahrscheinlichkeit von 95% mind. 28 rot sind

```
eq <- function(p){
pbinom(27,100,p, lower.tail = FALSE) - 0.95
}

uniroot(eq, 0:1)$root
```

Kapitel 2

Deskriptive Statistik

2.1 Grundlagen

2.1.1 Relative Häufigkeit

Die relative Häufigkeit eines Elements x_i aus einem Datensatz x mit n Elementen gibt an, wie häufig es prozentual gesehen a Ausprägungen desselben Wertes innerhalb des Datensatzes gab.

$$h_n = \frac{a}{n}$$

$$\frac{5}{20} = 0,25 \qquad \text{(Bsp.)}$$

2.1.2 Lagemaße

Arithmetisches Mittel / Mittelwert

Das arithmetische Mittel gibt den Durchschnittswert von Daten an.

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

R ANWENDUNG IN R: Mittelwert

- `mean(x)`

```
> x <- rnorm(1000)
> mean(x)
```

Median & Quantile

Der Median gibt den Wert an, der in sortierter Reihenfolge der Daten genau in der Mitte am 50%-Quantil liegt, sodass 50% der Daten unter und 50% der Daten über diesem Wert liegen.
Analog dazu sind weitere Quantile, bspw. das 5%-Quantil, bei dem 5% der Daten drunter liegen und 95% darüber.

R ANWENDUNG IN R: Median

- `median(x)`
- `quantile(x)`

```
> x <- rnorm(1000)
> median(x)
> quantile(x)
```

Modus

Der Modus ist der Wert, der am öftesten in den Daten vorkommt, der Wert mit den meisten gleichen Ausprägungen.

R ANWENDUNG IN R: Modus

```
- getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
  }
```

```
> x <- rnorm(1000)
> getmode(x)
```

2.1.3 Streuungsmaße

Varianz

Die Varianz gibt die mittlere quadratische Abweichung der Datenausprägungen vom Mittelwert an.

$$VAR(x) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

R ANWENDUNG IN R: Varianz

- `var(x)`

```
> x <- rnorm(1000)
> var(x)
```

Standardabweichung

Die Standardabweichung gibt an, wie weit die einzelnen Daten verteilt sind, bspw. wie weit die einzelnen Daten im Durchschnitt vom Mittelwert entfernt sind.

$$SD(x) = \sqrt{VAR(x)}$$

R ANWENDUNG IN R: Standardabweichung

- `sd(x)`

```
> x <- rnorm(1000)
> sd(x)
```

MAD

Der MAD gibt die mittlere Abweichung einer Stichprobe vom Median an.

$$MAD(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} |(x_i - \tilde{x})|$$

R ANWENDUNG IN R: MAD

- `mad(x)`

```
> x <- rnorm(1000)
> mad(x)
```

IQR (Interquantile Range)

Sortiert man eine Stichprobe nach Größe, so gibt der IQR an, wie breit das Intervall ist, in dem die mittleren 50% der Ausprägungen liegen.

$$IQR = x_{0.75} - x_{0.25}$$

R ANWENDUNG IN R: IQR

- `IQR(x)`

```
> x <- rnorm(1000)
> IQR(x)
```

2.2 Transformationen

Sowohl die lineare als auch die nicht lineare Transformationsfunktionen sind monoton wachsend, d.h. $x_1 \leq x_2 \Rightarrow f(x_1) \leq f(x_2)$. Die Transformation der Daten ändert nichts an deren Quantilen.

$$med(f(x)) = f(med(x))$$

2.2.1 Lineare Transformation

Manchmal kann es sein, dass Daten in einer anderen Einheit gemessen wurden, als der verwendete Standard. Wenn dies der Fall ist, nutzt man die lineare Transformation zur Umwandlung der Daten x in die Zielgröße. Sie hat ihren Namen, da die Daten aus dem Urbild durch eine lineare Funktion abgebildet werden. Die Transformation behält die Form der Verteilung bei, Die Transformationsfunktion besitzt folgende Form:

$$y = a \cdot x_i + b$$

Die Parameter a & b sind dabei von der Umrechnungsformel abhängig, es kann auch sein, dass $a = 1$ und/oder $b = 0$ und somit wegfallen.

$$x_F = 1,8 \cdot x_C + 32 \qquad \text{(Grad Celsius} \rightarrow \text{Grad Fahrenheit; } a = 1,8;\ b = 32)$$
$$x_K = x_C - 273,15 \qquad \text{(Grad Celsius} \rightarrow \text{Grad Kelvin; } a = 1;\ b = -273,15)$$
$$x_{USD} = 1,12 \cdot x_{EUR} \qquad \text{(Euro} \rightarrow \text{US-Dollar; } a = 1,12;\ b = 0)$$

Zur Rückumwandlung der transformierten Daten in die ursprüngliche Skala kann die Umkehrfunktion verwendet werden:

$$x = \frac{y - b}{a}$$

Es gelten folgende Regeln für Lage- und Streuungsmaße für linear transformierte Funktionen:

Zusammenfassung:

$$\bar{y} = a \cdot \bar{x} + b$$
$$\tilde{y} = \tilde{x}$$
$$VAR(y) = a^2 \cdot VAR(x)$$
$$SD(y) = |a| \cdot \sqrt{VAR(x)} = |a| \cdot SD(x)$$

Mittelwert

$$\bar{y} = a \cdot \bar{x} + b$$

BEWEIS:

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \qquad \text{(Einsetzen in Formel)}$$
$$= \frac{1}{n} \cdot \sum_{i=1}^{n} (a \cdot x_i + b) \qquad \text{(Resub.: } a \cdot x_i + b = y_i\text{)}$$
$$= a \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i + \frac{1}{n} \sum_{i=1}^{n} b \qquad \text{(Ausklammern von } a \text{ \& } b\text{)}$$
$$= a \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i + b \qquad \text{(Zusammenfassen zu } b\text{)}$$
$$= a \cdot \bar{x} + b \qquad \text{(Subst.: } \bar{x} = \tfrac{1}{n} \cdot \textstyle\sum_{i=1}^{n} x_i\text{)}$$
$$\Rightarrow \bar{y} = a \cdot \bar{x} + b$$

Varianz

$$VAR(y) = a^2 \cdot VAR(x)$$

BEWEIS:

$$VAR(y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{(Einsetzen in Formel)}$$
$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (a \cdot x_i + b - \bar{y})^2 \qquad \text{(Resub.: } a \cdot x_i + b = y_i\text{)}$$
$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (a \cdot x_i + b - a \cdot \bar{x} - b)^2 \qquad \text{(Resub.: } a \cdot \bar{x} + b = \bar{y}\text{)}$$
$$= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (a \cdot x_i - a \cdot \bar{x})^2 \qquad (b \text{ fällt weg, da } +b - b = 0)$$
$$= a^2 \cdot \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{(Ausklammern von } a\text{)}$$
$$= a^2 \cdot VAR(x) \qquad \text{(Subst.: } VAR(x) = \tfrac{1}{n-1} \cdot \textstyle\sum_{i=1}^{n} (x_i - \bar{x})^2\text{)}$$
$$\Rightarrow VAR(y) = a^2 \cdot VAR(x)$$

Standardabweichung

$$SD(y) = |a| \cdot SD(x)$$

BEWEIS:

$$\begin{aligned} SD(y) &= \sqrt{VAR(y)} \\ &= |a| \cdot \sqrt{VAR(x)} = |a| \cdot SD(x) \\ &\Rightarrow SD(y) = |a| \cdot SD(x) \end{aligned}$$

Z-Transformation / Standardisierung

Die Z-Transformation ist eine besondere Art der linearen Transformation. Sie wird benutzt, wenn skalenfreie Daten verwendet werden sollen, da die Z-Transformation die vorliegenden Daten in einheitenlose Daten umwandelt und somit Vergleichbarkeit schafft.
Die Transformationsfunktion ist wie folgt definiert:

$$z = \frac{x - \bar{x}}{SD(x)} = \frac{1}{SD(x)} \cdot x - \frac{\bar{x}}{SD(x)}$$

Dabei ist der Parameter $a = \frac{1}{SD(x)}$ und die Verschiebung $b = -\frac{\bar{x}}{SD(x)}$.

Da $\bar{y} = a \cdot \bar{x} + b$ gilt, gilt auch:

$$\begin{aligned} \bar{z} &= \frac{1}{SD(x)} \cdot \bar{x} - \frac{\bar{x}}{SD(x)} \\ &= \frac{\bar{x}}{SD(x)} - \frac{\bar{x}}{SD(x)} = 0 \end{aligned}$$

Da $VAR(y) = a^2 \cdot VAR(x)$ gilt, gilt auch:

$$VAR(z) = (\frac{1}{SD(x)})^2 \cdot VAR(x)$$

Da $SD(y) = |a| \cdot \sqrt{VAR(x)}$ gilt, gilt auch:

$$SD(z) = |\frac{1}{SD(x)}| \cdot SD(x) = 1$$

R ANWENDUNG IN R: Z-Transformation / Standardisierung

- `scale(x)`

```
> x <- rnorm(1000)
> scale(x)
```

2.2.2 Nicht-lineare Transformation

Bei der nichtlinearen Transformation werden die Daten durch nichtlineare Funktionen, bspw. den Logarithmus oder die Wurzelfunktion, transformiert.

log-Transformation

Die log-Transformation wirkt zwar nicht formerhaltend, transformiert die Daten jedoch so, dass multiplikative Erhöhungen in additive umgewandelt werden. Dies dient der besseren Darstellbarkeit.

R ANWENDUNG IN R: log-Transformation

- `log(x)`: *log* zur Basis e
- `log(x, base = b`: *log* zur Basis b

```
> x <- rnorm(1000)
> x <- log(x)
> y <- rnorm(1000)^2
> y <- log(y, base = 10)
```

2.3 Multivariate Merkmale

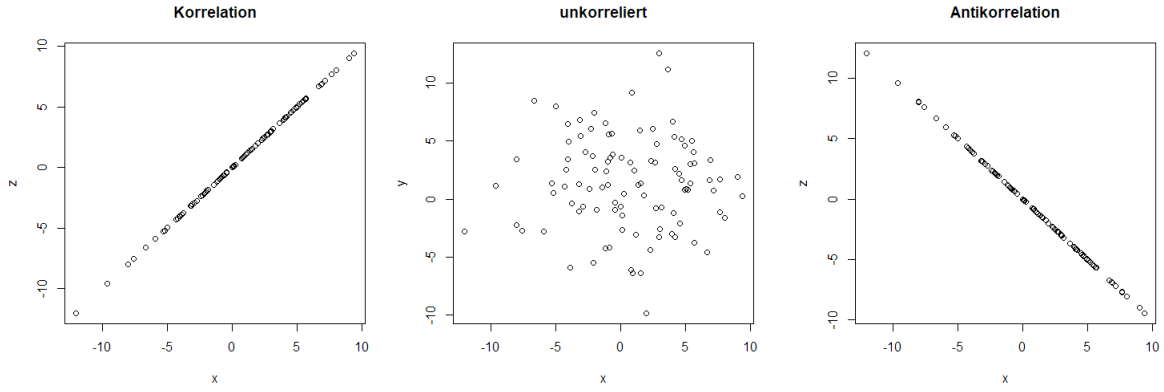
2.3.1 Korrelation

Die Korrelation ist ein Indiz dafür, dass Daten *möglicherweise* zusammenhängen. Die Korrelation beschreibt sowohl die Stärke dieser Zusammenhänge, als auch die Richtung.

$$\begin{aligned} cor > 0 &\Rightarrow \text{Korrelation} \\ cor \approx 0 &\Rightarrow \text{unkorreliert} \\ cor < 0 &\Rightarrow \text{Antikorrelation} \end{aligned}$$

Eine Korrelation gibt an, dass, wenn eine Variable größer bzw. kleiner wird, die andere ebenfalls größer bzw. kleiner wird. Analog dazu gibt eine Antikorrelation an, dass, wenn eine Variable größer bzw. kleiner wird, die andere kleiner bzw. größer wird.

Zur Veranschaulichung:



Pearson-Korrelationseffizient

Der Pearson-Korrelationseffizient evaluiert die linearen Zusammenhänge zweier Variablen. Ein Zusammenhang ist linear, wenn das Ändern einer Variable eine proportionale Änderung der anderen Variable bewirkt. Beim Pearson-Korrelationseffizient wird mit den Werten der Daten gearbeitet.

$$cor = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

R ANWENDUNG IN R: Pearson-Korrelationseffizient

- `cor(x, y, method = "pearson")`

```
> x <- rnorm(1000)
> y <- rnorm(1000)^2
> cor(x, y, method = "pearson")
```

Spearman-Korrelationseffizient

Der Spearman-Korrelationseffizient evaluiert die monotonen Zusammenhänge zweier Variablen. Ein Zusammenhang ist monoton, wenn durch Änderung einer Variable eine Änderung der anderen bewirkt, jedoch nicht immer mit einer konstanten Rate. Beim Spearman-Korrelationseffizienten wird mit der Rangfolge der Daten gearbeitet. Das heißt, dass beide Datensätze erst sortiert werden und dann werden die jeweiligen Ränge miteinander vergleichen.

$$cor = \frac{\sum_{i=1}^{n}(rang(x_i) - \bar{rang}(x)) \cdot (rang(y_i) - \bar{rang}(y))}{\sqrt{\sum_{i=1}^{n}(rang(x_i) - \bar{rang}(x))^2} \cdot \sqrt{\sum_{i=1}^{n}(rang(y_i) - \bar{rang}(y))^2}}$$

R ANWENDUNG IN R: Spearman-Korrelationseffizient

- `cor(x, y, method = "spearman")`

```
> x <- rnorm(1000)
> y <- rnorm(1000)^2
> cor(x, y, method = "spearman")
```

2.4 Darstellung & Plots

2.4.1 Darstellung kategorieller Daten

Kuchendiagramm

Eine Möglichkeit, kategorielle Daten zu visualisieren ist, ein Kuchendiagramm zu erstellen. Dabei entspricht die Größe eines Stücks der relativen Häufigkeit der jeweiligen Ausprägung.
Kuchendiagramme können manipulierend wirken, wenn sie dreidimensional dargestellt werden und dadurch die Größe eines Stücks eventuell größer wirkt, als es ist.

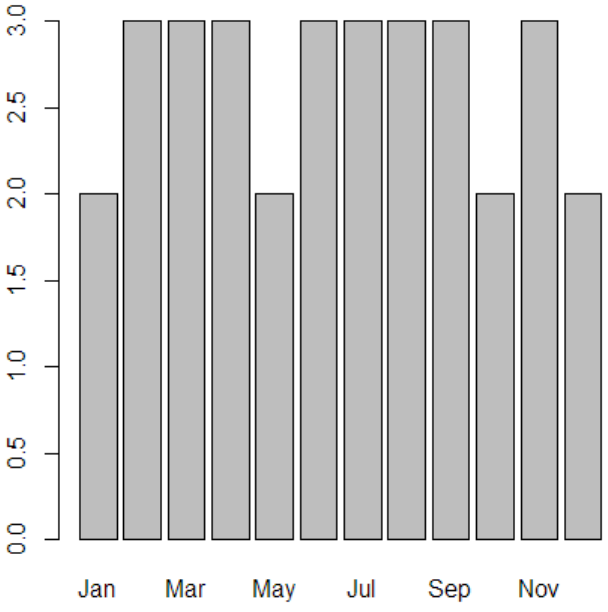
Balkendiagramm

Kategorielle Daten können auch als Balkendiagramm dargestellt werden, dabei entspricht die Höhe der Balken der relativen Häufigkeit der Ausprägungen.
Sie können in der Hinsiht manipuliert sein, dass die Balken eine verschiedene Breite besitzen. Durch die erhöhte Breite haben die Balken einen größen Flächeninhalt und wirken größer und somit, als ob sie eine höhere Häufigkeit der Ausprägung aufweisen.

R ANWENDUNG IN R: Balkendiagramme

- `barplot(x)`
- `barplot(table(df$col))`: Visualisierung kategorieller oder logischer Daten

```
> x <- rep(c(TRUE, TRUE, FALSE))
> barplot(table(x))
```



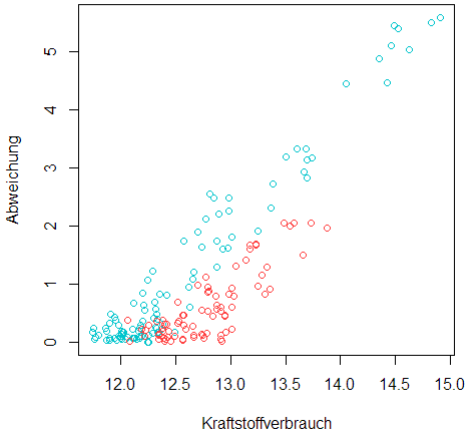
2.4.2 Plot

Es gibt verschiedene Möglichkeiten, Plots zu erstellen:

- Scatter Plots
- Line Plots
- Histogram-like Plots

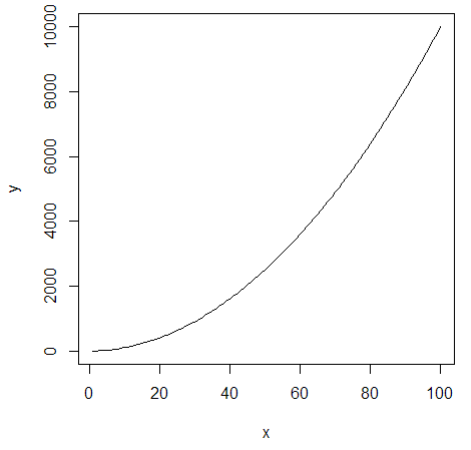
Scatter Plots

Scatter Plots stellen die genaue Position der Daten dar, entweder in Abhängigkeit zu einem gegebenen y-Vektor, wobei beide Vektoren gleich lang sein müssen, oder aber in Abhängigkeit vom Index.



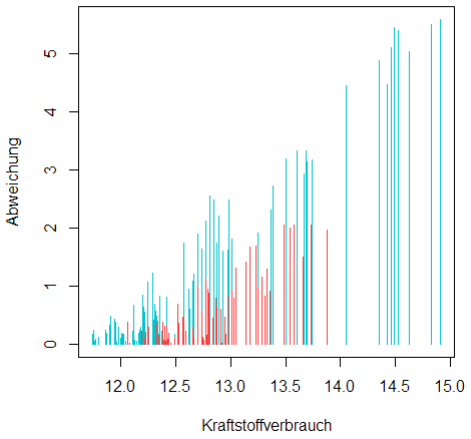
Line Plots

Line Plots stellen die Position der Daten dar, werden jedoch durch eine Linie miteinander verbunden bzw. interpoliert.



Histogram-like Plots

Histogram-like Plots stellen die genaue Position der Daten dar, werden dabei jedoch durch mit durchgezogenen vertikalen Linien zur x-Achse eingezeichnen.



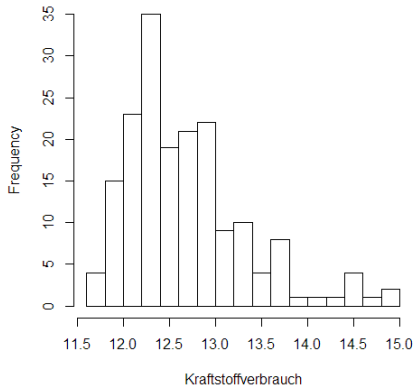
R ANWENDUNG IN R: Plots

- `plot(x)`:
  Scatter Plot von `x`; x-Achse = Index, y-Achse = `x`
- `plot(x, type = "l")`:
  Line Plot von `x`; x-Achse = Index, y-Achse = `x`
- `plot(x, type = "h")`:
  Histogram-like Plot von x; x-Achse = Index, y-Achse = `x`
- `plot(x, y)`:
  Scatter Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`
- `plot(x, y, type = "l")`:
  Line Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`
- `plot(x, y, type = "h")`:
  Histogram-like Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`
- `plot(y ~ x)`:
  Scatter Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`
- `plot(y ~ x, type = "l")`:
  Line Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`
- `plot(y ~ x, type = "h")`:
  Histogram-like Plot von `x` in Abhängigkeit von `y`; x-Achse = `x`, y-Achse = `y`

```
> x <- rnorm(1000)
> y <- rpois(1000, 8)
> plot(x, type = "h")
> plot(x, y, type = "l")
> plot(y ~ x)
```

2.4.3 Histogramm

Histogramme bilden die relative bzw. absolute Häufigkeitsverteilung eines Datensatzes ab. Dabei werden die dargestellten Daten in verschiedene Intervalle $I_1, \ldots, I_n$ unterteilt.



R ANWENDUNG IN R: Histogramme

- `hist(x)`: Histogramm von `x`
- `hist(x, freq = FALSE)`: Histogramm der relativen Häufigkeitsdichte

```
> x <- rnorm(1000)
> hist(x, freq = FALSE)
```

2.4.4 Hinzufügen von Elementen

Sowohl bei Plots als auch bei Histogrammen lassen sich Elemente hinzufügen.

Scatter Plot

R ANWENDUNG IN R: Hinzufügen von Scatter Plots

- `points(x,y)`

```
> x <- rnorm(1000)
> y <- rnorm(1000)
> plot(-4:4, -4:4)
> points(x, y, col = "blue")
```

Line Plot

R ANWENDUNG IN R: Hinzufügen von Line Plots

- `lines(x,y)`

```
> x <- rnorm(1000)
> plot(x, dnorm(x))
> lines(seq(-5, 5, by = 0.01), dnorm(seq(-5, 5, by = 0.01))
```

2.4.5 Boxplot

Die untere Grenze gibt das 25%-Quantil der Daten an.
Der mittlere Strich gibt den Median, das 50%-Quantil an.
Die obere Grenze gibt das 75%-Quantil an.
Elemente, die außerhalb der Box liegen, sind Ausreißer.

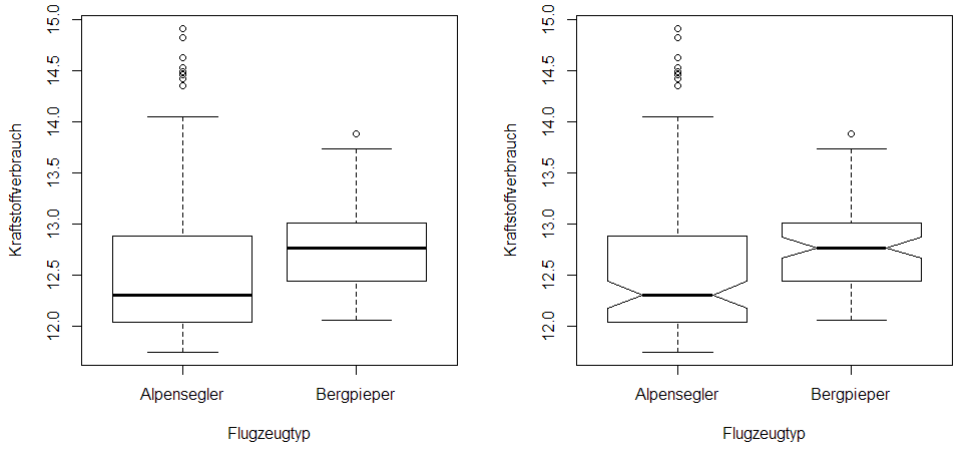
Boxplot mit Notches

Die Notches geben das 95%-Konfidenzintervall des Median an.
Bei überschneidenden Notches ist keine statistische Aussage möglich, da es möglicherweise ein Zufallseffekt ist.

R ANWENDUNG IN R: Boxplots

- `boxplot(x)`
- `boxplot(x, notch = TRUE)`
- `boxplot(x, y)`
- `boxplot(x, y, notch = TRUE)`
- `boxplot(y ~ x)`
- `boxplot(y ~ x, notch = TRUE)`

```
> x <- rnorm(1000)
> boxplot(x, notch = TRUE)
```



Kapitel 3

Wahrscheinlichkeitsrechnung

3.1 Definition der Wahrscheinlichkeitsrechnung

3.1.1 Definition: Laplace

Die Wahrscheinlichkeit eines Ereignisses ist gegeben durch die Rate zwischen der Anzahl der Ereignisse und der Gesamtanzahl der Ergebnisse.
Wenn $A = \Omega$, nennt man A das sichere Ergebnis.

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

$$P(A = \textit{gerade Augenzahl}) = \frac{3}{6}$$

3.1.2 Definition: Grenzwert der relativen Häufigkeit

Die Wahrscheinlichkeit eines Ereignisses ist gegeben durch den Grenzwert der relativen Häufigkeit, wenn man die Stichprobengröße gegen Unendlich laufen lässt.

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

3.1.3 Definition: Kolmogorov-Axiome

Eine Abbildung $P(A)$, die jedem Ereignis A eine reelle Zahl $p \in [0,1]$ zuordnet, nennt man Wahrscheinlichkeit, wenn folgendes zutrifft:

1. $P(A) \geq 0$: Nichtnegativität für jedes Ereignis A
2. $P(\Omega) = 1$: Die Summe aller Abbildungen aller Ergebnisse muss 1 ergeben.
3. $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ für $A_1 \cap A_2 = \varnothing$: Wenn zwei Ereignisse unabhängig sind, ist die Abbildung der Vereinigung beider Ereignisse gleich der Summe der Abbildungen beider Ereignisse.

Rechenregeln abgeleitet aus den Kolmogorov-Axiomen

1. $\bar{A} = \Omega \setminus A$
2. $P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1$
3. $P(\varnothing) = 1 - P(\Omega) = 0$
4. $A \subseteq B \Rightarrow P(A) \leq P(B)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3.2 Grundlagen

3.2.1 Zufallsexperiment

Ein Zufallsexperiment ist Versuch, bei dem das Ergebnis nicht deterministisch vorbestimmt ist. Sie können unter gleichen Bedingungen beliebig oft wiederholt werden.
Bsp.: Münzwurf

3.2.2 Zufallsvariable

Das Ergebnis eines Zufallsexperiments wird einer Zahl X zugeordnet. Die Realisation der Zufallsvariable meint den tatsächlich beobachteten Wert.
X ist also eine Abbildung $\Omega \to \mathbb{R}$: jedem Ergebnis wird eine Zahl zugeordnet.

3.2.3 Ergebnisraum

Die Menge aller möglichen Ergebnisse heißt Ergebnisraum Ω, die Ergebnisse sind Elemente von Ω.
Die Anzahl der möglichen Ergenisse ist gegeben durch $|\Omega|$.

Bsp.: Würfeln

$$\Omega = 1,2,3,4,5,6$$
$$|\Omega| = 6$$

3.2.4 Ereignis

Als Ereignis bezeichnet man eine Konstellation von Ergebnissen, wobei $A \subseteq \Omega$.
Ein Ereignis gilt als eingetroffen, wenn ein passendes Ergebnis aus dem Zufallsexperiment resultiert.

Bsp.: gerade Augenzahl beim Würfeln

$$A = 2,4,6$$
$$|A| = 3$$

3.2.5 Unabhängigkeit von Ereignissen

Zwei Ereignisse heißen unabhängig, wenn sie sich nicht beeinflussen.

3.2.6 Bedingte Wahrscheinlichkeit

$P(A|B)$ beschreibt die bedingte Wahrscheinlichkeit von A, gegeben, dass B bereits eingetreten ist.
Wenn gegeben ist, dass ein Ereignis B bereits eingetreten ist, geht man vom Wahrscheinlichkeitsraum Ω zu einem kleineren Wahrscheinlichkeitsraum B über. Dadurch werden alle Ergebnisse, die in B liegen, wahrscheinlicher und alle Ereignisse, die nicht in B liegen, unmöglich.

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

Satz von Bayes

Der Satz von Bayes beschreibt die Beziehung zwischen den zwei bedingten Wahrscheinlichkeiten $P(A|B)$ und $P(B|A)$ und erlaubt somit das Umkehren von bedingten Wahrscheinlichkeiten.

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

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B})}$$

$$P(B|A) = \frac{0,1 \cdot 0,55}{0,1 \cdot 0,55 + 0,14 \cdot 0,45} \qquad \text{(Bsp.)}$$

3.2.7 Gesetz der großen Zahlen

Das Gesetz der großen Zahlen sagt aus, dass, für eine steigende Anzahl an Stichproben, der Mittelwert immer weiter gegen den Erwartungswert geht.

$$\lim_{n \to \infty} \bar{X}_n \longrightarrow E(X)$$

3.2.8 Zentraler Grenzwerstsatz

Wird die Summe S aus vielen Zufallsvariablen gebildet, die alle aus derselben Verteilung stammen, so folgt S approximativ einer Normalverteilung, falls $E(X)$ und $VAR(X)$ existieren.

$$S_n = \sum_{i=1}^{n} X_i \overbrace{\sim}^{a} N(\mu = n \cdot E(X), \sigma^2 = n \cdot VAR(X)) \quad \text{asymptotisch}$$

3.3 Kombinatorik

3.3.1 Urnenmodelle

Ohne Ordnung & ohne Zurücklegen

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

$$\binom{49}{6} = \frac{49!}{(49-6)! \cdot 6!} = \frac{49!}{43! \cdot 6!} = 13983816 \qquad \text{(Lotto)}$$

R ANWENDUNG IN R: Ohne Ordnung & ohne Zurücklegen

- `choose(n, k)`
- `factorial(n)/(factorial(n - k)*factorial(k))`

```
> n <- 49
> k <- 6
> choose(n, k)
```

Bsp.: Geburtstagsproblem

```
> k <- 23
> n <- 365
> p <- 1
> for (i in 0:k-1){
> p <- p * (n - i) / k
> }
> 1 - p
```

Ohne Ordnung & mit Zurücklegen

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)! \cdot k!}$$

$$\binom{6+3-1}{3} = \frac{(6+3-1)!}{(6-1)! \cdot 3!} = 70 \qquad \text{(Möglichkeiten bei 3 Würfen)}$$

R ANWENDUNG IN R: Ohne Ordnung & mit Zurücklegen

- `choose(n + k - 1, k)`
- `factorial(n + k - 1)/(factorial(n-1)*factorial(k))`

```
> choose(49 - 6 - 1, 6)
```

Mit Ordnung & ohne Zurücklegen

$$\frac{n!}{(n-k)!}$$

$$\frac{8!}{(8-3)!} = 336 \qquad \text{(Bsp.: Gewinner Viertelfinales)}$$

R ANWENDUNG IN R: Mit Ordnung & ohne Zurücklegen

- `factorial(n)/(factorial(n-k)`

```
> factorial(8) / factorial(8 - 3)
```

Spezialfall, falls alle Kugeln gezogen werden, also $n = k$:

$$n!$$

$$26! \qquad \text{(Bsp.: Permatutationen des Alphabets)}$$

R ANWENDUNG IN R: Mit Ordnung & ohne Zurücklegen: Spezialfall 1

- `factorial(n)`

```
> factorial(26)
```

Spezialfall, falls es n Kugeln gibt, unter denen es k Duplikate gibt mit einer Häufigkeit von n_i:

$$\frac{n!}{\prod_{i=1}^{k} n_i!}$$

$$\frac{4!}{2! \cdot 2!} = 6 \qquad \text{(Bsp.: Permutationen von } \textit{ANNA}\text{)}$$

R ANWENDUNG IN R: Mit Ordnung & ohne Zurücklegen: Spezialfall 2

- `factorial(n)/(factorial(n-k)`

```
> factorial(4) / (factorial(2) * factorial(2))
```

Mit Ordnung & mit Zurücklegen

$$n^k$$

$$10^3 = 1000 \qquad \text{(Bsp.: Zahlenschloss)}$$

R ANWENDUNG IN R: Mit Ordnung & mit Zurücklegen

- `factorial(n)/(factorial(n-k)`

Bsp.: Anzahl Kombinationen eines dreistelligen Zahlenschlosses

```
> n <- 10
> k <- 3
> n^k
```

3.4 Diskrete Wahrscheinlichkeitsverteilungen

Wahrscheinlichkeitsfunktion

Die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariable gibt zu jedem Wert x_i die entsprechende Wahrscheinlichkeit p_i an.

Kumulative Verteilungsfunktion

Zu jeder diskreten Wahrscheinlichkeitsfunktion gibt es eine kumulative Verteilungsfunktion. Diese gibt an, wie groß die Wahrscheinlichkeit ist, dass $P(X \leq k)$ gilt.

R ANWENDUNG IN R: Kumulative Verteilungsfunktion

- `plot(ecdf(x)), verticals = TRUE)`

```
> x <- rbinom(10, 20, 0.5)
> plot(ecdf(x)), verticals = TRUE)
```

Erwartungswert

Der Erwartungswert gibt das Durchschnittsergebnis von Verteilungen an.

$$\mu = E(X) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \cdot p_i$$
$$E(a \cdot X + b) = a \cdot \bar{X} + b$$

Seien X und Y zwei unabhängige, unkorrelierte Zufallsvariablen, dann gilt:

$$E(X \pm Y) = E(X) \pm E(Y)$$

R ANWENDUNG IN R: Erwartungswert

- `mean(x)`

```
> x <- rnorm(1000)
> mean(x)
```

Varianz

Die Varianz gibt die mittlere quadratische Abweichung der Zufallsvariablen vom Erwartungswert an.

$$\sigma^2 = VAR(X) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X - E(X))^2$$
$$VAR(a \cdot X + b) = a^2 \cdot VAR(X)$$

Seien X und Y zwei unabhängige, unkorrelierte Zufallsvariablen, dann gilt:

$$VAR(X - Y) = VAR(X + Y) = VAR(X) + VAR(Y)$$

R ANWENDUNG IN R: Varianz

- `var(x)`

```
> x <- rnorm(1000)
> var(x)
```

Standardabweichung

Die Standardabweichung gibt an, wie weit die einzelnen Zufallsvariablen verteilt sind, also wie weit sie im Durchschnitt vom Mittelwert entfernt sind.

$$\sigma = SD(X) = \sqrt{VAR(X)}$$
$$SD(a \cdot X + b) = |a| \cdot SD(X)$$

R ANWENDUNG IN R: Standardabweichung

- `sd(x)`

```
> x <- rnorm(1000)
> sd(x)
```

3.4.1 Bernoulli-Verteilung

Die Bernoulli-Verteilung nutzt man zur Beschreibung von Zufallsereignissen, die nur zwei mögliche Ausgänge besitzen.
Bsp.: Geschlecht bei Neugeborenen

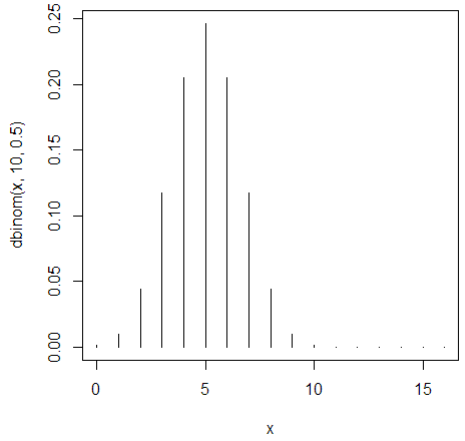
$$E(X) = p$$
$$VAR(X) = p - p^2$$

3.4.2 Binomialverteilung

Die Binomialverteilung beschreibt n unabhängig voneinander ausgeführte Wiederholungen eines Bernoulli-Versuchs mit Wahrscheinlichkeit p. Jedes dieser Zufallsexperimente hat also nur zwei mögliche Ausgänge, die immer dieselbe Wahrscheinlichkeit besitzen. Die Binomialverteilung wird also dann verwendet, wenn man die Wahrscheinlichkeit, dass ein bestimmtes Ereignis eintritt, berechnen will. Hierbei wird die Anzahl der Erfolge k bei n Versuchen gezählt.
Bsp.: Anzahl kaputter Schrauben bei einer Stichprobe in der Produktion

$$Binom_{n,p}(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$
$$E(X) = n \cdot p$$
$$VAR(X) = n \cdot p \cdot (1-p)$$

Die Verteilung hat folgende Form:



R ANWENDUNG IN R: Binomial-Verteilung

- `dbinom(x, n, p)`: $P(X = k)$, wobei k Zahl oder Vektor ist
- `pbinom(x, n, p)`: $P(X \leq k)$
- `pbinom(x, n, p, lower.tail = FALSE)`: $P(X > k)$
- `qbinom(p, n, p)`: p-Quantil der Verteilung, $p \in [0,1]$
- `rbinom(n, 1, p)`: Ziehung von x Zufallszahlen

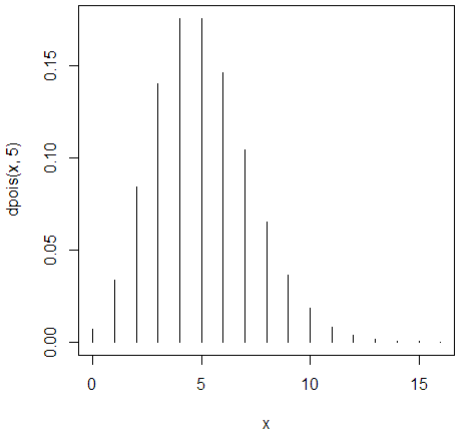
```
> n <- 100
> p <- 0.5
> dbinom(31, n, p)
> x <- rbinom(31, n, p)
> dbinom(x, n, p)
> pbinom(31, n, p)
> pbinom(31, n, p, lower.tail = FALSE)
```

3.4.3 Poisson-Verteilung

Die Poisson-Verteilung ist eine diskrete Verteilung, mit der die Anzahl eines Ereignisses modelliert werden kann, die bei konstanter mittlerer Rate unabhängig voneinander in einem festen Zeitintervall oder Gebiet eintreten. Sie wird verwendet, wenn ein durchschnittlicher Wert λ für ein eintretendes Ereignis bekannt ist und man herausfinden will, wie groß die Wahrscheinlichkeit ist, dass k Ereignisse eintreten.
Bei der Poissonverteilung gibt es keine klar definierte Obergrenze für die Anzahl an eintretenden Ereignissen.

$$P_\lambda(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$
$$E(X) = \lambda$$
$$VAR(X) = \lambda$$

Die Verteilung hat folgende Form:



R ANWENDUNG IN R: Poisson-Verteilung

- `dpois(x, lambda)`: $P(k = x)$
- `ppois(x, lambda)`: $P(k \leq x)$
- `ppois(x, lambda, lower.tail = FALSE)`: $P(k > x)$
- `qpois(p, x, lambda)`: p-Quantil der Verteilung
- `rpois(n, lambda)`: Ziehung von n poisson-verteilten Zufallsvariablen

```
> lambda <- 100/86400
> x <- rpois(86400, lambda)
> plot(1:86400, x)
> queries <- which(x == 1)
> ppois(5200, lambda)
> dpois(5200, lambda)
```

3.5 Stetige Wahrscheinlichkeitsverteilungen

Wahrscheinlichkeit

Um die Wahrscheinlichkeit stetiger Verteilungen berechnen zu können, muss man einen Bereich $[a,b]$ der Dichtefunktion integrieren.

$$p(x \leq b) \Rightarrow \int_{-\infty}^{b} f(x)dx$$

$$p(x > a) \Rightarrow 1 - \int_{-\infty}^{b} f(x)dx$$

$$p(a \leq x \leq b) \Rightarrow \int_{a}^{b} f(x)dx$$

Erwartungswert

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Varianz

$$\sigma^2 = VAR(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx$$

R ANWENDUNG IN R: Integration

1. univariate Funktion aufstellen:
```
f <- function(x){
return(1/6*x)
}
```

2. Integral bilden
```
integrate(f, lower = a, upper = b)
```

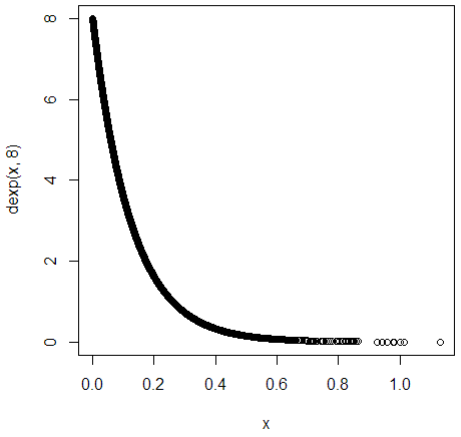
```
> f <- function(x){
> return(1/6*x)
> }
> integrate(f, lower = 1, upper = 4)
```

3.5.1 Exponentialverteilung

Die Exponentialverteilung ist ein Wahrscheinlichkeitsmodell für Zeitintervalle, bpsw. die Wartenschlangentheorie oder die Zeit zwischen Queries an einen Server. Weitere Anwendungen der Exponentialverteilung ist die Lebensdauer mechanischer Bauteile oder die Halbwertszeit von Radionukliden.

$$f_\lambda(x) = \lambda \cdot e^{-\lambda \cdot x}$$
$$E(x) = \frac{1}{\lambda}$$
$$VAR(X) = \frac{1}{\lambda^2}$$

Sie hat folgende Form:



R ANWENDUNG IN R: Exponentialverteilung

- `dexp(x, lambda)`: $P(X = k)$, wobei k Zahl oder Vektor ist
- `pexp(x, lambda)`: $P(X \leq k)$
- `pexp(x, lambda, lower.tail = FALSE)`: $P(X > k)$
- `qexp(p, lambda)`: p-Quantil der Verteilung, $p \in [0,1]$
- `rexp(n, lambda)`: Ziehung von n Zufallszahlen

```
> lambda <- 12
> dexp(31, lambda)
> x <- rexp(42, lambda)
> dexp(x, lambda)
> pexp(31, lambda)
> pexp(31, lambda, lower.tail = FALSE)
```

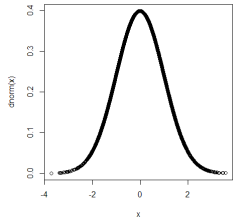
3.5.2 Normalverteilung

Die Normalverteilung ist eine der wichtigsten Verteilungen der Wahrscheinlichkeitstheorie. Sie wird benutzt, um die zufällige Streuung von Messwerten oder die zufällige Abweichung vom Sollmaß zu beschreiben. Zudem kommt sie sehr oft in der Natur vor, in der Biologie bei der Verteilung menschlicher Körpergrößen, des IQ und des EQ, bei physikalischen Sachverhalten wie der Sonnendauer an einem Tag pro Jahr oder bei statistischen Fehlern. Eine besondere Bedeutung hat die Normalverteilung durch den zentralen Grenzwertsatz, dem zufolge die additive Überlagerung vieler kleiner unabhängiger Zufallseffekte zu einem Gesamteffekt zumindest annähernd eine Normalverteilung ergibt.

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma}} \cdot e^{-\frac{1}{2} \cdot (\frac{x-\mu}{\sigma})^2}$$
$$E(X) = \mu$$
$$VAR(X) = \sigma^2$$

Verändert man den Parameter μ, so verschiebt sich die Funktion entlang der x-Achse. Verändert man den Parameter σ, so wird die Funktion gestreckt bzw. gestaucht.

Form der Verteilung:



R ANWENDUNG IN R: Normalverteilung

- `dnorm(x)`: $P(X = x)$, Standardnormalverteilung
- `dnorm(x, mu, sigma)`: $P(X = k)$, wobei k Zahl oder Vektor ist
- `pnorm(x)`: $P(X = x)$, Standardnormalverteilung
- `pnorm(x, mu, sigma)`: $P(X \leq k)$
- `pnorm(x, lower.tail = FALSE)`: $P(X > k)$, Standardnormalverteilung
- `pnorm(x, mu, sigma, lower.tail = FALSE)`: $P(X > k)$
- `qnorm(p)`: p-Quantil der Standardnormalverteilung, $p \in [0,1]$
- `qnorm(p, mu, sigma)`: p-Quantil der Verteilung, $p \in [0,1]$
- `rnorm(n)`: Ziehung von n Zufallszahlen
- `rnorm(n, mu, sigma)`: Ziehung von n Zufallszahlen

```
> mu <- 100
> sigma <- 0.5
> dnorm(31, mu, sigma)
> x <- rnorm(31, mu, sigma)
> dnorm(x, mu, sigma)
> pnorm(31)
> pnorm(31, lower.tail = FALSE)
```

Kapitel 4

Induktive Statistik

4.1 Schätzer

In der Statistik hat man sehr oft Daten gegeben, von denen man die Verteilung kennt, nicht jedoch die theoretischen Parameter, sodass man keine genauen Werte hat. Diesen wahren Wert θ kennt man meist nicht. Dafür muss man aus der Verteilung auf einen sogenannten Schätzer schließen, der der Verteilung zugrunde liegt. Der Schätzer ergibt sich aus einer Berechnungsmethode aus den Daten. Schätzer sind dementsprechend wieder Zufallsvariablen, die jedoch aus der Standardnormalverteilung gezogen werden.
Beispiele für Schätzfunktionen wären:

- Mittelwert der Daten
- Standardabweichung der Daten
- Varianz der Daten
- $(min + max)/2$

In weiteren Berechnungen werden die theoretischen Werte der Verteilung durch die geschätzten Parameter ersetzt.

Schätzer haben folgende Eigenschaften:

Konsistenz

Bei Erhöhung des Stichprobenumfangs liegt der Schätzer immer näher am wahren Wert θ des zu schätzenden Parameter.

$$T_n \to \theta$$

Bei steigendem Stichprobenumfang verringert sich also die Streuung der Ziehungen.

Erwartungstreue

Beim Bilden von n Schätzern aus n Stichproben soll gelten, dass der Mittelwert der Schätzer dem wahren Wert θ entspricht, sodass gilt:

$$Bias_\theta = E(\hat{\theta}) - \theta = 0$$

Es tauchen also keine systematischen Abweichungen vom wahren Wert auf.

Mean Squared Error (MSE)

$$MSE = VAR(\hat{\theta}) + Bias_\theta^2$$

4.2 Schließen auf eine Verteilung

Da die Normalverteilung so oft in der Natur vorkommt und viele, wenn nicht alle, Verteilungen für große n in die Normalverteilung übergehen, kann man erstmal annehmen, dass die gezogene Stichprobe normalverteilt ist. Will man aber trotzdem überprüfen, ob eine Stichprobe einer bestimmten Verteilung entspricht, kann man das relativ einfach überprüfen, indem man die Quantile der Daten mit denen einer Verteilung vergleicht. Gibt es systematische Abweichungen, kann man davon ausgehen, dass die Daten dann nicht der angenommenen Verteilung entspringen.

R ANWENDUNG IN R: Schließen auf eine Verteilung

```
1. install.packages("car")
2. library("car")
3. x <- rnorm(100)
4. qqPlot(x, distribution = "norm")
```

4.3 Momentenmethode

Die Momentenmethode ist eine Methode zur Schätzung von Parametern einer Verteilung. Das Ziel der Momentenmethode ist das Schätzen der Parameter und Setzen dieser Parameter als theoretische Parameter einer Wahrscheinlichkeitsverteilung.

4.3.1 Momentenmethode anhand der Binomialverteilung

Angenommen, die Wahrscheinlichkeit p wird gesucht:

$$E(X) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = k \cdot p$$
$$\Rightarrow \hat{p} = \frac{\bar{X}}{k}$$

Wird zusätzlich noch die Anzahl der Treffer k gesucht, dann:

$$\Rightarrow \hat{k} = \frac{\bar{X}^2}{\bar{X} - \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

4.3.2 Momentenmethode anhand der Poissonverteilung

Angenommen, die Rate λ wird gesucht:

$$E(X) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \lambda$$
$$\Rightarrow \hat{\lambda} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \bar{X}$$

4.3.3 Momentenmethode anhand der Exponentialverteilung

Angenommen, die Rate λ wird gesucht:

$$E(X) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \frac{1}{\lambda}$$
$$\Rightarrow \hat{\lambda} = \frac{1}{\bar{X}}$$

4.3.4 Momentenmethode anhand der Normalverteilung

Angenommen, es werden sowohl μ als auch σ^2 gesucht:

$$E(X) = \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \mu$$
$$\frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2$$
$$\vdots$$
$$\Rightarrow \hat{\mu} = \bar{X}$$
$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$$

4.4 Maximum Likelihood-Methode

Die zentrale Verwendung der ML-Methode ist die Konstruktion bzw. Schätzung noch unbekannter Parameter einer durch Daten gegebenen Verteilung. Dies hat Anwendung, wenn eine Messung getätigt wurde, man weiß, dass sie bspw. normalverteilt ist, den zugrunde liegenden Erwartungswert aber noch nicht weiß.
Das Vorgehen ist so, dass ein oder mehrere Parameter einer Verteilung nicht bekannt sind und man aus einer Stichprobe deswegen darauf schließen muss. Es werden alle Möglichkeiten durchprobiert und der Schätzer, der mit größter Wahrscheinlichkeit dem wahren Wert entspricht, wird endgültig als Parameter übernommen.

4.4.1 Grundlegende Definitionen

Symbol	Bedeutung
$\sim$	in Abhängigkeit von
ϑ	Tupel von Parametern einer Verteilung BINOM: $\vartheta = (n, p)$ POIS: $\vartheta = \lambda$ EXP: $\vartheta = \lambda$ NORM: $\vartheta = (\mu, \sigma^2)$
k	Variable einer diskreten Verteilungsfunktion
x	Variable einer stetigen Verteilungsfunktion
X	Zufallsvariable einer bestimmten Verteilung
$f_\vartheta(x)$	Wahrscheinlichkeits(dichte)-Funktion mit gegebenen Parametern ϑ in Abhängigkeit von x BINOM: $B(n, p, k)$ POIS: $P(\lambda, k)$ EXP: $f(\lambda, x)$ NORM: $f(\mu, \sigma^2, x)$
$L_x(\vartheta)$	Likelihood-Funktion mit gegebenen x-Werten und variablen ϑ
$l_x(\vartheta)$	Log-Likelihood-Funktion

4.4.2 Likelihood-Funktion

Die Likelihood-Funktion ist eine Funktion, die aus einer Wahrscheinlichkeitsdichtefunktion (bei stetigen Verteilungen) bzw. aus einer Wahrscheinlichkeitsfunktion (bei diskreten Verteilungen) gewonnen wird.
Die Wahrscheinlichkeits(dichte)-Funktion $f(x)$ wird zur Likelihood-Funktion, nur das die Parameter ϑ als Variable und die Variable x als Parameter aufgefasst werden.

$$L_x(\vartheta) = f_\vartheta(x)$$
$$L_x(\mu, \sigma^2) = f_{\mu,\sigma^2}(x)$$

Das heißt, wenn beide Funktionen die selben Parameter- & Variablenausprägungen besitzen, besitzen sie den selben Funktionswert.

4.4.3 Log-Likelihood-Funktion

Die Log-Likelihood-Funktion ist einfach die logarithmierte Likelihood-Funktion. Dies ist ein Vorteil, wenn man eine Funktionsgleichung nach einer Variablen umstellen muss, da aufgrund der Logarithmusgesetze multiplikative Faktoren zu additiven Faktoren werden, da gilt: $log(a \cdot b) = log(a) + log(b)$. Da der Logarithmus eine strong monotone Funktion ist, ist das Maximum der Log-Likelihood-Funktion an der gleichen Stelle wie das der Likelihood-Funktion.

4.4.4 ML-Methode anhand der Binomialverteilung

$$L(p) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

$$l(p) = ln(\binom{n}{k}) + k \cdot ln(p) + (n-k) \cdot ln(1-p)$$

$$argmax\ l(p) \Rightarrow l'(p) \stackrel{!}{=} 0$$

$$\vdots$$

$$\hat{p} = \frac{k}{n}$$

R ANWENDUNG IN R: ML-Methode anhand der Binomialverteilung

Bei gesuchtem p:

1. `xi <- rbinom(100, 100, 0.5)`
2. Funktion zur Berechnung des Log-Likelihood-Funktionswerts aufstellen:
```
l <- function(p) {
return(-sum(log(dbinom(xi, 1, p))))
}
```
3. Optimum der Funktion bestimmen:
```
optimise(l, interval = c(0,1))$minimum
```

4.4.5 ML-Methode anhand der Poissonverteilung

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda}$$

$$l(\lambda) = \sum_{i=1}^{n} (ln(\lambda) + (-\lambda \cdot x_i))$$

$$argmax\ l(\lambda) \Rightarrow l'(\lambda) \stackrel{!}{=} 0$$

$$\vdots$$

$$\hat{\lambda} = \bar{x}$$

R ANWENDUNG IN R: ML-Methode anhand der Poissonverteilung

1. `xi <- rpois(100, 4.3)`
2. Funktion zur Berechnung des Log-Likelihood-Funktionswerts aufstellen:
```
l <- function(lambda) {
return(-sum(log(dpois(xi, lambda))))
}
```
3. Optimum der Funktion bestimmen:
```
optimise(l, interval = c(0,5))$minimum
```

4.4.6 ML-Methode anhand der Exponentialverteilung

$$L(\lambda) = \prod_{i=1}^{n} \lambda \cdot e^{-\lambda \cdot x_i}$$

$$l(\lambda) = \sum_{i=1}^{n} (ln(\lambda) + (-\lambda \cdot x_i))$$

$$argmax\ l(\lambda) \Rightarrow l'(\lambda) \stackrel{!}{=} 0$$

$$\vdots$$

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

R ANWENDUNG IN R: ML-Methode anhand der Exponentialverteilung

1. `xi <- rexp(100, 4.3)`
2. Funktion zur Berechnung des Log-Likelihood-Funktionswerts aufstellen:
```
l <- function(lambda) {
return(-sum(log(dexp(xi, lambda))))
}
```
3. Optimum der Funktion bestimmen:
```
optimise(l, interval = c(0,5))$minimum
```

4.4.7 ML-Methode anhand der Normalverteilung

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\cdot\pi\cdot\sigma^2}} \cdot e^{-\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2} = \frac{1}{(2\cdot\pi\cdot\sigma^2)^{\frac{n}{2}}} \cdot e^{-\frac{1}{2\cdot\sigma^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$l(\mu,\sigma^2) = -\frac{n}{2}\cdot log(2\cdot\pi\cdot\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$argmax\ l(\mu,\sigma^2) \Rightarrow grad\ l \overset{!}{=} \begin{pmatrix}0\\0\end{pmatrix}$$

$$\vdots$$

$$\hat{\mu} = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

R ANWENDUNG IN R: ML-Methode anhand der Normalverteilung

1. `xi <- rnorm(1000)`
2. Funktion zur Berechnung des Log-Likelihood-Funktionswerts aufstellen:
```
l <- function(xi, mu, sigma) {
logl <- sum(log(dnorm(xi, mu, sigma)))
return(logl)
}
```
3. Funktion mit einem Parameter für die Log-Likelihood-Funktion aufstellen :
```
lik <- function(x) {
return (-l(xi, mu = x[1], sigma = x[2]))
}
```
4. Optimum der Funktion bestimmen:
```
optim(c(0,1), lik)$par
```

4.5 Konfidenzintervalle

Ein Konfidenzintervall stellt sicher, dass bei n unabhängigen Stichproben ein gewisser Prozentteil der Mittelwerte sicher in diesem Bereich liegen. Ist bspw. das Konfidenzniveau $\alpha = 0.1$, liegen 90% der gezogenen Mittelwerte der einzelnen Stichproben innerhalb des Intervalls.

$$[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}]$$

Angenommen, alle anderen Parameter bleiben identisch und

- die Stichprobe wird größer, so wird das Konfidenzintervall schmaler
- die Streuung (Standardabweichung / Varianz) wird größer, so wird das Konfidenzintervall breiter
- das Konfidenzniveau α wird schmaler, so wird das Konfidenzintervall breiter

R ANWENDUNG IN R: Konfidenzintervalle

Wenn σ bekannt:

1. `alpha <- a`: Signifikanzniveau festlegen
2. `error <- qnorm(1-alpha/2) * sigma / sqrt(n)`: Bestimmen der Abweichung
3. `leftBoundary <- mu - error`: Linke Grenze bestimmen
4. `rightBoundary <- mu + error`: Rechte Grenze bestimmen

Wenn σ unbekannt:

- `t.test(x)`

```
> mu <- 6.048
> sigma <- 0.02
> alpha <- 0.1
> error <- qnorm(1-alpha/2) * sigma / sqrt(n)
> leftBoundary <- mu - error
> rightBoundary <- mu + error
> leftBoundary; rightBoundary
> x <- rnorm(1000)
> t.test(x, conf.level = 0.9)
```

4.5.1 Länge des Konfidenzintervalls

Will man ein Konfidenzintervall so bestimmen, dass es eine bestimmte Länge y besitzt, muss man das algebraisch lösen und die Gleichung nach der Größe der Stichprobe n umstellen.

$$2 \cdot q_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = y$$

R ANWENDUNG IN R: Länge des Konfidenzintervalls bestimmen

1. Funktion in Abhängigkeit von der Stichprobengröße n abzüglich der gewünschten Länge y aufstellen
2. Auflösen der Gleichung nach n

```
f <- function(x,y){
2*(qnorm(1-alpha/2) * sigma / sqrt(x)) - y
}
uniroot(f, y=0.001, lower=0, upper=20000)$root
```

4.5.2 Statistische Aussagen mit Konfidenzintervallen

Angenommen wir haben aus vorherigen Messungen einen Durchschnittswert θ gegeben, bspw. das Produktionsniveau einer Maschine, und wollen, nachdem wir den Prozess optimiert haben, eine Aussage treffen, ob das Produktionsniveau tatsächlich gestiegen ist.

- Liegt das neue Konfidenzintervalle der einzelnen Mittelwerte unter θ, so kann man sagen, dass es keine Verbesserung gab.
- Liegt das neue Konfidenzintervalle zwischen θ und einer signifikanten Erhöhung $\theta + \Delta$, kann man sagen, dass es sich um eine *signifikante* Erhöhung handelt, es ist aber nicht sicher sagbar, ob es nur aus Zufall so hoch ist.
- Liegen das neue Konfidenzintervalle über $\theta + \Delta$, so kann man sicher sagen, dass es sich um eine *relevante*, also dauerhafte Erhöhung des Produktionsniveaus handelt.
- Liegt die obere Grenze des Konfidenzintervalls über θ bzw. $\theta + \Delta$, die untere Grenze jedoch unter θ, so kann man keine Aussage treffen. Man muss dann die Stichprobe erhöhen, sodass das Konfidenzintervall klein genug ist, sodass es über bzw. unter θ bzw. $\theta + \Delta$ liegt.

4.6 Hypothesentests

Wurde nun eine statistische Aussage getätigt, kann man überprüfen, ob diese denn auch stimmt, es könnte ja die Möglichkeit bestehen, dass diese nicht richtig ist. Dabei wird immer erstmal davon ausgegangen, dass derjenige, der die Aussage getroffen hat, nicht lügt (*in dubio pro reo*).

Bsp: Oktoberfest
Bei der Nullhypothese H_0 wird nun angenommen, dass der durchschnittliche Inhalt eines Maßkrugs $1l$ beträgt.
Die Alternativhypothese H_A ist, dass im Durchschnitt weniger als $1l$ ausgeschenkt wird.

- H_0: Der Angeklagte ist unschuldig
- H_A: Der Angeklagte ist schuldig

Somit stehen sich zwei Hypothesen gegenüber. Die Begriffe sind so zu verstehen, dass geschaut wird, ob H_A überhaupt bewiesen werden kann. Bei ergebnisloser Suche kann weiter H_0 als gültig erachtet werden. Der Hypothesentest dient nun dazu, anhand eines Ergebnisses einer Stichprobe zu einer Entscheidung darüber zu kommen, welcher der beiden Hypothesen man eher zu glauben bereit ist bzw. welche Hypothese angenommen und welche verworfen wird. Eine 100%-ige Sicherheit, dass die angenommene Hypothese tatsächlich wahr ist, kann der Hypothesentest naturgemäß niemals bieten, da wir von einer Stichprobe auf die Grundgesamtheit schließen.

Nun gibt es mehrere Möglichkeiten, die man überprüfen kann:

- $H_0 : \mu = 1l$ gegen $H_A : \mu < 1l$ (linksseitiger Test)
- $H_0 : \mu = 1l$ gegen $H_A : \mu > 1l$ (rechtsseitiger Test)
- $H_0 : \mu = 1l$ gegen $H_A : \mu \neq 1l$ (beidseitiger Test)

VORGEHENSWEISE:

1. Signifikanzniveau α wählen
2. Wahl des Modells & Formulierung der Nullhypothese H_0
3. Formulierung der Alternativhypothese H_A
4. Bestimmung der Teststatistik T
5. Verteilung von T unter H_0
6. Überprüfung der Gültigkeit der Nullhypothese H_0
7. Testentscheidung

MÖGLICHKEITEN ZUR ÜBERPRÜFUNG DER GÜLTIGKEIT DER HYPOTHESEN:

- Möglichkeit 1 (leichter)
    1. Berechnung des realisierten Wertes p
- Möglichkeit 2 (Alternative)
    1. Bestimmung des Annahme- & Verwerfungsbereichs
    2. Bestimmung des realisierten Wertes t

Insgesamt können bei einem Test vier Fälle auftreten:

- Wir verwerfen H_0, also nehmen H_A an.
  - In Wirklichkeit stimmt H_0: Hier lehnen wir H_0 fälschlicherweise ab. Das ist der α-Fehler, auch Fehler 1. Art genannt. Dieser Fall tritt genau mit einer Wahrscheinlichkeit von α auf – weil ein Test genau so konstruiert ist. Das Niveau α regelt also, wie sicher man sich sein kann dass H_A tatsächlich wahr ist, gegeben man lehnt H_0 auch ab.
  - In Wirklichkeit stimmt H_A: H_A stimmt und wir nehmen H_A an.
- Wir behalten H_0 bei.
  - In Wirklichkeit stimmt H_0: H_0 stimmt und wir nehmen H_0 an.
  - In Wirklichkeit stimmt H_A: In diesem Fall ist unsere Vermutung wahr (d.h. H_A , die wir ja nachweisen möchten, stimmt), aber durch den Test konnte sie nicht bestätigt werden, da wir H_0 beibehalten. Dies ist der Fehler 2. Art. Diese Wahrscheinlichkeit können wir nicht kontrollieren, sie ist abhängig von der Art des Tests und des Signifikanzniveaus α.

ANWENDUNG IN R: Tests

- `t.test(x)`
- `t.test(x, mu = a)`
- `t.test(x, alternative = "less")`
- `t.test(x, alternative = "greater")`
- `t.test(x, alternative = "two.sided")`

```
> x <- rnorm(1000)
> t.test(x)
> t.test(x, mu = 8.5)
> t.test(x, conf.level = 0.9)
> t.test(x, alternative = "less")
> t.test(x, alternative = "greater")
> t.test(x, alternative = "two.sided")
```

Das Signifikanzniveau α

Das Signifikanzniveau gibt eine Grenze für die Wahrscheinlichkeit, dass die Nullhypothese H_0 fälschlicherweise verworfen werden kann, obwohl sie eigentlich richtig ist, an. Das Signifikanzniveau wird auch deshalb auch *Irrtumswahrscheinlichkeit* genannt.

Nullhypothese H_0 & Alternativhypothese H_A

Die Nullhypothese H_0 ist eine Behauptung, die getestet werden muss. Sie wird immer so aufgestellt, dass die Grundlage der Verteilung bekannt ist. Wird beispielsweise behauptet, dass sich durch Veränderungen etwas am Mittelwert geändert hat, wird diese Behauptung als Alternativhypothese aufgestellt, da man die Form der neuen Verteilung nicht kennt, die der alten jedoch schon. Die Alternativhypothese ist die, die man eigentlich beweisen will.

Annahmebereich

Der Annahmebereich enthält alle gültigen Daten, bei denen die Nullhypothese beibehalten wird.

R ANWENDUNG IN R: Annahmebereich

```
> alpha <- 0.1
> qpois(1 - alpha), lambda
```

Ablehnungsbereich

Analog zum Annahmebereich enthält der Ablehnungsbereich all jene Daten, bei denen die Nullhypothese nicht mehr gültig ist, sodass diese verworfen und die Alternativhypothese H_A angenommen wird.

Teststatistik T

μ_0	in der Nullhypothese angenommener Mittelwert
$\hat{\sigma}$	Standardabweichung der Stichprobe

Zur Überprüfung, ob eine Stichprobe tatsächlich eine Abweichung zeigt, muss sie vorher standardisiert werden. Dies hat den Sinn, dass es egal ist, wie groß die Stichprobe ist, da mit $\sqrt{n}$ multipliziert wird, welcher Mittelwert als Nullhypothese festgelegt wurde, da dieser Wert wieder abgezogen wird, und welche Streuung die Daten aufweisen, da durch die Standardabweichung der Daten, $\hat{\sigma}$, geteilt wird.
Bsp.: Stichprobe von $n = 10$ ergibt, dass durchschnittlich $984,4ml$ ausgeschenkt werden, mit einer Standardabweichung von $\hat{\sigma} = 16,057$.

$$T = \sqrt{n} \cdot \frac{\bar{x} - \mu_0}{\hat{\sigma}}$$

$$T = \sqrt{10} \cdot \frac{984,4 - 1000}{16,057} = \sqrt{10} \cdot \frac{-15,6}{16,057} = -3,072 \qquad \text{(Bsp.)}$$

R ANWENDUNG IN R: Teststatistik T

```
t.standardize <- function(x, mu.0) {
n <- length(x)
return(sqrt(n) * (mean(x) - mu.0) / sd(x))
}
```

4.6.1 Der p-Wert

Der p-Wert ist die Wahrscheinlichkeit, dass die zu prüfende Größe T, auch Teststatistik genannt, - unter der Annahme, dass die Nullhypothese H_0 gilt - mindestens den in der Stichprobe berechneten Wert t annimmt.

$$p = P(T \geq t)$$

Bei statistischen Tests wird vor dem eigentlichen Test ein Signifikanzniveau α gewählt. Die Nullhypothese H_0 wird dann abgelehnt, wenn $p \leq \alpha$. Ist der Wert für die Teststatistik so unwahrscheinlich, dass er kleiner ist als das zuvor festgelegt Signifikanzniveau, so entscheidet man sich, die Nullhypothese abzulehnen, man nimmt H_A an. Ist die Wahrscheinlichkeit dagegen größer als das zuvor festgelegte Signifikanzniveau, so kann man die Nullhypothese nicht ablehnen.

R ANWENDUNG IN R: Testentscheidung durch den p-Wert

```
> alpha <- 0.1
> t.test(x, ...)$p.value < alpha
```

- `FALSE` ⇒ Nullhypothese wird nicht abgelehnt
- `TRUE` ⇒ Alternativhypothese wird angenommen

4.6.2 t-Verteilung

Ist eine Stichprobe eigentlich normalverteilt, aber die Standardabweichung ist nicht gegeben, so ist die Stichprobe tatsächlich t-verteilt. In diesem Fall muss die Teststatistik T nach der obigen Formel standardisiert werden, bevor mit ihr gerechnet werden kann.

R ANWENDUNG IN R: t-Verteilung

- `dt(x, df = n - 1)`: $P(X = k)$, wobei k Zahl oder Vektor ist
- `pt(x, df = n - 1)`: $P(X \leq k)$
- `pt(x, df = n - 1, lower.tail = FALSE)`: $P(X > k)$
- `qt(x, df = n - 1)`: p-Quantil der Verteilung, $p \in [0,1]$
- `rt(x, df = n - 1)`: Ziehung von x Zufallszahlen
- `t.test(x, mu = a, alternative = "two.sided")`

```
> x <- rt(42, 10)
> n <- length(n)
> dt(x, df = n - 1)
> pt(31, df = n - 1)
> qt(0.9, df = n - 1)
> pt(31, df = n - 1, lower.tail = FALSE)
```