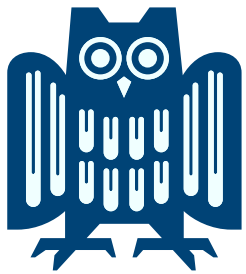
Schließende Statistik

Vorlesung an der Universität des Saarlandes

PD Dr. Martin Becker

Wintersemester 2020/21



Organisatorisches I

- Vorlesung: voraussichtlich nur online, Inhalte jederzeit abrufbar
- Übungen: voraussichtlich nur online, Inhalte jederzeit abrufbar
- Prüfung: 2-stündige Klausur nach Semesterende (1. Prüfungszeitraum)

Wichtig:

Anmeldung (ViPa) vom 24. November – 08. Dezember (bis 15 Uhr) möglich
Abmeldung bis 21. Januar 2021 (12 Uhr) möglich

- Hilfsmittel für Klausur
  - „Moderat“ programmierbarer Taschenrechner, auch mit Grafikfähigkeit
  - 2 *beliebig gestaltete* DIN A 4–Blätter (bzw. 4, falls nur einseitig)
  - Benötigte Tabellen werden gestellt, aber **keine weitere Formelsammlung!**
- Durchgefallen — was dann?
  - „Wiederholungskurs“ im kommenden (Sommer-)Semester
  - „Nachprüfung“ (voraussichtlich) erst September/Oktober 2021 (2. Prüfungszeitraum)
  - „Reguläre“ Vorlesung/Übungen wieder im Wintersemester 2021/22

Organisatorisches II

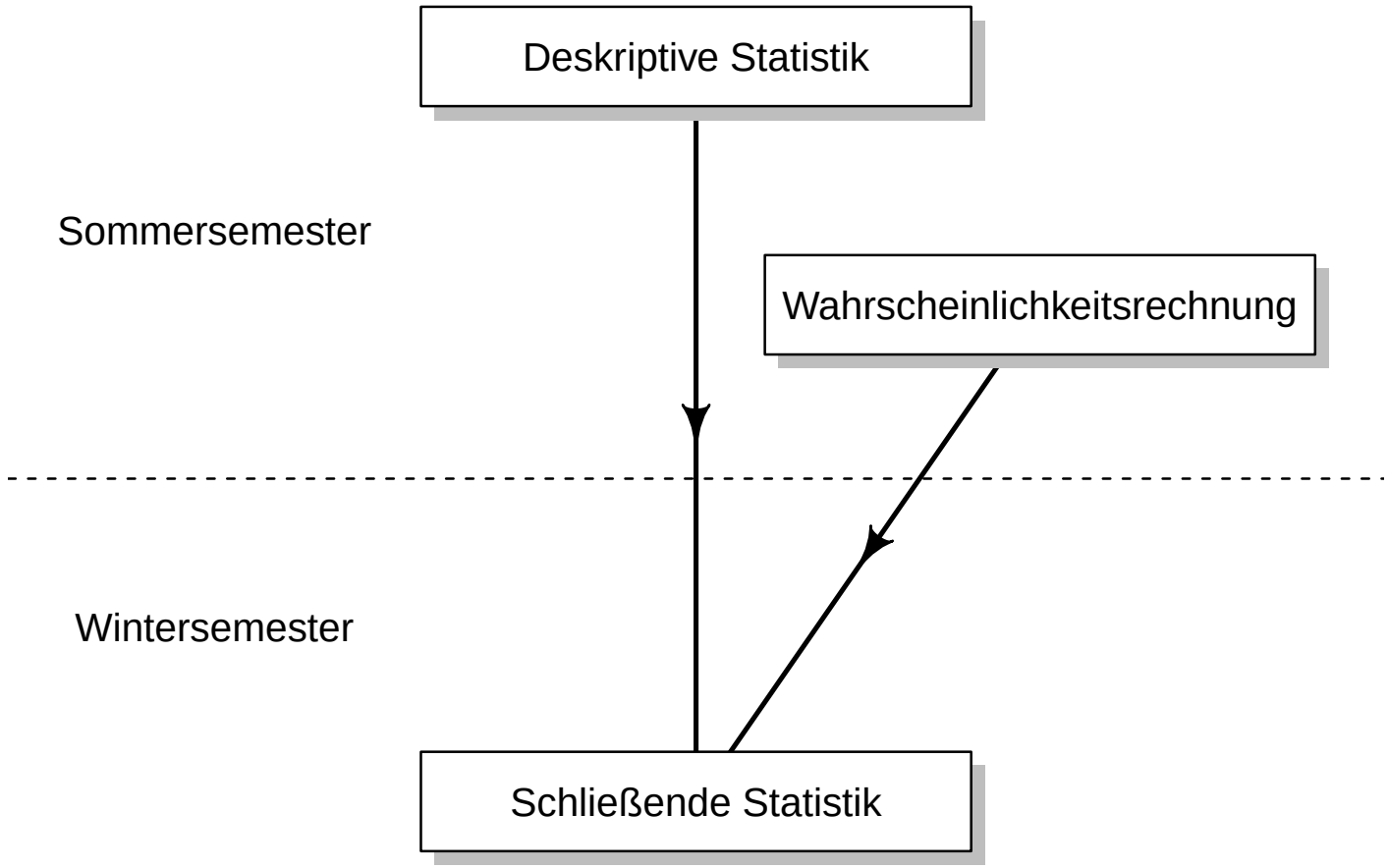
- Kontakt: PD Dr. Martin Becker
  Geb. C3 1, 2. OG, Zi. 2.17
  e-Mail: `martin.becker@mx.uni-saarland.de`
- Sprechstunde (via MS Teams) nach Terminabstimmung per e-Mail
- Informationen und Materialien im (UdS-)Moodle und auf Homepage:

  `http://www.lehrstab-statistik.de`
- Material zu dieser Veranstaltung: Vorlesungsfolien *i.d.R. vor Vorlesung* zum Download (inklusive Drucker-freundlicher 2-auf-1 bzw. 4-auf-1 Versionen)
- Wie in „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“:
  - Neben theoretischer Einführung der Konzepte auch einige Beispiele auf Vorlesungsfolien
  - Einige wichtige Grundlagen werden gesondert als „Definition“, „Satz“ oder „Bemerkung“ hervorgehoben
  - **Aber:** Auch vieles, was nicht formal als „Definition“, „Satz“ oder „Bemerkung“ gekennzeichnet ist, ist wichtig!

Organisatorisches III

- Übungsblätter i.d.R. zusammen mit neuen Vorlesungsunterlagen zum Download
- Ergebnisse *(keine Musterlösungen!)* zu den meisten Aufgaben ebenfalls unmittelbar verfügbar
- Ausführlichere Lösungen zu den Übungsaufgaben (Online-Skript + noch ausführlichere Erklärvideos) einige Tage später, *damit Sie nicht zu sehr in Versuchung geraten, sich die Lösung* **vor** *der eigenen Bearbeitung der Übungsblätter anzuschauen!*
- Eigene Bearbeitung der Übungsblätter (**vor** Betrachten der bereitgestellten Lösungen) wichtigste Klausurvorbereitung (eine vorhandene Lösung zu verstehen etwas **ganz** anderes als eine eigene Lösung zu finden!).

Organisation der Statistik-Veranstaltungen



Benötigte Konzepte

aus den mathematischen Grundlagen

- Rechnen mit Potenzen

$$a^m \cdot b^m = (a \cdot b)^m \qquad a^m \cdot a^n = a^{m+n} \qquad \frac{a^m}{a^n} = a^{m-n} \qquad (a^m)^n = a^{m \cdot n}$$

- Rechnen mit Logarithmen

$$\ln(a \cdot b) = \ln a + \ln b \qquad \ln\left(\frac{a}{b}\right) = \ln a - \ln b \qquad \ln\left(a^r\right) = r \cdot \ln a$$

- Rechenregeln auch mit Summen-/Produktzeichen, z.B.

$$\ln\left(\prod_{i=1}^{n} x_i^{r_i}\right) = \sum_{i=1}^{n} r_i \ln(x_i)$$

- Maximieren differenzierbarer Funktionen
  - Funktionen (ggf. partiell) ableiten
  - Nullsetzen von Funktionen (bzw. deren Ableitungen)
- „Unfallfreies“ Rechnen mit 4 Grundrechenarten und Brüchen...

Benötigte Konzepte

aus Veranstaltung „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“

- Diskrete und stetige Zufallsvariablen X, Verteilungsfunktionen, Wahrscheinlichkeitsverteilungen, ggf. Dichtefunktionen
- Momente (Erwartungswert $\mathrm{E}(X)$, Varianz $\mathrm{Var}(X)$, höhere Momente $\mathrm{E}(X^k)$)
- „Einbettung“ der deskriptiven Statistik in die Wahrscheinlichkeitsrechnung
  - Ist Ω die (endliche) Menge von Merkmalsträgern einer deskriptiven statistischen Untersuchung, $\mathcal{F} = \mathcal{P}(\Omega)$ und P die Laplace-Wahrscheinlichkeit

    $$P : \mathcal{P}(\Omega) \to \mathbb{R}; B \mapsto \frac{\#B}{\#\Omega} ,$$

    so kann jedes numerische Merkmal X als Zufallsvariable $X : \Omega \to \mathbb{R}$ verstanden werden.
  - Der Träger von X entspricht dann dem Merkmalsraum $A = \{a_1, \ldots, a_m\}$, die Punktwahrscheinlichkeiten den relativen Häufigkeiten, d.h. es gilt $p(a_j) = r(a_j)$ bzw. — äquivalent — $P_X(\{a_j\}) = r(a_j)$ für $j \in \{1, \ldots, m\}$.
- Verteilung von $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ für unabhängig identisch verteilte X_i
  - falls X_i normalverteilt
  - falls $n \to \infty$ (Zentraler Grenzwertsatz!)

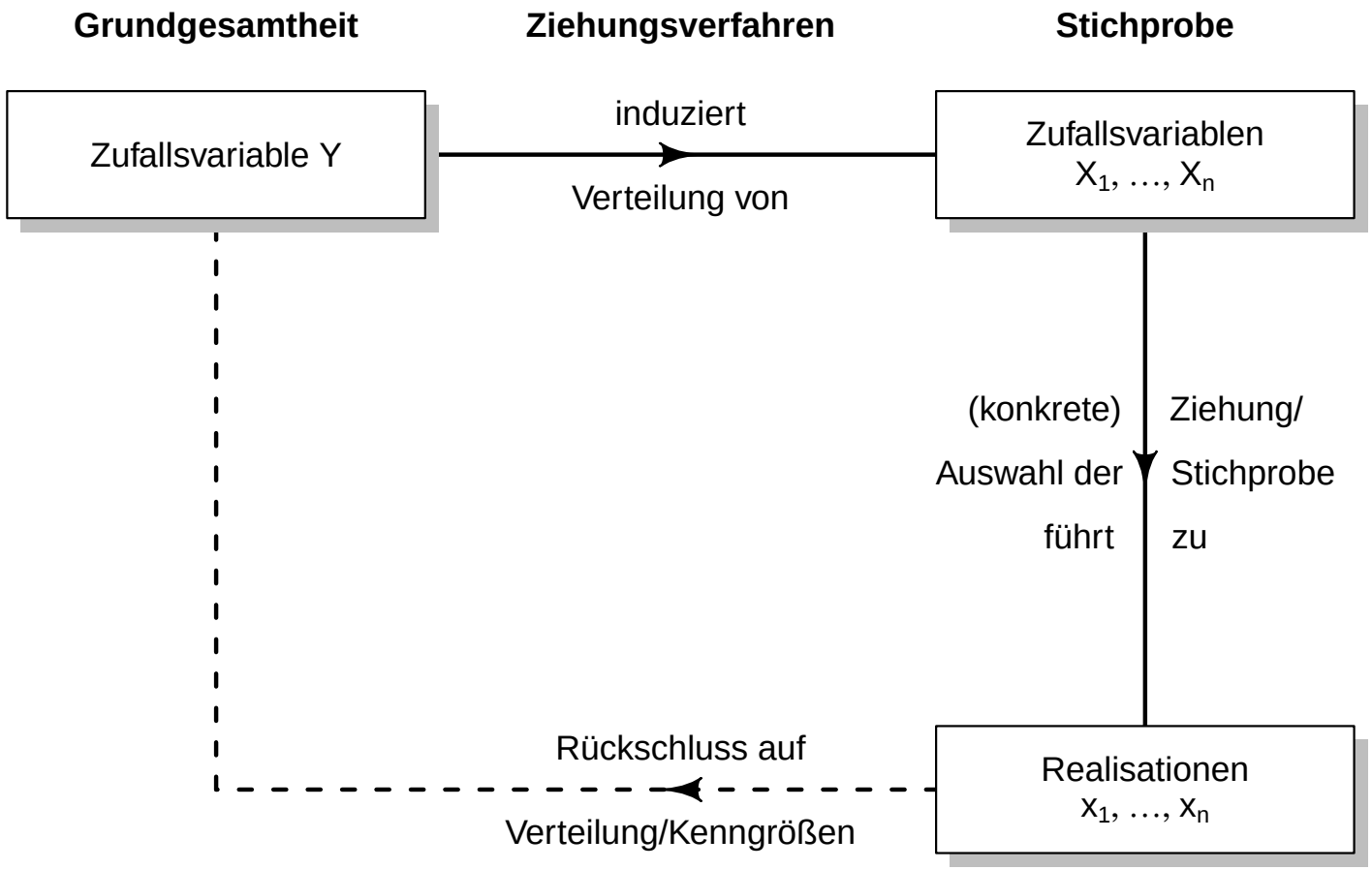
Grundidee der schließenden Statistik

- Ziel der schließenden Statistik/induktiven Statistik:

  Ziehen von Rückschlüssen auf die Verteilung einer (größeren) Grundgesamtheit auf Grundlage der Beobachtung einer (kleineren) Stichprobe.

- Rückschlüsse auf die Verteilung können sich auch beschränken auf spezielle Eigenschaften/Kennzahlen der Verteilung, z.B. den Erwartungswert.
- „Fundament“: **Drei Grundannahmen**
  1. Der interessierende Umweltausschnitt kann durch eine (ein- oder mehrdimensionale) Zufallsvariable Y beschrieben werden.
  2. Man kann eine *Menge* W von Wahrscheinlichkeitsverteilungen angeben, zu der die *unbekannte* wahre Verteilung von Y gehört.
  3. Man beobachtet Realisationen $x_1, \ldots, x_n$ von (Stichproben-)Zufallsvariablen $X_1, \ldots, X_n$, deren gemeinsame Verteilung *in vollständig bekannter Weise* von der Verteilung von Y abhängt.
- Ziel ist es also, aus der Beobachtung der n Werte $x_1, \ldots, x_n$ mit Hilfe des bekannten Zusammenhangs zwischen den Verteilungen von $X_1, \ldots, X_n$ und Y Aussagen über die Verteilung von Y zu treffen.

„Veranschaulichung“ der schließenden Statistik



Bemerkungen zu den 3 Grundannahmen

- Die 1. Grundannahme umfasst insbesondere die Situation, in der die Zufallsvariable Y einem (ein- oder mehrdimensionalen) Merkmal auf einer *endlichen* Menge von Merkmalsträgern entspricht, vgl. die Einbettung der deskriptiven Statistik in die Wahrscheinlichkeitsrechnung auf Folie 7.
  In diesem Fall interessiert man sich häufig für Kennzahlen von Y, z.B. den Erwartungswert von Y (als Mittelwert des Merkmals auf der Grundgesamtheit).
- Die Menge W von Verteilungen aus der 2. Grundannahme ist häufig eine *parametrische* Verteilungsfamilie, zum Beispiel die Menge aller Exponentialverteilungen oder die Menge aller Normalverteilungen mit Varianz $\sigma^2 = 2^2$.
  In diesem Fall ist die Menge der für die Verteilung von Y denkbaren Parameter interessant (später mehr!). Wir betrachten dann nur solche Verteilungsfamilien, in denen verschiedene Parameter auch zu verschiedenen Verteilungen führen („Parameter sind *identifizierbar.*“).
- Wir beschränken uns auf *sehr* einfache Zusammenhänge zwischen der Verteilung der interessierenden Zufallsvariablen Y und der Verteilung der Zufallsvariablen $X_1, \ldots, X_n$.

Beispiel I

Stichprobe aus endlicher Grundgesamtheit Ω

- Grundgesamtheit: $N = 4$ Kinder (**A**nna, **B**eatrice, **C**hristian, **D**aniel) gleichen Alters, die in derselben Straße wohnen: $\Omega = \{A, B, C, D\}$
- Interessierender Umweltausschnitt: monatliches Taschengeld Y (in €) bzw. später spezieller: Mittelwert des monatlichen Taschengelds der 4 Kinder (entspricht $\mathrm{E}(Y)$ bei Einbettung wie beschrieben)
- (Verteilungsannahme:) Verteilung von Y unbekannt, aber sicher in der Menge der diskreten Verteilungen mit maximal $N = 4$ (nichtnegativen) Trägerpunkten und Punktwahrscheinlichkeiten, die Vielfaches von $1/N = 1/4$ sind.
  Im Beispiel nun: Zufallsvariable Y nehme Werte

ω	A	B	C	D
$Y(\omega)$	15	20	25	20

an, habe also folgende zugehörige Verteilung:

y_i	15	20	25	Σ
$p_Y(y_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Beispiel II

Stichprobe aus endlicher Grundgesamtheit Ω

- **Beachte:** Verteilung von Y nur im Beispiel bekannt, in der Praxis: Verteilung von Y natürlich unbekannt!
- Einfachste Möglichkeit, um Verteilung von Y bzw. deren Erwartungswert zu ermitteln: alle 4 Kinder nach Taschengeld befragen!
- Typische Situation in schließender Statistik: nicht alle Kinder können befragt werden, sondern nur eine kleinere Anzahl $n < N = 4$, beispielsweise $n = 2$. Erwartungswert von Y (mittleres Taschengeld aller 4 Kinder) kann dann nur noch **geschätzt** werden!
- Ziel: Rückschluss aus der Erhebung von $n = 2$ Taschengeldhöhen auf die größere Grundgesamtheit von $N = 4$ Kindern durch
  - Schätzung des mittleren Taschengeldes aller 4 Kinder
  - Beurteilung der Qualität der Schätzung (mit welchem „Fehler“ ist zu rechnen)
- (Qualität der) Schätzung hängt ganz entscheidend vom Ziehungs-/Auswahlverfahren ab!

Beispiel III

Stichprobe aus endlicher Grundgesamtheit Ω

- Erhebung von 2 Taschengeldhöhen führt zu Stichprobenzufallsvariablen X_1 und X_2.
- X_1 bzw. X_2 entsprechen in diesem Fall dem Taschengeld des 1. bzw. 2. befragten Kindes
- **Sehr wichtig** für Verständnis:
  X_1 und X_2 sind Zufallsvariablen, da ihr Wert (Realisation) davon abhängt, **welche Kinder** man zufällig ausgewählt hat!
- Erst **nach Auswahl** der Kinder (also nach „Ziehung der Stichprobe“) steht der Wert (die Realisation) x_1 von X_1 bzw. x_2 von X_2 fest!

Variante A

- Naheliegendes Auswahlverfahren: nacheinander **rein zufällige** Auswahl von 2 der 4 Kinder, d.h. **zufälliges Ziehen ohne Zurücklegen mit Berücksichtigung der Reihenfolge**
- Alle $(4)_2 = 12$ Paare (A, B); (A, C); (A, D); (B, A); (B, C); (B, D); (C, A); (C, B); (C, D); (D, A); (D, B); (D, C) treten dann mit der gleichen Wahrscheinlichkeit $(1/12)$ auf und führen zu den folgenden „Stichprobenrealisationen“ (x_1, x_2) der Stichprobenvariablen (X_1, X_2):

Beispiel IV

Stichprobe aus endlicher Grundgesamtheit Ω

- Realisationen (x_1, x_2) zur Auswahl von 1. Kind (Zeilen)/2. Kind (Spalten):

	A	B	C	D
A	unmöglich	(15,20)	(15,25)	(15,20)
B	(20,15)	unmöglich	(20,25)	(20,20)
C	(25,15)	(25,20)	unmöglich	(25,20)
D	(20,15)	(20,20)	(20,25)	unmöglich

- Resultierende gemeinsame Verteilung von (X_1, X_2):

$x_1 \backslash x_2$	15	20	25	Σ
15	0	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$
20	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
25	$\frac{1}{12}$	$\frac{1}{6}$	0	$\frac{1}{4}$
Σ	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

- Es fällt auf (**Variante A**):
  - X_1 und X_2 haben die gleiche Verteilung wie Y.
  - X_1 und X_2 sind **nicht** stochastisch unabhängig.

Beispiel V

Stichprobe aus endlicher Grundgesamtheit Ω

- Naheliegend: Schätzung des Erwartungswertes $\mathrm{E}(Y)$, also des mittleren Taschengeldes aller 4 Kinder, durch den (arithmetischen) Mittelwert der erhaltenen Werte für die 2 befragten Kinder.
- **Wichtig: Nach** Auswahl der Kinder ist dieser Mittelwert eine Zahl, es ist aber sehr nützlich, den Mittelwert schon **vor** Auswahl der Kinder (dann) als Zufallsvariable (der Zufall kommt über die zufällige Auswahl der Kinder ins Spiel) zu betrachten!
- Interessant ist also die Verteilung der **Zufallsvariable** $\overline{X} := \frac{1}{2}(X_1 + X_2)$, also des Mittelwerts der Stichprobenzufallsvariablen X_1 und X_2.
  Die (hiervon zu unterscheidende!) **Realisation** $\overline{x} = \frac{1}{2}(x_1 + x_2)$ ergibt sich erst (als Zahlenwert) nach Auswahl der Kinder (wenn die Realisation (x_1, x_2) von (X_1, X_2) vorliegt)!
- Verteilung von $\overline{X}$ hier (**Variante A**):

$\overline{x}_i$	17.5	20	22.5	Σ
$p_{\overline{X}}(\overline{x}_i)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

Beispiel VI

Stichprobe aus endlicher Grundgesamtheit Ω

Variante B

- Weiteres mögliches Auswahlverfahren: 2-fache **rein zufällige** und **voneinander unabhängige** Auswahl eines der 4 Kinder, wobei erlaubt ist, dasselbe Kind mehrfach auszuwählen, d.h. **zufälliges Ziehen mit Zurücklegen und Berücksichtigung der Reihenfolge**
- Alle $4^2 = 16$ Paare (A, A); (A, B); (A, C); (A, D); (B, A); (B, B); (B, C); (B, D); (C, A); (C, B); (C, C); (C, D); (D, A); (D, B); (D, C); (D, D) treten dann mit der gleichen Wahrscheinlichkeit $(1/16)$ auf und führen zu den folgenden „Stichprobenrealisationen“ (x_1, x_2) der Stichprobenvariablen (X_1, X_2) (zur Auswahl von 1. Kind (Zeilen)/2. Kind (Spalten)):

	A	B	C	D
A	(15,15)	(15,20)	(15,25)	(15,20)
B	(20,15)	(20,20)	(20,25)	(20,20)
C	(25,15)	(25,20)	(25,25)	(25,20)
D	(20,15)	(20,20)	(20,25)	(20,20)

Beispiel VII

Stichprobe aus endlicher Grundgesamtheit Ω

- Resultierende gemeinsame Verteilung von (X_1, X_2):

$x_1 \backslash x_2$	15	20	25	Σ
15	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
20	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
25	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
Σ	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

- Es fällt auf (**Variante B**):
  - X_1 und X_2 haben die gleiche Verteilung wie Y.
  - X_1 und X_2 **sind** stochastisch unabhängig.
- Verteilung von $\overline{X}$ hier (**Variante B**):

$\overline{x}_i$	15	17.5	20	22.5	25	Σ
$p_{\overline{X}}(\overline{x}_i)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	1

Zufallsstichprobe

- Beide Varianten zur Auswahl der Stichprobe führen dazu, dass alle Stichprobenzufallsvariablen X_i $(i = 1, 2)$ **identisch** verteilt sind wie Y.
- Variante **B** führt außerdem dazu, dass die Stichprobenzufallsvariablen X_i $(i = 1, 2)$ **stochastisch unabhängig** sind.

Definition 2.1 ((Einfache) Zufallsstichprobe)

Seien $n \in \mathbb{N}$ und $X_1, \ldots, X_n$ Zufallsvariablen einer Stichprobe vom Umfang n zu Y. Dann heißt $(X_1, \ldots, X_n)$

- **Zufallsstichprobe** vom Umfang n zu Y, falls die Verteilungen von Y und X_i für alle $i \in \{1, \ldots, n\}$ übereinstimmen, alle X_i also identisch verteilt sind wie Y,
- **einfache (Zufalls-)Stichprobe** vom Umfang n zu Y, falls die Verteilungen von Y und X_i für alle $i \in \{1, \ldots, n\}$ übereinstimmen und $X_1, \ldots, X_n$ außerdem stochastisch unabhängig sind.

- (X_1, X_2) ist in Variante A des Beispiels also eine Zufallsstichprobe vom Umfang 2 zu Y, in Variante B sogar eine einfache (Zufalls-)Stichprobe vom Umfang 2 zu Y.

- $X_1, \ldots, X_n$ ist also nach Definition 2.1 auf Folie 18 genau dann eine **Zufallsstichprobe**, falls für die Verteilungsfunktionen zu $Y, X_1, \ldots, X_n$

$$F_Y = F_{X_1} = \cdots = F_{X_n}$$

gilt.

- Ist $X_1, \ldots, X_n$ eine **einfache Stichprobe** vom Umfang n zu Y, so gilt für die *gemeinsame* Verteilungsfunktion von $(X_1, \ldots, X_n)$ sogar

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_Y(x_1) \cdot \ldots \cdot F_Y(x_n) = \prod_{i=1}^{n} F_Y(x_i) \ .$$

Ist Y diskrete Zufallsvariable gilt also insbesondere für die beteiligten Wahrscheinlichkeitsfunktionen

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_Y(x_1) \cdot \ldots \cdot p_Y(x_n) = \prod_{i=1}^{n} p_Y(x_i) \ ,$$

ist Y stetige Zufallsvariable, so existieren Dichtefunktionen von Y bzw. $(X_1, \ldots, X_n)$ mit

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_Y(x_1) \cdot \ldots \cdot f_Y(x_n) = \prod_{i=1}^{n} f_Y(x_i) \ .$$

Stichprobenrealisation/Stichprobenraum

Definition 2.2 (Stichprobenrealisation/Stichprobenraum)

Seien $n \in \mathbb{N}$ und $X_1, \ldots, X_n$ Zufallsvariablen einer Stichprobe vom Umfang n zu Y. Seien $x_1, \ldots, x_n$ die beobachteten Realisationen zu den Zufallsvariablen $X_1, \ldots, X_n$. Dann heißt

- $(x_1, \ldots, x_n)$ **Stichprobenrealisation** und
- die Menge $\mathcal{X}$ aller möglichen Stichprobenrealisationen **Stichprobenraum**.

- Es gilt offensichtlich immer $\mathcal{X} \subseteq \mathbb{R}^n$.
- „Alle möglichen Stichprobenrealisationen“ meint alle Stichprobenrealisationen, die für *irgendeine* der möglichen Verteilungen W von Y aus der Verteilungsannahme möglich sind.
- Wenn man davon ausgeht, dass ein Kind „schlimmstenfalls“ 0 € Taschengeld erhält, wäre im Beispiel also $\mathcal{X} = \mathbb{R}_+^2$ (Erinnerung: $\mathbb{R}_+ := \{x \in \mathbb{R} \,|\, x \geq 0\}$).
- Meist wird die Information der Stichprobenzufallsvariablen bzw. der Stichprobenrealisation weiter mit sog. „Stichprobenfunktionen“ aggregiert, die oft (große) Ähnlichkeit mit Funktionen haben, die in der deskriptiven Statistik zur Aggregierung von Urlisten eingesetzt werden.

Stichprobenfunktion/Statistik

Definition 2.3 (Stichprobenfunktion/Statistik)

Seien $n \in \mathbb{N}$ und $X_1, \ldots, X_n$ Zufallsvariablen einer Stichprobe vom Umfang n zu Y mit Stichprobenraum $\mathcal{X}$. Dann heißt eine Abbildung

$$T : \mathcal{X} \to \mathbb{R}; (x_1, \ldots, x_n) \mapsto T(x_1, \ldots, x_n)$$

Stichprobenfunktion oder **Statistik**.

- Stichprobenfunktionen sind also Abbildungen, deren Wert mit Hilfe der Stichprobenrealisation bestimmt werden kann.
- Stichprobenfunktionen müssen (geeignet, z.B. $\mathcal{B}^n$-$\mathcal{B}$-) messbare Abbildungen sein; diese Anforderung ist aber für alle hier interessierenden Funktionen erfüllt, Messbarkeitsüberlegungen bleiben also im weiteren Verlauf außen vor.
- Ebenfalls als Stichprobenfunktion bezeichnet wird die (als Hintereinanderausführung zu verstehende) Abbildung $T(X_1, \ldots, X_n)$, wegen der Messbarkeitseigenschaft ist dies immer eine **Zufallsvariable**.
  Die Untersuchung der zugehörigen Verteilung ist für viele Anwendungen von **ganz wesentlicher** Bedeutung.

- Wenn man sowohl die Zufallsvariable $T(X_1, \ldots, X_n)$ als auch den aus einer vorliegenden Stichprobenrealisation $(x_1, \ldots, x_n)$ resultierenden Wert $T(x_1, \ldots, x_n)$ betrachtet, so bezeichnet man $T(x_1, \ldots, x_n)$ oft auch als **Realisation** der Stichprobenfunktion.
- Im Taschengeld-Beispiel war die betrachtete Stichprobenfunktion das arithmetische Mittel, also konkreter

$$T : \mathbb{R}^2 \to \mathbb{R};\ T(x_1, x_2) = \overline{x} := \frac{1}{2}(x_1 + x_2)$$

bzw. — als Zufallsvariable betrachtet —

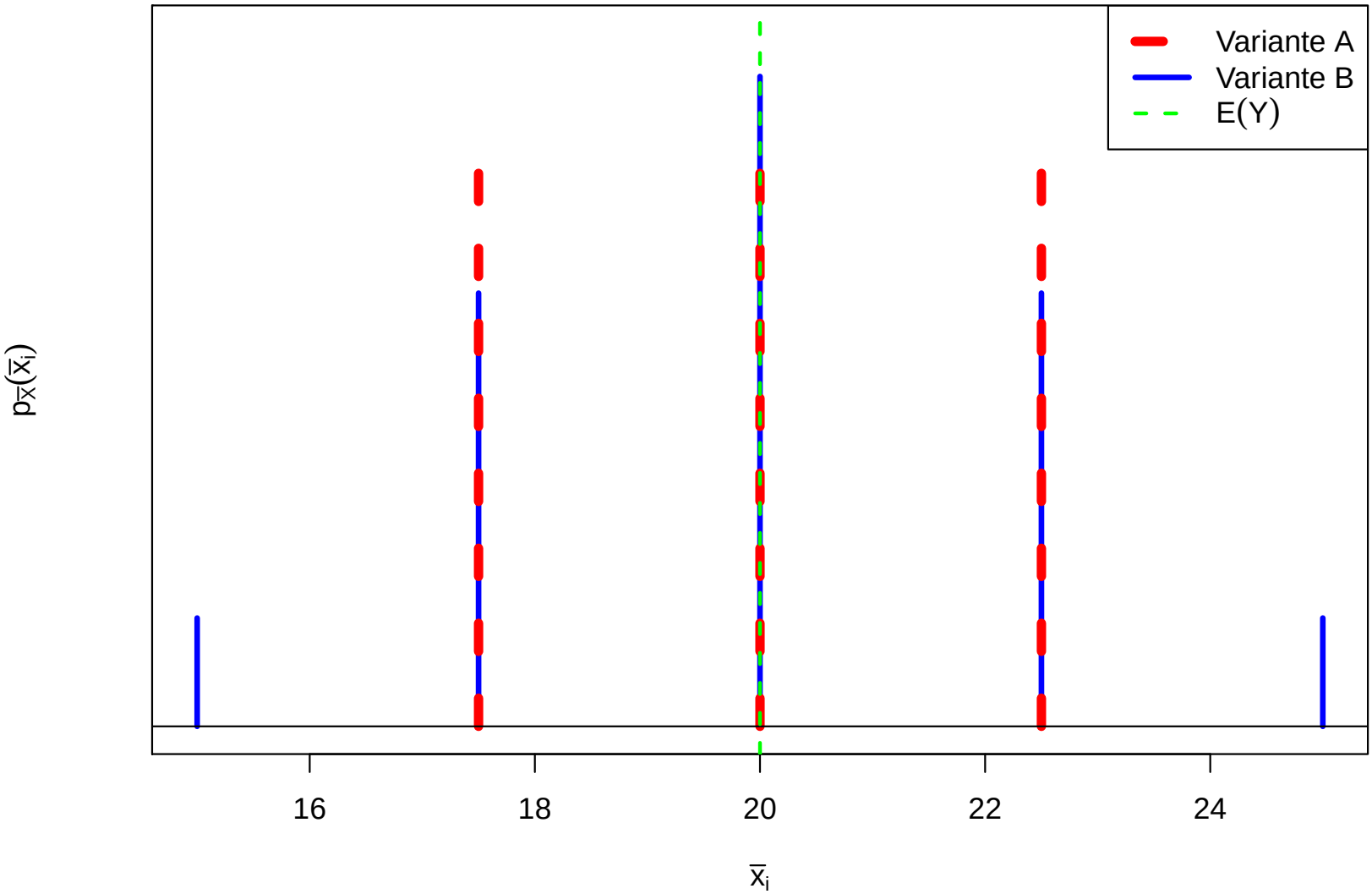
$$T(X_1, X_2) = \overline{X} := \frac{1}{2}(X_1 + X_2) \ .$$

- Je nach Anwendung erhalten Stichprobenfunktionen auch speziellere Bezeichnungen, z. B.
  - **Schätzfunktion** oder **Schätzer**, wenn die Stichprobenfunktion zur Schätzung eines Verteilungsparameters oder einer Verteilungskennzahl verwendet wird (wie im Beispiel!),
  - **Teststatistik**, wenn auf Grundlage der Stichprobenfunktion Entscheidungen über die Ablehnung oder Annahme von Hypothesen über die Verteilung von Y getroffen werden.

Beispiel VIII

Stichprobe aus endlicher Grundgesamtheit Ω

Vergleich der Verteilungen von $\overline{X}$ in beiden Varianten:



Beispiel IX

Stichprobe aus endlicher Grundgesamtheit Ω

- Verteilung von Y

y_i	15	20	25	Σ
$p_Y(y_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

  hat Erwartungswert $\mathrm{E}(Y) = 20$ und Standardabweichung $\mathrm{Sd}(Y) \approx 3.536$.
- Verteilung von $\overline{X}$ (Variante **A**):

$\overline{x}_i$	17.5	20	22.5	Σ
$p_{\overline{X}}(\overline{x}_i)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

  hat Erwartungswert $\mathrm{E}(\overline{X}) = 20$ und Standardabweichung $\mathrm{Sd}(\overline{X}) \approx 2.041$.
- Verteilung von $\overline{X}$ (Variante **B**):

$\overline{x}_i$	15	17.5	20	22.5	25	Σ
$p_{\overline{X}}(\overline{x}_i)$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	1

  hat Erwartungswert $\mathrm{E}(\overline{X}) = 20$ und Standardabweichung $\mathrm{Sd}(\overline{X}) = 2.5$.

Beispiel X

Stichprobe aus endlicher Grundgesamtheit Ω

- In beiden Varianten schätzt man das mittlere Taschengeld $\mathrm{E}(Y) = 20$ also „im Mittel“ richtig, denn es gilt für beide Varianten $\mathrm{E}(\overline{X}) = 20 = \mathrm{E}(Y)$.
- Die Standardabweichung von $\overline{X}$ ist in Variante A kleiner als in Variante B; zusammen mit der Erkenntnis, dass beide Varianten „im Mittel“ richtig liegen, schätzt also Variante A „genauer“.
- In beiden Varianten hängt es vom Zufall (genauer von der konkreten Auswahl der beiden Kinder — bzw. in Variante B möglicherweise zweimal desselben Kindes — ab), ob man *nach Durchführung der Stichprobenziehung* den tatsächlichen Mittelwert als Schätzwert erhält oder nicht.
- Obwohl $\overline{X}$ in Variante A die kleinere Standardabweichung hat, erhält man in Variante B den tatsächlichen Mittelwert $\mathrm{E}(Y) = 20$ mit einer größeren Wahrscheinlichkeit (3/8 in Variante B gegenüber 1/3 in Variante A).

Parameterpunktschätzer

- Im Folgenden: Systematische Betrachtung der Schätzung von Verteilungsparametern, wenn die Menge W der (möglichen) Verteilungen von Y eine **parametrische** Verteilungsfamilie gemäß folgender Definition ist: *(Z.T. Wdh. aus „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“)*

Definition 3.1 (Parametrische Verteilungsfamilie, Parameterraum)

1. Eine Menge von Verteilungen W heißt **parametrische Verteilungsfamilie**, wenn jede Verteilung in W durch einen endlich-dimensionalen Parameter $\theta = (\theta_1, \ldots, \theta_K) \in \Theta \subseteq \mathbb{R}^K$ charakterisiert wird.

   Um die Abhängigkeit von θ auszudrücken, notiert man die Verteilungen, Verteilungsfunktionen sowie die Wahrscheinlichkeits- bzw. Dichtefunktionen häufig als

   $$P(\cdot|\theta_1, \ldots, \theta_K),\ F(\cdot|\theta_1, \ldots, \theta_K) \text{ sowie } p(\cdot|\theta_1, \ldots, \theta_K) \text{ bzw. } f(\cdot|\theta_1, \ldots, \theta_K)\ .$$

2. Ist W die Menge von Verteilungen aus der 2. Grundannahme („Verteilungsannahme“), so bezeichnet man W auch als **parametrische Verteilungsannahme**. Die Menge Θ heißt dann auch **Parameterraum**.

Bemerkungen

- Wir betrachten nur „identifizierbare“ parametrische Verteilungsfamilien, das heißt, unterschiedliche Parameter aus dem Parameterraum Θ müssen auch zu unterschiedlichen Verteilungen aus W führen.
- Die Bezeichnung θ dient lediglich zur vereinheitlichten Notation. In der Praxis behalten die Parameter meist ihre ursprüngliche Bezeichnung.
- In der Regel gehören alle Verteilungen in W zum gleichen Typ, zum Beispiel als
  - Bernoilliverteilung $B(1, p)$: Parameter $p \equiv \theta$, Parameterraum $\Theta = [0, 1]$
  - Poissonverteilung Pois(λ): Parameter $\lambda \equiv \theta$, Parameterraum $\Theta = \mathbb{R}_{++}$
  - Exponentialverteilung Exp(λ): Parameter $\lambda \equiv \theta$, Parameterraum $\Theta = \mathbb{R}_{++}$
  - Normalverteilung $N(\mu, \sigma^2)$: Parameter**vektor** $(\mu, \sigma^2) \equiv (\theta_1, \theta_2)$, Parameterraum $\mathbb{R} \times \mathbb{R}_{++}$

  (mit $\mathbb{R}_{++} := \{x \in \mathbb{R} \,|\, x > 0\}$).
- Suche nach **allgemein anwendbaren** Methoden zur Konstruktion von Schätzfunktionen für unbekannte Parameter θ aus parametrischen Verteilungsannahmen.
- Schätzfunktionen für einen Parameter(vektor) θ *sowie deren Realisationen (!)* werden üblicherweise mit $\widehat{\theta}$, gelegentlich auch mit $\widetilde{\theta}$ bezeichnet.
- Meist wird vom Vorliegen einer einfachen Stichprobe ausgegangen.

Methode der Momente (Momentenmethode)

- Im Taschengeldbeispiel: Schätzung des Erwartungswerts $\mathrm{E}(Y)$ *naheliegenderweise* durch das arithmetische Mittel $\overline{X} = \frac{1}{2}(X_1 + X_2)$.
- Dies entspricht der Schätzung des 1. (theoretischen) Moments von Y durch das 1. empirische Moment der Stichprobenrealisation (aufgefasst als Urliste im Sinne der deskriptiven Statistik).
- Gleichsetzen von theoretischen und empirischen Momenten bzw. Ersetzen theoretischer durch empirische Momente führt zur gebräuchlichen **(Schätz-)Methode der Momente** für die Parameter von parametrischen Verteilungsfamilien.
- Grundlegende Idee: Schätze Parameter der Verteilung so, dass zugehörige theoretische Momente $\mathrm{E}(Y)$, $\mathrm{E}(Y^2)$, ... mit den entsprechenden empirischen Momenten $\overline{X}$, $\overline{X^2}$, ... der Stichprobenzufallsvariablen $X_1, \ldots, X_n$ (bzw. deren Realisationen) übereinstimmen.
- Es werden dabei (beginnend mit dem ersten Moment) gerade so viele Momente einbezogen, dass das entstehende Gleichungssystem für die Parameter eine eindeutige Lösung hat.
  Bei eindimensionalen Parameterräumen genügt *i.d.R.* das erste Moment.

Momente von Zufallsvariablen

- Bereits aus „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“ bekannt ist die folgende Definition für die (theoretischen) Momente von Zufallsvariablen:

Definition 3.2 (k-te Momente)

Es seien Y eine (eindimensionale) Zufallsvariable, $k \in \mathbb{N}$.
Man bezeichnet den Erwartungswert $\mathrm{E}(Y^k)$ (falls er existiert) als das **(theoretische) Moment k-ter Ordnung** von Y, oder auch das **k-te (theoretische) Moment** von Y und schreibt auch kürzer

$$\mathrm{E}\, Y^k := \mathrm{E}(Y^k).$$

- Erinnerung (unter Auslassung der Existenzbetrachtung!):
  Das k-te Moment von Y berechnet man für diskrete bzw. stetige Zufallsvariablen Y durch

$$\mathrm{E}(Y^k) = \sum_{y_i} y_i^k \cdot p_Y(y_i) \qquad \text{bzw.} \qquad \mathrm{E}(Y^k) = \int_{-\infty}^{\infty} y^k \cdot f_Y(y)dy \ ,$$

  wobei y_i (im diskreten Fall) alle Trägerpunkte von Y durchläuft.

Empirische Momente von Stichproben

- Analog zu empirischen Momenten von Urlisten in der deskriptiven Statistik definiert man empirische Momente von Stichproben in der schließenden Statistik wie folgt:

Definition 3.3 (empirische Momente)

Ist $(X_1, \ldots, X_n)$ eine (einfache) Zufallsstichprobe zu einer Zufallsvariablen Y, so heißt

$$\overline{X^k} := \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

das **empirische k-te Moment**, oder auch das **Stichprobenmoment der Ordnung** k. Zu einer Realisation $(x_1, \ldots, x_n)$ von $(X_1, \ldots, X_n)$ bezeichnet

$$\overline{x^k} := \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

entsprechend die zugehörige **Realisation** des k-ten empirischen Moments.

Durchführung der Momentenmethode

- Zur Durchführung der Momentenmethode benötigte Anzahl von Momenten meist gleich der Anzahl der zu schätzenden Verteilungsparameter.
- Übliche Vorgehensweise:
  - Ausdrücken/Berechnen der theoretischen Momente in Abhängigkeit der Verteilungsparameter
  - Gleichsetzen der theoretischen Momente mit den entsprechenden empirischen Momenten und Auflösen der entstehenden Gleichungen nach den Verteilungsparametern.
- Alternativ, falls Verteilungsparameter Funktionen theoretischer Momente sind: Ersetzen der theoretischen Momente in diesen „Formeln“ für die Verteilungsparameter durch die entsprechenden empirischen Momente.
- Nützlich ist für die alternative Vorgehensweise gelegentlich der Varianzzerlegungssatz

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \ .$$

Beispiele (Momentenmethode) I

1. Schätzung des Parameters p einer Alternativ-/Bernoulliverteilung:
   - Verteilungsannahme: $W = \{B(1,p) \mid p \in \Theta = [0,1]\}$
   - Theoretisches 1. Moment: $\mathrm{E}(Y) = p$ (bekannt aus W'rechnung)
   - Gleichsetzen (hier besonders einfach!) von $\mathrm{E}(Y)$ mit 1. empirischen Moment $\overline{X}$ liefert sofort Momentenmethodenschätzer (Methode 1) $\widehat{p} = \overline{X}$.

   Der Schätzer $\widehat{p}$ für die Erfolgswahrscheinlichkeit p nach der Methode der Momente entspricht also gerade dem Anteil der Erfolge in der Stichprobe.
2. Schätzung des Parameters λ einer Exponentialverteilung:
   - Verteilungsannahme: $W = \{\mathrm{Exp}(\lambda) \mid \lambda \in \Theta = \mathbb{R}_{++}\}$
   - Theoretisches 1. Moment: $\mathrm{E}(Y) = \frac{1}{\lambda}$ (bekannt aus W'rechnung)
   - Gleichsetzen von $E(Y)$ mit 1. empirischen Moment $\overline{X}$ liefert (Methode 1)

$$\overline{X} \stackrel{!}{=} \mathrm{E}(Y) = \frac{1}{\lambda} \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{\overline{X}} \, .$$

(Vorsicht bei Berechnung der Realisation: $\frac{1}{\overline{x}} \neq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}$)

Beispiele (Momentenmethode) II

3. Schätzung der Parameter (μ, σ^2) einer Normalverteilung:
   - Verteilungsannahme: $W = \{N(\mu, \sigma^2) \,|\, (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_{++}\}$
     Hier bekannt: $E(Y) = \mu$ und $\mathrm{Var}(Y) = \sigma^2$.
     $\rightsquigarrow$ Alternative Methode bietet sich an (mit Varianzzerlegungssatz):
   - Verteilungsparameter $\mu = \mathrm{E}(Y)$
     Verteilungsparameter $\sigma^2 = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2$
   - Einsetzen der empirischen Momente anstelle der theoretischen Momente liefert $\widehat{\mu} = \overline{X}$ sowie $\widehat{\sigma^2} = \overline{X^2} - \overline{X}^2$ als Schätzer nach der Momentenmethode.
   - Am Beispiel der Realisation

     $$8.75, 10.37, 8.33, 13.19, 10.66, 8.36, 10.97, 11.48, 11.15, 9.39$$

     einer Stichprobe vom Umfang 10 erhält man mit

     $$\overline{x} = 10.265 \qquad \text{und} \qquad \overline{x^2} = 107.562$$

     die realisierten Schätzwerte

     $$\widehat{\mu} = 10.265 \qquad \text{und} \qquad \widehat{\sigma^2} = 107.562 - 10.265^2 = 2.192\ .$$

Maximum-Likelihood-Methode (ML-Methode)

- Weitere geläufige Schätzmethode: **Maximum-Likelihood-Methode**
- **Vor** Erläuterung der Methode: einleitendes Beispiel

Beispiel: ML-Methode durch Intuition (?)

Ein „fairer“ Würfel sei auf einer unbekannten Anzahl $r \in \{0, 1, 2, 3, 4, 5, 6\}$ von Seiten rot lackiert, auf den übrigen Seiten andersfarbig.
Der Würfel wird 100-mal geworfen und es wird festgestellt, wie oft eine rote Seite (oben) zu sehen war.

- Angenommen, es war 34-mal eine rote Seite zu sehen; wie würden Sie die Anzahl der rot lackierten Seiten auf dem Würfel schätzen?
- Angenommen, es war 99-mal eine rote Seite zu sehen; wie würden Sie nun die Anzahl der rot lackierten Seiten auf dem Würfel schätzen?

Welche Überlegungen haben Sie insbesondere zu dem zweiten Schätzwert geführt?

Erläuterung Beispiel I

- Bei der Bearbeitung des obigen Beispiels wendet man (zumindest im 2. Fall) vermutlich intuitiv die Maximum-Likelihood-Methode an!
- Prinzipielle Idee der Maximum-Likelihood-Methode:

  Wähle denjenigen der möglichen Parameter als Schätzung aus, bei dem die beobachtete Stichprobenrealisation am plausibelsten ist!
- Im Beispiel interessiert die (unbekannte) Anzahl der roten Seiten.
- Kenntnis der Anzahl der roten Seiten ist (Würfel ist „fair“!) gleichbedeutend mit der Kenntnis der Wahrscheinlichkeit, dass eine rote Seite oben liegt; offensichtlich ist diese Wahrscheinlichkeit nämlich $\frac{r}{6}$, wenn $r \in \{0, \ldots, 6\}$ die Anzahl der roten Seiten bezeichnet.
- Interessierender Umweltausschnitt kann also durch die Zufallsvariable Y beschrieben werden, die den Wert 1 annimmt, falls bei einem Würfelwurf eine rote Seite oben liegt, 0 sonst.
- Y ist dann offensichtlich $B(1, p)$-verteilt mit unbekanntem Parameter $p \in \{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1\}$, die 2. Grundannahme ist also erfüllt mit

$$W = \left\{ B(1,p) \,|\, p \in \left\{ 0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, 1 \right\} \right\} .$$

Erläuterung Beispiel II

- 100-maliges Werfen des Würfels und jeweiliges Notieren einer 1, falls eine rote Seite oben liegt, einer 0 sonst, führt offensichtlich zu einer Realisation $x_1, \ldots, x_n$ einer einfachen Stichprobe $X_1, \ldots, X_n$ vom Umfang $n = 100$ zu Y, denn $X_1, \ldots, X_n$ sind als Resultat wiederholter Würfelwürfe offensichtlich unabhängig identisch verteilt wie Y.
- Wiederum (vgl. Taschengeldbeispiel) ist es aber nützlich, sich schon *vorher* Gedanken über die Verteilung der Anzahl der (insgesamt geworfenen) Würfe mit obenliegender roten Seite zu machen!
- Aus Veranstaltung „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“ bekannt: Für die Zufallsvariable Z, die die Anzahl der roten Seiten bei 100-maligem Werfen beschreibt, also für

$$Z = \sum_{i=1}^{100} X_i = X_1 + \ldots + X_{100} \ ,$$

  gilt $Z \sim B(100, p)$, falls $Y \sim B(1, p)$.
- Ziel: Aus Stichprobe $X_1, \ldots, X_{100}$ bzw. der Realisation $x_1, \ldots, x_{100}$ (über die Stichprobenfunktion Z bzw. deren Realisation $z = x_1 + \ldots + x_{100}$) auf unbekannten Parameter p und damit die Anzahl der roten Seiten r schließen.

Erläuterung Beispiel III

- Im Beispiel: Umsetzung der ML-Methode besonders einfach, da Menge W der möglichen Verteilungen (aus Verteilungsannahme) **endlich**.
- „Plausibilität“ einer Stichprobenrealisation kann hier direkt anhand der Eintrittswahrscheinlichkeit der Realisation gemessen und für alle möglichen Parameter p bestimmt werden.
- Wahrscheinlichkeit (abhängig von p), dass Z Wert z annimmt:

$$P\{Z = z|p\} = \binom{100}{z} \cdot p^z \cdot (1-p)^{100-z}$$

- Für die erste Realisation $z = 34$ von Z:

r	0	1	2	3	4	5	6
p	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
$P\{Z = 34\|p\}$	0	$1.2 \cdot 10^{-5}$	$8.31 \cdot 10^{-2}$	$4.58 \cdot 10^{-4}$	$1.94 \cdot 10^{-11}$	$5.17 \cdot 10^{-28}$	0

- Für die zweite Realisation $z = 99$ von Z:

r	0	1	2	3	4	5	6
p	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
$P\{Z = 99\|p\}$	0	$7.65 \cdot 10^{-76}$	$3.88 \cdot 10^{-46}$	$7.89 \cdot 10^{-29}$	$1.23 \cdot 10^{-16}$	$2.41 \cdot 10^{-7}$	0

Bemerkungen zum Beispiel

- Die angegebenen Wahrscheinlichkeiten für Z fassen jeweils mehrere mögliche Stichprobenrealisationen zusammen (da für den Wert von Z irrelevant ist, *welche* der Stichprobenzufallsvariablen X_i den Wert 0 bzw. 1 angenommen haben), für die ML-Schätzung ist aber eigentlich die Wahrscheinlichkeit einer einzelnen Stichprobenrealisation maßgeblich. Die Wahrscheinlichkeit einer einzelnen Stichprobenrealisation erhält man, indem der Faktor $\binom{100}{z}$ entfernt wird; dieser ist jedoch in jeder der beiden Tabellen konstant und beeinflusst daher die Bestimmung des Maximums nicht.
- Eher untypisch am Beispiel (aber umso geeigneter zur Erklärung der Methode!) ist die Tatsache, dass W eine endliche Menge von Verteilungen ist. In der Praxis wird man in der Regel unendlich viele Möglichkeiten für die Wahl des Parameters haben, z.B. bei Alternativverteilungen $p \in [0, 1]$. Dies ändert zwar *nichts* am Prinzip der Schätzung, wohl aber an den zur Bestimmung der „maximalen Plausibilität“ nötigen (mathematischen) Techniken.
- Dass die „Plausibilität“ hier genauer einer Wahrscheinlichkeit entspricht, hängt an der diskreten Verteilung von Y. Ist Y eine stetige Zufallsvariable, übernehmen Dichtefunktionswerte die Messung der „Plausibilität“.

Maximum-Likelihood-Methode (im Detail)

Schritte zur ML-Schätzung

Die Durchführung einer ML-Schätzung besteht aus folgenden Schritten:

1. Aufstellung der sog. **Likelihood-Funktion** $L(\theta)$, die *in Abhängigkeit des (unbekannten) Parametervektors* θ die Plausibilität der beobachteten Stichprobenrealisation misst.
2. Suche des (eines) Parameters bzw. Parametervektors $\widehat{\theta}$, der den (zu der beobachteten Stichprobenrealisation) maximal möglichen Wert der Likelihoodfunktion liefert.
   Es ist also *jeder* Parameter(vektor) $\widehat{\theta}$ ein ML-Schätzer, für den gilt:

$$L(\widehat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

- Je nach Anwendungssituation unterscheidet sich die Vorgehensweise in beiden Schritten erheblich.
- Wir setzen bei der Durchführung von ML-Schätzungen **stets** voraus, dass eine **einfache (Zufalls-)Stichprobe** vorliegt!

1. Schritt: Aufstellen der Likelihoodfunktion

- „Plausibilität“ oder „Likelihood“ der Stichprobenrealisation wird gemessen
  - mit Hilfe der **Wahrscheinlichkeit**, die Stichprobenrealisation $(x_1, \ldots, x_n)$ zu erhalten, d.h. dem Wahrscheinlichkeitsfunktionswert

    $$L(\theta) := p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \,|\, \theta) \ ,$$

    falls Y diskrete Zufallsvariable ist,
  - mit Hilfe der **gemeinsamen Dichtefunktion** ausgewertet an der Stichprobenrealisation $(x_1, \ldots, x_n)$,

    $$L(\theta) := f_{X_1,\ldots,X_n}(x_1, \ldots, x_n \,|\, \theta) \ ,$$

    falls Y stetige Zufallsvariable ist.
- Bei Vorliegen einer einfachen Stichprobe lässt sich die Likelihoodfunktion für diskrete Zufallsvariablen Y **immer** darstellen als

$$\begin{aligned} L(\theta) &= p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \,|\, \theta) \\ &\overset{X_i \text{ unabhängig}}{=} \prod_{i=1}^{n} p_{X_i}(x_i|\theta) \\ &\overset{X_i \text{ verteilt wie } Y}{=} \prod_{i=1}^{n} p_Y(x_i|\theta) \ . \end{aligned}$$

- Analog erhält man bei Vorliegen einer einfachen Stichprobe für stetige Zufallsvariablen Y **immer** die Darstellung

$$\begin{aligned} L(\theta) &= f_{X_1,\ldots,X_n}(x_1,\ldots,x_n \mid \theta) \\ &\overset{X_i \text{ unabhängig}}{=} \prod_{i=1}^{n} f_{X_i}(x_i|\theta) \\ &\overset{X_i \text{ verteilt wie } Y}{=} \prod_{i=1}^{n} f_Y(x_i|\theta)\ . \end{aligned}$$

  für die Likelihoodfunktion.
- Ist der Parameterraum Θ endlich, kann im Prinzip $L(\theta)$ für alle $\theta \in \Theta$ berechnet werden und eines der θ als ML-Schätzwert $\widehat{\theta}$ gewählt werden, für das $L(\theta)$ maximal war.
  Für diese (einfache) Situation wird Schritt 2 nicht weiter konkretisiert.
- Ist der Parameterraum Θ ein Kontinuum (z.B. ein Intervall in $\mathbb{R}^K$), müssen für den 2. Schritt i.d.R. Maximierungsverfahren aus der Analysis angewendet werden.

2. Schritt: Maximieren der Likelihoodfunktion

(falls Θ ein Intervall in $\mathbb{R}^K$ ist)

- Wichtige Eigenschaft des Maximierungsproblems aus Schritt 2:

  Wichtig ist nicht der **Wert** des Maximums $L(\widehat{\theta})$ der Likelihoodfunktion, sondern die **Stelle** $\widehat{\theta}$, an der dieser Wert angenommen wird!
- Aus Gründen (zum Teil ganz erheblich) vereinfachter Berechnung:
  - Bilden der **logarithmierten** Likelihoodfunktion (Log-Likelihoodfunktion) $\ln L(\theta)$.
  - Maximieren der Log-Likelihoodfunktion $\ln L(\theta)$ **statt** Maximierung der Likelihoodfunktion.
- Diese Änderung des Verfahrens ändert nichts an den Ergebnissen, denn
  - $\ln : \mathbb{R}_{++} \to \mathbb{R}$ ist eine streng monoton wachsende Abbildung,
  - es genügt, die Likelihoodfunktion in den Bereichen zu untersuchen, in denen sie *positive* Werte annimmt, da nur dort das Maximum angenommen werden kann. Dort ist auch die log-Likelihoodfunktion definiert.

- Maximierung von $\ln L(\theta)$ kann *oft* (aber nicht immer!) auf die aus der Mathematik bekannte Art und Weise erfolgen:
  1. Bilden der ersten Ableitung $\frac{\partial \ln L}{\partial \theta}$ der log-Likelihoodfunktion.
     (Bei mehrdimensionalen Parametervektoren: Bilden der partiellen Ableitungen
     $$\frac{\partial \ln L}{\partial \theta_1}, \ldots, \frac{\partial \ln L}{\partial \theta_K}$$
     der log-Likelihoodfunktion.)
  2. Nullsetzen der ersten Ableitung, um „Kandidaten“ für Maximumsstellen von $\ln L(\theta)$ zu finden:
     $$\frac{\partial \ln L}{\partial \theta} \stackrel{!}{=} 0 \qquad \rightsquigarrow \qquad \widehat{\theta}$$
     (Bei mehrdimensionalen Parametervektoren: Lösen des Gleichungssystems
     $$\frac{\partial \ln L}{\partial \theta_1} \stackrel{!}{=} 0, \qquad \ldots \qquad , \frac{\partial \ln L}{\partial \theta_K} \stackrel{!}{=} 0$$
     um „Kandidaten“ $\widehat{\theta}$ für Maximumsstellen von $\ln L(\theta)$ zu finden.)
  3. Überprüfung anhand des Vorzeichens der 2. Ableitung $\frac{\partial^2 \ln L}{(\partial \theta)^2}$ (bzw. der Definitheit der Hessematrix), ob tatsächlich eine Maximumsstelle vorliegt:
     $$\frac{\partial^2 \ln L}{(\partial \theta)^2}(\widehat{\theta}) \stackrel{?}{<} 0$$

- Auf die Überprüfung der 2. Ableitung bzw. der Hessematrix verzichten wir häufig, um nicht durch mathematische Schwierigkeiten von den statistischen abzulenken.
- Durch den Übergang von der Likelihoodfunktion zur log-Likelihoodfunktion erhält man gegenüber den Darstellungen aus Folie 40 und 41 im diskreten Fall nun

$$\ln L(\theta) = \ln \left(\prod_{i=1}^{n} p_Y(x_i|\theta) \right) = \sum_{i=1}^{n} \ln \left(p_Y(x_i|\theta) \right)$$

  und im stetigen Fall

$$\ln L(\theta) = \ln \left(\prod_{i=1}^{n} f_Y(x_i|\theta) \right) = \sum_{i=1}^{n} \ln \left(f_Y(x_i|\theta) \right) \ .$$

- Die wesentliche Vereinfachung beim Übergang zur log-Likelihoodfunktion ergibt sich meist dadurch, dass die Summen in den obigen Darstellungen deutlich leichter abzuleiten sind als die Produkte in den Darstellungen der Likelihoodfunktion auf Folie 40 und Folie 41.
- Falls „Standardverfahren“ keine Maximumsstelle liefert $\rightsquigarrow$ „Gehirn einschalten“

Beispiel: ML-Schätzung für Exponentialverteilung

Erinnerung: $f_Y(y|\lambda) = \lambda e^{-\lambda y}$ für $y > 0$, $\lambda > 0$

1. Aufstellen der Likelihoodfunktion (im Fall $x_i > 0$ für alle i):

$$L(\lambda) = \prod_{i=1}^{n} f_Y(x_i|\lambda) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda x_i}\right)$$

2. Aufstellen der log-Likelihoodfunktion (im Fall $x_i > 0$ für alle i):

$$\ln L(\lambda) = \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda x_i}\right) = \sum_{i=1}^{n} (\ln\lambda + (-\lambda x_i)) = n \cdot \ln\lambda - \lambda \cdot \sum_{i=1}^{n} x_i$$

3. Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \stackrel{!}{=} 0$$

liefert

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}}$$

als ML-Schätzer (2. Ableitung $\frac{\partial^2 \ln L}{(\partial\lambda)^2} = -\frac{n}{\lambda^2} < 0$).

Bemerkungen

- Häufiger wird die Abhängigkeit der Likelihoodfunktion von der Stichprobenrealisation auch durch Schreibweisen der Art $L(\theta; x_1, \ldots, x_n)$ oder $L(x_1, \ldots, x_n|\theta)$ ausgedrückt.
- Vorsicht geboten, falls Bereich positiver Dichte bzw. der Träger der Verteilung von Y von Parametern abhängt!
  Im Beispiel: Bereich positiver Dichte $\mathbb{R}_{++}$ *unabhängig* vom Verteilungsparameter λ, Maximierungsproblem unter Vernachlässigung des Falls *„mindestens ein x_i kleiner oder gleich 0“* betrachtet, da dieser Fall **für keinen der möglichen Parameter** mit positiver Wahrscheinlichkeit eintritt.
  Dieses „Vernachlässigen“ ist nicht immer unschädlich!
- Bei diskreten Zufallsvariablen mit „wenig“ verschiedenen Ausprägungen oft Angabe der absoluten Häufigkeiten für die einzelnen Ausprägungen in der Stichprobe statt Angabe der Stichprobenrealisation $x_1, \ldots, x_n$ selbst.
  Beispiel: Bei Stichprobe vom Umfang 25 zu alternativverteilter Zufallsvariablen Y häufiger Angabe von „18 Erfolge in der Stichprobe der Länge 25“ als Angabe der Stichprobenrealisation

$$0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1 \ .$$

Beispiel: ML-Schätzung für Alternativverteilungen I

- Verteilungsannahme: $Y \sim B(1,p)$ für $p \in \Theta = [0,1]$ mit

$$p_Y(y|p) = \left\{ \begin{array}{cc} p & \text{falls } y = 1 \\ 1-p & \text{falls } y = 0 \end{array} \right\} = p^y \cdot (1-p)^{1-y} \text{ für } y \in \{0,1\} \ .$$

1. Aufstellen der Likelihoodfunktion:

$$L(p) = \prod_{i=1}^{n} p_Y(x_i|p) = \prod_{i=1}^{n} \left(p^{x_i} \cdot (1-p)^{1-x_i} \right) = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n-\sum_{i=1}^n x_i}$$

   bzw. — wenn $n_1 := \sum_{i=1}^n x_i$ die Anzahl der „Einsen“ (Erfolge) in der Stichprobe angibt —

$$L(p) = p^{n_1} \cdot (1-p)^{n-n_1}$$

2. Aufstellen der log-Likelihoodfunktion:

$$\ln L(p) = n_1 \ln(p) + (n-n_1) \ln(1-p)$$

Beispiel: ML-Schätzung für Alternativverteilungen II

3. Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\begin{aligned} & \frac{\partial \ln L}{\partial p} &=& \frac{n_1}{p} - \frac{n - n_1}{1-p} \stackrel{!}{=} 0 \\ \Leftrightarrow \quad & n_1 - n_1 p &=& np - n_1 p \\ \Rightarrow \quad & \widehat{p} &=& \frac{n_1}{n} \end{aligned}$$

Die 2. Ableitung $\frac{\partial^2 \ln L}{(\partial p)^2} = -\frac{n_1}{p^2} - \frac{n-n_1}{(1-p)^2}$ ist negativ für $0 < p < 1$, der Anteil der Erfolge in der Stichprobe $\widehat{p} = n_1/n$ ist also der ML-Schätzer.

Bemerkungen:

- Es wird die Konvention $0^0 := 1$ verwendet.
- Die Bestimmung des ML-Schätzers in Schritt 3 ist so nur für $n_1 \neq 0$ und $n_1 \neq n$ korrekt.
- Für $n_1 = 0$ und $n_1 = n$ ist die (log-) Likelihoodfunktion jeweils streng monoton, die ML-Schätzer sind also Randlösungen (später mehr!).
- Für $n_1 = 0$ gilt jedoch $\widehat{p} = 0 = \frac{0}{n}$, für $n_1 = n$ außerdem $\widehat{p} = 1 = \frac{n}{n}$, die Formel aus Schritt 3 bleibt also gültig!

Beispiel: ML-Schätzung für Poissonverteilungen I

- Verteilungsannahme: $Y \sim \text{Pois}(\lambda)$ für $\lambda \in \Theta = \mathbb{R}_{++}$ mit

$$p_Y(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

für $k \in \mathbb{N}_0$.

1. Aufstellen der Likelihoodfunktion:

$$L(\lambda) = \prod_{i=1}^{n} p_Y(x_i|\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right)$$

(falls alle $x_i \in \mathbb{N}_0$)

2. Aufstellen der log-Likelihoodfunktion:

$$\ln L(\lambda) = \sum_{i=1}^{n} (x_i \ln(\lambda) - \ln(x_i!) - \lambda) = \left(\sum_{i=1}^{n} x_i \right) \ln(\lambda) - \left(\sum_{i=1}^{n} \ln(x_i!) \right) - n\lambda$$

Beispiel: ML-Schätzung für Poissonverteilungen II

3. Ableiten und Nullsetzen der log-Likelihoodfunktion:

$$\begin{aligned} \frac{\partial \ln L}{\partial \lambda} &= \frac{\sum_{i=1}^{n} x_i}{\lambda} - n \stackrel{!}{=} 0 \\ \Rightarrow \quad \widehat{\lambda} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \end{aligned}$$

mit $\frac{\partial^2 \ln L}{(\partial \lambda)^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2} < 0$ für alle $\lambda > 0$, $\widehat{\lambda} = \overline{x}$ ist also der ML-Schätzer für λ.

- Aus Wahrscheinlichkeitsrechnung bekannt: $Y \sim \text{Pois}(\lambda) \Rightarrow E(Y) = \lambda$, also ergibt sich (hier) auch für den Schätzer nach der Momentenmethode offensichtlich $\widehat{\lambda} = \overline{X}$.
- Wird (ähnlich zur Anzahl n_1 der Erfolge in einer Stichprobe zu einer alternativverteilten Grundgesamtheit) statt der (expliziten) Stichprobenrealisation $x_1, \ldots, x_n$ eine „Häufigkeitsverteilung“ der in der Stichprobe aufgetretenen Werte angegeben, kann $\overline{x}$ mit der aus der deskriptiven Statistik bekannten „Formel“ ausgerechnet werden.

Beispiel: ML-Schätzung bei diskreter Gleichverteilung

- Verteilungsannahme: für ein (unbekanntes) $M \in \mathbb{N}$ nimmt Y die Werte $\{1, \ldots, M\}$ mit der gleichen Wahrscheinlichkeit von jeweils $1/M$ an, d.h.:

$$p_Y(k|M) = \begin{cases} \frac{1}{M} & \text{falls } k \in \{1, \ldots, M\} \\ 0 & \text{falls } k \notin \{1, \ldots, M\} \end{cases}$$

1. Aufstellen der Likelihoodfunktion:

$$L(M) = \prod_{i=1}^{n} p_Y(x_i|M) = \begin{cases} \frac{1}{M^n} & \text{falls } x_i \in \{1, \ldots, M\} \text{ für alle } i \\ 0 & \text{falls } x_i \notin \{1, \ldots, M\} \text{ für mindestens ein } i \end{cases}$$

$$= \begin{cases} \frac{1}{M^n} & \text{falls } \max\{x_1, \ldots, x_n\} \leq M \\ 0 & \text{falls } \max\{x_1, \ldots, x_n\} > M \end{cases} \quad \text{(gegeben } x_i \in \mathbb{N} \text{ für alle } i\text{)}$$

2. Maximieren der Likelihoodfunktion:

   Offensichtlich ist $L(M)$ für $\max\{x_1, \ldots, x_n\} \leq M$ streng monoton fallend in M, M muss also **unter Einhaltung der Bedingung** $\max\{x_1, \ldots, x_n\} \leq M$ möglichst klein gewählt werden. Damit erhält man den ML-Schätzer als $\widehat{M} = \max\{x_1, \ldots, x_n\}$.

Beurteilung von Schätzfunktionen

- *Bisher:* Zwei Methoden zur Konstruktion von Schätzfunktionen bekannt.
- *Problem:*

  Wie kann Güte/Qualität dieser Methoden bzw. der resultierenden Schätzfunktionen beurteilt werden?

- *Lösung:*

  Zu gegebener Schätzfunktion $\widehat{\theta}$ für θ: Untersuchung des **zufälligen** Schätzfehlers $\widehat{\theta} - \theta$ (bzw. dessen Verteilung)

- Naheliegende Forderung für „gute“ Schätzfunktionen:

  Verteilung des Schätzfehler sollte möglichst „dicht“ um 0 konzentriert sein (d.h. Verteilung von $\widehat{\theta}$ sollte möglichst „dicht“ um θ konzentriert sein)

- Aber:
  - Was bedeutet das?
  - Wie vergleicht man zwei Schätzfunktionen $\widehat{\theta}$ und $\widetilde{\theta}$? Wann ist Schätzfunktion $\widehat{\theta}$ „besser“ als $\widetilde{\theta}$ (und was bedeutet „besser“)?
  - Was ist zu beachten, wenn Verteilung des Schätz**fehlers** noch vom zu schätzenden Parameter abhängt?

Bias, Erwartungstreue

- Eine offensichtlich gute Eigenschaft von Schätzfunktionen ist, wenn der zu schätzende (wahre) Parameter zumindest *im Mittel* getroffen wird, d.h. der *erwartete* Schätzfehler gleich Null ist:

Definition 3.4 (Bias, Erwartungstreue)

Seien W eine parametrische Verteilungsannahme mit Parameterraum Θ, $\widehat{\theta}$ eine Schätzfunktion für θ. Dann heißt

1. der erwartete Schätzfehler

$$\text{Bias}(\widehat{\theta}) := \text{E}(\widehat{\theta} - \theta) = \text{E}(\widehat{\theta}) - \theta$$

   die **Verzerrung** oder der **Bias** von $\widehat{\theta}$,

2. die Schätzfunktion $\widehat{\theta}$ **erwartungstreu für** θ oder auch **unverzerrt für** θ, falls $\text{Bias}(\widehat{\theta}) = 0$ bzw. $\text{E}(\widehat{\theta}) = \theta$ für alle $\theta \in \Theta$ gilt.

3. Ist allgemeiner $g : \Theta \to \mathbb{R}$ eine (messbare) Abbildung, so betrachtet man auch Schätzfunktionen $\widehat{g(\theta)}$ für $g(\theta)$ und nennt diese **erwartungstreu für** $g(\theta)$, wenn $\text{E}(\widehat{g(\theta)} - g(\theta)) = 0$ bzw. $\text{E}(\widehat{g(\theta)}) = g(\theta)$ für alle $\theta \in \Theta$ gilt.

Bemerkungen

- Obwohl Definition 3.4 auch für mehrdimensionale Parameterräume Θ geeignet ist („0“ entspricht dann ggf. dem Nullvektor), betrachten wir zur Vereinfachung im Folgenden meist nur noch **eindimensionale** Parameterräume $\Theta \subseteq \mathbb{R}$.
- Ist beispielsweise W als Verteilungsannahme für Y die Menge aller Alternativverteilungen $B(1,p)$ mit Parameter $p \in \Theta = [0,1]$, so ist der ML-Schätzer $\widehat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ bei Vorliegen einer Zufallsstichprobe $X_1, \ldots, X_n$ zu Y erwartungstreu für p, denn es gilt:

$$\begin{aligned} \mathrm{E}(\widehat{p}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &\overset{\text{E linear}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) \\ &\overset{F_{X_i}=F_Y}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(Y) \\ &= \frac{1}{n} \cdot n \cdot p = p \text{ für alle } p \in [0,1] \end{aligned}$$

- Allgemeiner gilt, dass $\overline{X}$ bei Vorliegen einer Zufallsstichprobe stets erwartungstreu für $\mathrm{E}(Y)$ ist, denn es gilt analog zu oben:

$$\begin{aligned}\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &\overset{\text{E linear}}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i)\\ &\overset{F_{X_i}=F_Y}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y)\\ &= \frac{1}{n}\cdot n\cdot \mathrm{E}(Y) = \mathrm{E}(Y)\end{aligned}$$

- Genauso ist klar, dass man für beliebiges k mit dem k-ten empirischen Moment $\overline{X^k}$ bei Vorliegen einer Zufallsstichprobe stets erwartungstreue Schätzer für das k-te theoretische Moment $\mathrm{E}(Y^k)$ erhält, denn es gilt:

$$\mathrm{E}(\overline{X^k}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i^k) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Y^k) = \mathrm{E}(Y^k)$$

- Der nach der Methode der Momente erhaltene Schätzer

$$\widehat{\sigma^2} = \overline{X^2} - \overline{X}^2 \quad \overset{\text{Verschiebungssatz}}{=} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

für den Parameter σ^2 einer normalverteilten Zufallsvariable ist **nicht** erwartungstreu für σ^2.
Bezeichnet $\sigma^2 := \mathrm{Var}(Y)$ nämlich die (unbekannte) Varianz der Zufallsvariablen Y, so kann gezeigt werden, dass für $\widehat{\sigma^2}$ generell

$$\mathrm{E}(\widehat{\sigma^2}) = \frac{n-1}{n}\sigma^2$$

gilt. Einen erwartungstreuen Schätzer für σ^2 erhält man folglich mit der sogenannten **Stichprobenvarianz**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}\widehat{\sigma^2} \ ,$$

denn es gilt offensichtlich

$$\mathrm{E}(S^2) = \mathrm{E}\left(\frac{n}{n-1}\widehat{\sigma^2}\right) = \frac{n}{n-1}\,\mathrm{E}\left(\widehat{\sigma^2}\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2 \ .$$

Vergleich von Schätzfunktionen

- Beim Vergleich von Schätzfunktionen: **oft** Beschränkung auf erwartungstreue Schätzfunktionen
- In der Regel: viele erwartungstreue Schätzfunktionen denkbar.
- Für die Schätzung von $\mu := E(Y)$ beispielsweise alle *gewichteten* Mittel

$$\widehat{\mu}_{w_1,\ldots,w_n} := \sum_{i=1}^{n} w_i \cdot X_i$$

mit der Eigenschaft $\sum_{i=1}^{n} w_i = 1$ erwartungstreu für μ, denn es gilt dann offensichtlich

$$\mathrm{E}\left(\widehat{\mu}_{w_1,\ldots,w_n}\right) = \mathrm{E}\left(\sum_{i=1}^{n} w_i \cdot X_i\right) = \sum_{i=1}^{n} w_i \,\mathrm{E}(X_i) = \mathrm{E}(Y) \cdot \sum_{i=1}^{n} w_i = \mathrm{E}(Y) = \mu \ .$$

- Problem: Welche Schätzfunktion ist „die beste“?
- Übliche Auswahl (bei Beschränkung auf erwartungstreue Schätzfunktionen!): Schätzfunktionen mit geringerer **Streuung (Varianz)** bevorzugen.

Wirksamkeit, Effizienz

Definition 3.5 (Wirksamkeit, Effizienz)

Sei W eine parametrische Verteilungsannahme mit Parameterraum Θ.

1. Seien $\widehat{\theta}$ und $\widetilde{\theta}$ erwartungstreue Schätzfunktionen für θ. Dann heißt $\widehat{\theta}$ **mindestens so wirksam** wie $\widetilde{\theta}$, wenn

$$\text{Var}(\widehat{\theta}) \leq \text{Var}(\widetilde{\theta}) \text{ für alle } \theta \in \Theta$$

   gilt. $\widehat{\theta}$ heißt **wirksamer** als $\widetilde{\theta}$, wenn *außerdem* $\text{Var}(\widehat{\theta}) < \text{Var}(\widetilde{\theta})$ für mindestens ein $\theta \in \Theta$ gilt.
2. Ist $\widehat{\theta}$ mindestens so wirksam wie alle (anderen) Schätzfunktionen einer Klasse mit erwartungstreuen Schätzfunktionen für θ, so nennt man $\widehat{\theta}$ **effizient** in dieser Klasse erwartungstreuer Schätzfunktionen.

- Die Begriffe „Wirksamkeit“ und „Effizienz“ betrachtet man analog zu Definition 3.5 ebenfalls, wenn Funktionen $g(\theta)$ von θ geschätzt werden.
- $\text{Sd}(\widehat{\theta}) = \sqrt{\text{Var}(\widehat{\theta})}$ wird auch **Standardfehler** oder **Stichprobenfehler** von $\widehat{\theta}$ genannt.

Beispiel: Effizienz

- Betrachte Klasse der (linearen) erwartungstreuen Schätzfunktionen

$$\widehat{\mu}_{w_1,\ldots,w_n} := \sum_{i=1}^{n} w_i \cdot X_i$$

  mit $\sum_{i=1}^{n} w_i = 1$ für den Erwartungswert $\mu := E(Y)$ aus Folie 57.
- Für welche $w_1, \ldots, w_n$ erhält man (bei Vorliegen einer einfachen Stichprobe) die in dieser Klasse **effiziente** Schätzfunktion $\widehat{\mu}_{w_1,\ldots,w_n}$?

⇝ Suche nach den Gewichten $w_1, \ldots, w_n$ (mit $\sum_{i=1}^{n} w_i = 1$), für die $\mathrm{Var}(\widehat{\mu}_{w_1,\ldots,w_n})$ möglichst klein wird.

- Man kann zeigen, dass $\mathrm{Var}(\widehat{\mu}_{w_1,\ldots,w_n})$ minimal wird, wenn

$$w_i = \frac{1}{n} \text{ für alle } i \in \{1, \ldots, n\}$$

  gewählt wird.
- Damit ist $\overline{X}$ also effizient in der Klasse der linearen erwartungstreuen Schätzfunktionen für den Erwartungswert μ einer Verteilung!

Mittlerer quadratischer Fehler (MSE)

- Wenn Erwartungstreue im Vordergrund steht, ist Auswahl nach minimaler Varianz der Schätzfunktion sinnvoll.
- Ist Erwartungstreue nicht das „übergeordnete“ Ziel, verwendet man zur Beurteilung der Qualität von Schätzfunktionen häufig auch den sogenannten mittleren quadratischen Fehler (mean square error, MSE).

Definition 3.6 (Mittlerer quadratischer Fehler (MSE))

Sei W eine parametrische Verteilungsannahme mit Parameterraum Θ, $\widehat{\theta}$ eine Schätzfunktion für $\theta \in \Theta$. Dann heißt $\text{MSE}(\widehat{\theta}) := \text{E}\left[(\widehat{\theta} - \theta)^2\right]$ der **mittlere quadratische Fehler (mean square error, MSE)** von $\widehat{\theta}$.

- Mit dem (umgestellten) Varianzzerlegungssatz erhält man direkt

$$\text{E}\left[(\widehat{\theta} - \theta)^2\right] = \underbrace{\text{Var}(\widehat{\theta} - \theta)}_{=\text{Var}(\widehat{\theta})} + \underbrace{\left[\text{E}(\widehat{\theta} - \theta)\right]^2}_{=(\text{Bias}(\widehat{\theta}))^2},$$

  für erwartungstreue Schätzfunktionen stimmt der MSE einer Schätzfunktion also gerade mit der Varianz überein!

Konsistenz im quadratischen Mittel

- Basierend auf dem MSE ist ein „minimales“ Qualitätskriterium für Schätzfunktionen etabliert.
- Das Kriterium fordert (im Prinzip), dass man den MSE durch Vergrößerung des Stichprobenumfangs beliebig klein bekommen muss.
- Zur Formulierung des Kriteriums müssen Schätzfunktionen $\widehat{\theta}_n$ für „variable“ Stichprobengrößen $n \in \mathbb{N}$ betrachtet werden.

Definition 3.7 (Konsistenz im quadratischen Mittel)

Seien W eine parametrische Verteilungsannahme mit Parameterraum Θ, $\widehat{\theta}_n$ eine Schätzfunktion für $\theta \in \Theta$ zum Stichprobenumfang $n \in \mathbb{N}$.
Dann heißt die (Familie von) Schätzfunktion(en) $\widehat{\theta}_n$ **konsistent im quadratischen Mittel für** θ, falls

$$\lim_{n \to \infty} \mathrm{MSE}(\widehat{\theta}_n) = \lim_{n \to \infty} \mathrm{E}\left[(\widehat{\theta}_n - \theta)^2\right] = 0$$

für alle $\theta \in \Theta$ gilt.

- Mit der (additiven) Zerlegung des MSE in Varianz und quadrierten Bias aus Folie 60 erhält man sofort:

Satz 3.8

Seien W eine parametrische Verteilungsannahme mit Parameterraum Θ, $\widehat{\theta}_n$ eine Schätzfunktion für $\theta \in \Theta$ zum Stichprobenumfang $n \in \mathbb{N}$. Dann ist die Familie $\widehat{\theta}_n$ von Schätzfunktionen genau dann konsistent im quadratischen Mittel für θ, wenn sowohl

1. $\lim\limits_{n\to\infty} \mathrm{E}(\widehat{\theta}_n - \theta) = 0$ *bzw.* $\lim\limits_{n\to\infty} \mathrm{E}(\widehat{\theta}_n) = \theta$ *als auch*
2. $\lim\limits_{n\to\infty} \mathrm{Var}(\widehat{\theta}_n) = 0$

für alle $\theta \in \Theta$ gilt.

- Eigenschaft 1 aus Satz 3.8 wird auch **asymptotische Erwartungstreue** genannt; asymptotische Erwartungstreue ist offensichtlich schwächer als Erwartungstreue.
- Es gibt also auch (Familien von) Schätzfunktionen, die für einen Parameter θ zwar konsistent im quadratischen Mittel sind, aber nicht erwartungstreu.

Beispiel: Konsistenz im quadratischen Mittel

- Voraussetzung (wie üblich): $X_1, \ldots, X_n$ einfache Stichprobe zu Y.
- Bekannt: Ist $\mu := \mathrm{E}(Y)$ der unbekannte Erwartungswert der interessierenden Zufallsvariable Y, so ist $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ für alle $n \in \mathbb{N}$ erwartungstreu.
- Ist $\sigma^2 := \mathrm{Var}(Y)$ die Varianz von Y, so erhält man für die Varianz von $\overline{X}_n$ (vgl. Beweis der Effizienz von $\overline{X}$ unter allen linearen erwartungstreuen Schätzfunktionen für μ):

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}(X_i)}_{=\sigma^2} = \frac{\sigma^2}{n}$$

- Es gilt also $\lim_{n\to\infty} \mathrm{Var}(\overline{X}_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$, damit folgt zusammen mit der Erwartungstreue, dass $\overline{X}_n$ konsistent im quadratischen Mittel für μ ist.

Verteilung des Stichprobenmittels $\overline{X}$

- **Bisher:** Interesse meist an einigen *Momenten* (Erwartungswert und Varianz) von Schätzfunktionen, insbesondere des Stichprobenmittels $\overline{X}$.
- Bereits bekannt: Ist $\mu := E(Y)$, $\sigma^2 := \mathrm{Var}(Y)$ und $X_1, \ldots, X_n$ eine einfache Stichprobe zu Y, so gilt

$$\mathrm{E}(\overline{X}) = \mu \qquad \text{sowie} \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \ .$$

- Damit Aussagen über Erwartungstreue, Wirksamkeit, Konsistenz möglich.
- **Jetzt:** Interesse an ganzer **Verteilung** von Schätzfunktionen, insbesondere $\overline{X}$.
- Verteilungsaussagen entweder
  - auf Grundlage des Verteilungstyps von Y aus der Verteilungsannahme in speziellen Situationen **exakt** möglich oder
  - auf Grundlage des zentralen Grenzwertsatzes (bei genügend großem Stichprobenumfang!) allgemeiner **näherungsweise (approximativ)** möglich.
- Wir unterscheiden im Folgenden nur zwischen:
  - Y normalverteilt $\rightsquigarrow$ Verwendung der exakten Verteilung von $\overline{X}$.
  - Y nicht normalverteilt $\rightsquigarrow$ Verwendung der Näherung der Verteilung von $\overline{X}$ aus dem zentralen Grenzwertsatz.

Aus „Deskriptive Statistik und Wahrscheinlichkeitsrechnung“:

1. Gilt $Y \sim N(\mu, \sigma^2)$, so ist $\overline{X}$ **exakt** normalverteilt mit Erwartungswert μ und Varianz $\frac{\sigma^2}{n}$, es gilt also
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \ .$$
2. Ist Y beliebig verteilt mit $\mathrm{E}(Y) =: \mu$ und $\mathrm{Var}(Y) =: \sigma^2$, so rechtfertigt der zentrale Grenzwertsatz **für ausreichend große Stichprobenumfänge** n die Näherung der tatsächlichen Verteilung von $\overline{X}$ durch eine Normalverteilung mit Erwartungswert μ und Varianz $\frac{\sigma^2}{n}$ (wie oben!), man schreibt dann auch
$$\overline{X} \overset{\bullet}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$
und sagt „$\overline{X}$ ist approximativ (näherungsweise) $N\left(\mu, \frac{\sigma^2}{n}\right)$-verteilt“.

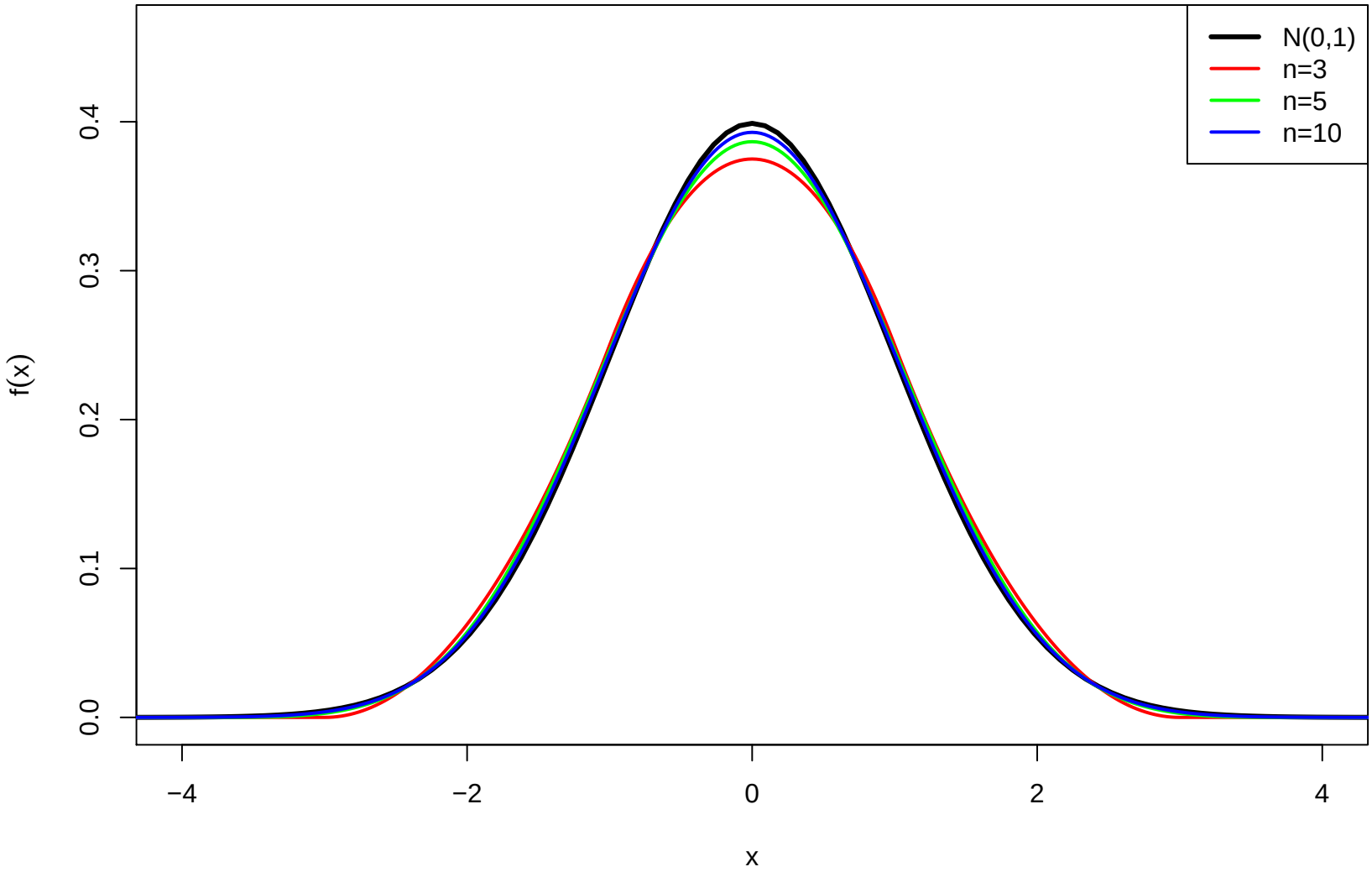
Der Standardabweichung $\mathrm{Sd}(\overline{X}) = \sqrt{\mathrm{Var}(\overline{X})}$ von $\overline{X}$ (also der Standardfehler der Schätzfunktion $\overline{X}$ für μ) wird häufig mit $\sigma_{\overline{X}} := \frac{\sigma}{\sqrt{n}}$ abgekürzt.

- Die Qualität der Näherung der Verteilung im Fall ❷ wird mit zunehmendem Stichprobenumfang höher, hängt aber **ganz entscheidend** vom Verteilungstyp (und sogar der konkreten Verteilung) von Y ab!
- Pauschale Kriterien an den Stichprobenumfang n („Daumenregeln“, z.B. $n \geq 30$) finden sich häufig in der Literatur, sind aber nicht ganz unkritisch.
- Verteilungseigenschaft $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ bzw. $\overline{X} \overset{\bullet}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ wird meistens (äquivalent!) in der (auch aus dem zentralen Grenzwertsatz bekannten) Gestalt

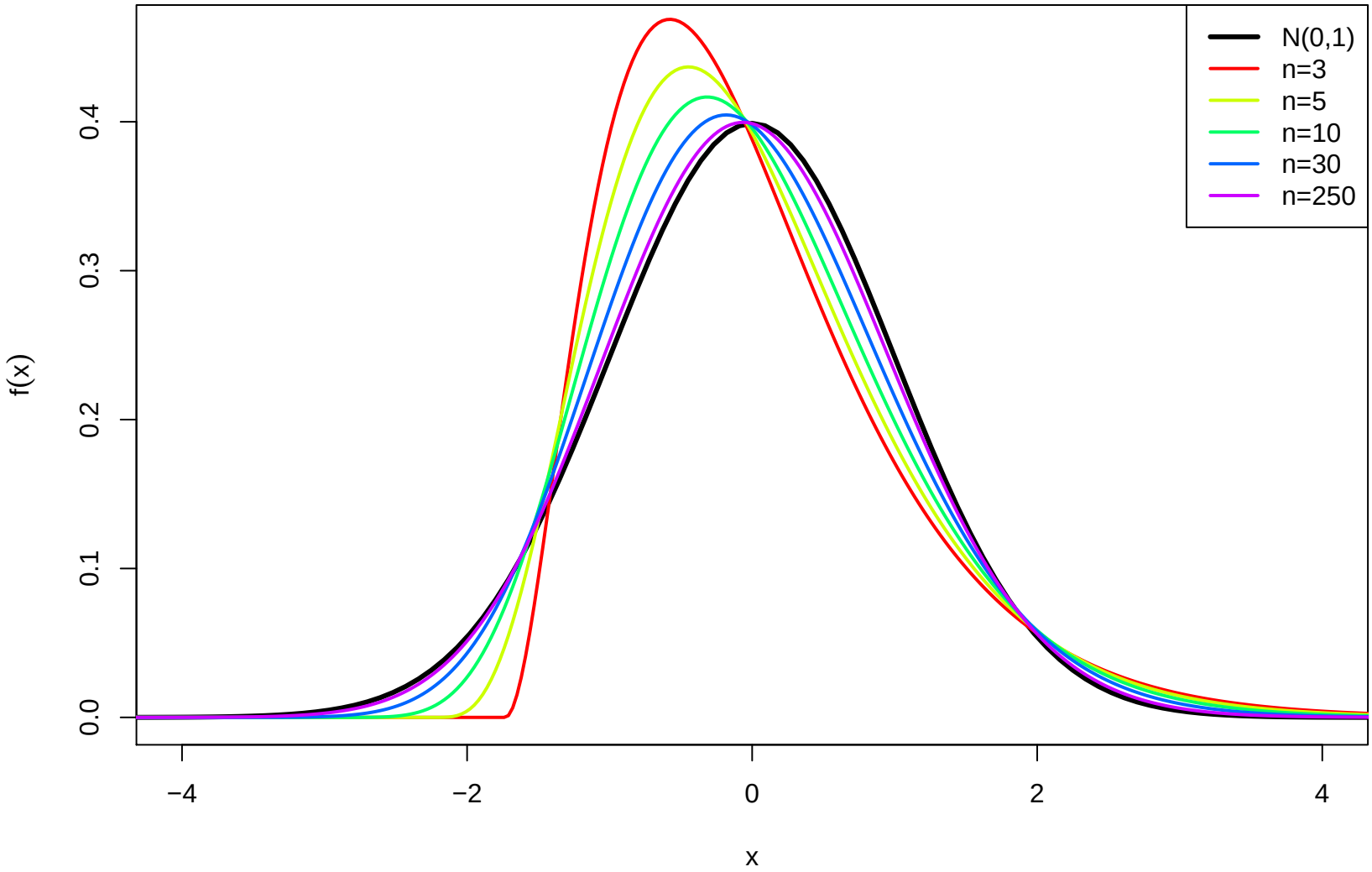
$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1) \qquad \text{bzw.} \qquad \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \overset{\bullet}{\sim} N(0,1)$$

  verwendet, da dann Verwendung von Tabellen zur Standardnormalverteilung möglich.
- Im Folgenden: Einige Beispiele für Qualität von Näherungen durch Vergleich der Dichtefunktion der Standardnormalverteilungsapproximation mit der tatsächlichen Verteilung von $\frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ für unterschiedliche Stichprobenumfänge n.

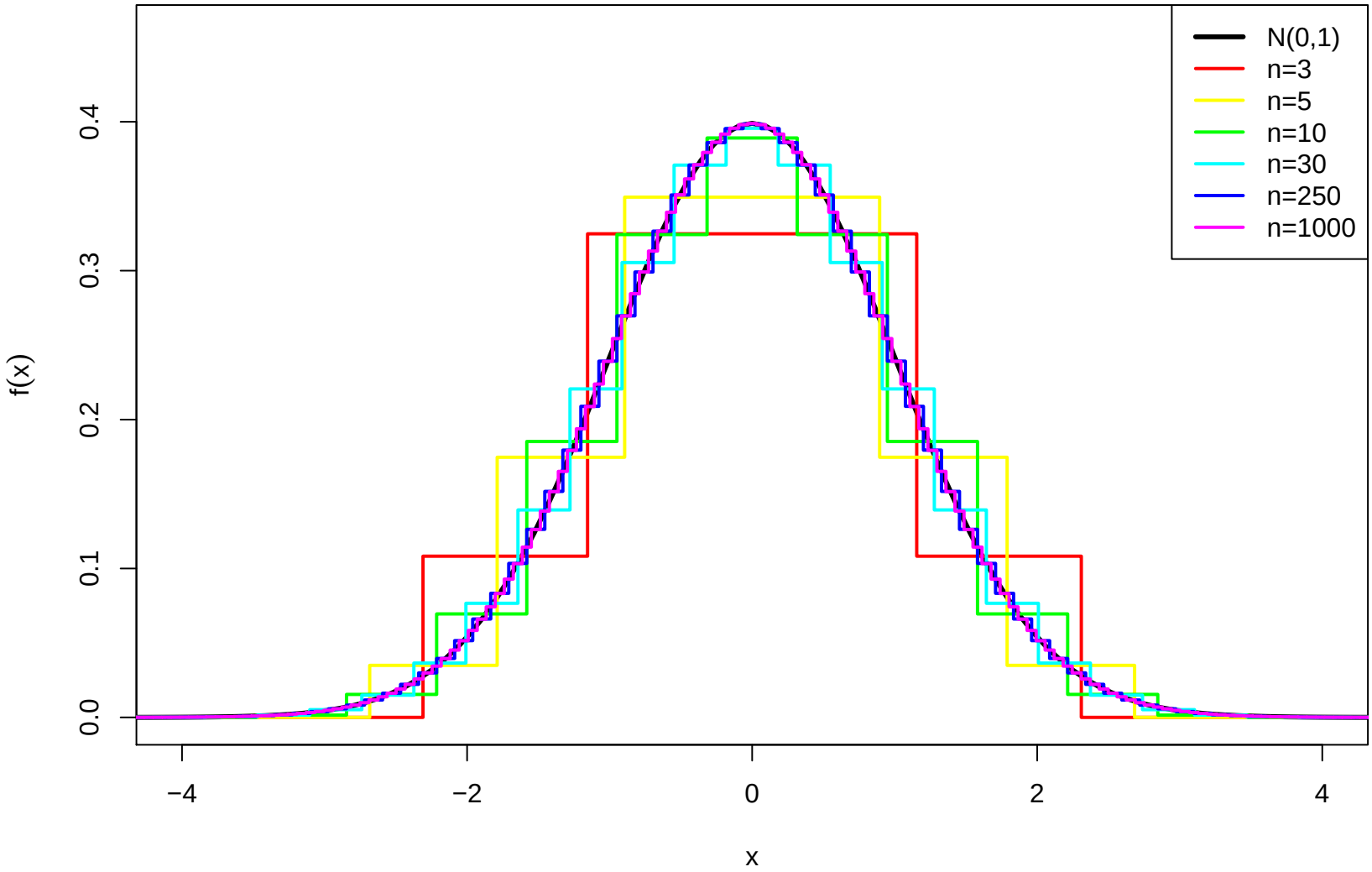
Beispiel: Näherung, falls $Y \sim \text{Unif}(20, 50)$



Beispiel: Näherung, falls $Y \sim \text{Exp}(2)$



Beispiel: Näherung, falls $Y \sim B(1, 0.5)$



Beispiel: Näherung, falls $Y \sim B(1, 0.05)$