Gesetz der grossen Zahlen
Zentraler Grenzwertsatz

Peter Büchel

HSLU I

ASTAT: Block 08

Funktionen von mehreren Zufallsvariablen

- Bis jetzt: Verteilung *einer* Zufallsvariable (ZV)
- Aber: üblicherweise wird die *gleiche* Grösse mehrmals gemessen
- Bsp: Misst Gewicht mehrmals
- Allgemein: Messungen $x_1, x_2, \ldots, x_n$ als Realisierungen der ZV auffassen:

$$X_1, \ldots, X_n$$

- X_i: Die i-te Wiederholung von Zufallsexperiment

Beispiel

- 20 Messungen der Wasserverschmutzung in einem See
- Messungen (konkrete Werte):

$$x_1,\ x_2,\ \ldots,\ x_{20}$$

- Realisierungen der ZV:

$$X_1,\ X_2,\ \ldots,\ X_{20}$$

- Annahme: 20 ZV mit gleichen W'keitsverteilungen
- Wasserproben: Alle aus demselben See mit identischer Methode gemessen
- Interessant: *Durchschnitt* dieser Messungen und die Verteilung der zugehörigen ZV

Summe und Durchschnitt

- Gegeben ZV:

$$X_1, \ldots, X_n$$

- *Summe*:

$$S_n = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

- *Arithmetisches Mittel*:

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} S_n$$

Beispiel: Warum ist der Durchschnitt wichtig?

- Untersuchen, ob Angabe 500 ml Inhalt einer Pet-Flasche gilt
- Kaufen *eine* Flasche: Messen Inhalt 495.21 ml
- Das ist weniger als 500 ml, aber ist es zuwenig?
- Bei einer Flasche möglich
- Idee: Kaufen 100 Flaschen und messen Inhalt
- Durchschnitt 463.21 ml
- Scheint eindeutig zu wenig: Angabe 500 ml kann nicht stimmen
- Genaues Vorgehen: Siehe Hypothesentest

Kennzahlen von S_n und $\overline{X}_n$

- Annahme:

$$X_1, \ldots, X_n \text{ i.i.d.}$$

- Zweites „i" in i.i.d.: X_i dieselbe Verteilung mit denselben Kennzahlen:

$$\mathrm{E}(X_i) = \mu \quad \text{und} \quad \mathrm{Var}(X_i) = \sigma_X^2$$

- Gesucht: Erwartungswert und Varianz für:
  - Summe S_n:

$$S_n = X_1 + X_2 + \ldots + X_n$$

  - Durchschnitt $\overline{X}_n$:

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

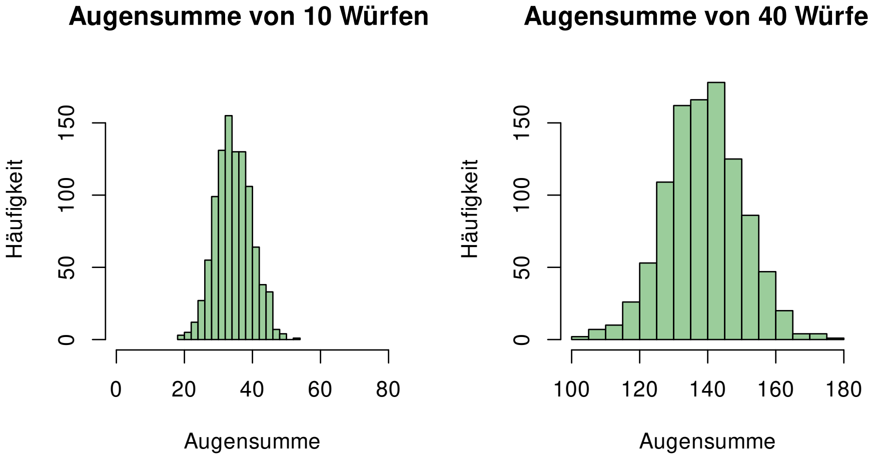
Graphisches Beispiel

- Werfen einen fairen Würfels
- X: ZV für geworfene Augenzahl
- Erwartungswert:
$$\mathrm{E}(X) = \mu = \frac{1}{6}(1+2+3+4+5+6) = 3.5$$
- Varianz:
$$\begin{aligned}\mathrm{Var}(X) &= \frac{1}{6}\left((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2\right)\\ &= 2.92\end{aligned}$$
- R:

```
x <- c(1, 2, 3, 4, 5, 6)

ave <- mean(x)

var <- mean((x - ave)^2)

var
## [1] 2.916667
```

- Würfeln 10mal
- ZV's:
$$X_1, X_2, \ldots, X_{10} \quad \text{i.i.d.}$$
- X_i: Augenzahl im i-ten Wurf
- Erwartungswert und Varianz: Werte der ZV's X_i oben
- Notieren Augensumme s_{10} dieser 10 Würfe
- 1000 mal machen: Histogramm aller vorkommenden Augensummen
- Dasselbe mit 40 Würfen
- Simulation mit `R`

Histogramme



Feststellungen

- Mittlere Augensumme verschiebt sich, wenn mehr Würfe gemacht werden

- Abbildung links: Grösste Häufigkeit bei etwa 35, also

$$10 \cdot 3.5 = 10 \cdot \mu$$

- $\mu = 3.5$: Erwartungswert für einen Wurf

- Abbildung rechts: Grösste Häufigkeit bei etwa 140

$$40 \cdot 3.5 = 40 \cdot \mu$$

- Vermutung:

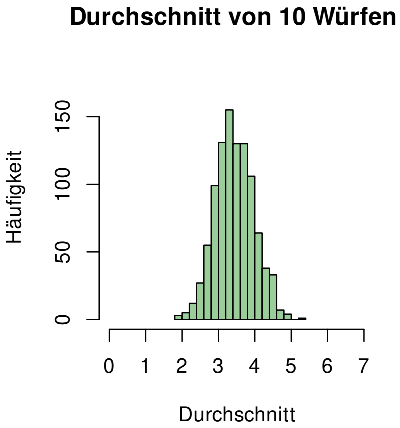
$$\mathrm{E}(S_n) = n\mu$$

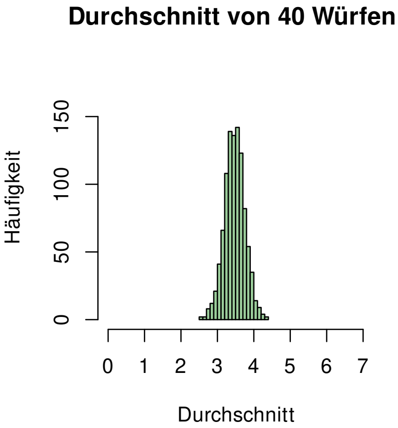
- Stimmt auch (ohne Beweis)

- Varianz/Standardabw. nimmt mit zunehmender Anzahl Würfen *zu*
- Gesetz (ohne Beweis) angegeben:

$$\mathrm{Var}(S_n) = n\,\mathrm{Var}(X), \qquad \sigma_{S_n} = \sqrt{n}\sigma_X$$

- Dasselbe mit dem Durchschnitt $\overline{X}_n$
- Histogramme:





Feststellungen

- Beide Histogramme: Grösste Häufigkeit bei 3.5, also μ
- Vermutung (stimmt, aber ohne Beweis):

$$\mathrm{E}(\overline{X}_n) = \mu$$

- Varianz/Standardabw. nimmt mit zunehmender Anzahl Würfen *ab*
- Gesetz (ohne Beweis):

$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n}; \qquad \sigma_{\overline{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

- Oben gemachte Beobachtungen gelten allgemein

Allgemein

- Annahme:

$$X_1, \ldots, X_n \text{ i.i.d.}$$

- Es gilt:

Kennzahlen von S_n

$$\begin{aligned} \mathrm{E}(S_n) &= n\mu \\ \mathrm{Var}(S_n) &= n\,\mathrm{Var}(X_i) \\ \sigma(S_n) &= \sqrt{n}\sigma_X \end{aligned}$$

Kennzahlen von $\overline{X}_n$

$$\begin{aligned} \mathrm{E}(\overline{X}_n) &= \mu \\ \mathrm{Var}(\overline{X}_n) &= \frac{\sigma_X^2}{n} \\ \sigma(\overline{X}_n) &= \frac{\sigma_X}{\sqrt{n}} \end{aligned}$$

Bemerkungen

- Standardabweichung von $\overline{X}_n$: *Standardfehler* des arithmetischen Mittels
- Standardabweichung der Summe: Wächst mit wachsendem n, aber langsamer als die Anzahl Beobachtungen n
- D. h.: Kleinere Streuung für wachsendes n
- Erwartungswert von $\overline{X}_n$: Gleich demjenigen einer einzelnen ZV X_i, die *Streuung nimmt jedoch ab mit wachsendem n*

Standardfehler

Standardfehler

Standardabweichung des arithmetischen Mittels (*Standardfehler*) ist *nicht* proportional zu $1/n$, sondern nimmt ab mit dem Faktor $1/\sqrt{n}$:

$$\sigma_{\overline{X}_n} = \frac{1}{\sqrt{n}}\,\sigma_X$$

Um *Standardfehler* zu halbieren, braucht man also *viermal* so viele Beobachtungen

Dies nennt man auch das $\sqrt{n}$-*Gesetz*

Zentraler Grenzwertsatz

- Bekannt: Kennzahlen von S_n und $\overline{X}_n$
- Unbekannt: Verteilung von S_n und $\overline{X}_n$
- Würfelbeispiel: X_i gleichverteilt:

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- *Wie* sind S_n und $\overline{X}_n$ verteilt?
- Vermutung wegen Slides 9 und 12: Beide normalverteilt
- Dies ist die Aussage des *Zentralen Grenzwertsatzes*
- Simulation der Aussage, kein Beweis

Simulation von $\overline{X}_n$

- Ergebnismenge
$$\Omega = \{0, 10, 11\}$$
- Ziehen eine Zahl
- ZV X: Wert der gezogenen Zahl
- Es gilt:
$$P(X = 0) = P(X = 10) = P(X = 11) = \frac{1}{3}$$
- Erwartungswert von X:
$$\mathrm{E}(X) = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 10 + \frac{1}{3} \cdot 11 = 7$$

```r
werte <- c(0, 10, 11)
ew <- sum(werte * 1/3)
ew
## [1] 7
```

- Varianz von X:

$$\text{Var}(X) = \frac{1}{3} \cdot (0-7)^2 + \frac{1}{3} \cdot (10-7)^2 + \frac{1}{3} \cdot (11-7)^2 = 24.6667$$

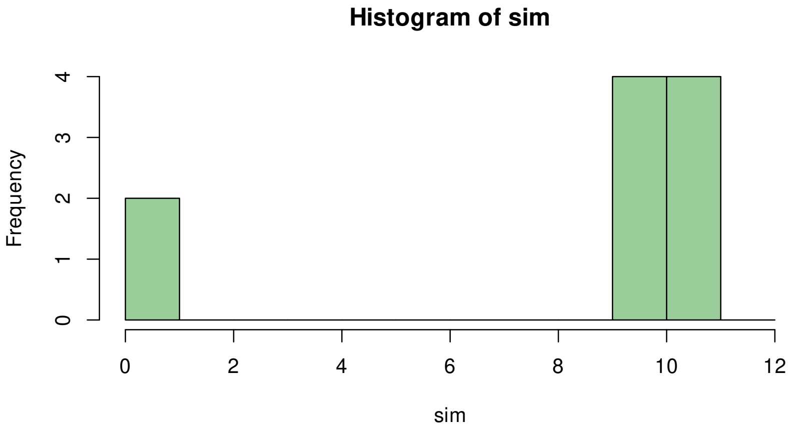
```r
var.X <- sum((werte - ew)^2 * 1/3)

var.X
## [1] 24.66667
```

- Jetzt 10 Ziehungen
- Anzahl Ziehungen zu klein, aber man „sieht" besser, was passiert
- Ein Versuch (10 Ziehungen):

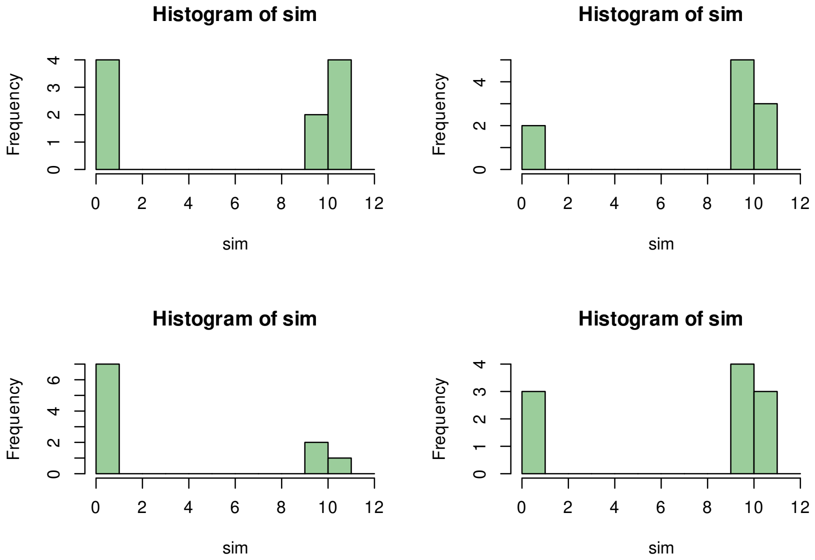
```
# Zieht 10-mal aus der Menge {0,10,11} einen Wert mit
# gleicher W'keit
sim <- sample(werte, 10, replace = T)

# Vektor mit 10 Werten
sim
##  [1]  0 10 11 11 11 11 10 10  0 10
```

- `sample`: Zieht zufällig Zahlen aus `werte`
- `replace = T`: Legt die Zahl nach dem Ziehen wieder zurück

```r
# Histogramm mit diesen 10 Werten
hist(sim, col = "darkseagreen3", breaks = 0:12)
```



- Bei jedem Versuch: Anderes Histogramm
- Histogramme von 4 Versuchen (10 Ziehungen):



- Offensichtlich keine Normalverteilung
- Bis jetzt: Kommen nur die Zahlen 0, 10, 11 vor
- Nun: Zwei solche Versuche (je 10 Ziehungen) hintereinander ausführen
- *Durchschnitt* aus beiden Versuchen berechnen:

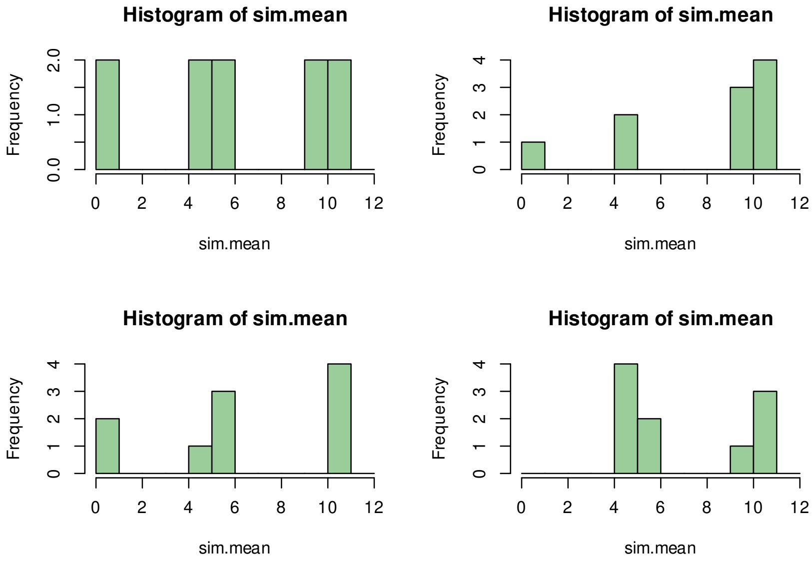
```r
sim.1 <- sample(werte, 10, replace = T)
sim.1
##  [1]  0 11  0 10  0 11 11 10 10 11
sim.2 <- sample(werte, 10, replace = T)
sim.2
##  [1] 11  0  0  0 10 10 10 10 11  0
sim.mean <- (sim.1 + sim.2)/2
sim.mean
##  [1]  5.5  5.5  0.0  5.0  5.0 10.5 10.5 10.0 10.5  5.5
```

- Neben Zahlen 0, 10, 11: Auch Zahlen 5, 5.5 und 10.5 können vorkommen

- Histogramm:

```
hist(sim.mean, col = "darkseagreen3", breaks = 0:12)
```

- 4 Histogramme: Alle verschieden



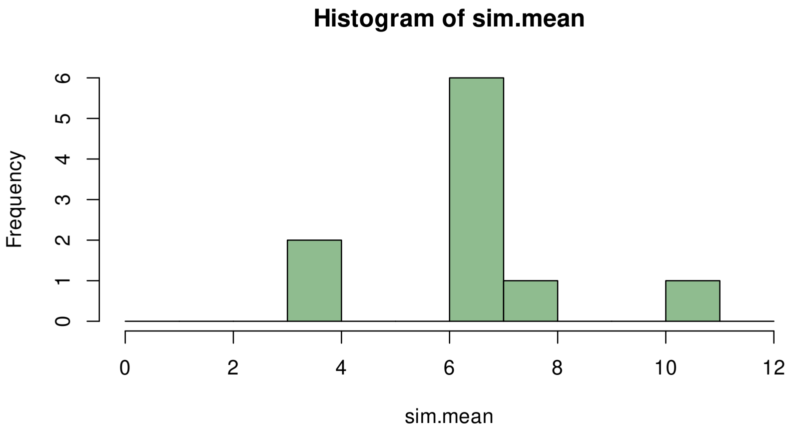
- Jeder Versuch sieht anders aus
- Aber: Tendenzen zeichnen sich ab
- 0 weniger oft vertreten, da doppelte 0 nur mit W'keit $\frac{1}{9}$ vorkommt

- Nun 3 Versuche wiederholen und Durchschnitt nehmen:

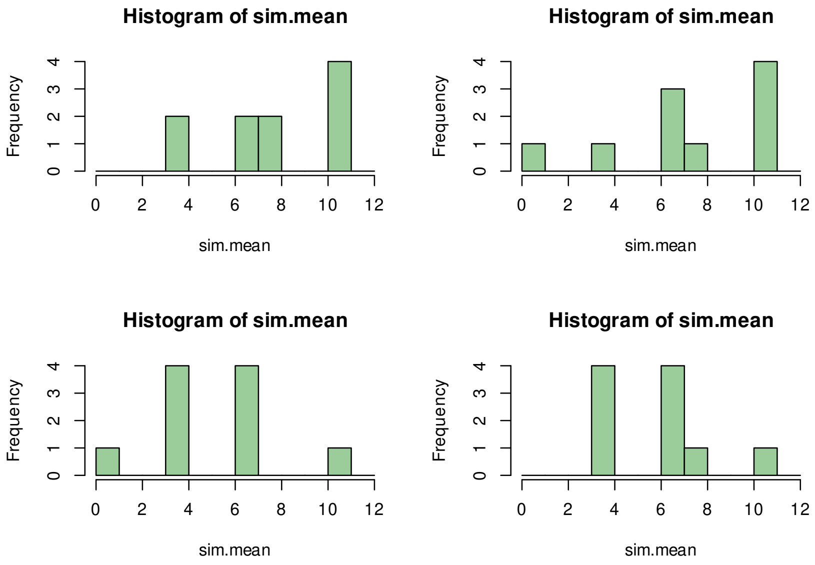
```
sim.1 <- sample(werte, 10, replace = T)
sim.1

##  [1] 10 10  0 11 11 10 11 10  0 10

sim.2 <- sample(werte, 10, replace = T)
sim.2

##  [1]  0 11 11  0 10  0 11 10 11 11

sim.3 <- sample(werte, 10, replace = T)
sim.3

##  [1]  0  0 10 10 10  0  0  0 10  0

sim.mean <- (sim.1 + sim.2 + sim.3)/3
round(sim.mean, 2)

##  [1]  3.33  7.00  7.00  7.00 10.33  3.33  7.33  6.67
##  [9]  7.00  7.00
```

- Histogramm:

```r
hist(sim.mean, col = "darkseagreen", breaks = 0:12)
```

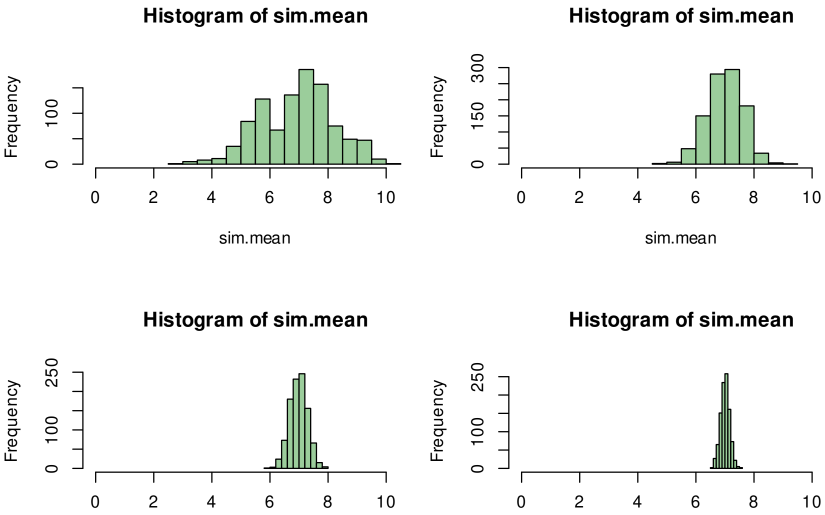


- Mehrere Versuche:



- Tendenz gegen den Erwartungswert 7
- Beim *Durchschnitt* gibt es immer mehr Werte
- Häufung um Erwartungswert 7
- Warum ist dies so?
- Zahl 0 im Durchschnitt kommt praktisch nicht mehr vor: W'keit, dass 3 mal an gleicher Stelle eine 0 vorkommt, ist nur noch $\frac{1}{27}$
- Dasselbe für Zahl 11

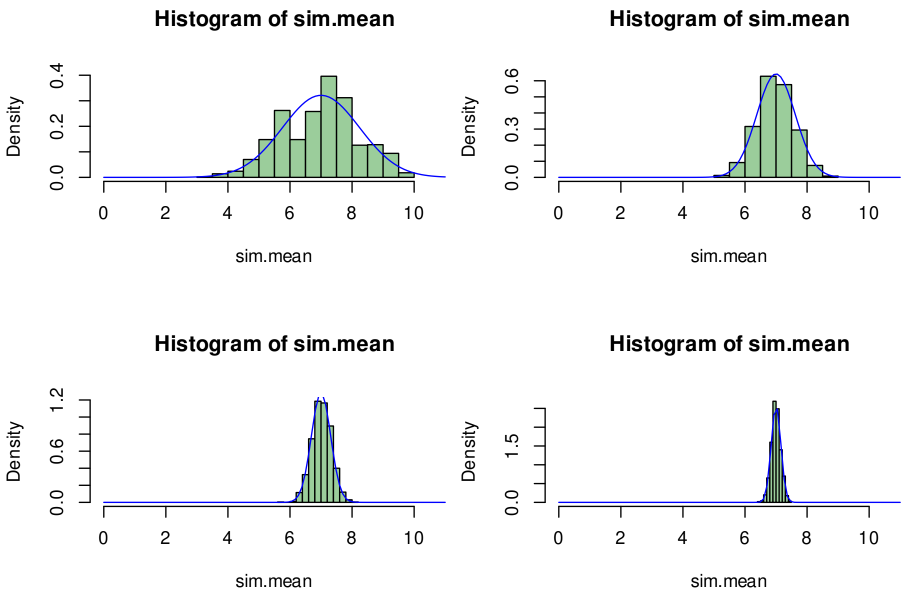
- Nun 16, 64, 256 und 1024 solche Versuche mit jeweils 1000 Ziehungen
- Nehmen jeweils den Durchschnitt wie in den Beispielen vorher
- Histogramme:



- Bei genauerem Hinsehen fällt auf:
  - Werte häufen sich um den Erwartungswert 7
  - Standardabweichung wird kleiner: Halbiert sie sich etwa beim Vervierfachen der Anzahl Versuche
  - Histogramme scheinen einer Normalverteilung zu folgen
- Zeichnen noch die jeweiligen Dichtekurven für

$$\mathcal{N}\left(7, \frac{24.6667}{n}\right)$$

Histogramme



- Fällt auf: Dichtekurven für grössere n passen immer besser zu den Histogrammen
- Nochmals: Begannen mit Verteilung, die *nichts* mit einer Normalverteilung zu tun hat
- Aber: Verteilung *Mittelwerte* $\overline{X}_n$ (oder Summen) nähert sich mit wachsendem n einer Normalverteilung an

Zentraler Grenzwertsatz

- X_i's i.i.d. (nicht notwendig normalverteilt), dann gilt der berühmte

Zentraler Grenzwertsatz

$X_1, \ldots, X_n$ i.i.d. mit irgendeiner Verteilung mit Erwartungswert μ und Varianz σ^2, dann gilt (ohne Beweis):

$$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

- Approximation wird mit grösserem n i.A. besser
- Approximation besser, je näher die Verteilung von X_i bei der Normalverteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist

Beispiel

- Strassenverkehrsamt hat genug Streusalz gelagert, um mit einem Schneefall von insgesamt 80 cm pro Jahr fertigzuwerden
- Täglich fallen im Mittel 1.5 cm mit einer Standardabw. von 0.3 cm
- Wie gross ist W'keit, dass das gelagerte Salz für die nächsten 50 Tage ausreicht?

Lösung

- X_i: ZV für die gefallene Menge Schnee am Tag i
- Annahme: i.i.d. $\quad\rightarrow\quad$ gerechtfertigt?
- Es gilt $\mu = 1.5$ und $\sigma_X = 0.3$
- Schneemenge (Summe) S_{50} der nächsten 50 Tage
- Soll 80 nicht übersteigen
- Es gilt annähernd:

$$S_{50} \sim \mathcal{N}\left(50 \cdot \mu, 50 \cdot \sigma_X^2\right) = \mathcal{N}(75, 4.5)$$

- Gesucht:

$$P(S_n \leq 80) = 0.991$$

```
pnorm(q = 80, mean = 50 * 1.5, sd = sqrt(50) * 0.3)
## [1] 0.9907889
```

Beispiel

- Die Lebensdauer eines bestimmten elektrischen Teils ist durchschnittlich 100 Stunden mit Standardabweichung von 20 Stunden
- Testen 16 solcher Teile
- Wie gross ist W'keit, dass das Stichprobenmittel
  - unter 104 Stunden oder
  - zwischen 98 und 104 Stunden liegt?

Lösung

- X_i: Zufallsvariable für die Lebensdauer des Teils i
- Es gilt $\mu = 100$ und $\sigma_X = 20$
- Annahme i.i.d.
- Betrachten durchschnittliche Lebensdauer $\overline{X}_{16}$
- Annähernd verteilt wie:

$$\overline{X}_{16} \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) = \mathcal{N}\left(100, \frac{20^2}{16}\right) = \mathcal{N}(100, 25)$$

- Gesucht:

$$P(\overline{X}_{16} \leq 104) = 0.788$$

```r
pnorm(q = 104, mean = 100, sd = 20/sqrt(16))
## [1] 0.7881446
```

- Gesucht:

$$P(98 \leq \overline{X}_{16} \leq 104) = 0.444$$

```r
pnorm(q = 104, mean = 100, sd = 20/sqrt(16)) - pnorm(98,
    100, 20/sqrt(16))
## [1] 0.4435663
```