1.7.4 Der zentrale Grenzwertsatz

- Gibt es für große Stichprobenumfänge Regelmäßigkeiten im Verteilungstyp?
- Gibt es eine Standardverteilung, mit der man oft bei großen empirischen Untersuchungen rechnen kann? Damit kann man dann insbesondere Fehlermargen einheitlich behandeln.

Satz 1.86. [Zentraler Grenzwertsatz]

Seien $X_1, \ldots, X_n$ i.i.d. mit (existierendem) Erwartungswert $\mathrm{E}(X_i)$ und (existierender) Varianz $\mathrm{Var}(X_i) > 0$ sowie

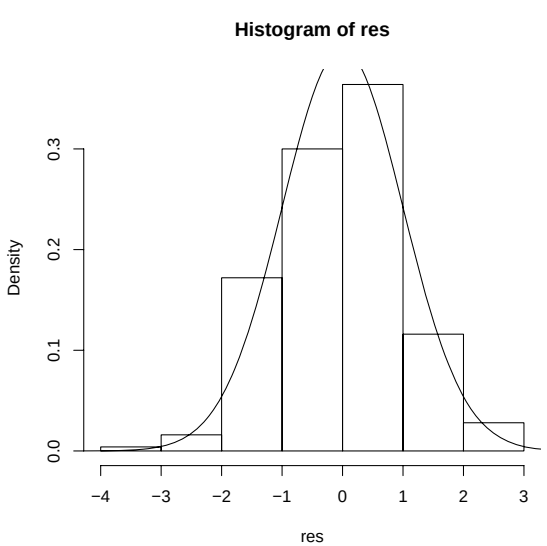
$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mathrm{E}(X_i)}{\sqrt{\mathrm{Var}(X_i)}} \right).$$

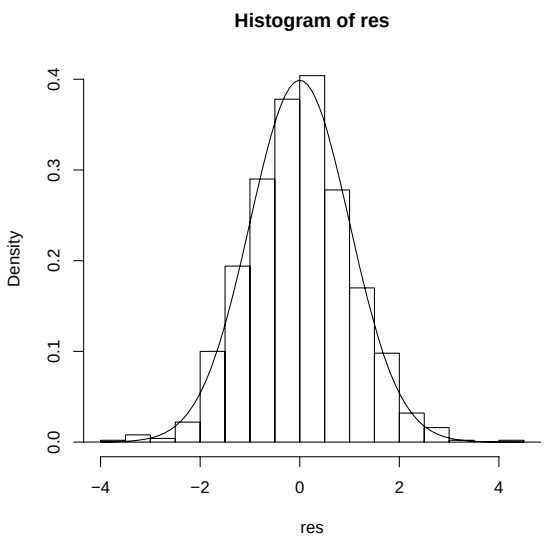
Dann gilt: Z_n ist *asymptotisch standardnormalverteilt*, in Zeichen: $Z_n \overset{a}{\sim} \mathcal{N}(0;1)$, d.h. es gilt für jedes z

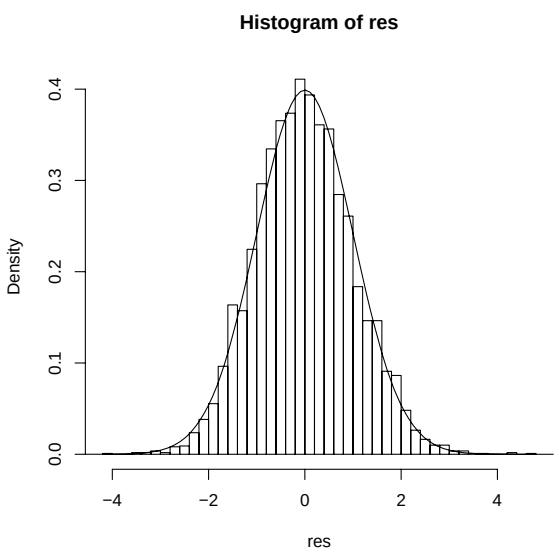
$$\lim_{n\to\infty} \underbrace{P(Z_n \leq z)}_{\substack{\text{Verteilungsfunktion} \\ \text{von } Z_n}} = \underbrace{\Phi(z).}_{\substack{\text{Verteilungsfunktion} \\ \text{der Standardnormalverteilung}}}$$

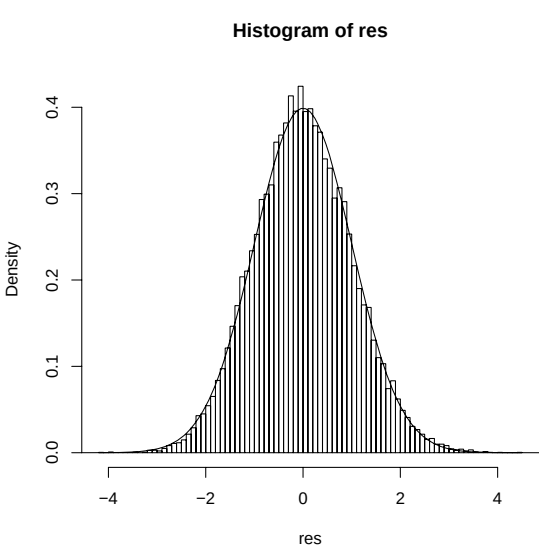
Für die Eingangsfragen gilt also:

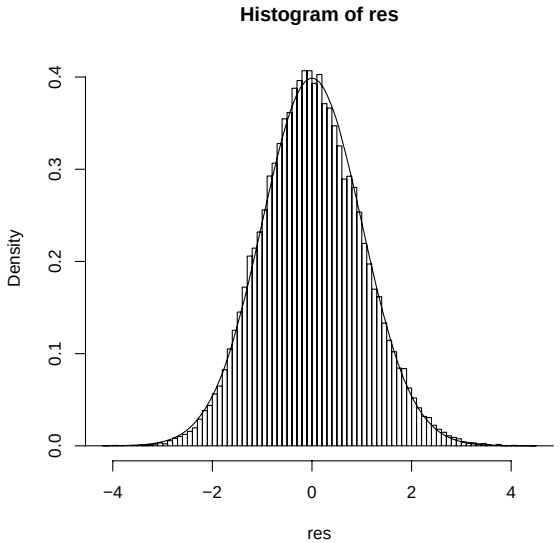
- Ja, wenn man die Variablen geeignet mittelt, standardisiert und mit $\frac{1}{\sqrt{n}}$ reskaliert, dann kann man bei großem n näherungsweise mit der Normalverteilung rechnen. Dabei ist für festes n die Approximation umso besser, je „symmetrischer“ die ursprüngliche Verteilung ist.











Anwendung des zentralen Grenzwertsatz auf $\bar{X}$:

Gemäß dem Gesetz der großen Zahlen weiß man: $\bar{X}_n \longrightarrow \mathrm{E}(X_i)$

Für die Praxis ist es aber zudem wichtig, die konkreten Abweichungen bei großem aber endlichem n zu quantifizieren, etwa zur Beantwortung folgender Fragen:

- Gegeben eine Fehlermarge ε und Stichprobenumfang n: Wie groß ist die Wahrscheinlichkeit, dass $\bar{X}$ höchstens um ε von μ abweicht?
- Gegeben eine Fehlermarge ε und eine „Sicherheitswahrscheinlichkeit“ γ: Wie groß muss man n mindestens wählen, damit mit mindestens Wahrscheinlichkeit γ das Stichprobenmittel höchstens um ε von μ abweicht *(Stichprobenplanung)*?

Aus dem zentralen Grenzwertsatz folgt mit $\mu = \mathrm{E}(X_i)$ und $\sigma^2 = \mathrm{Var}(X_i)$:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right) &= \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n} \cdot \sigma} \\ &= \frac{n\bar{X}_n - n\mu}{\sqrt{n} \cdot \sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} \mathcal{N}(0,1) \end{aligned}$$

oder auch („Standardisierung rüchwärts angewandt“)

$$\bar{X}_n \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n} \right).$$

Wichtige Anwendung: Approximation der Binomialverteilung

Sei $X \sim B(n, \pi)$. Kann man die Verteilung von X (jetzt ja nur eine Beobachtung!) approximieren?

Hier hat man zunächst nur ein X. Der zentrale Grenzwertsatz gilt aber für eine Summe vieler Glieder. Idee: Schreibe X als Summe von binären Zufallsvariablen.

X ist die Anzahl der Treffer in einer $i.i.d.$ Folge $Y_1, \ldots, Y_n$ von Einzelversuchen, wobei

$$Y_i = \begin{cases} 1 & \text{Treffer} \\ 0 & \text{kein Treffer} \end{cases}$$

Die Y_i sind i.i.d. Zufallsvariablen mit $Y_i \sim Bin(1, \pi)$ und es gilt

$$X = \sum_{i=1}^{n} Y_i, \quad \mathrm{E}(Y_i) = \pi, \quad \mathrm{Var}(Y_i) = \pi \cdot (1 - \pi).$$

Damit lässt sich der zentrale Grenzwertsatz anwenden:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \mathrm{E}(Y_i)}{\sqrt{\mathrm{Var}(Y_i)}}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{Y_i - \pi}{\sqrt{\pi(1-\pi)}}\right) &= \frac{1}{\sqrt{n}}\frac{\sum_{i=1}^{n} Y_i - n\cdot\pi}{\sqrt{\pi(1-\pi)}} \\ &= \frac{\sum_{i=1}^{n} Y_i - n\cdot\pi}{\sqrt{n\cdot\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)\end{aligned}$$

und damit

$$\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

so dass

$$P(X \leq x) \;\approx\; \Phi\left(\frac{x - n \cdot \pi}{\sqrt{n \cdot \pi(1-\pi)}}\right)$$

falls n groß genug ist.

Man beachte: Ob eine Approximation gut genug ist, ist eine *inhaltliche* Entscheidung. Trotzdem werden oft Faustregeln angegeben, ab wann diese Approximation als gut gelte, z.B.

$$n \cdot \pi \geq 5 \;\text{ und }\; n \cdot (1-\pi) \geq 5$$

$$n \cdot \pi(1-\pi) \geq 9$$

Stetigkeitskorrektur: Durch die Approximation der $diskreten$ Binomialverteilung durch die $stetige$ Normalverteilung geht der diskrete Charakter verloren. Man erhält als Approximation $P(X = x) \approx 0$ für jedes $x \in \mathbb{N}$, was gerade für mittleres n unerwünscht ist.

Benutze deshalb

$$P(X \leq x) = P(X \leq x + 0.5)$$

bei ganzzahligem $x \in \mathbb{N}$.

Man erhält als bessere Approximation

$$P(X \leq x) \approx \Phi\left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

und damit

$$\begin{aligned} P(X = x) &= P(X \leq x + 0.5) - P(X \leq x - 0.5) \\ &\approx \Phi\left(\frac{x + 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) - \Phi\left(\frac{x - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) \end{aligned}$$

Bsp. 1.87. *Fiktives Beispiel*

Ein Politiker ist von einer gewissen, umstrittenen Maßnahme in seiner Partei überzeugt und überlegt, ob es taktisch geschickt ist, zur Unterstützung der Argumentation eine Mitgliederbefragung zu dem Thema durchzuführen. Er wählt als "Probelauf" 200 Mitglieder zufällig aus und beschließt, eine Mitgliederbefragung zu „riskieren", falls er in der Stichprobe mindestens 52% Zustimmung erhält.

Wie groß ist die Wahrscheinlichkeit, in der Stichprobe mindestens 52% Zustimmung zu erhalten, obwohl der wahre Anteil nur 48% beträgt?

1.8 Mehrdimensionale Zufallsvariablen

Im Folgenden Beschränkung auf den diskreten Fall und zweidimensionale Zufallsvariablen.
„Schnelldurchgang unter Bezug auf das Eindimensionale, Statistik I und auf die Tatsache, dass $X = x_i$ und $Y = y_i$ Ereignisse sind“

Das Hauptinteresse gilt (entsprechend der Kontingenztafel in Statistik I) der gemeinsamen Verteilung

$$P(\{X = x_i\} \cap \{Y = y_j\})$$

Definition 1.88.

Betrachtet werden zwei eindimensionale diskrete Zufallselemente X und Y (zu demselben Zufallsexperiment, also über denselben Grundraum). Die Wahrscheinlichkeit

$$P(X = x_i, Y = y_j) := P(\{X = x_i\} \cap \{Y = y_j\})$$

in Abhängigkeit von x_i und y_j heißt $gemeinsame\ Verteilung$ der mehrdimensionalen Zufallsvariable $\binom{X}{Y}$ bzw. der Variablen X und Y.

Randwahrscheinlichkeiten:

$$p_{i\bullet} = P(X = x_i) = \sum_{j=1}^{m} P(X = x_i, Y = y_j)$$

$$p_{\bullet j} = P(Y = y_j) = \sum_{i=1}^{k} P(X = x_i, Y = y_j)$$

Bedingte Verteilungen:

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

Stetiger Fall (nicht klausurrelevant): Zufallsvariable mit zweidimensionaler Dichtefunktion $f(x, y)$:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \left(\int_c^d f(x, y)\ dy \right) dx$$

Definition 1.89.

Seien X und Y zwei Zufallsvariablen. Dann heißt

$$\sigma_{X,Y} := \mathrm{Cov}(X,Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y)))$$

Kovarianz von X und Y.

Rechenregeln:

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist

$$\mathrm{Cov}(\tilde{X},\tilde{Y}) = a_X \cdot a_Y \cdot \mathrm{Cov}(X,Y)$$

- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y)$

Definition 1.90.

Zwei Zufallsvariablen X und Y mit $\mathrm{Cov}(X, Y) = 0$ heißen *unkorreliert*.

Satz 1.91.

Stochastisch unabhängige Zufallsvariablen sind unkorreliert. Die Umkehrung gilt jedoch im allgemeinen nicht.

Definition 1.92.

Gegeben seien zwei Zufallsvariablen X und Y. Dann heißt

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Korrelationskoeffizient von X und Y.

Eigenschaften des Korrelationskoeffizienten:

- Mit $\tilde{X} = a_X X + b_X$ und $\tilde{Y} = a_Y Y + b_Y$ ist

$$|\rho(\tilde{X}, \tilde{Y})| = |\rho(X, Y)|.$$

- $-1 \leq \rho(X, Y) \leq 1$.

- $|\rho(X, Y)| = 1 \Longleftrightarrow Y = aX + b$

- Sind $\mathrm{Var}(X) > 0$ und $\mathrm{Var}(Y) > 0$, so gilt $\rho(X, Y) = 0$ genau dann, wenn $\mathrm{Cov}(X, Y) = 0$.

Bsp. 1.93. [Chuck-a-Luck:]

X_1 Gewinn, wenn beim ersten Wurf ein Einsatz auf 1 gesetzt wird.

X_6 Gewinn, wenn beim ersten Wurf ein Einsatz auf 6 gesetzt wird.

Kovarianz zwischen X_1 und X_6:

$$Cov(X_1, X_6) = \mathrm{E}(X_1 \cdot X_6) - \mathrm{E}(X_1) \cdot \mathrm{E}(X_6)$$

Zur Berechnung von $\mathrm{E}(X_1 \cdot X_6)$ Hilfsvariable Z denken: $Z = X_1 \cdot X_6$;
dann: $\mathrm{E}(X_1 \cdot X_6) = \mathrm{E}(Z) = \sum_{z \in \mathcal{Z}} z \cdot P(Z = z)$

(x_1, x_6)	$P(X_1 = x_1, X_6 = x_6)$	$x_1 \cdot x_6$	(x_1, x_6)	$P(X_1 = x_1, X_6 = x_6)$	$x_1 \cdot x_6$
$(-1,-1)$	$\frac{64}{216}$	1	$(-1,3)$	$\frac{1}{216}$	-3
$(-1,1)$	$\frac{48}{216}$	-1	$(3,-1)$	$\frac{1}{216}$	-3
$(1,-1)$	$\frac{48}{216}$	-1	$(1,1)$	$\frac{24}{216}$	1
$(-1,2)$	$\frac{12}{216}$	-2	$(1,2)$	$\frac{3}{216}$	2
$(2,-1)$	$\frac{12}{216}$	-2	$(1,2)$	$\frac{3}{216}$	2

z	$P(Z=z)\cdot 216$
-3	$1+1=2$
-2	$12+12=24$
-1	$48+48=96$
1	$64+24=88$
2	$3+3=6$

$$
\begin{aligned}
\mathrm{E}(Z) &= \sum_{z\in\mathcal{Z}} z\cdot P(Z=z),\\
\text{also } \mathrm{E}(X_1\cdot X_6) &= ((-3)\cdot 2+(-2)\cdot 24+(-1)\cdot 96+1\cdot 88+2\cdot 6)\cdot\frac{1}{216}\\
&= \frac{-50}{216}=-0.23148\\
\Rightarrow \quad \mathrm{Cov}(X_1,X_6) &= \mathrm{E}(X_1\cdot X_6)-\mathrm{E}(X_1)\cdot\mathrm{E}(X_6)\\
&= -0.23148-(-0.0787)\cdot(-0.0787)=-0.23768
\end{aligned}
$$

X_1 und X_6 sind negativ korreliert.