# Stochastik für Informatik und Ingenieurswissenschaften

Noemi Kurt

Institut für Mathematik
TU Berlin

Vorlesungsskript[1]

`Vorläufige Version, unvollständig und fehlerhaft!`


*Date*: 17. Juni 2016.

## Inhaltsverzeichnis

**Teil 2. Einführung in die Statistik** 3
Einleitung 3
8. Beschreibende Statistik 4
8.1. Daten, Klassen, graphische Darstellungen 4
8.2. Mittelwert, Median, empirische Varianz 8
8.3. Empirische Korrelation 10
8.4. Lineare Regression 11
9. Grundlagen: Grenzwertsätze 13
9.1. Gesetz der großen Zahlen 13
9.2. Zentraler Grenzwertsatz und Anwendungen 14
10. Parameterschätzung 18
10.1. Beispiele und Eigenschaften von Schätzern 18
10.2. Maximum Likelihood-Schätzung 20
10.3. Der Expectation-Maximization-Algorithmus 24

# Teil 2. Einführung in die Statistik

## Einleitung

Ganz zu Beginn der Vorlesung hatten wir den Begriff "Stochastik" als Überbegriff für die beiden Gebiete *Wahrscheinlichkeitstheorie* und *Statistik* definiert. Im ersten Teil der Vorlesung hatten wir uns mit der Wahrscheinlichkeitstheorie beschäftigt, und deren mathematische Begriffsbildung als theoretische Grundlage kennen gelernt. Im zweiten Teil wenden wir uns nun der Statistik zu, das bedeutet, der Analyse und Interpretation von Messwerten und Daten aus zufälligen Vorgängen mittels wahrscheinlichkeitstheoretischen Methoden.

Die Wahrscheinlichkeitstheorie bildet dafür das Fundament. In Kapitel 9 werden wir als wichtige theoretische Grundlage einige *Grenzwertsätze* kennen lernen.

Während die Wahrscheinlichkeitstheorie von Axiomen und mathematischen Modellen ausgeht, basiert die Statistik in erster Linie auf *Daten* oder *Messwerten*. Grundsätzlich geht man in der Statistik davon aus, dass man eine große Anzahl von solchen Messwerten gegeben hat. Für diese Daten soll ein geeignetes mathematischen Modell aufgestellt werden, welches diese Daten beschreibt, und welches mit wahrscheinlichkeitstheoretischen Methoden untersucht werden kann.

Das ist nun erst einmal recht vage, was daran liegt, dass es eine Unzahl möglicher Fragestellungen gibt, die jeweils mit verschiedenen Methoden behandelt werden können. In dieser Vorlesung werden wir die wichtigsten Grundprobleme behandeln:

- Schätzer, Bestimmung von Kenngrößen
- Hypothesentests
- Konfidenzintervalle

Betreffend der gemessenen Daten gehen wir dabei stets von einer der beiden (sich nicht ausschließenden) Grundannahmen aus:

- Die gemessenen Daten sind einzelne *Realisierungen* von (unabhängigen, identisch verteilten) *Zufallsvariablen*
- Die gemessenen Daten stellen eine *Stichprobe* aus einer (noch viel größeren) *Population* dar.

Unter diesen Prämissen sollen mittels der Stichprobe Aussagen über die zugrundeliegende Zufallsvariablen bzw. über die gesamte Population gemacht werden.

Wir werden nun in einem ersten Kapitel kurz die Darstellung von statistischen Daten behandeln, und anschließend mit den Grenzwertsätzen die noch fehlenden wahrscheinlichkeitstheoretischen Grundlagen vervollständigen, bevor wir uns den oben genannten Fragestellungen zuwenden.

# 8. Beschreibende Statistik

In diesem Kapitel diskutieren wir kurz mögliche Formen der Präsentation von Daten, z.B. als Häufigkeiten, Histogramme, oder mittels Kenngrößen. Die **Lernziele** dieses Kapitel sind:

- Verschiedenen Formen der Präsentation von Daten kennen
- Kenngrößen aus Daten bestimmen können

## 8.1. **Daten, Klassen, graphische Darstellungen.**

**Beispiel 8.1** (Eine Messreihe)**.** In der Produktion eines elektronischen Geräts wird in einer Testreihe die Zeit (in Sekunden, gerundet auf eine Nachkommastelle) vom Start des Geräts bis zur Betriebsbereitschaft gemessen. Die folgende Tabelle gibt 14 Messwerte aus dieser Testreihe an:

Ursprüngliche Messreihe

| Messung | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wert | 10.9 | 6.8 | 9.5 | 6.9 | 8.2 | 3.4 | 6.2 | 8.6 | 5.3 | 10.7 | 8.1 | 8.0 | 8.9 | 10.7 |

Wir wollen uns in diesem Kapitel folgenden Fragestellungen widmen:

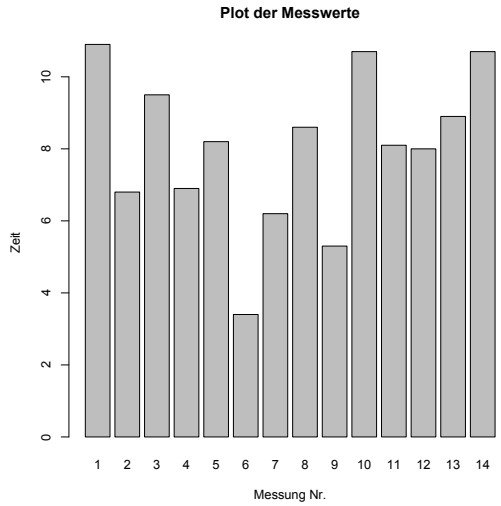
- Wie können solche Daten geeignet dargestellt werden?
- Welche Informationen können aus diesen Daten abgelesen werden?
- Um welche "Art" von Daten handelt es sich hier, und inwiefern sind sie mit unseren Grundannahmen (siehe Einleitung) kompatibel?
- Welche weiterführenden Fragestellungen ergeben sich möglicherweise?

Eine mögliche Art der Darstellung ist bereits die obenstehende Tabelle. Alternativ kann man auch einfach den Vektor der Daten, ohne die Nummerierung der Messungen, angeben:

$$(10.9, 6.8, 9.5, 6.9, 8.2, 3.4, 6.2, 8.6, 5.3, 10.7, 8.1, 8.0, 8.9, 10.7)$$

Dieser enthält dieselbe Information wie die Tabelle. Wir werden in dieser Vorlesung Daten meistens als Vektor angeben.

Die Tabelle oder der Vektor könnte man auch direkt in eine graphische Darstellung übersetzen, z.B. in einen Plot:




Hier haben wir einen Barplot (Balkendiagramm) gewählt, man könnte auch einen Punktplot, ein Kuchendiagramm... wählen. In einer graphischen Darstellung fallen möglicherweise Dinge auf, die aus der Tabelle weniger leicht ersichtlich sind, z.B. sieht der 6. Wert wie ein Ausreißer nach unten aus. Auch scheinen Werte um die 8 herum besonders häufig vorzukommen. Vielleicht fällt auch auf, dass – mit Ausnahme des Werts 10.7 – alle Messwerte genau einmal auftreten. Letztere Beobachtung hängt damit zusammen, dass die gemessene Größe – eine zeitliche Dauer – eine *stetige* Größe ist, aber die Feststelung hat auch etwas mit der Messgenauigkeit zu tun: Runden wir nämlich die Messungen auf ganze Sekunden, so sieht die Tabelle so aus:

Gerundete Werte

| Messung | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wert | 11 | 7 | 10 | 7 | 8 | 3 | 6 | 9 | 5 | 11 | 8 | 8 | 9 | 11 |

In dieser Tabelle kommen nun etliche Werte mehrfach vor. Wir stellen die Häufigkeiten der vorkommenden Werte in einer neuen Tabelle zusammen:

Häufigkeiten der gerundeten Werte

| Wert | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|
| Häufigkeit | 1 | 0 | 1 | 1 | 2 | 3 | 2 | 0 | 3 |

Auch diese Tabelle kann graphisch dargestellt werden, wir erhalten damit das *Histogramm* der gerundeten Werte.

Die Wahl einer geeigneten Darstellung hängt immer auch von der Frage ab, die damit beantwortet oder zumindest diskutiert werden soll. In diesem Beispiel könnte z.B. die Firma, welche die betreffenden Geräte herstellt, damit werben, dass ihre Geräte in höchstens 10 Sekunden betriebsbereit sind. Im Rahmen einer Überprüfung dieser Behauptung könnten dann die Daten erhoben worden sein. Eine andere Fragestellung könnte lauten: Wie lange dauert es im Mittel, bis ein Gerät betriebsbereit ist? Auch diese Frage kann an Hand der Daten untersucht werden.

In beiden Fällen wird man statistische Verfahren anwenden, die wir in dieser Vorlesung noch kennen lernen werden. Dabei sind jedoch im Allgemeinen nicht die 14 konkret getesteten Geräte von Interesse, sondern die Gesamtheit der von der Firma produzierten Geräte dieser Art. Um statistische Fragen zu beantworten, können nicht alle jemals produzierten Geräte untersucht werden. Deshalb hat man eine *Stichprobe* von 14 genommen und diese untersucht, in der (impliziten oder expliziten) Grundannahmen, dass die Lebensdauer der Geräte Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, und somit mit HIlfe der Stichprobe eine valide Aussage über die Gesamtheit (oder *Population*) aller solcher Geräte gemacht werden kann.

Einige Begriffe die in diesem Beispiel benutzt wurden, benötigen noch einer genauen Definition, obwohl sie umgangssprachlich mehrheitlich klar sein dürften.

**Definition 8.2** (Absolute und relative Häufigkeiten)**.** Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor von Messwerten. Sei $x \in \mathbb{R}$. Die *absolute Häufigkeit* von $x$ ist

$$H(x) := |\{i : x_i = x\}|,$$

d.h. sie gibt an wie oft der Wert $x$ im Vektor $(x_1, ..., x_n)$ vorkommt. Die *relative Häufigkeit* von $x$ ist

$$h(x) := \frac{H(x)}{n}.$$

**Beispiel 8.3** (Forsetzung von Beispiel 8.1)**.** Wir berechnen die absoluten und relativen Häufigkeiten für einige Werte aus dem Beispiel 8.1. Absolute Häufigkeiten:

$$H(10.9) = 1, \quad H(10.7) = 2, \quad H(4.0) = 0.$$

Relative Häufigkeiten:

$$h(10.9) = \frac{1}{14}, \quad h(10.7) = \frac{1}{7}, \quad h(4.0) = 0.$$

**Definition 8.4** (Klasseneinteilung, Häufigkeiten von Klassen)**.** Sein $(x_1, ..., x_n)$ ein Vektor von Daten. Eine *Klasseneinteilung* ist eine Zusammenfassung der $x_i$ zu disjunkten Mengen $A_1, ..., A_m \subseteq \mathbb{R}$ mit $m \leq n$, so dass jedes $x_i$ zu genau einem $A_j$ gehört.
Die *absolute Häufigkeit* einer Klasse $A$ ist

$$H(A) := |\{i : x_i \in A\}|$$

und die *relative Häufigkeit* der Klasse $A$ bei einer Einteilung in $m$ verschiedene Klassen ist

$$h(A) := \frac{H(A)}{n}.$$

**Beispiel 8.5** (Fortsetzung von Beispiel 8.1)**.** In Beispiel 8.1 haben wir den Datenvektor $(10.9, 6.8, 9.5, 6.9, 8.2, 3.4, 6.2, 8.6, 5.3, 10.7, 8.1, 8.0, 8.9, 10.7)$ durch Runden auf ganze Zahlen zu den Klassen

$$[3, 4[, \quad [4, 5[, \quad [5, 6[, \quad [6, 7[, \quad [7, 6[, \quad [8, 9[, \quad [9, 10[, \quad [10, 11[$$

zusammengefasst. Eine andere mögliche Klasseneinteilung mit zugehörigen Häufigkeiten ist

| Klasse | [2,4[ | [4,6[ | [6,8[ | [8,10[ | [10,12[ |
|---|---|---|---|---|---|
| Häufigkeit | 1 | 1 | 3 | 5 | 3 |

oder

| Klasse | $< 6$ | [6,7[ | [7,8[ | [8,9[ | $> 9$ |
|---|---|---|---|---|---|
| Häufigkeit | 2 | 3 | 0 | 5 | 3 |

Wir erinnern daran, dass wir immer von der Gültigkeit mindestens einer der beiden folgenden Aussagen ausgehen:

*Bemerkung* (Grundannahme).
- Die gemessenen Daten sind einzelne *Realisierungen* von (unabhängigen, identisch verteilten) *Zufallsvariablen*
- Die gemessenen Daten stellen eine *Stichprobe* aus einer (noch viel größeren) *Population* dar.

**Definition 8.6** (Stetige und diskrete Merkmale)**.** Sei $(x_1, ..., x_n)$ ein Vektor von Daten. Falls diese Daten Realisierungen von *diskreten* Zufallsvariablen sind, so beschreiben die Daten ein *diskretes Merkmal*. Sind es Realisierungen von Zufallsvariablen, welche im Prinzip sämtliche Werte von $\mathbb{R}$ oder zumindest sämtliche Werte mindestens eines Teilintervalles von $\mathbb{R}$ annehmen können, so beschreiben die Daten ein *stetiges Merkmal*.

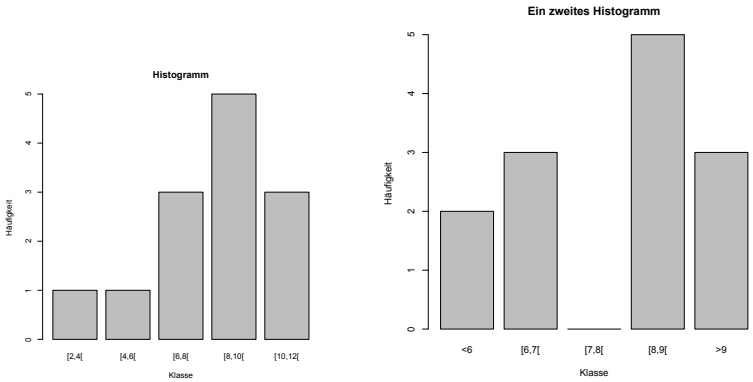
Stetige Merkmale machen üblicherweise eine Klasseneinteilung notwendig, da normalerweise sehr viele verschiedene Werte gemessen werden.

**Beispiel 8.7.** Typische Beispiele für stetige Merkmale sind Längen, Gewicht, Zeitdauer, Spannung, usw. Diskrete Merkmale sind z.B. Teilnehmerzahlen, Anzahl Probanden mit einer bestimmten Krankheit, oder auch bereits diskretisierte, aber ursprünglich stetige Merkmale, z.B. Lebensalter in Jahren, Gewicht in kg.

**Definition 8.8** (Histogramm)**.** Ein *Histogramm* der Daten ist ein Plot der Funktion $x \mapsto H(x)$ oder $x \mapsto h(x)$, oder, im Falle einer Einteilung in Klassen, der Funktion $A \mapsto H(A)$ bzw. $A \mapsto h(A)$.

Histogramme stellen also immer Häufigkeiten dar, hingegen können Plots oder Diagramme beliebige Arten von Daten darastellen.

**Beispiel 8.9** (Fortsetzung von Beispiel 8.5)**.** Die Klasseneinteilungen in den beiden Tabellen von Beispiel 8.5 liefern folgende Histogramme:




Wichtig ist dabei immer, dass bei Histogrammen die Klasseneinteilung sowie die Skalen klar angegeben sind. Mit dem Befehl `hist` erzeugt R aus einem Datenvektor ein Histogramm.

8.2. **Mittelwert, Median, empirische Varianz.** Wie bei Zufallsvariablen gibt es diverse Kenngrößen, welche Aufschluss über gewisse Eigenschaften von Datenmengen geben.

**Definition 8.10.** Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Das *empirische Mittel* von $(x_1, ..., x_n)$ ist definiert als

$$\bar{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} x_i$$

**Definition 8.11.** Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Der *Median* von $(x_1, ..., x_n)$ ist definiert als der Wert in der Mitte der geordneten Liste. Falls $n$ gerade ist, wird der Durchschnitt aus den beiden mittleren Werten gebildet.

**Definition 8.12.** Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor (von Messwerten/Daten). Die *empirische Varianz* von $(x_1, ..., x_n)$ ist definiert als

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

Das empirische Mittel ist also nichts anderes als das arithmetische Mittel oder der Durchschnittswert der Daten. Die empirische Varianz ist, analog zur Varianz von Zufallsvariablen, ein Maß für die Streuung, bzw. für die durchschnittliche Abweichung vom Mittelwert. Zu beachten ist, dass zwar $n$ Werte aufsummiert werden, jedoch nur durch $n-1$ dividiert wird. Den Grund dafür werden wir später sehen.

**Beispiel 8.13** (Fortsetzung von Beispiel 8.1)**.** Wir betrachten noch einmal die Daten von Beispiel 8.1. Einsetzen in die Definitionen ergibt für das empirisches Mittel

$$\bar{\mu}_x = 8.014,$$

und für die empirische Varianz

$$\bar{\sigma}_x^2 = 4.677.$$

Zur Bestimmung des Medians ordnen wir erst einmal die Daten in aufsteigender Reihenfolge. Als Vektor geschrieben erhalten wir

$$(3.4, 5.3, 6.2, 6.8, 6.9, 8.0, 8.1, 8.2, 8.6, 8.9, 9.5, 10.7, 10.7, 10.9$$

Insgesamt haben wir 14 Werte, also eine gerade Zahl. Die beiden Werte in der Mitte der geordneten Liste sind 8.1 und 8.2. Somit ist der Median

$$\frac{8.1 + 8.2}{2} = 8.15.$$

Es gibt Situationen, in denen das empirische Mittel durch einige wenige Ausreißer, also extrem hohe oder extrem tiefe Werte, dominiert wird, aber die Mehrheit der Werte weit vom Durchschnitt entfernt sind. In solchen Fällen ist der Median oft aussagekräftig.

**Beispiel 8.14** (Einkommensverteilung)**.** `siehe Vorlesung`

Zum Schluss dieses Abschnitts hier noch einige R-Befehle, welche im Zusammenhang mit der Darstellung von Daten bzw. deren Kenngrößen nützlich sind.

- mean() empirisches Mittel
- var() empirische Varianz
- median() Median
- sort() Liste aufsteigend sortieren
- hist() Zeichnet Histogramm.

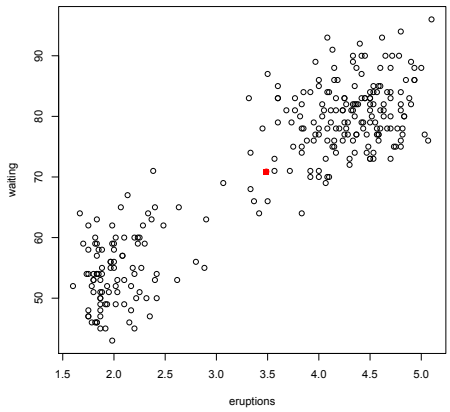
**Beispiel 8.15** (R-Datensatz, Fortsetzung von Beispiel 5.19)**.** In Kapitel 5 hatten wir bereits den R-Datensatz des Yellowstone-Geysirs "Old faithful" kennengelernt. Dieser enthält Daten zweier gleichzeitig gemessener Größen, d.h. Paare von Daten $(x_1, y_1), (x_2, y_2), ...., (x_n, y_n)$, welche die Wartezeit und die Dauer eines Ausbruchs angeben. Wir hatten bereits in Kapitel 5 (implizit) die Grundannahme der Statistik verwendet, nämlich dass es sich hierbei um die Realisierung von (unabhängigen Kopien) zweier Zufallsvariablen $X$ und $Y$ handelt. Mit dem Befehl `faithful` können die 272 Datenpaare in R aufgerufen werden. Wir können diese Datenpaare anders sortieren, und statt $(x_1, y_1), ..., (x_n, y_n)$ in der Form $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ angeben. Folgendermaßen können wir die empirischen Mittel und die empirische Varianz jeweils für den Datenvektor $x$ und für den Datenvektor $y$ berechnen.

```
> data<-faithful
> x<-faithful$eruptions
> y<-faithful$waiting
> mx<-mean(x)
> my<-mean(y)
> vx<-var(x)
> vy<-var(y)
```

Als Ergebnis erhalten wir damit

$$\bar{\mu}_x = 3.487783, \quad \bar{\mu}_y = 70.89706, \quad \bar{\sigma}_x^2 = 1.302728, \quad \bar{\sigma}_y^2 = 184.8233.$$

Das Paar $(\bar{\mu}_x, \bar{\mu}_y)$ wird auch *Schwerpunkt* der Daten genannt, und kann in den Plot der Datenpaare als Punktwolke eingezeichnet werden (rotes Quadrat):




## 8.3. **Empirische Korrelation.**

**Beispiel 8.16** (Abhängigkeit von Zufallsvariablen, Fortsetzung von Beispiel 8.15)**.** Die Form der Punktwolke aus Beispiel 8.15 weist darauf hin, dass eine gewisse Abhängigkeit zwischen den Daten vorhanden ist. In diesem Kapitel werden wir ein Maß für die Abhängigkeit zwischen zwei Vektoren von Daten angeben.

**Definition 8.17** (Empirische Kovarianz)**.** Seien $x = (x_1, ..., x_n)$ und $y = (y_1, ..., y_n)$ gegeben. Die *empirische Kovarianz* ist definiert als

$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$

wobei $\bar{\mu}_x$ das empirische Mittel der $x_i$ und $\bar{\mu}_y$ das empirische Mittel der $y_i$ ist.
Der *empirische Korrelationskoeffizient* ist definiert als

$$r_{xy} = \frac{c_{xy}}{\bar{\sigma}_x \bar{\sigma}_y},$$

wobei $\bar{\sigma}_x := \sqrt{\bar{\sigma}_x^2}$ die *empirische Standardabweichung* von $x$ ist (und analog $\bar{\sigma}_y$ für $y$).

Zur Erinnerung: Die Kovarianz von zwei Zufallsvariablen $X$ und $Y$ war definiert als $\mathrm{cov}(X, Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$, und der Korrelationskoeffizient von $X$ und $Y$ als $\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}}$. Wie im Fall von Zufallsvariablen gilt für den Korrelationskoeffizienten

$$-1 \leq r_{x,y} \leq 1,$$

und wiederum ist der Korrelationskoeffizient ein Maß für die Stärke des linearen Zusammenhangs zwischen $x$ und $y$.

**Beispiel 8.18** (Fortsetzung von Beispiel 8.15)**.** Im Beispieldatensatz erhalten wir durch einsetzen in die Definition, bzw. durch Aufrufen des R-Befehls `cov(x,y)` für die empirische Kovarianz

$$c_{x,y} = 13.97781,$$

und mit `cor(x,y)` für die empirische Korrelation

$$r_{x,y} = 0.9008112.$$

Dieser Wert ist relativ nah an 1, es gibt also einen relativ starken linearen Zusammenhang zwischen den beiden Werten.

8.4. **Lineare Regression.** Auf der Verbindung zwischen empirischer Kovarianz und linearem Zusammenhang beruht auch das Prinzip der linearen Regression. Diese kommt zur Anwendung, wenn Daten $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ gegeben sind, zwischen denen ein linearer Zusammenhang vermutet wird. Dann möchte man eine Gerade

$$y = ax + b,$$

welche durch den Schwerpunkt $(\bar{\mu}_x, \bar{\mu}_y)$ der Daten verläuft, und welche die Abstände der Datenpunkte $(x_1, y_1), ..., (x_n, y_n)$ von der Geraden minimiert (unter allen Geraden durch den Schwerpunkt).

$$y = ax + b$$

zwei Bedingungen erfüllt:

(LR1) $(\bar{\mu}_x, \bar{\mu}_y)$ liegt auf der Geraden, also $\bar{\mu}_y = a \cdot \bar{\mu}_x + b$,
(LR2) Der Ausdruck

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ist minimal, wobei $\hat{y}_i := ax_i + b$ die $y-$Koordinaten des Punktes auf der Geraden ist, welcher $x-$Koordinate $x_i$ hat.

**Satz 8.19** (Lineare Regression)**.** *Die beiden Bedingungen (LR1) und (LR2) legen a und b eindeutig fest, und zwar als*

$$a = \frac{c_{xy}}{\bar{\sigma}_x^2}, \quad b = \bar{\mu}_y - a \cdot \bar{\mu}_x.$$

Wir beweisen diesen Satz nicht, er basiert auf Methoden der Linearen Algebra, und wurde vermutlich in der entsprechenden Vorlesung geführt. Zu beachten ist, dass die Rollen von $x$ und $y$ nicht symmetrisch sind.

**Beispiel 8.20** (Lineare Regression, Fortsetztung von Beispiel 8.15)**.** In R kann mit dem Befehl `lm(y∼ x)` eine lineare Regression direkt berechnet werden. Angewandt auf die Daten aus Beispiel 8.15 erhält man mittels

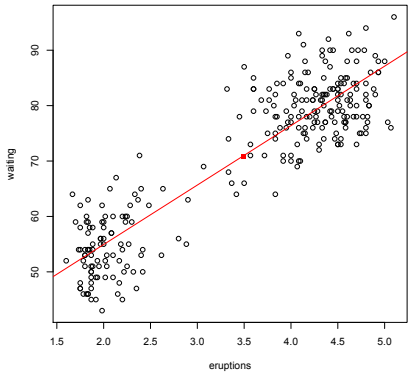
```
> reg<-lm(y x)
> summary(reg)
```

die Ausgabe

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122 | 3.9246  15.9719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
x            10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared: 0.8115, Adjusted R-squared: 0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Daraus kann man den $y-$Achsenabschnitt $b \approx 33.4744$ und die Steigung $a \approx 10.7296$ der Geraden ablesen. Außerdem werden mehrere weitere Werte angegeben, welche Auskunft über die Güte der Approximation machen. Einige davon werden wir in späteren Kapiteln noch genauer kennen lernen.

Ein Plot der resultierenden Regressionsgeraden durch die Punktwolke visualisiert die Approximation.



# 9. Grundlagen: Grenzwertsätze

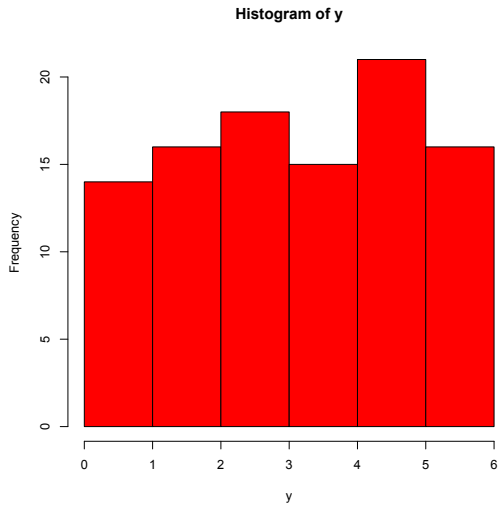
In diesem Kapitel beschäftigen wir uns noch einmal mit einem Thema der Wahrscheinlichkeitstheorie, bevor wir uns wieder der Statistik zuwenden. Wir formulieren einige der wichtigsten sogenannten *Grenzwertsätze*, welche eine wichtige theoretische Grundlage für die Statistik bilden. Dabei betrachten wir nun nicht mehr, wie bisher meistens, eine oder endlich viele Zufallsvariablen, sondern *Folgen* von Zufallsvariablen, und untersuchen deren Konvergenzverhalten.

Die **Lernziele** dieses Kapitel sind:

- Das Gesetz der großen Zahlen und seine Bedeutung kennen
- Den zentralen Grenzwertsatz und seine Bedeutung kennen
- Normalapproximation für die Binomialverteilung berechnen können.

## 9.1. **Gesetz der großen Zahlen.**

**Beispiel 9.1** (Mehrfaches Würfeln, Fortsetzung von Beispiel 5.1)**.** In Beispiel 5.1 hatten wir einen fairen Würfel betrachtet, und mit $X$ =Ergebnis eines Wurfs bezeichnet. Das Histogramm zeigt die Häufigkeitsverteilung der Ergebnisse beim hundertfachen Würfeln. Mit $y_i$ bezeichnen wir Ergebnis des $i-$ten Wurfs.




Wir hatten bereits in Kapitel 5 gesehen, dass

$$\frac{1}{100}\sum_{i=1}^{100} y_i = 3.61 \approx 3.5 = \mathbb{E}[X]$$

gilt, d.h. das *empirische Mittel* der hundert Würfe liegt nah am berechneten Erwartungswert. Wenn wenn wir nun 1000 statt 100 Würfe ausführen, erwarten wir, dass das Ergebnis noch näher am Erwartungswert liegt. Dies kann man tatsächlich beweisen. Das bedeutet also, dass man den Erwartungswert als Nëherung für das empirische Mittel verwenden

kann, und umgekehrt. Interessant ist dabei auch die Frage nach der Güte dieser Approximation.

**Theorem 9.2** (Gesetz der großen Zahlen)**.** *Sei $(X_i)_{i\in\mathbb{N}}$ eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$ mit $\mathbb{V}(X_i) < \infty$. Sei*

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

*das* empirische Mittel *der ersten $n$ dieser Zufallsvariablen. Die Folge $(S_n)_{n\in\mathbb{N}}$ konvergiert gegen den Erwartungswert $\mathbb{E}[X_1]$, in dem Sinne dass für alle $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mathbb{E}[X_1]\big| > \varepsilon) = 0$$

*gilt.*

Wenn wir also die $y_i$ aus Beispiel 9.1 als Realisierungen der unabhängigen, identisch verteilten Zufallsvariablen $X_i$ auffassen, dann bestätigt dieser Satz unsere Vermutung aus Beispiel 9.1, nämlich dass das empirische Mittel gegen den Erwartungswert konvergiert.

*Beweis des Gesetzes der großen Zahlen.* Sei $Y = S_n - \mathbb{E}[S_n]$. Nach den Rechenregeln für die Varianz (Satz 5.15) gilt

$$\mathbb{V}(Y) = \mathbb{V}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[S_n]\Big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) = \frac{1}{n^2} n \cdot \mathbb{V}(X_1) = \frac{\mathbb{V}(X_1)}{n}.$$

Somit folgt aus der Chebyshev-Ungleichung (Satz 5.16) für jedes $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mathbb{E}[S_n]| > \varepsilon) \le \lim_{n\to\infty} \frac{\mathbb{V}(X_1)}{n\varepsilon^2} = 0.$$

□

*Bemerkung.* Theorem 9.2 wird normalerweise *schwaches Gesetz der großen Zahlen* genannt, wobei sich das Wort *schwach* auf die genaue Formulierung der Konvergenz bezieht. Es gelten auch ähnliche Aussagen für andere Formen der Konvergenz von Zufallsvariablen, auf die wir hier jedoch nicht eingehen. Für unsere Zwecke wichtig ist die Tatsache, dass die Folge der empirischen Mittel von unabhängigen und identisch verteilten Zufallsvariablen in einem geeigneten Sinne gegen den Erwartungswert konvergieren.

9.2. **Zentraler Grenzwertsatz und Anwendungen.** Aus dem Gesetz der großen Zahlen wissen wir, dass für unabhängige, identisch verteilte Zufallsvariablen $(X_i)_{i\in\mathbb{N}}$ gilt:

$$\sum_{i=1}^{n} X_i \approx n \cdot \mathbb{E}[X_1].$$

In diesem Abschnitt wollen wir diese Approximation verbessern, und eine Aussage über den “Approximationsfehler” machen.

**Theorem 9.3** (Zentraler Grenzwertsatz)**.** *Sei $(X_i)$ eine Folge von unabhängigen, identisch verteilten Zufallsvariablen auf einem Wahrscheinlichkeitsraum$(\Omega, \mathbb{P})$, mit $\mathbb{E}[X_1] = \mu, \mathbb{V}(X_1) = \sigma^2 \in (0, \infty)$. Dann gilt für alle $x \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \leq x\Big) = \Phi_{0,1}(x).$$

*Dabei ist $\Phi_{0,1}$ die Verteilungsfunktion der Standardnormalverteilung, vgl. Kapitel 6.3.*

*Bewis.* Siehe z.B. Satz 15.37 in Klenke, Wahrscheinlichkeitstheorie (Springer, 2. Auflage 2008). □

Der zentrale Grenzwertsatz besagt, dass für unabhängige, identisch verteilte Zufallsvariablen $X_i$ mit endlicher Varianz die davon abgeleitete Zufallsvariable

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{\sqrt{n}}{\sigma}(S_n - \mu)$$

für große $n$ ungefähr *standardnormalverteilt* ist, wobei $S_n$ definiert ist wie in Theorem 9.2 Dabei spielt die genaue Verteilung der $X_i$ überhaupt keine Rolle! Diese können eine beliebige Verteilung haben, sofern alle $X_i$ dieselbe Verteilung besitzen, und die Varianz endlich ist. In jedem Fall ist der Grenzwert von $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$ durch die Standardnormalverteilung gegeben. Man spricht deshalb auch von einem *universellen Limes*, da er in einer großen Zahl von Situationen auftritt.

Aus dem Gesetz der großen Zahlen wissen wir, dass für unabhängige, identisch verteilte Zufallsvariablen $X_i$ und für große $n$ die Approximation

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx \mathbb{E}[X_1]$$

gilt, bzw.

$$\sum_{i=1}^{n} X_i \approx n \cdot \mathbb{E}[X_1].$$

Wir haben also eine Approximation für die Summe von $n$ unabhängigen, identisch verteilten Zufallsvariablen. Der zentrale Grenzwertsatz gibt uns nun eine weitere Information, die diese Approximation genauer werden lässt, nämlich dass gilt

$$\sum_{i=1}^{n} X_i \approx n \cdot \mathbb{E}[X_i] + \sqrt{n} \cdot \sigma \cdot Y,$$

wobei $Y$ eine standardnormalverteilte Zufallsvariable ist.

**Beispiel 9.4** (Normalverteilungsannahme der Regressionsfehler)**.** `siehe Vorlesung`

**Beispiel 9.5** (Zentraler Grenzwertsatz und Binomialverteilung)**.** Sei $X$ eine binomialverteilte Zufallsvariable mit Parametern $p = 0.4$ und $n = 20$. In Kapitel 3 hatten wir gesehen,

dass man $X$ mit Hilfe von unabhängigen, zum Parameter $p = 0.4$ Bernoulli-verteilten Zufallsvariablen $Y_i, i = 1, ..., 20$ schreiben kann als

$$X = \sum_{i=1}^{20} Y_i.$$

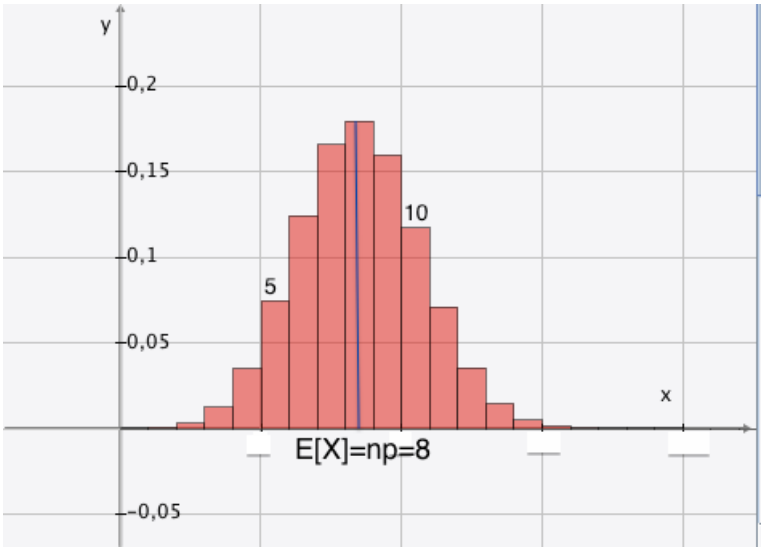
In anderen Worten, $X$ kann als Summe von unabhängigen, identisch verteilten Zufallsvariablen geschrieben werden – also genau die Größe, über die uns das Gesetz der großen Zahlen und der zentrale Grenzwertsatz Auskunft geben. Wir haben nämlich gesehen, dass nach diesen Sätzen gilt, dass

$$X =\approx 20 \cdot \mathbb{E}[Y_1] + \sqrt{20} \cdot \sigma \cdot Y,$$

wobei $\mathbb{E}[Y_1] = p = 0.4$ und $\sigma = \sqrt{\mathbb{V}(Y_1)} = \sqrt{p(1-p)} = \sqrt{0.24}$ sind, und $Y$ eine standardnormalverteilte Zufallsvariable ist.

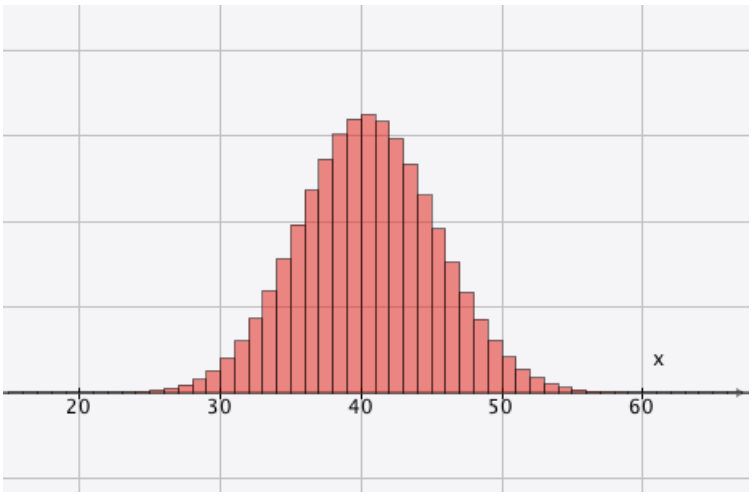
Dies kann man erkennen, wenn man sich die Verteilungsfunktion der Binomialverteilung aufzeichnet, und die Achsen etwas streckt bzw. staucht:

$X \sim$Bin$(0.4, 20)$ (skaliert)




Dabei wird sichtbar, dass die Form der Binomialverteilung schon recht nah an der Gauß'schen Glockenkurve, d.h. der Dichte der Normalverteilung, liegt. Noch deutlicher wird es, wenn statt $n = 20$ nun $n = 100$ gewählt wird:

$X \sim$Bin$(0.4, 100)$ (skaliert)




**Satz 9.6** (Normalapproximation der Binomialverteilung)**.** *Sei $X \sim Bin(n,p)$. Dann gilt für $a,b \in \{0,...,n\}$ mit $a<b$*

$$\mathbb{P}(a \leq X \leq b) \approx \Phi_{0,1}\Big(\frac{b+1/2-np}{\sqrt{np(1-p)}}\Big) - \Phi_{0,1}\Big(\frac{a-1/2-np}{\sqrt{np(1-p)}}\Big).$$

Wir geben in der Vorlesung eine Herleitung dieses Ergebnisses aus dem zentralen Grenzwertsatz an, die im Wesentlichen nur eine Umformung ist, unter Berücksichtigung der Tatsache, dass die Binomialverteilung eine diskrete Verteilung ist, die Normalverteilung jedoch stetig. Für eine diskrete Verteilung gilt für $a,b \in \mathbb{N}$ dass

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a-1/2 \leq X \leq b+1/2)$$

ist, was die Approximation durch die stetige Normalverteilung verbessert.

Wir wissen, dass die Approximation durch den zentralen Grenzwertsatz besser wird, je größer $n$ ist. Im Zusammenhang mit der Binomialverteilung spielt insbesondere die Größe von $n$ in Abhängigkeit von $p$ eine Rolle. Als Faustregel kann verwendet werden, dass die Approximation gut ist, d.h. bis auf 2-3 Nachkommastellen mit dem tatsächlichen Wert übereinstimmt, falls $np \geq 5$ und $n(1-p) \geq 5$ erfüllt sind.

**Beispiel 9.7** (Normalapproximation der Binomialverteilung)**.** Siehe Vorlesung.

# 10. Parameterschätzung

In diesem Kapitel widmen wir uns dem ersten der in der Einleitung erwähnten Grundprobleme der Statistik, der Schätzung von Parametern. Das bedeutet, dass aus gemessenen Daten relevante Kenngrößen der zugrundeliegenden Verteilung *geschätzt* oder approximiert werden, wobei normalerweise gewisse Annahmen getroffen werden müssen, z.B. Unabhängigkeit.

Es gibt viele verschiedene Methoden für die Parameterschätzung. Wir lernen klassische Schätzer für wichtige Kenngrößen kennen, sowie die Methode der "Maximum Likelihood"-Schätzung.

Die **Lernziele** dieses Kapitel sind:

- Klassische Schätzer für wichtige Kenngrößen kennen
- ML-Schätzer berechnen können
- Den EM-Algorithmus kennen

## 10.1. **Beispiele und Eigenschaften von Schätzern.**

**Beispiel 10.1** (Empirisches Mittel)**.** In Definition 8.10 hatten wir das empirische Mittel $\bar{\mu}_x$ eines Datenvektors $x = (x_1, ..., x_n) \in \mathbb{R}^n$ betrachtet, definiert als

$$\bar{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Wir schreiben dafür auch

$$\bar{\mu}_n(x_1, ..., x_n) := \bar{\mu}_x,$$

wobei einerseits die Länge $n$ des Datenvektors betont wird, andererseits die Tatsache, dass es sich dabei um eine *Funktion* der Messwerte $x_1, ..., x_n$ handelt.

Falls die Daten Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit $\mathbb{E}[X_1] = ... = \mathbb{E}[X_n] = \mu$ sind, so gilt

$$\bar{\mu}_n(x_1, ..., x_n) \approx \mu,$$

zumindest für hinreichend große $n$, da laut dem Gesezt der großen Zahlen

$$\bar{\mu}_n(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) \to \mathbb{E}[X_i] = \mu$$

gilt. Somit ist $\bar{\mu}_x$ eine Größe, welche direkt aus den Daten $x_1, ..., x_n$ berechnet werden kann, und einen Schätzwert für den Erwartungswert der zugrundeliegenden Zufallsvariablen liefert.

**Beispiel 10.2** (Empirische Varianz)**.** Die empirische Varianz von $x = (x_1, ..., x_n)$ wurde definiert (vgl. Def. 8.12) als

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

Wieder schreiben wir auch

$$\bar{\sigma}_n^2(x_1, ..., x_n) := \bar{\sigma}_x^2.$$

Falls die $x_i$ Realisierungen von unabhängigen identisch verteilten Zufallsvariablen $X_1, ..., X_n$ sind, d.h. so kann man zeigen (Hausaufgaben), dass

$$\mathbb{E}[\bar{\sigma}_n^2(X_1, ..., X_n) = \mathbb{E}\big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\mu}_n(X_1, ..., X_n)^2\big] = \mathbb{V}(X_1)$$

gilt. Auch hier haben wir es wieder mit einer aus den Daten berechenbaren Größe zu tun, welche zumindest in einigen Fällen ungefähr gleich der Varianz der zugrundeliegenden Zufallsvariablen ist. Die empirische Varianz ist deshalb ein Schätzer für die Varianz.

**Definition 10.3** (Schätzfunktion)**.** Seien $X_1, ..., X_n$ Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Eine *Schätzfunktion* zur Stichprobengröße $n$ ist eine Funktion $\theta_n : \mathbb{R}^n \to \mathbb{R},\ (X_1, ..., X_n) \mapsto \theta_n(X_1, ..., X_n)$.

Eine Schätzfunktion ist also erst einmal nur eine Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$. Üblicherweise setzt man aber eine Schätzfunktion in Verbindung zu einem (in der konkreten Situation normalerweise unbekannten) Parameter der Verteilung der $X_i$. Dabei wird normalerweise angenommen, dass die $X_1, ..., X_n$ alle dieselbe Verteilung haben, welche von einem Parameter $\theta$ abhängen. Dann soll $\theta_n(X_1, ..., X_n)$ als Approximation, oder eben als *Schätzer* für $\theta$ dienen. Damit letzteres der Fall ist, sollte eine Schätzfunktion eine oder mehrere günstige Eigenschaften haben.

**Definition 10.4** (Schätzer mit günstigen Eigenschaften)**.** Sei $(x_1, ..., x_n) \in \mathbb{R}^n$ ein Vektor von Daten, welche als Realisierungen von identisch verteilten Zufallsvariablen $X_1, ..., X_n$ mit Parameter $\theta$ auf einem Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$. Sei $\theta_n$ eine Schätzfunktion zur Stichprobengröße $n$. Die Schätzfunktion $\theta_n$ ist ein *erwartungstreuer Schätzer* für $\theta$, falls

$$\mathbb{E}[\theta_n(X_1, ..., X_n)] = \theta$$

ist. Statt erwartungstreu verwendet man auch die Bezeichnungen *unverzerrt* oder (englisch) *unbiassed*. Die Schätzfunktion $\theta_n$ ist ein *konsistenter Schätzer* für $\theta$, falls

$$\lim_{n\to\infty} \theta_n(X_1, ..., X_n) = \theta$$

gilt. Die Schätzfunktion $\theta_n$ ist ein *effizienter Schätzer* für $\theta$, falls

$$\lim_{n\to\infty} \mathbb{V}(\theta_n(X_1, ..., X_n)) = 0$$

gilt.

Bei der Parameterschätzung steht man also normalerweise vor der Situation, dass man Daten $x_1, ..., x_n$ gemessen hat, von denen man weiß oder annimmt, dass sie einer gewissen Verteilung mit einem unbekannten Parameter $\theta$ folgen. Gesucht ist dann eine Schätzfunktion $\theta_n$, welche als Schätzer für $\theta$ dient, wenn möglich mit einer oder mehreren günstigen Eigenschaften. Hat man $\theta_n$ konstruiert, kann

$$\hat{\theta} := \theta_n(x_1, ..., x_n)$$

als Schätzung oder Approximation für den unbekannten Parameter $\theta$ dienen.

**Beispiel 10.5** (Empirisches Mittel und empirische Varianz)**.** In den Beispielen 10.1 und 10.2 haben wir gesehen, dass das empirische Mittel ein erwartungstreuer und konsistenter Schätzer für den Erwartungswert einer (beliebigen) Verteilung ist, und die empirische Varianz ein erwartungstreuer Schätzer für die Varianz.

**Beispiel 10.6** (Erwartungstreu impliziert nicht konsistent)**.** (siehe Vorlesung)

**Beispiel 10.7** (Konsitent impliziert nicht erwartungstreu)**.** (siehe Vorlesung)

10.2. **Maximum Likelihood-Schätzung.** Die Maximum Likelihood-Schätzung ist eine spezielle Methode der Parameterschätzung, bei der unter allen theoretisch möglichen Werten eines Parameters einer Verteilung derjenige gesucht wird, für den die Wahrscheinlichkeit maximal wird, die beobachteten Messwerte $x_1, ..., x_n$ zu finden. Dazu wird die sogenannte Likelihood-Funktion maximiert. Dabei unterscheiden wir zwei Fälle. Im ersten Fall sind die $x_i$ Realisierungen von *diskreten* Zufallsvariablen $X_i$, wobei wir annehmen dass die Verteilung der $X_i$ bekannt ist, jedoch von einem unbekannten Parameter $\theta$ abhängt. Wir schreiben für die Verteilung

$$p_\theta(k) = \mathbb{P}_\theta(X_i = k),$$

wobei wir die Abhängigkeit von $\theta$ in der Notation betonen. Im zweiten Fall sind die $x_i$ Realisierungen von Zufallsvariablen $X_i$ welche eine Dichte

$$f_\theta$$

in Abhängigkeit von $\theta$ besitzen.

**Definition 10.8** (Likelihood-Funktion)**.** Seien $x_1, ..., x_n \in \mathbb{R}$ Messwerte, welche als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen $X_1, ..., X_n$ aufgefasst werden können. Dabei nehmen wir an, dass die Zufallsvariablen entweder diskret sind oder eine Dichte besitzen, wobe Verteilung bzw. Dichte jeweils bekannt sind, aber von einem (unbekannten) Parameter $\theta$ abhängen. Die *Likelihood-Funktion* ist die Funktion $L((x_1, ..., x_n); \theta)$ von $\mathbb{R}^n \times \mathbb{R}$ nach $[0, 1]$, für welche gilt:

- Falls $X_i$ eine diskrete Verteilung $p_theta$ mit Parameter $\theta$ besitzt, so ist

$$L((x_1, ..., x_n); \theta) = p_\theta(x_1) \cdot ... \cdot p_\theta(x_n) = \prod_{i=1}^{n} p_\theta(x_i).$$

- Falls $X_i$ eine Dichte $f_\theta$ mit Parameter $\theta$besitzts, so ist

$$L((x_1, ..., x_n); \theta) = f_\theta(x_1) \cdot ... \cdot f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Eine Likelihood-Funktion hängt also von $n+1$ Werten ab: Den $n$ Messwerten, sowie dem (normalerweise unbekannten) Parameter $\theta$. In der Notation fassen wir dabei die Messwerte zu einem Vektor zusammen.

**Definition 10.9** (Maximum Likelihood-Schätzer, MLE)**.** Seien $x_1, ..., x_n$ Messwerte, welche Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, welche entweder der diskreten Verteilung $p_\theta$ folgen, oder die Dichte $f_\theta$ haben. Der *Maximum Likelihood-Schätzer* für $\theta$ ist $\theta_* = \theta_*(x_1, ..., x_n)$, welches die Likelihood-Funktion $L((x_1, ..., x_n); \theta)$ unter allen $\theta \in \mathbb{R}$ maximiert. Formal gesprochen ist

$$\theta_*(x_1, ..., x_n) = \text{argmax}_{\theta \in \mathbb{R}} L((x_1, ..., x_n); \theta).$$

Der Maximum Likelihood-Schätzer wird dabei manchmal auch kurz als MLE bezeichnet, für das englische "maximum likelihood estimator".

Um einen MLE zu bestimmen, geht man also folgendermaßen vor: Gegeben sind üblicherweise die Daten $x_1, ..., x_n$, von denen man weiß oder annimmt, dass sie Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen sind, deren Verteilung man bis auf den unbekannten Parameter $\theta$ kennt. Danach macht man die folgenden Schritte: Vorgehen:

- Likelihood-Funktion aufstellen (nach Definition 10.8, abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt)
- Den Parameter $\theta$ so bestimmen, dass $L((x_1, ..., x_n); \theta)$ für das vorgegebene $x = (x_1, ..., x_n)$ maximal ist. D.h. man maximiert die Funktion $\theta \mapsto L((x_1, ..., x_n); \theta)$.

Aus der Analysis sollte bekannt sein, dass Maxima einer differenzierbaren Funktion durch Bestimmung der Nullstellen der ersten Ableitung dieser Funktion bestimmt werden kann, wobei jeweils noch überprüft werden muss, dass die gefundenen Werte tatsächlich Maxima sind (z.B. durch den bekannten Test mit der zweiten Ableitung). Um einen MLE-Schätzer zu bestimmen, muss also die Likelihood-Funktion *nach* $\theta$ abgeleitet werden (und nicht etwa nach den $x_i$). In vielen Fällen ist es jedoch so, dass die Likelihood-Funktion etwas aufwändig abzuleiten ist, hingegen eine einfache Transformation davon viel leichter zu behandeln ist.

**Definition 10.10.** Sei $L((x_1, ..., x_n); \theta)$ eine Likelihood-Funktion. Die *Log-Likelihood-Funktion* ist definiert als

$$l((x_1, ..., x_n); \theta) = \ln L((x_1, ..., x_n); \theta) (= \ln L((x_1, ..., x_n); \theta).$$

Hierbei bezeichnet ln den natürliche Logarithmus.

**Satz 10.11.** *$l((x_1, ..., x_n); \theta)$ ist genau dann maximal, wenn $L((x_1, ..., x_n); \theta)$ maximal ist.*

*Beweis.* Wir verzichten auf einen formalen Beweis, dieser ist nicht schwer und beruht darauf, dass der Logarithmus eine monoton wachsende Funktion $(0, 1] \to (-\infty, 0]$ ist. □

Somit wird für praktische Zwecke das obige Vorgehen oft zu

- Likelihood-Funktion aufstellen (nach Definition 10.8, abhängig davon ob die Verteilung diskret ist oder eine Dichte besitzt)
- Das Maximum der log-Likelihood-Funktion $l((x_1, ..., x_n); \theta)$ durch Ableiten nach $\theta$ bestimmen, und überprüfen ob die gefundene(n) Nullstelle(n) tatsächlich Maxima der Likelihood-Funktion sind.

**Beispiel 10.12** (Poisson-Verteilung)**.** Wir haben Messwerte $(x_1, ..., x_n)$ von denen wir wissen, dass sie Poisson-verteilt sind, wobei wir jedoch den Parameter $\lambda$ nicht kennen. Somit haben wir die Verteilung

$$p_\lambda(k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

der zu schätzende Parameter $\theta$ ist also einfach der Parameter der Poisson-Verteilung. Entsprechend dem oben beschriebenen Vorgehen stellen wir die Likelihood-Funktion auf:

$$\begin{aligned} L((x_1, ..., x_n); \lambda) &= \prod_{i=1}^{n} p_\lambda(x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \left(e^{-\lambda}\right)^n \cdot \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}. \end{aligned}$$

Diese Funktion sollen wir jetzt bezüglich $\lambda$ maximieren. Wir berechnen zuerst die log-Likelihood-Funktion, und erhalten nach Anwendung der Logarithmengesetze

$$\begin{aligned} l((x_1, ..., x_n); \lambda) &= \ln L((x_1, ..., x_n); \lambda) = \ln\left[\left(e^{-\lambda}\right)^n \cdot \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}\right] \\ &= \ln e^{-n\cdot\lambda} + \sum_{i=1}^{n} (x_i \cdot \ln\lambda - \ln(x_i!)) \\ &= -n \cdot \lambda + \ln\lambda \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!). \end{aligned}$$

Diese Funktion ist nun leicht nach $\lambda$ abzuleiten, unter anderem weil die Summen jeweils nicht von $\lambda$ abhängt. Wir erhalten

$$\frac{\partial l((x_1, ..., x_n); \lambda)}{\partial\lambda} = -n + \frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i$$

Setzen wir die Ableitung gleich null und lösen nach $\lambda$ auf, so finden wir die einzige Nullstelle

$$\lambda_0 = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Dies ist also ein Kandidat für das Maximum der Likelihood-Funktion. Wir müssen nun noch überprüfen, dass es sich hier tatsächlich um ein Maximum (und nicht etwa um ein Minimum oder einen Sattelpunkt) handelt. Berechnen der zweiten Ableitung ergibt

$$\frac{\partial^2 l((x_1, ..., x_n); \lambda)}{\partial\lambda^2} = -\frac{1}{\lambda^2} \cdot \sum_{i=1}^{n} x_i.$$

Setzt man $\lambda_0 = \frac{n}{\sum_{i=1}^{n} x_i}$ für $\lambda$ ein, sieht man sofort dass die zweite Ableitung an dieser Stelle negativ ist, somit handelt es sich um ein Maximum.

Es fällt auf, dass der Maximum-Likelihood-Schätzer gegeben ist durch das empirische Mittel der Messwerte:

$$\lambda_* = \lambda_0 = \frac{\sum_{i=1}^n x_i}{n} = \bar{\mu}_n(x_1, ..., x_n).$$

Dies überrascht möglicherweise nicht, wenn man sich daran erinnert, dass der Erwartungswert der Poisson-Verteilung durch den Verteilungsparameters gegeben ist.

Sind die Messwerte nun konkret gegeben, z.B. $(x_1, ..., x_10) = (5, 3, 4, 3, 2, 9, 3, 1, 7, 3)$, so ergibt Einsetzen den konkreten Maximum-Likelihood-Schätzer $\lambda_* = 4$.

**Beispiel 10.13** (Normalverteilung)**.** Wir nehmen an, dass unsere Messwerte $(x_1, ..., x_n)$ normalverteilt sind. Bekanntlich hat die Normalverteilung die Dichte

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Insbesondere haben wir es hier mit zwei Parametern zu tun, $\mu$ und $\sigma^2$. Obwohl es durchaus auch Möglichkeiten gibt, zwei Parameter gleichzeitig zu schätzen, nehmen wir für dieses Beispiel an, dass $\mu = 0$ gilt (z.B. weil wir wissen dass die gemessenen Daten symmetrisch um 0 verteilt sind). Somit haben wir die Dichte

$$\varphi_{0,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

mit einem unbekannten Parameter $\sigma^2$, den wir mit Hilfe der Maximum-Likelihood-Methode schätzen wollen. Für die Likelihood-Funktion erhalten wir

$$L((x_1, ..., x_n); \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_i^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}},$$

wobei wir im letzten Schritt die Potenzgesetze verwendet haben. Logarithmieren führt auf

$$l((x_1, ..., x_n); \sigma^2) = -n \cdot \ln(\sqrt{2\pi\sigma^2}) - \frac{\sum_{i=1}^n x_i^2}{2\sigma^2} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^n x_i^2}{2\sigma^2}.$$

Hier ist es wichtig zu beachten, dass $\sigma^2$ eine Variable ist – das Quadrat kann möglicherweise verwirren, insbesondere wenn wir gleich ableiten, deshalb schreiben wir überall $\theta = \sigma^2$, so dass die Log-Likelihood-Funktion zu

$$l((x_1, ..., x_n); \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta - \frac{\sum_{i=1}^n x_i^2}{2\theta}$$

wird. Dies leiten wir nun nach $\theta$ ab und erhalten

$$\frac{\partial l((x_1, ..., x_n); \theta)}{\partial\theta} = -\frac{n}{2} \cdot \frac{1}{\theta} - \frac{\sum_{i=1}^n x_i^2}{2} \cdot \frac{-1}{\theta^2}.$$

Die (einzige) Nullstelle der ersten Ableitung ist

$$\theta_0 = \frac{\sum_{i=1}^n x_i^2}{n} = \frac{n}{2\theta^2}\big( - \theta + \frac{1}{n}\sum_{i=1}^n x_i^2.$$

Der Test mit der zweiten Ableitung ergibt

$$\frac{\partial^2 l((x_1,...,x_n);\theta)}{\partial\theta^2} = \frac{1}{\theta^2}\cdot\frac{n}{2} - \frac{\sum_{i=1}^n x_i^2}{2}\frac{2}{\theta^3},$$

und an der Stelle $\theta = \theta_0$ wird dieser Ausdruck zu

$$\frac{n^3}{2(\sum_{i=1}^n x_i^2)^2} - \frac{n^3}{(\sum_{i=1}^n x_i^2)^2} < 0.$$

Somit handelt es sich tatsächlich um ein Maximum.

Wir haben also als Maximum-Likelihood-Schätzer für die Varianz der Normalverteilung

$$\sigma_*^2 = \frac{\sum_{i=1}^n x_i^2}{n}.$$

Dabei fällt auf, dass dies *nicht* die empirische Varianz ist – der MLE ist also insbesondere nicht erwartungstreu (jedoch konsistent).