universität wien

Stetige Zufallsvariablen

Univ.Prof. Dr. Marcus Hudec

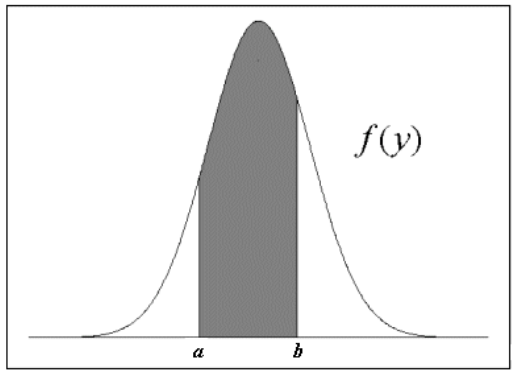
Stetige Zufalls-Variable

- Erweitert man den Begriff der diskreten Zufallsvariable für stetige Merkmale gibt es einige technische Probleme
- Die Wahrscheinlichkeit einen bestimmten konkreten Wert zu beobachten ist null, da es ja unendlich viele unterschiedliche Wert gibt.
- Eine stetige Zufallsvariable liefert daher Wahrscheinlichkeitswerte immer nur für Intervalle.
- Man erhält Wahrscheinlichkeiten indem man eine Fläche evaluiert. Konkret betrachtet man das Integral unter der **Dichtefunktion**, die das stetige Analogon zur Wahrscheinlichkeitsfunktion bildet.

Dichtefunktion

- f(x) … Dichtefunktion

$$\int_{\text{all } x} f(x)\,dx = 1$$



- 1) f(x) > 0 für alle x
  2) Gesamte Fläche unter der Kurve ist 1
- Einzelne Werte von f(x) können größer als 1 sein!
- f(x) ist eine Dichte aber keine Wahrscheinlichkeit
- Vergleiche dazu das Histogramm, wo auch die Fläche als Maß für die Häufigkeit fungiert

Stetige Verteilungsfunktion

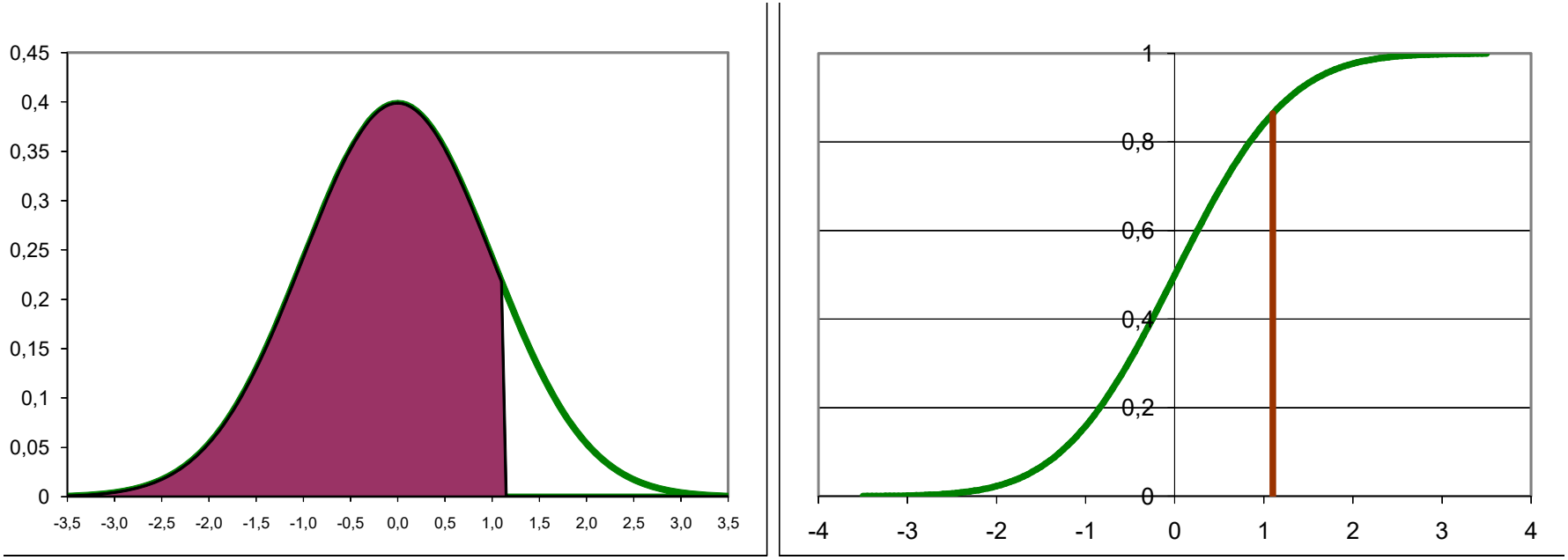
- Die theoretische Verteilungsfunktion einer stetigen Zufallsvariablen X mit Dichtefunktion f(x) bezeichnen wir mit F(x)
- Die theoretische Verteilungsfunktion wird durch das Integral (stetiges Analogon zur Summe) definiert

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

Beziehung zwischen Dichte- und Verteilungsfunktion

Dichtefunktion f(x)

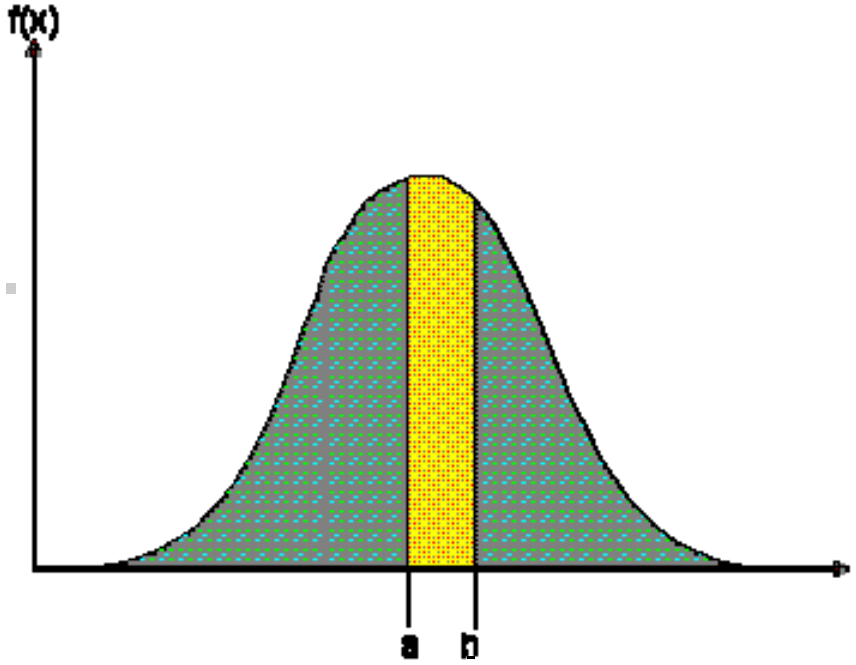
Verteilungsfunktion F(x)



Wahrscheinlichkeiten als Integral

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$



Erwartungswert und Varianz

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

$$V(X) = \sigma^2 = \int_{-\infty}^{+\infty} [x - E(X)]^2 \cdot f(x)dx$$

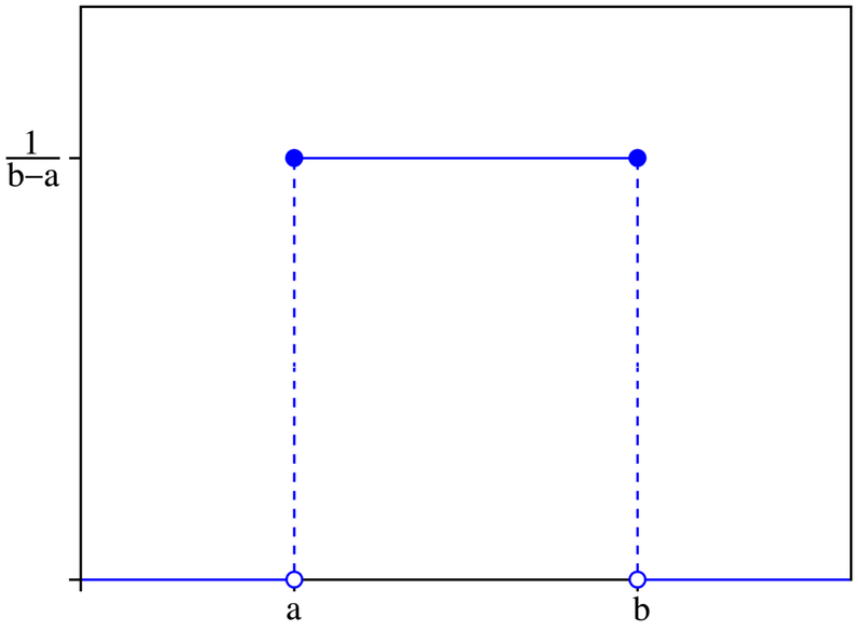
$$Var(X) = E(X^2) - [E(X)]^2$$

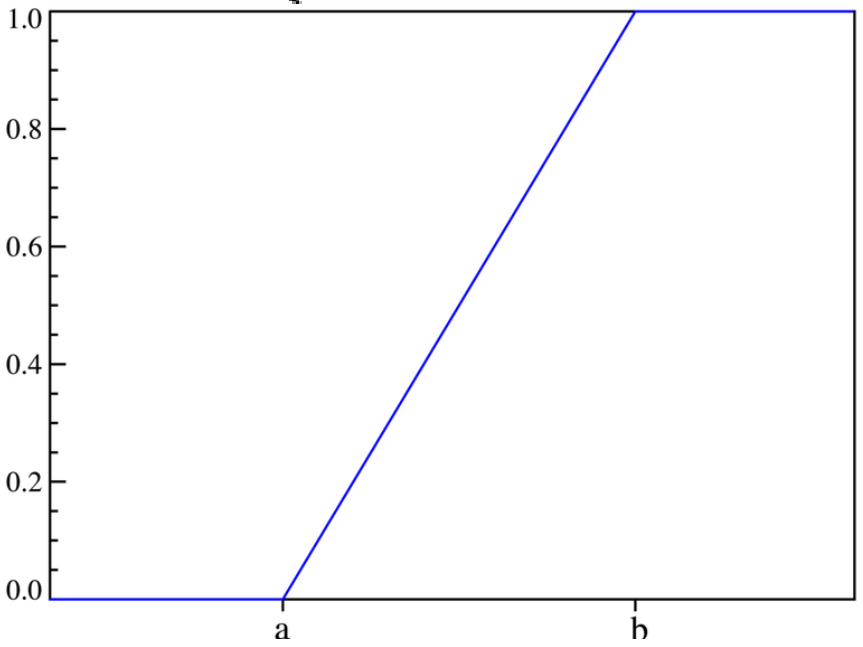
Continuous Uniform Distribution

- The probability density and the distribution function of the continuous uniform distribution are:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x < b \\ 1 & \text{for } x \ge b \end{cases}$$



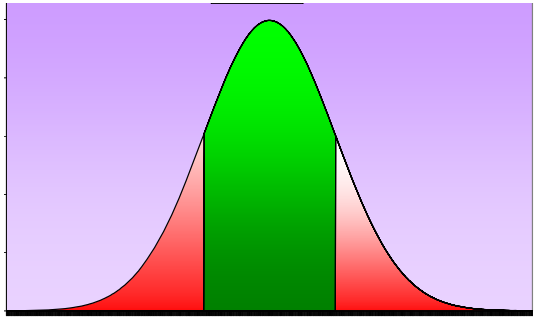


Moments

- For U~Continuous Uniform Distribution on [a,b]
- Expectation: $\frac{a+b}{2}$
- Variance: $\frac{1}{12}(b-a)^2$

The Normal Distribution

- The **normal distribution**, also called the **Gaussian distribution**, is the most important family of continuous probability distributions.



- It is widely applicable to many fields and a lot of statistical methods are based on this distributional model.
- Each member of the family may be defined by two parameters characterizing *location* and *scale*:
  - the mean ("average", μ)
  - and the variance ("variability", σ^2)

Standardnormalverteilung

- 1720 erstmals von Abraham de Moivre beschrieben
- 1809 und 1816 grundlegende Arbeiten von Carl Friedrich Gauß
- 1870 von Adolphe Quetelet als "ideales" Histogramm verwendet
- alternative Bezeichnungen: Gaußsche Glockenkurve;Fehlerkurve
- Natürliche Prozesse
  - Körpergröße, Gewicht von Lebewesen
- Messung von physikalischen Größen
  - Messfehlermodell
- Variable, die sich aus der Summe von vielen zufälligen Einzelwerten ergeben
  - zentraler Grenzwertsatz

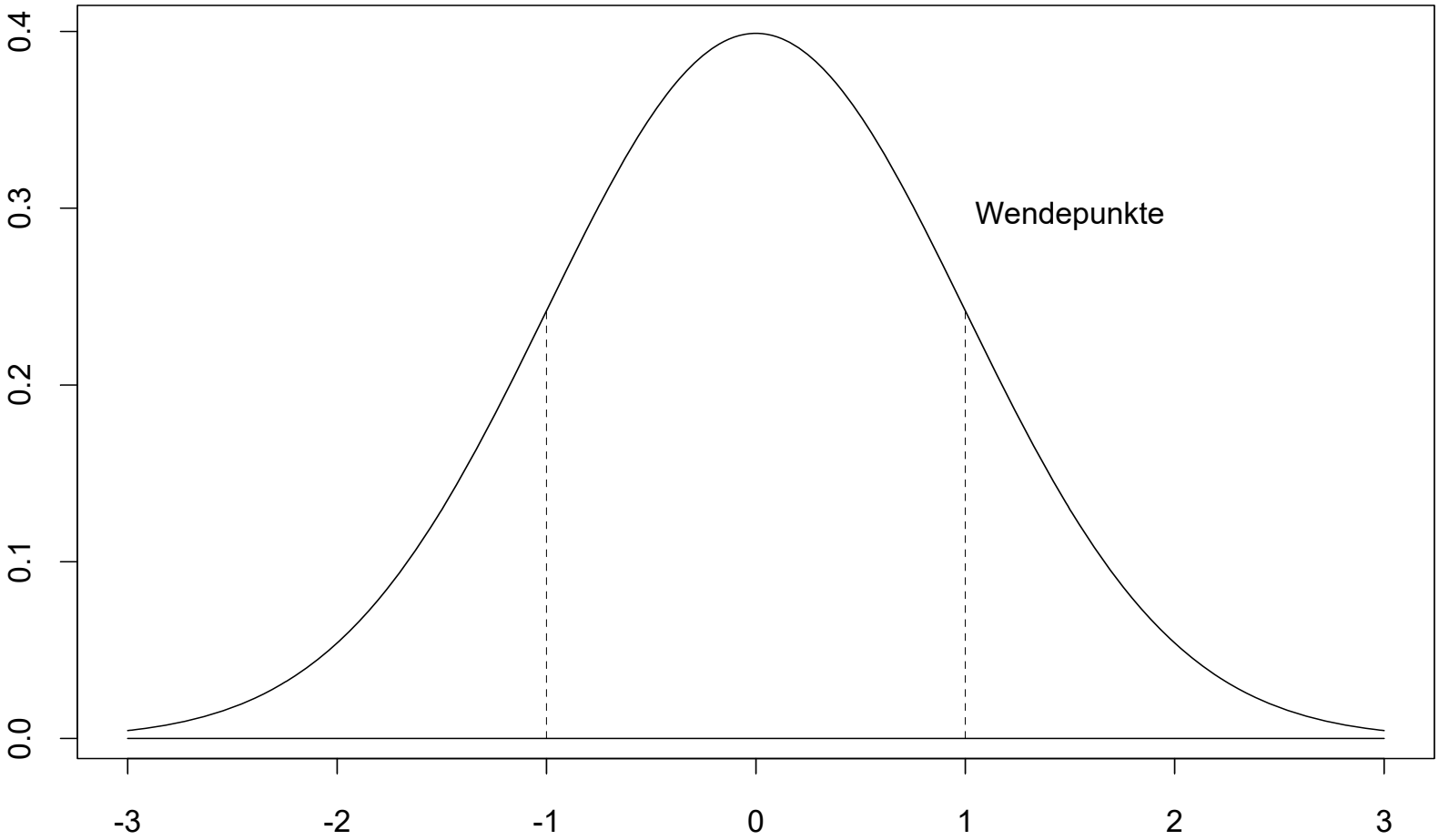
Dichtefunktion

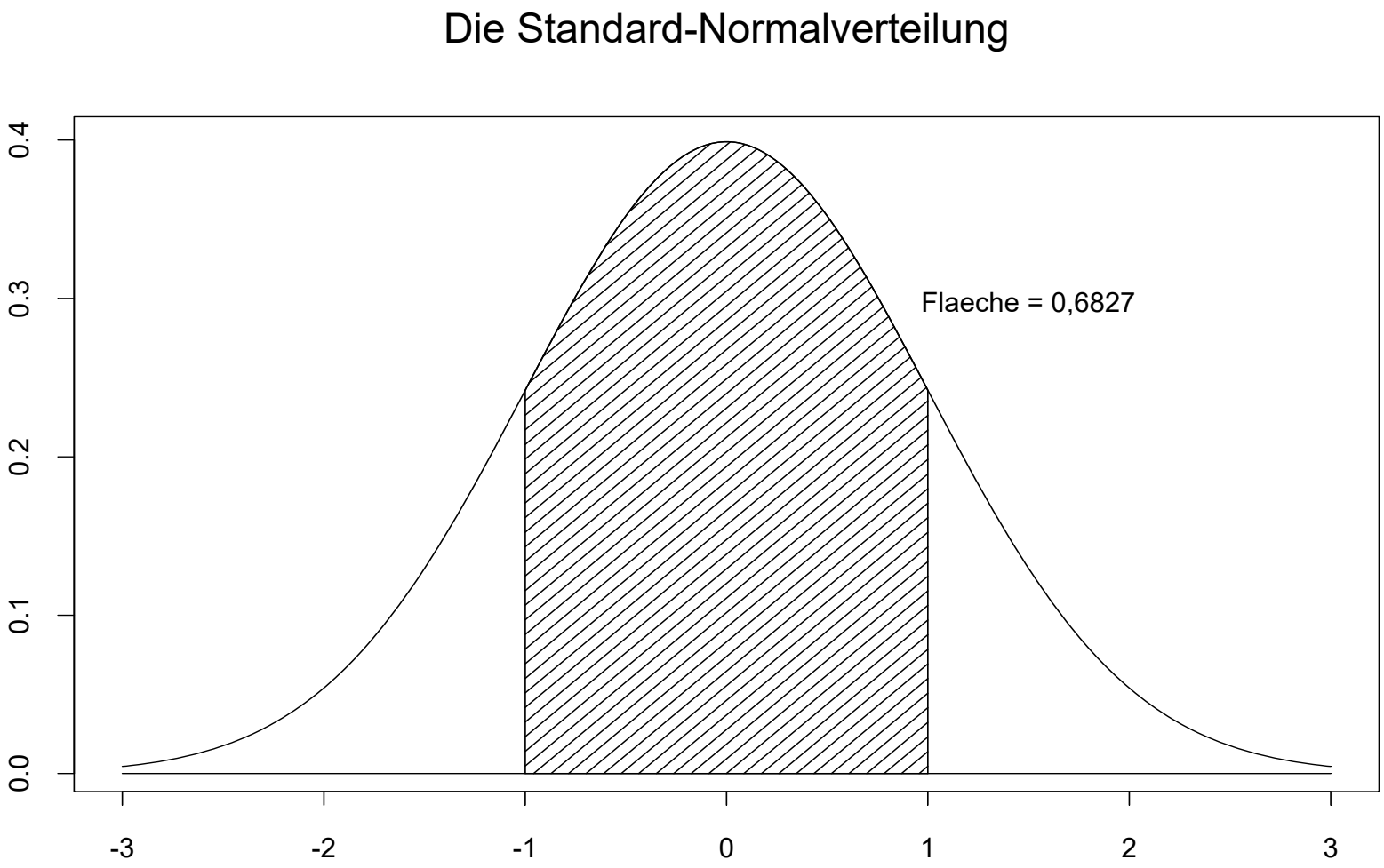
In der einfachsten Form:

- Standardnormalverteilung
- X~N(0; 1)
- E(X)=0 Erwartungswert = 0
- V(X)=1 Varianz bzw. Standard-Abweichung =1

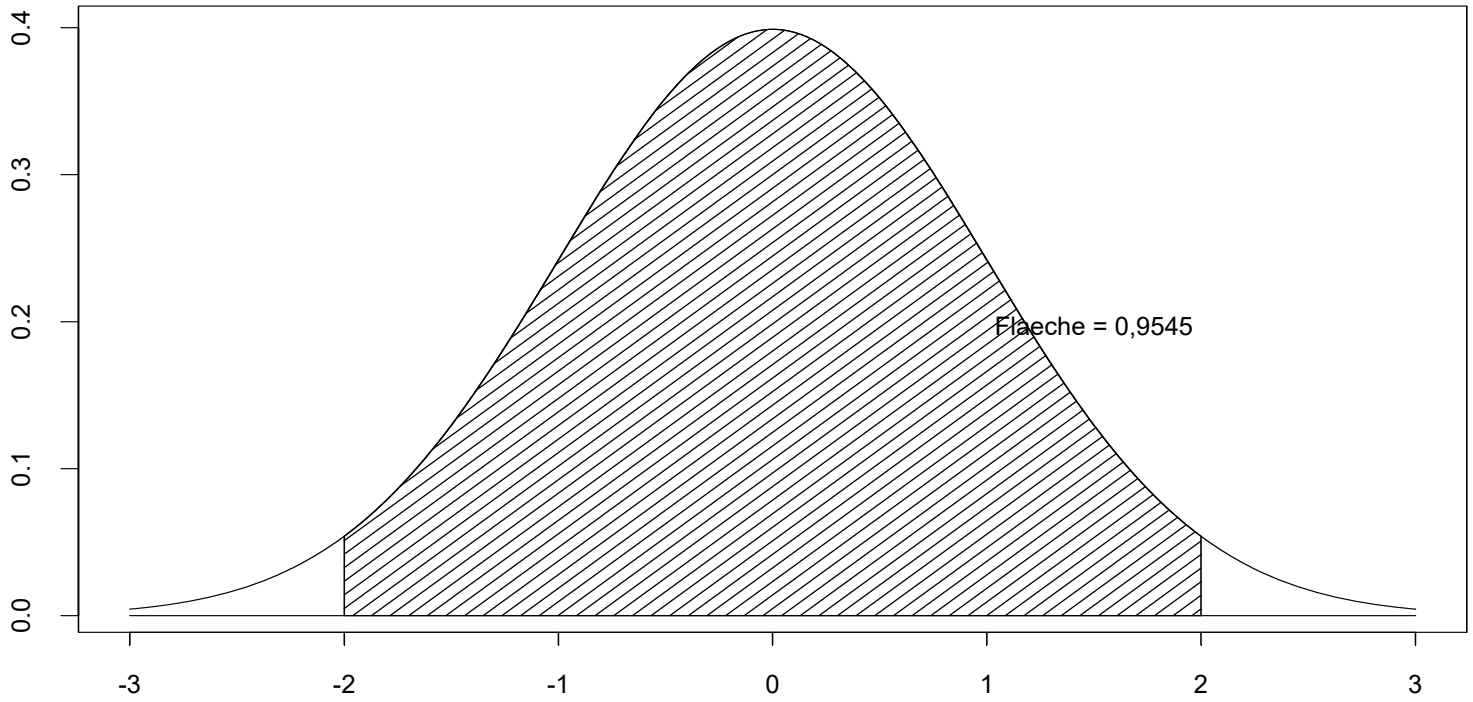
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Die Standard-Normalverteilung





Die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert im Bereich -1 bis +1 annimmt ist 68,27%



Die Wahrscheinlichkeit, dass die Zufallsvariable einen Wert im Bereich -2 bis +2 annimmt, ist rund 95%

Allgemein:

Bei einem normalverteilten Merkmal liegen rund 95% der Beobachtungen liegen im Bereich Erwartungswert plus/minus 2*Standardabweichung

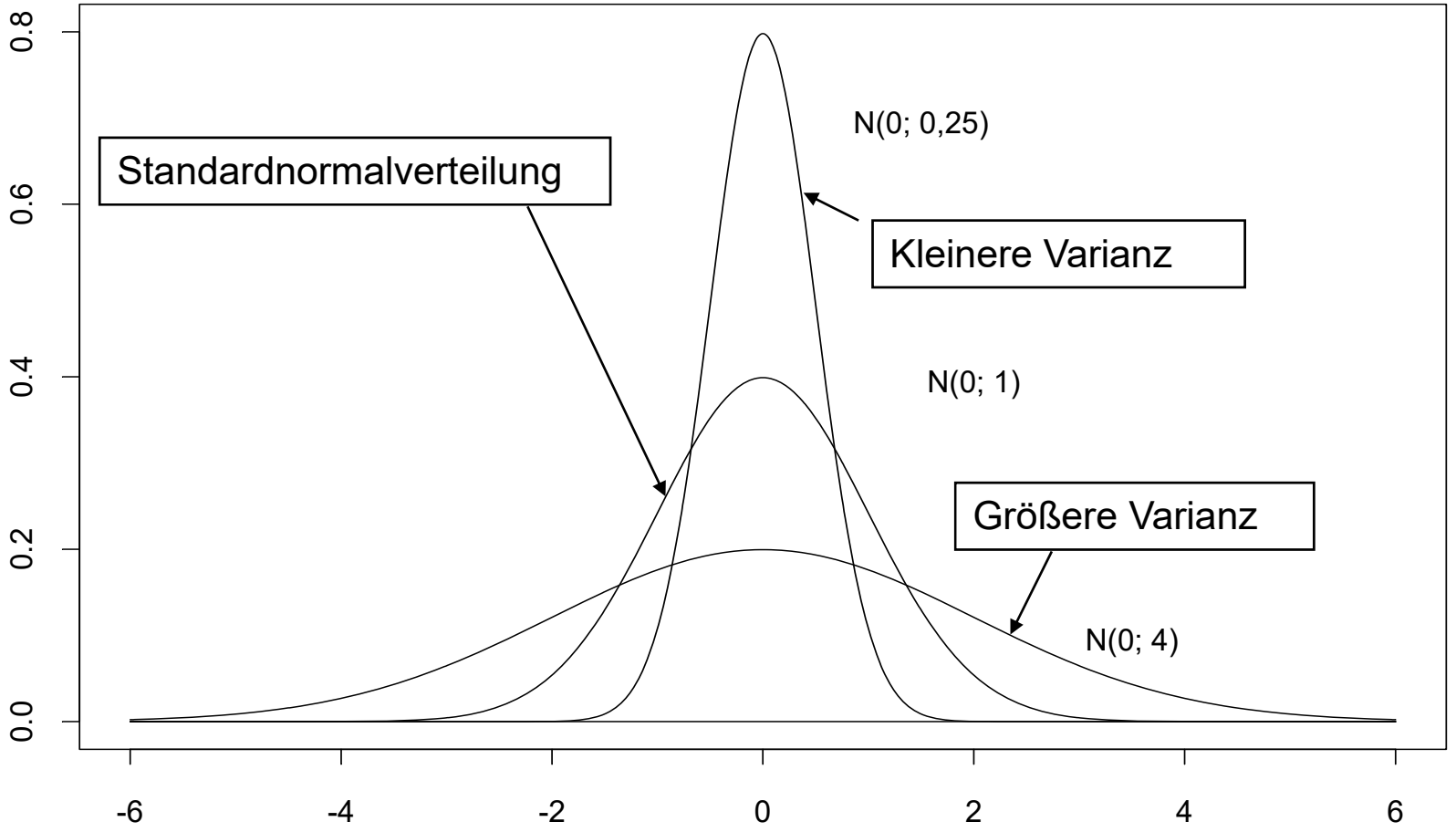
Allgemeine Form der Normalverteilung

Im allgemeinen:

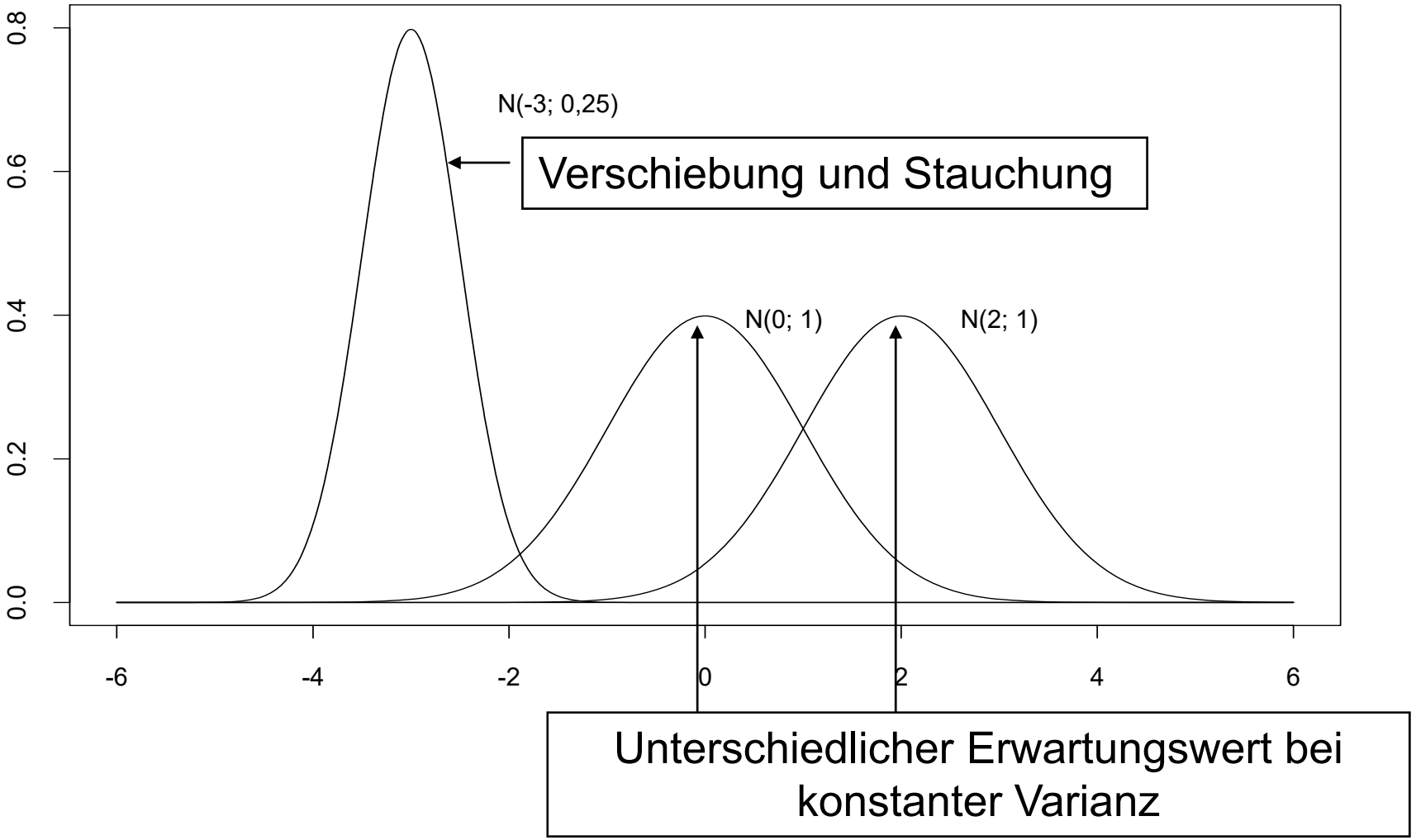
- Normalverteilung mit Erwartungswert μ und Varianz σ^2
- $X \sim N(\mu; \sigma^2)$
- $E(X) = \mu$
- $V(X) = \sigma^2$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Verschiedene Normalverteilungen



Verschiedene Normalverteilungen

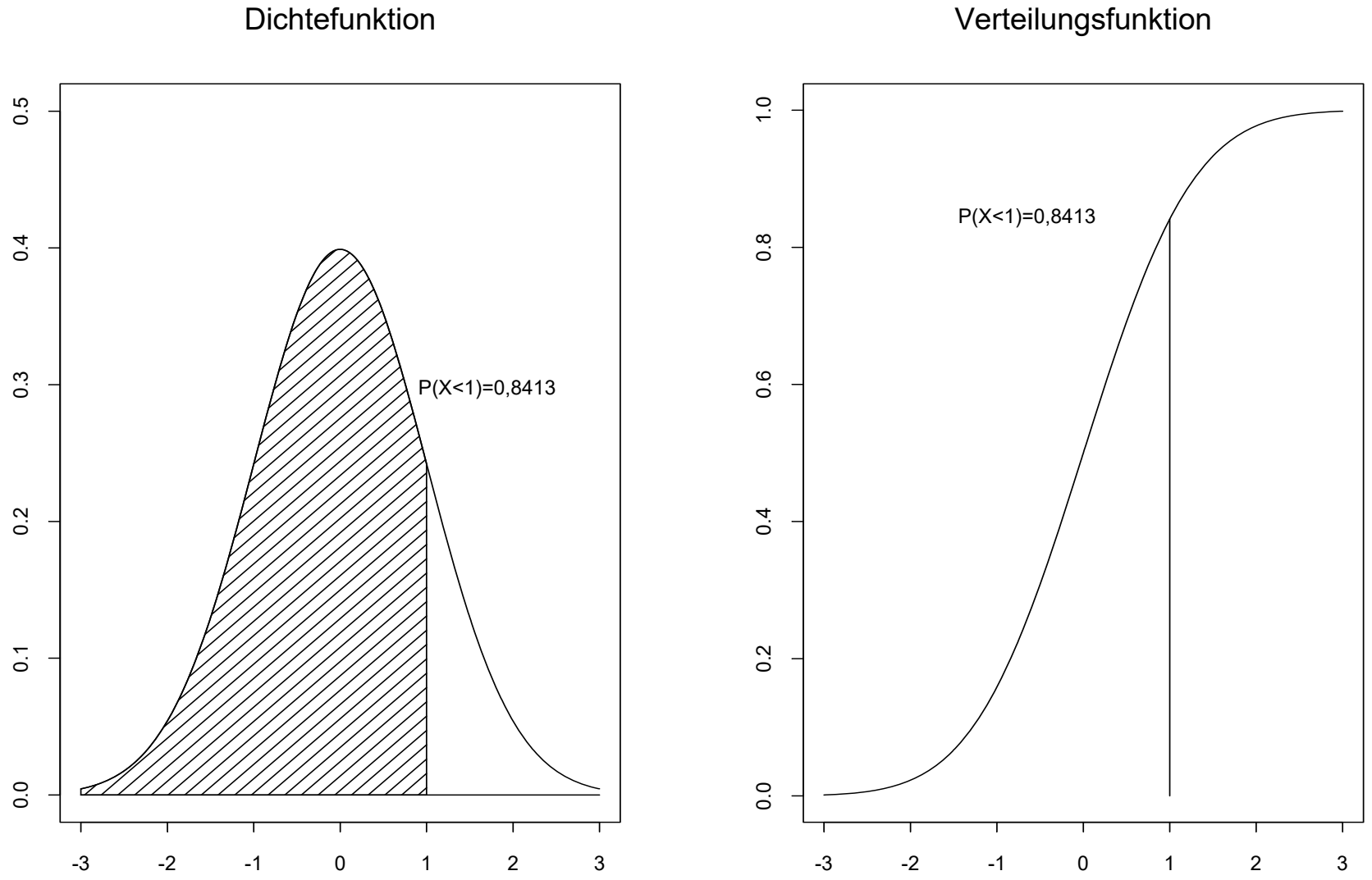


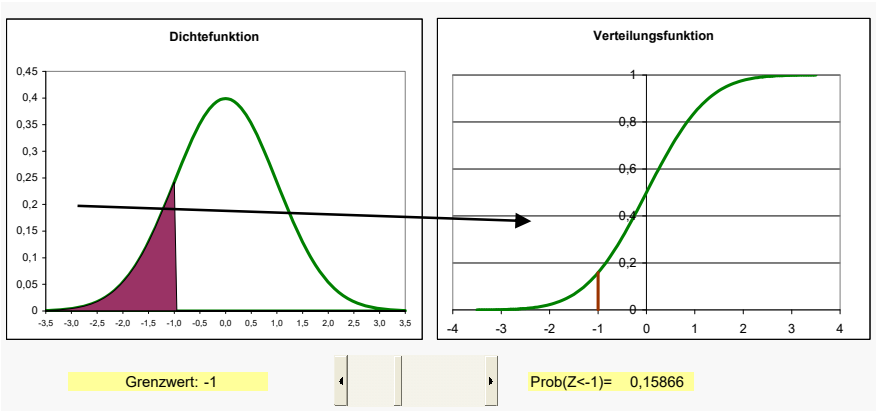
Bestimmung von Wahrscheinlichkeiten

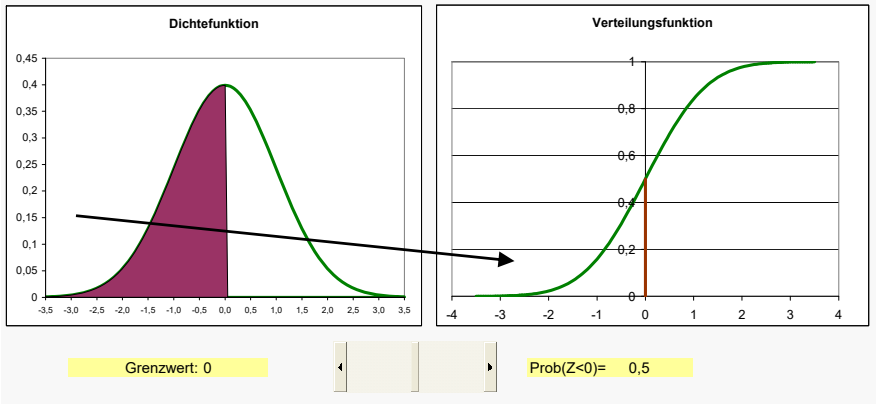
- Die Verteilungsfunktion der Standardnormalverteilung ergibt sich durch Integration der Dichtefunktion
- Man findet Tabellen in fast allen Lehrbüchern
- Unterschiedliche Notation:
  - Bleymüller: $F_N(z)$
  - Schlittgen: $\phi(z)$

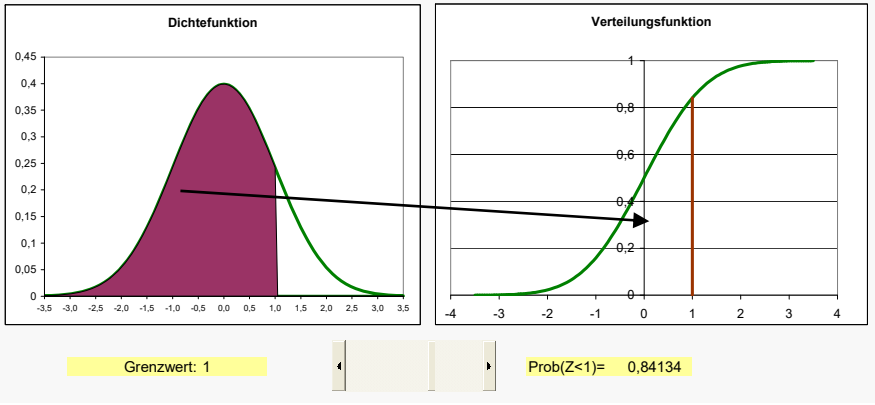
$$F_N(z) = \phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Von der Dichte zur Verteilungsfunktion




Dichtefunktion
Verteilungsfunktion
Grenzwert: -1
Prob(Z<-1)= 0,15866


Dichtefunktion
Verteilungsfunktion
Grenzwert: 0
Prob(Z<0)= 0,5


Dichtefunktion
Verteilungsfunktion
Grenzwert: 1
Prob(Z<1)= 0,84134

Lineartransformation

- Wenn X eine normalverteilte Zufallsvariable ist, dann ist auch $Y=a+bX$ normalverteilt.
- $E(Y)=E(a+bX)=a+bE(X)$
- $V(Y)=V(a+bX)=b^2V(X)$
- Knapp formuliert:
- Sei $X \sim N(\mu; \sigma^2)$
- und $Y=a+bX$ dann gilt $Y \sim N(a+b\mu; b^2\sigma^2)$
- Änderung des Erwartungswertes: Verschiebung (Translation)
- Änderung der Varianz: Dehnung oder Stauchung der Verteilungsform
- Prinzipielle Gestalt der Glockenkurve bleibt erhalten

Standardisierung

- Aus dem vorigen folgt:
- Sei $X \sim N(\mu; \sigma^2)$
- dann gilt für
- $Z=(X- \mu)/\sigma$ … standardisierte Variable
- $Z \sim N(0;1)$
- Durch Anwendung der Standardisierung lässt sich jede Normalverteilung in die Standardnormalverteilung überführen.
- Daher reichen Algorithmen bzw. Tabellen für Wahrscheinlichkeiten der Standardnormalverteilung für alle Fragestellungen

Standardisierung

Anwendungsbeispiel:

X sei die Körpergröße in cm von einer bestimmten Population

Es sei $X \sim N(175; 64)$

dann ist $Z=(X-175)/8$

Frage: $P(167<X<183)= ?$

$P(167<X<183)=$

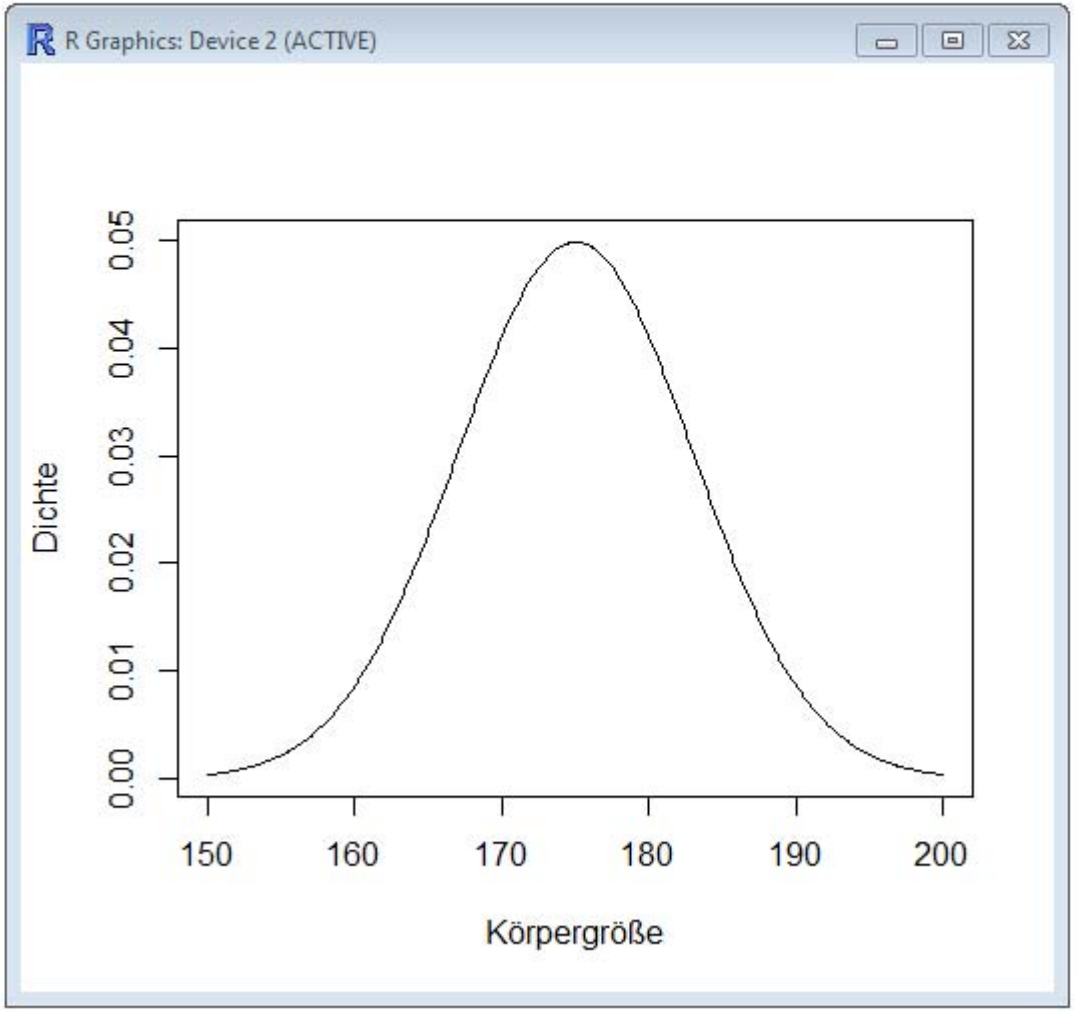
$=P((167-175)/8<Z<(183-175)/8)=$

$=P(-1<Z<1)=0{,}6826$

Bei Kenntnis des Mittelwertes und der Varianz lassen sich unter der Modellannahme, dass das Merkmal normalverteilt ist, die Wahrscheinlichkeit für alle denkmöglichen Fragestellungen mit der Standard-normalverteilung ermitteln.

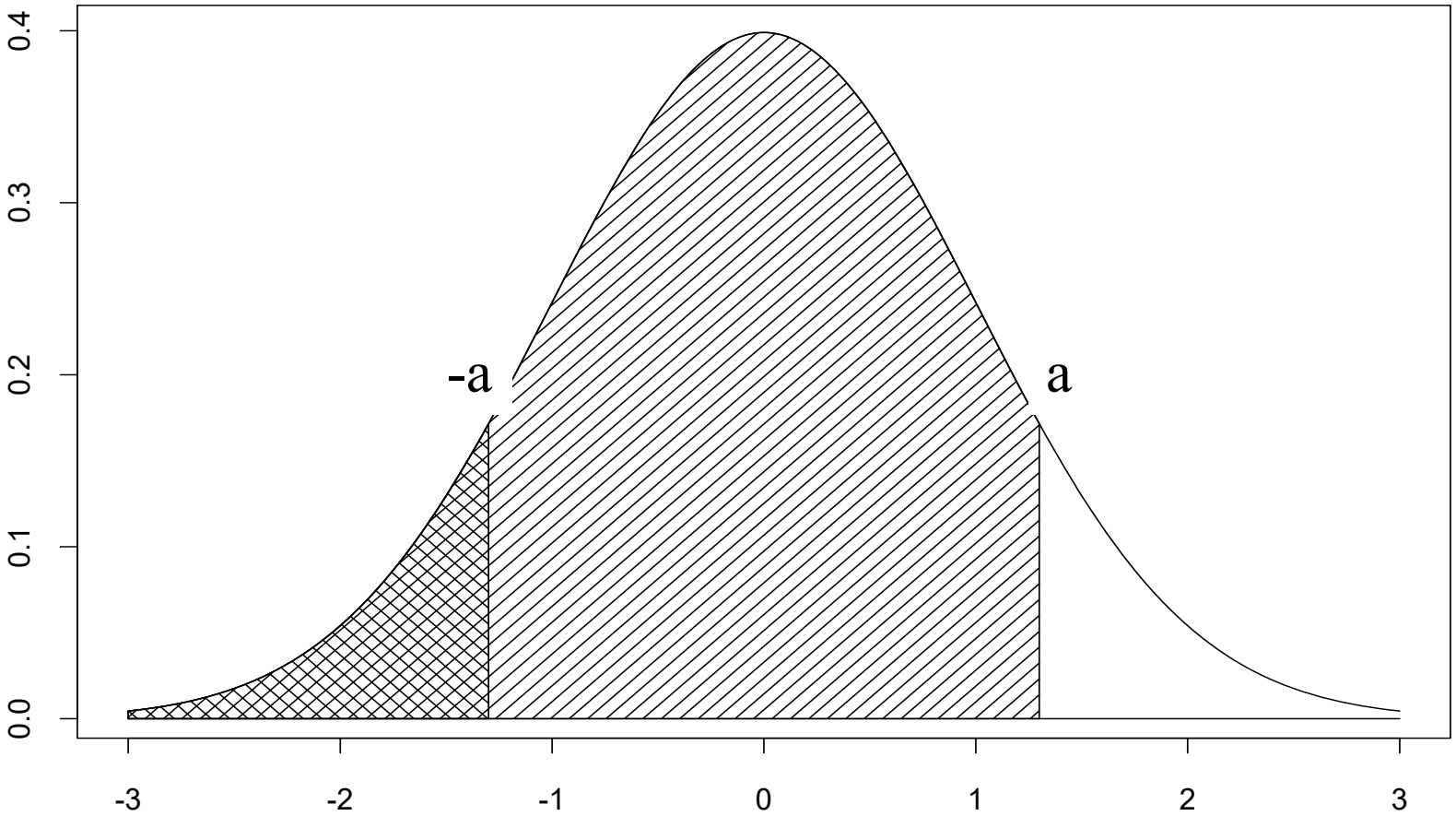
R Console

```
> # Beispiele zur Normalverteilung
>
> xi <- seq(from=150, to=200, length=400)
> plot(xi, dnorm(xi, 175, 8), type="l", xlab="Körpergröße", ylab="Dichte")
> pnorm(183, 175, 8)
[1] 0.8413447
> pnorm(1,0, 1)
[1] 0.8413447
> pnorm(1)
[1] 0.8413447
>
> pnorm(183, 175, 8)-pnorm(167, 175, 8)
[1] 0.6826895
>
```



Ausnützung der Symmetrie um Null

$P(Z < a) = 1 - P(Z < -a)$ oder $P(Z < -a) = 1 - P(Z < a)$

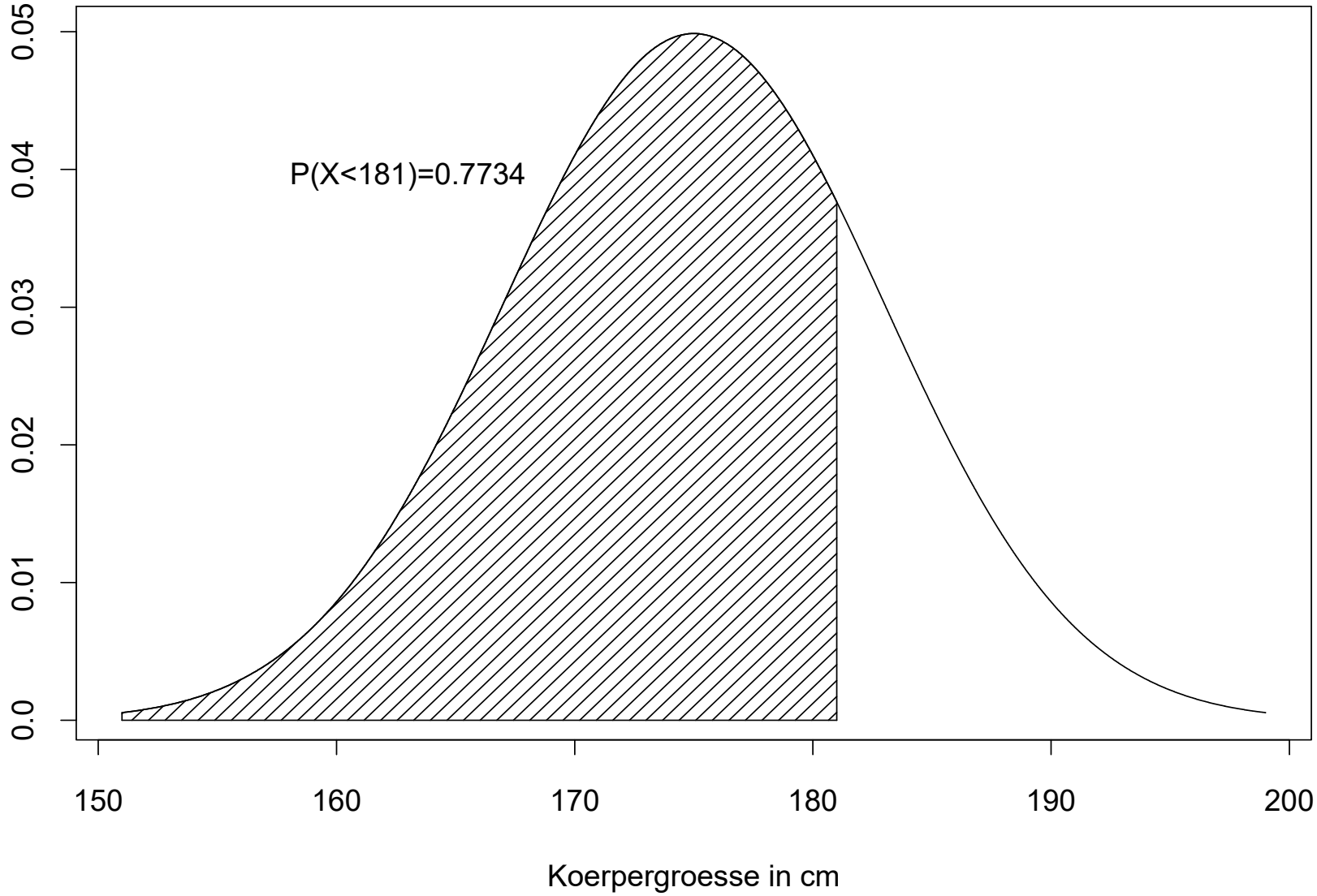


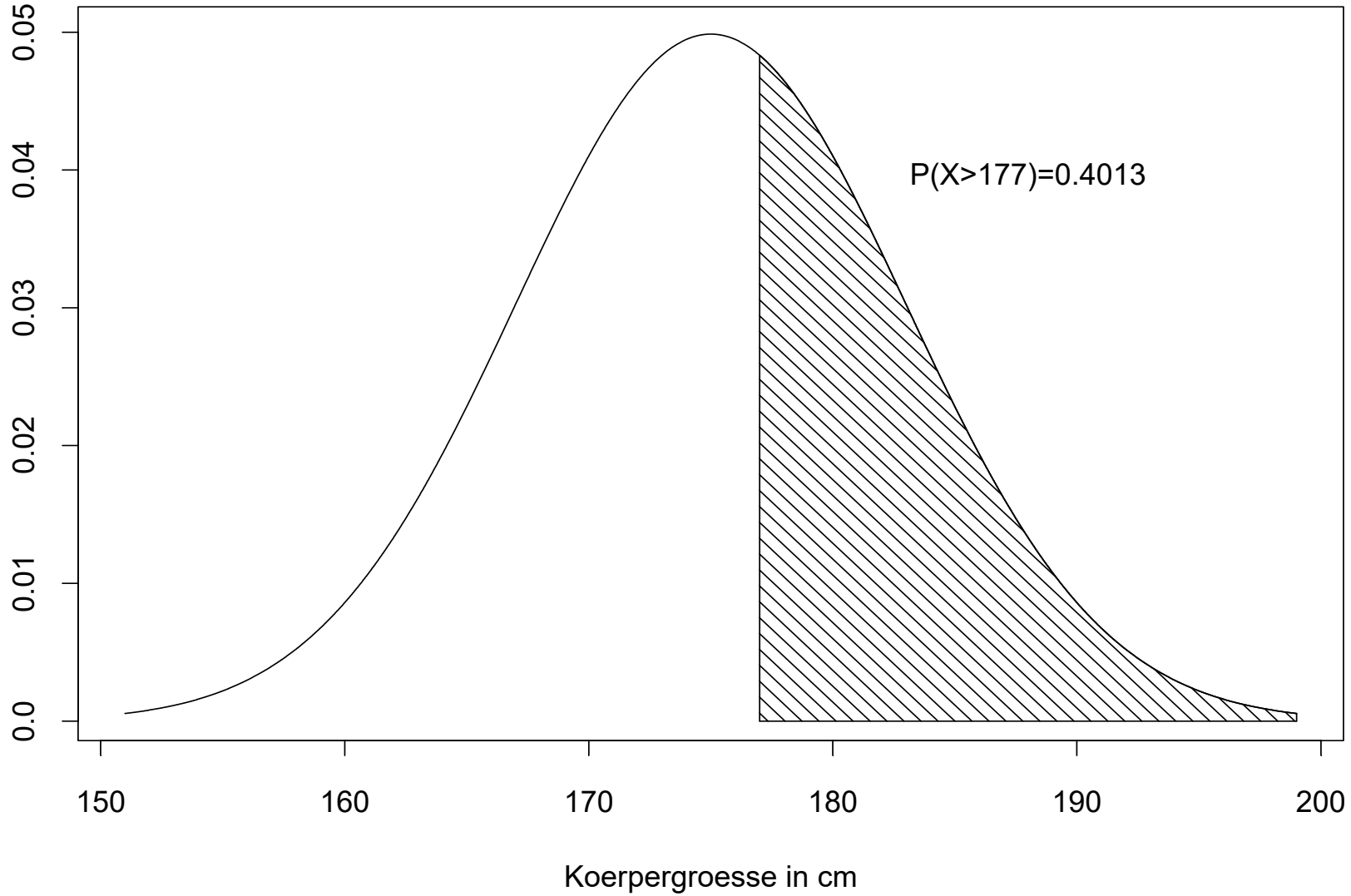
$P(Z>a) = P(Z<-a)$

Beispiel

- Wir wollen für eine Normalverteilung mit Erwartungswert 170 und Standardabweichung 16 die Wahrscheinlichkeit einen Wert kleiner als 180 zu erhalten ermitteln.
- $P(X<180) = $
  $=P(Z<(180-170)/16)=P(Z<0{,}625)=F_N(0{,}625)=$
  $=0{,}734$

```
> pnorm(180, 170, 16)   #Probability X <180 (mu=170, sigma=16)
[1] 0.7340145
> 1-pnorm(180, 170, 16)    # Probability X >180 (mu=170,
  sigma=16)
[1] 0.2659855
```


P(X<181)=0.7734
0.0
0.01
0.02
0.03
0.04
0.05
150
160
170
180
190
200
Koerpergroesse in cm


P(X>177)=0.4013
0.0
0.01
0.02
0.03
0.04
0.05
150
160
170
180
190
200
Koerpergroesse in cm

Beispiel

IQ-Test

E(X)= 100
V(X)= 225
σ(X)= 15

4-Sigma Gesellschaft Personen mit einem IQ über 160

P(X>100+4*σ)=?

100+4*s= 160
P(X>160)= 0,00317%

Bei 100.000 3,17
1 von 31.574

```
> # Über 4 sigma
> 1-pnorm(4)
[1] 3.167124e-05
>
```

In Österreich leben: 8.032.926 Menschen
Österreicher in 4-Sigma 254

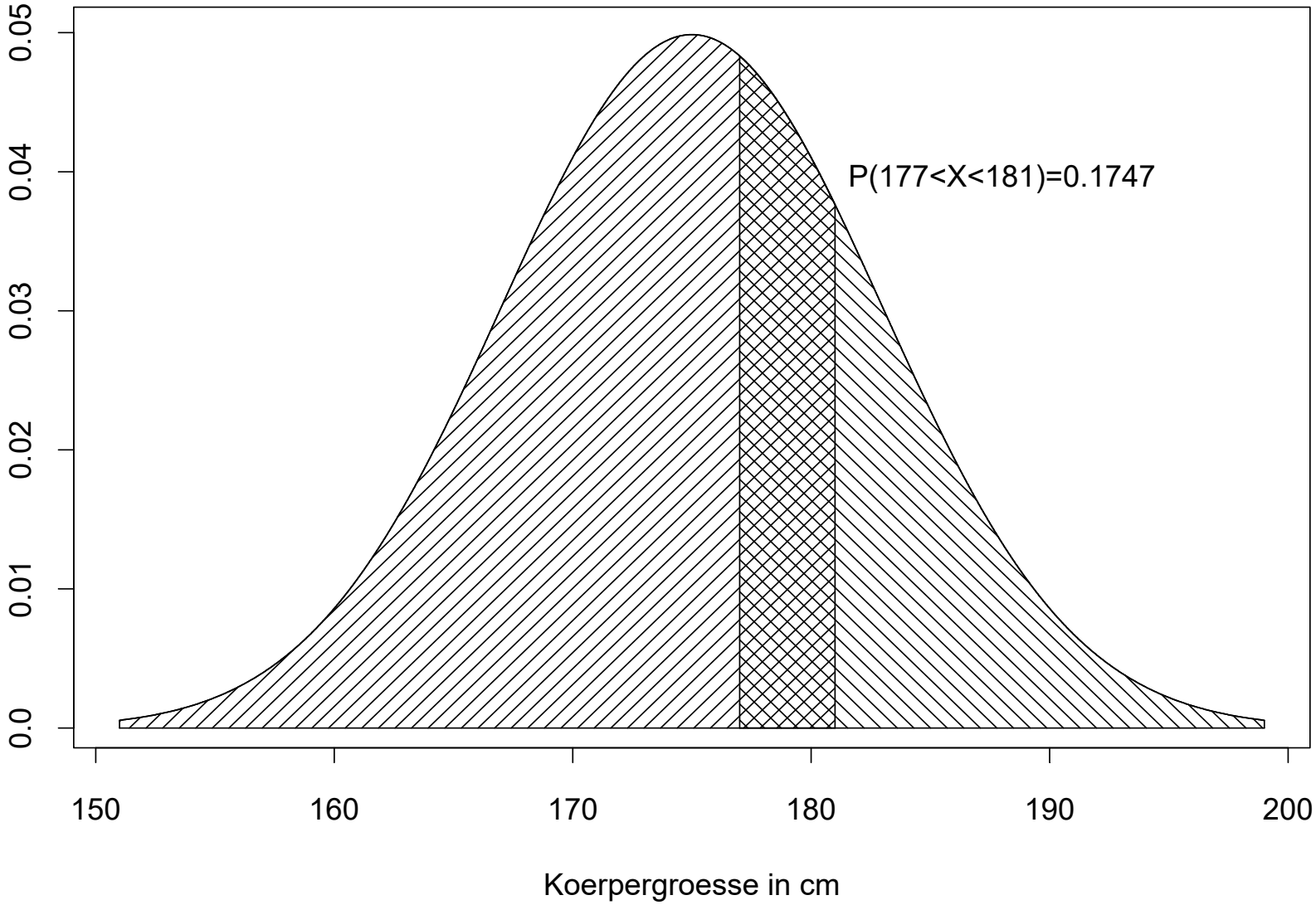
Wahrscheinlichkeiten für Intervalle

- P(a<X<b) = P(X<b) – P(X<a)
- X~N(175; 64)
- P(177<X<181) = ?
- P(X<181) - P(X<177) =
- =P(Z<(181-175)/8) - P(Z<(177-175)/8) =

= ϕ(0,75) - ϕ(0,25) =

- = 0,7734 - 0,5987 = 0,1747

```
> #Wahrscheinlichkeiten für Intervalle
> pnorm(181, 175, 8)-pnorm(177, 175, 8)
[1] 0.1746663
>
```


0.05
0.04
0.03
0.02
0.01
0.0
P(177<X<181)=0.1747
150
160
170
180
190
200
Koerpergroesse in cm

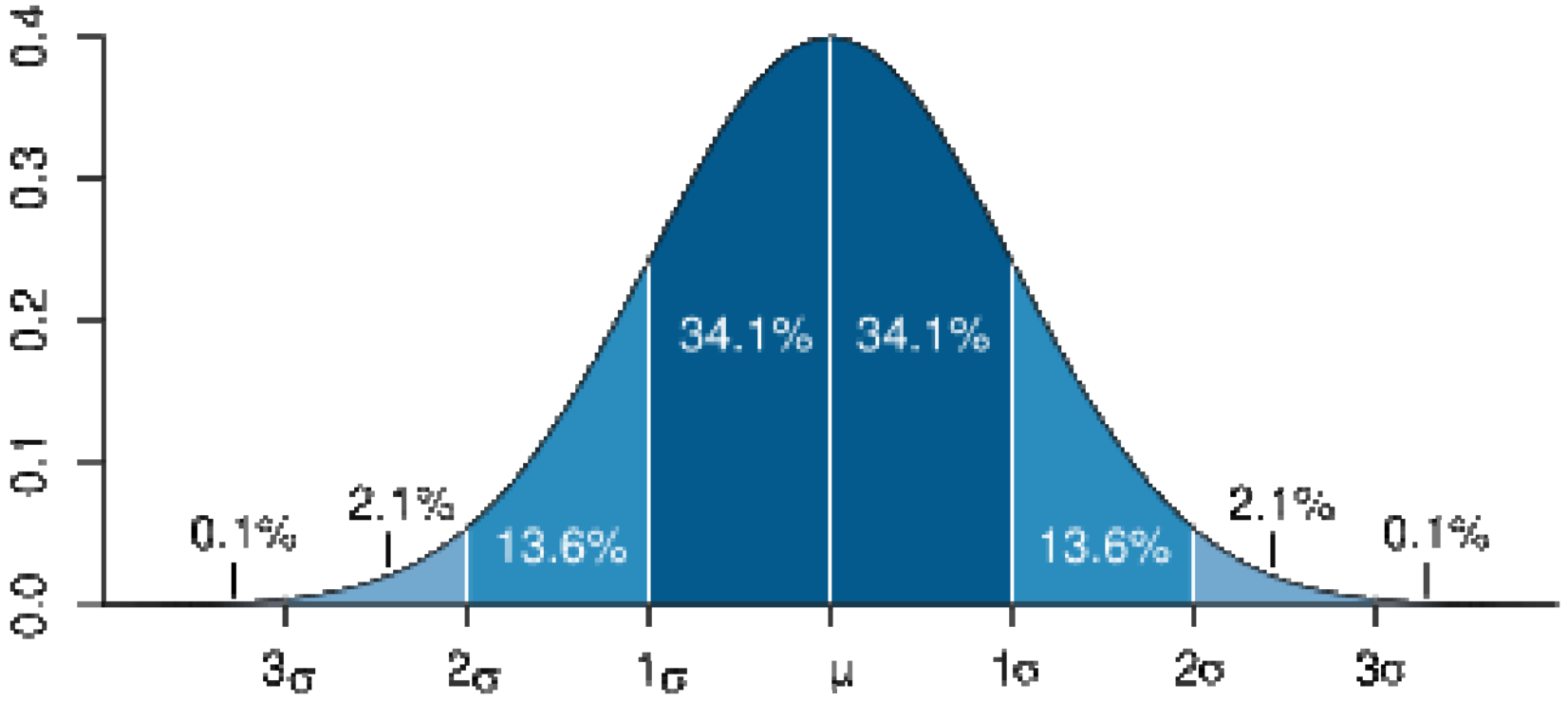
Symmetrische Intervalle

- $P(-1<Z<1)=?$
- $P(-1<Z<1)= \phi(1) - \phi(-1)= 0{,}8413-0{,}1587 = 0{,}6826$
- $P(-a<Z<a)=P(Z<a)-P(Z<-a)=\phi(a)-(1-\phi(a))=2\phi(a)-1$

$$P(-a<Z<a)=2\phi(a)-1$$

- $P(-1<Z<1)= 2*\phi(1) -1=2*0{,}8413-1=0{,}6826$

Probability of Symmetric Intervals



General guideline for data analysis:

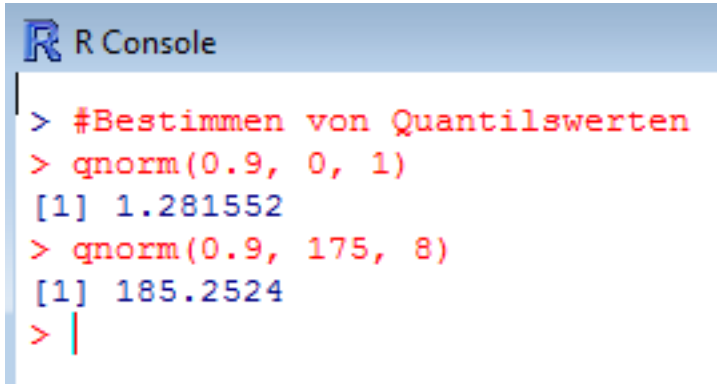
If the distribution is bell-shaped about 95% of the observations lie within the range of mean plus/minus two times the standard deviation

X~N(175; 64)

- Wie groß ist die Wahrscheinlichkeit, dass eine Person maximal 8 cm vom Erwartungswert abweicht?
- $P(167<X<183) = \phi((183-175)/8) - \phi((167-175)/8)$
- $= \phi(1) - \phi(-1) = 0{,}8413-0{,}1587 = 0{,}6827$
- $P(167<X<183) = 2* \phi((183-175)/8) -1=$
  $= 2* \phi(1)-1 = 2*0{,}8413-1 = 0{,}6827$

Inverse Fragestellung X~N(175; 64)

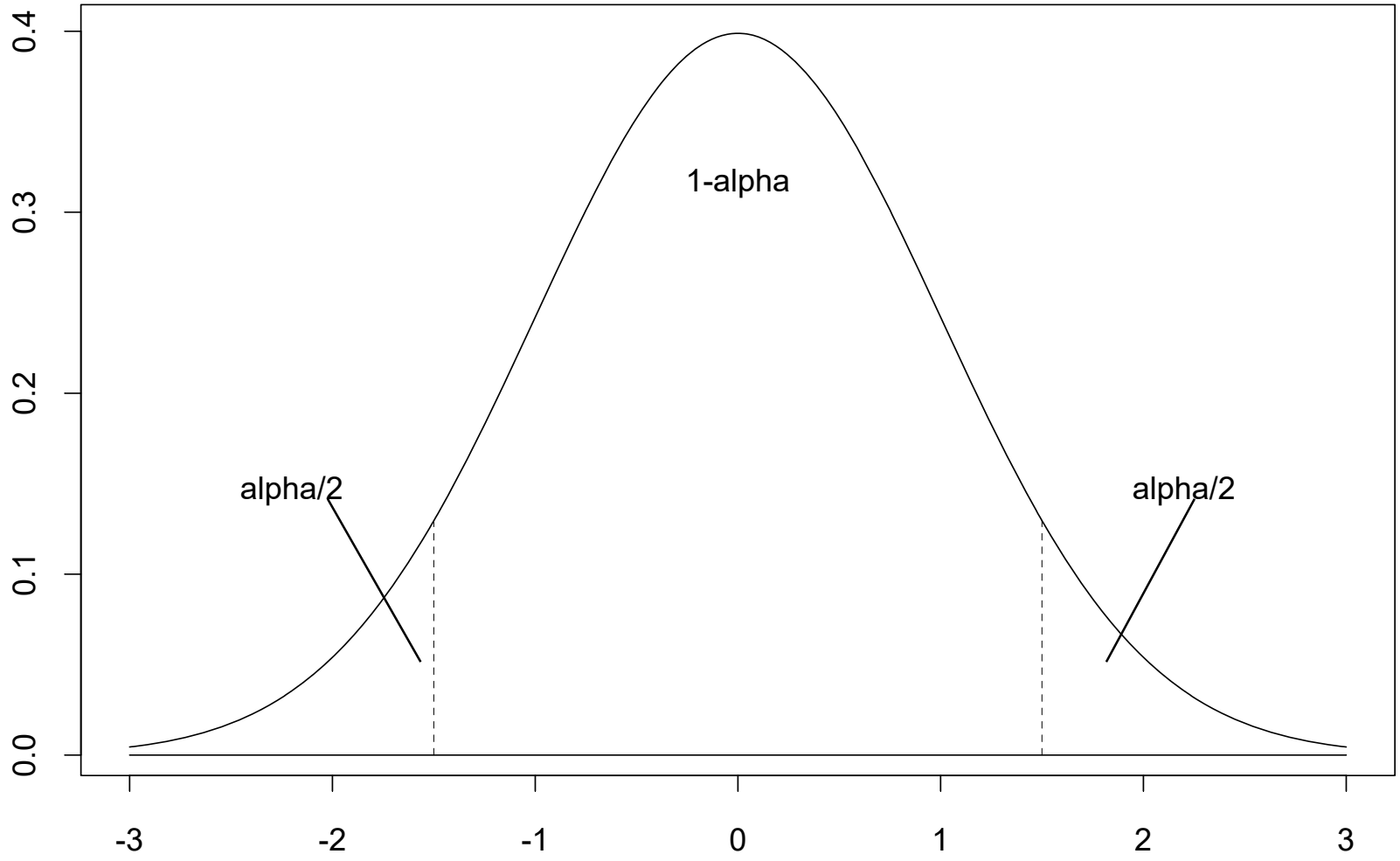
- Gesucht sind Quantilswerte z_α für die bestimmte Wahrscheinlichkeitsaussagen gelten:
- $P(Z< z_\alpha) = \phi(z_\alpha) = \alpha$
- $P(Z<z_{0,9}) = 0,9 \implies z_\alpha = ?$
- Nachschlagen in der Tabelle: ==> $z_{0,9} = 1,2816$
- Gesucht ist jene Körpergröße x_α für die gilt, dass die Wahrscheinlichkeit $P(X<x_\alpha)=0,9$
- Lösung:
- $x_\alpha = \mu + \sigma z_\alpha$
- $x_\alpha = 175 + 1,2816*8 = 185,25$



Zentrale Schwankungsintervalle (Streubereiche)

- symmetrische Intervalle um den Erwartungswert $[\mu-c;\mu+c]$
- Von Interesse sind Aussagen der Form
  a) $P(\mu-c < X < \mu+c) = ?$
  b) $P(\mu-? < X < \mu+?) = 1-\alpha$
- Beispiel für a) Wie groß ist die Wahrscheinlichkeit, dass eine Person maximal 8 cm vom Erwartungswert abweicht?
- Beispiel für b) Wie groß ist das symmetrische Intervall in welchem Personen mit einer Wahrscheinlichkeit $1-\alpha$ liegen?
- Wir ordnen dem zentralen Schwankungsbereich die Wahrscheinlichkeit $1-\alpha$ zu. Dadurch kommt außerhalb des Bereichs an jedem Ende eine Randwahrscheinlichkeit von $\alpha/2$ zustande.

Konzept zentraler Schwankungsintervalle



Zentrale Schwankungsintervalle

- Sei $X \sim N(\mu,\sigma^2)$ so ergibt sich das zentrale Schwankungsintervall,welches eine Wahrscheinlichkeit von $1-\alpha$ abdeckt durch:
- $[\mu - z_{1-\alpha/2}\sigma ; \mu + z_{1-\alpha/2}\sigma]$
- bzw.
- $P(\mu - z_{1-\alpha/2}\sigma < X < \mu + z_{1-\alpha/2}\sigma) = 1-\alpha$
- Für $\alpha=0,1$ ($\alpha=0,05$; $\alpha=0,01$) ergibt sich aus der Tabelle für $z_{1-\alpha/2} = 1,6449\ (1,96;\ 2,5758)$
- d.h. $P(\mu - 1,6449 < Z < \mu + 1,6449) = 0,9$
  $P(\mu - 1,96 < Z < \mu + 1,96) = 0,95$
  $P(\mu - 2,5758 < Z < \mu + 2,5758) = 0,99$

X~N(175; 64)

- Gesucht ist ein zentrales Schwankungsintervall, das eine Wahrscheinlichkeit von 0,95 aufweist
- $P(\mu - z_{1-\alpha/2}\sigma < X < \mu + z_{1-\alpha/2}\sigma) = 1-\alpha$
- $\alpha = 0{,}05 \quad 1-\alpha/2 = 0{,}975$
- P(175-1,96*8 < X < 175 + 1,96*8) = 0,95
- P(159,32 < X < 190,68) = 0,95
- Falls man eine höhere Wahrscheinlichkeit anstrebt wird das Intervall größer:
- P(175- 2,5758 *8 < X < 175 + 2,5758 *8) = 0,99
- P(154,39 < X < 195,61) = 0,99

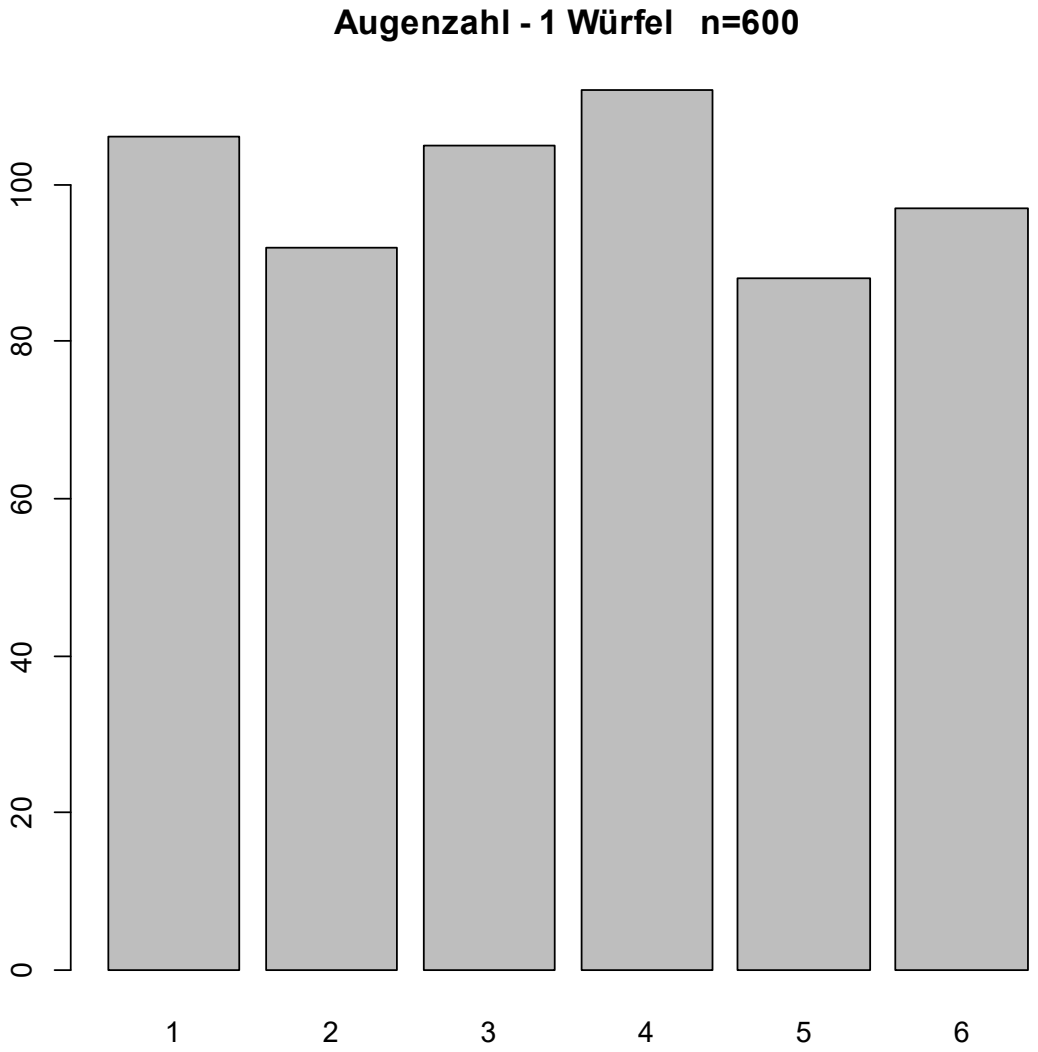
Zentraler Grenzwertsatz

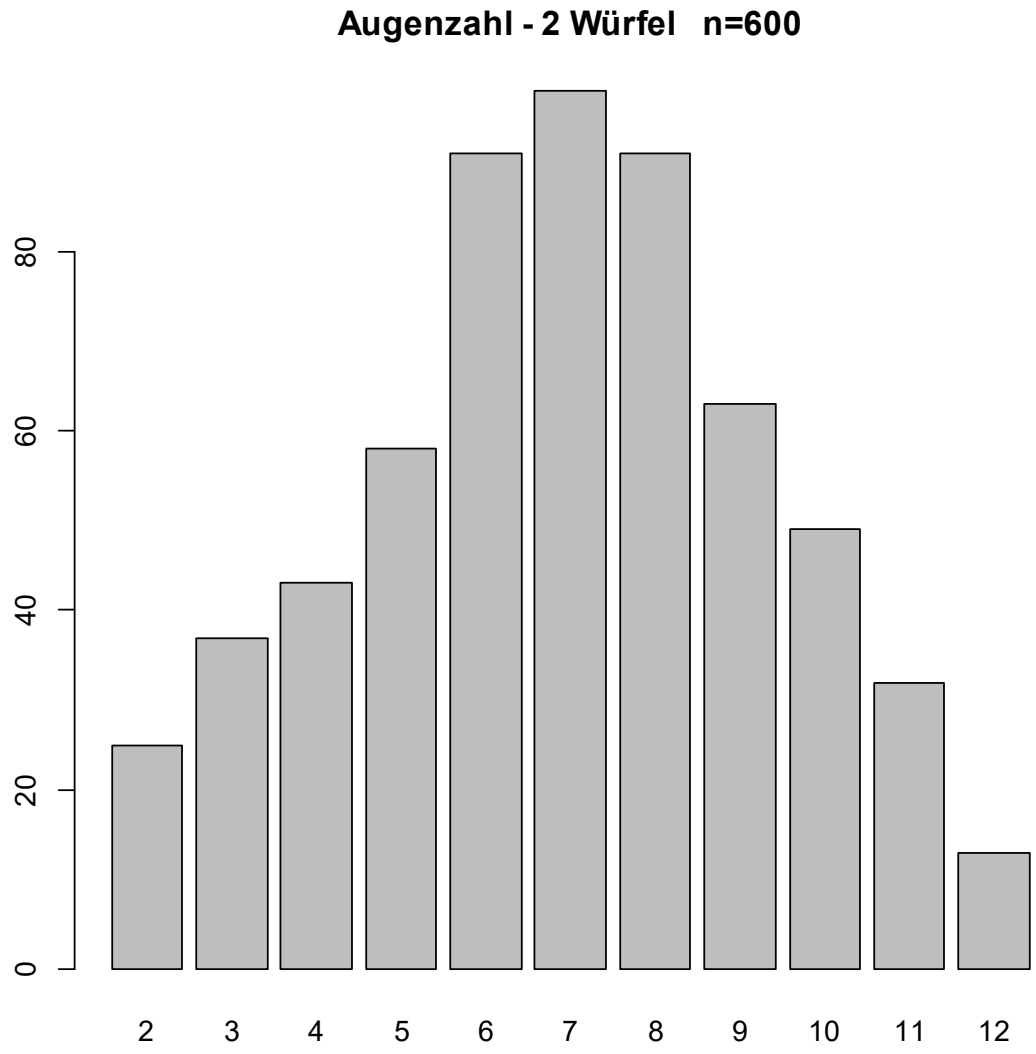
- Die Normalverteilung verdankt ihre universelle theoretische und praktische Bedeutung dem zentralen Grenzwertsatz. Unabhängig von der konkreten Ausgangsverteilung konvergiert nämlich die Verteilungsfunktion einer Summe gegen die Normalverteilung. (sehr grob formuliert)
- Ist die Anzahl der Summanden (n) hinreichend groß, so kann in der Praxis die Verteilung einer Summe durch die Normalverteilung approximiert werden.
- Die Frage, ab wann n hinreichend groß ist, hängt von der gewünschten Genauigkeit und der Form der Ausgangsverteilung ab.

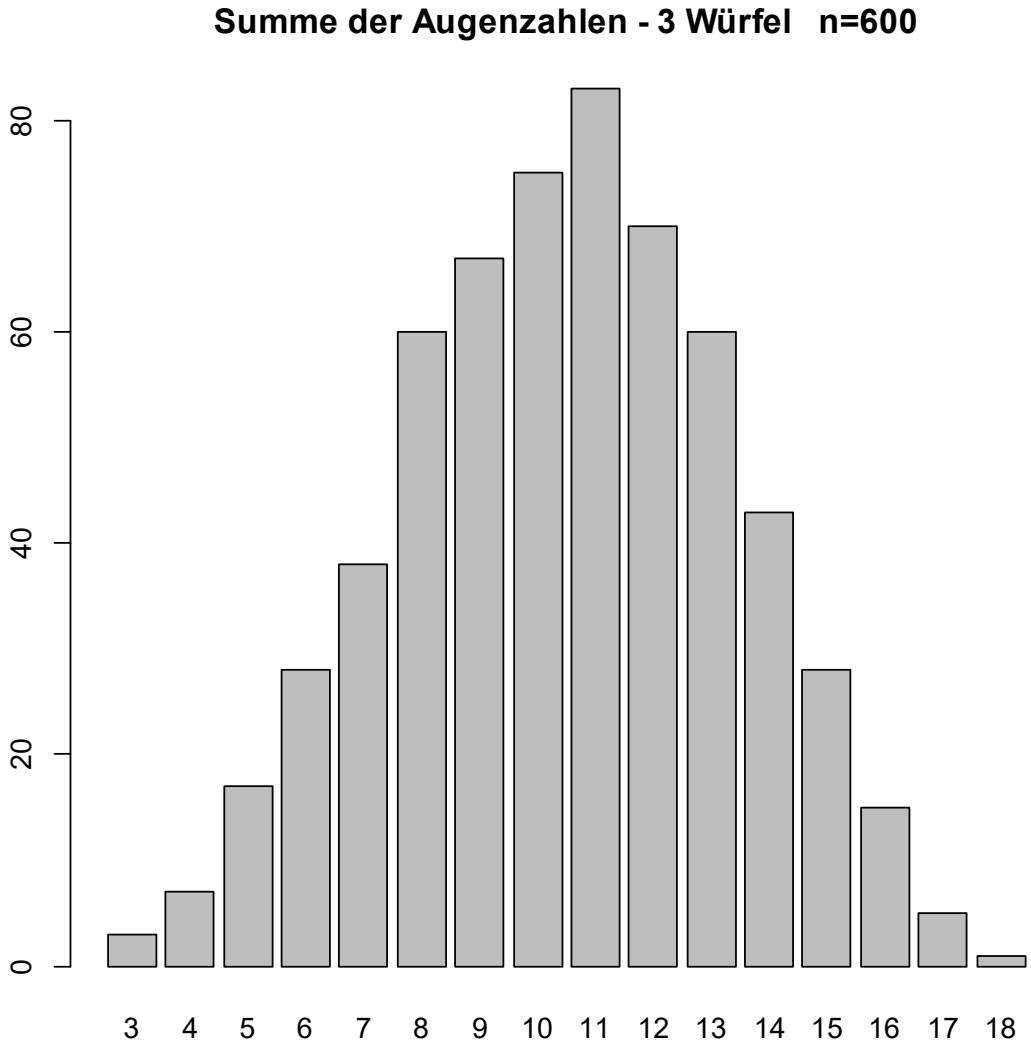
Verteilung von Summen

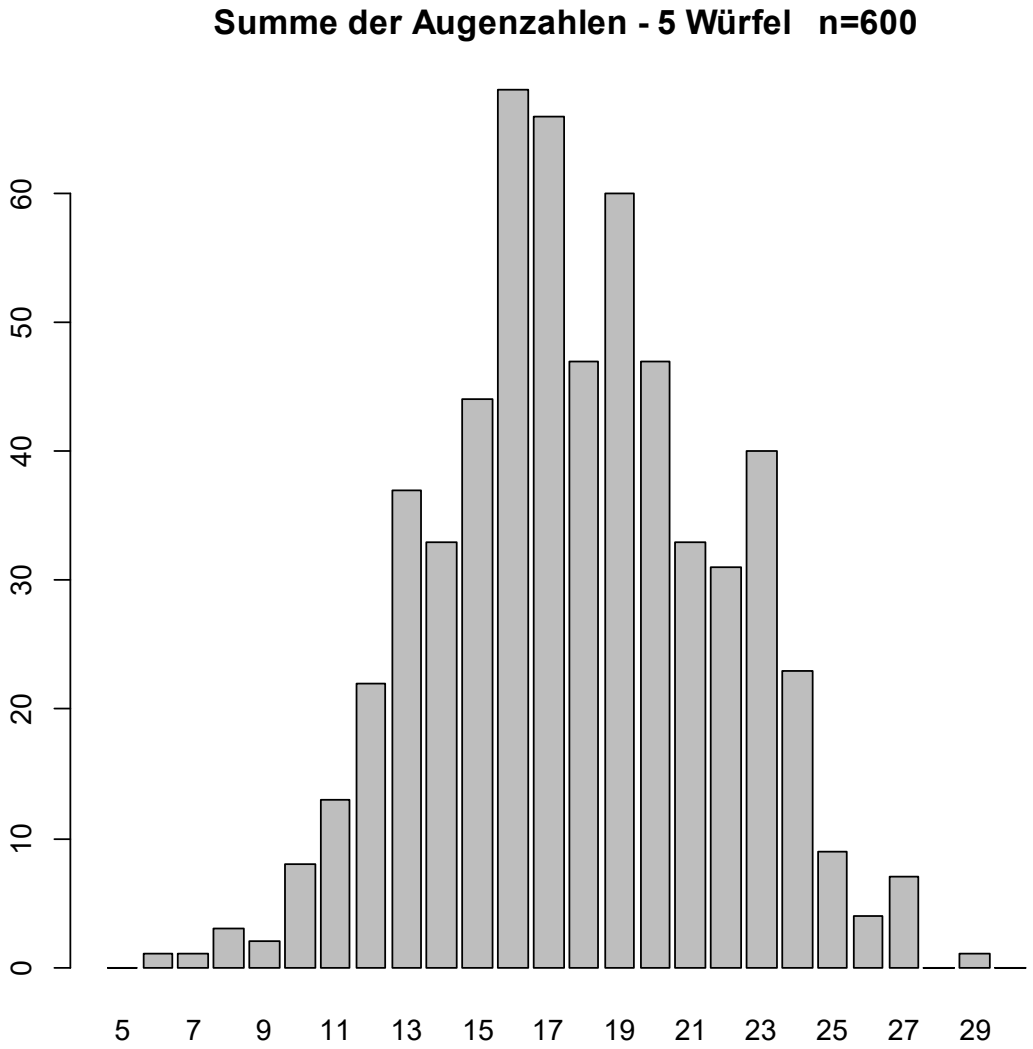
- Beispiel: Würfelwurf
- Frage:

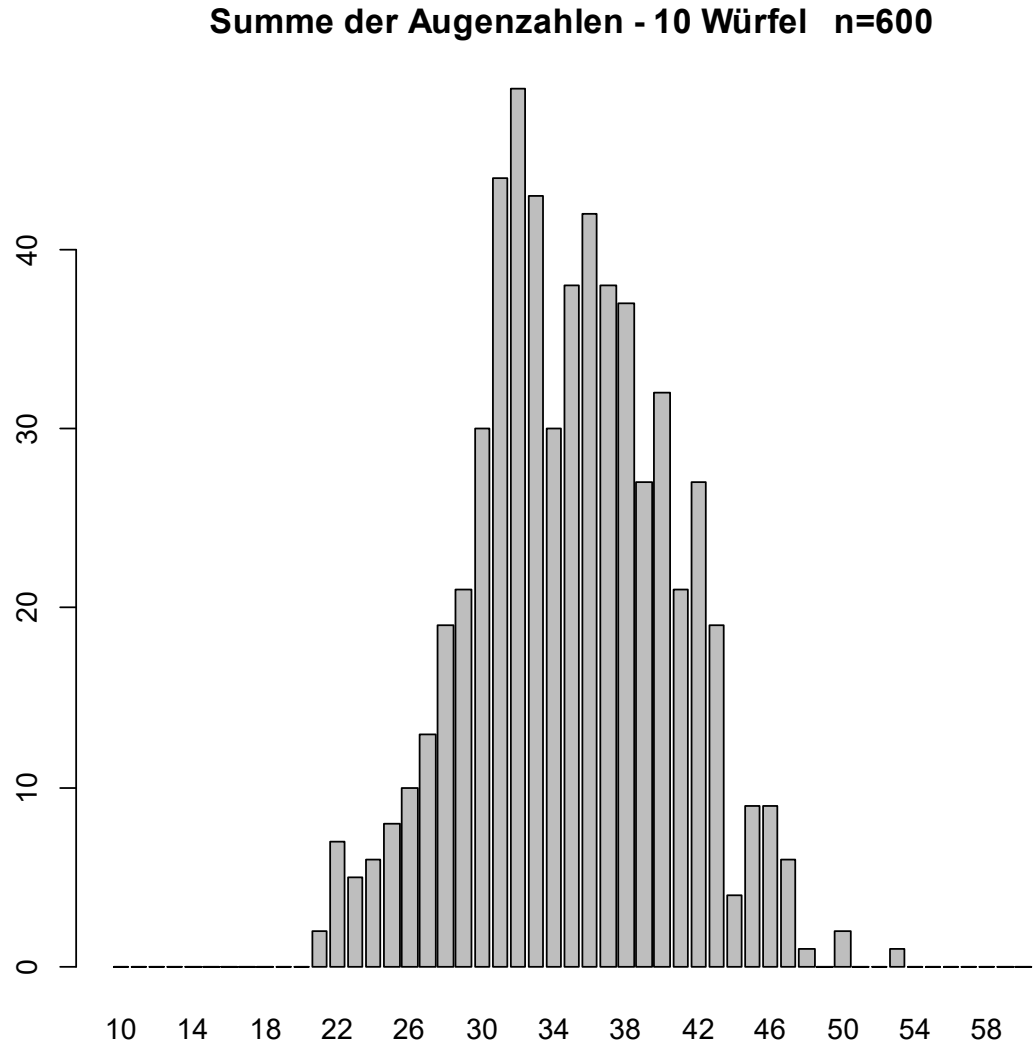
  Wie verhält sich die Verteilung der Augensumme von x-Würfeln bei wachsendem n?
- Zur Beantwortung führen wir ein Simulationsexperiment durch.
  - 600 Würfe mit 1 Würfel
  - 600 Würfe mit 2 Würfel
  - 600 Würfe mit 3 Würfel
  - etc.

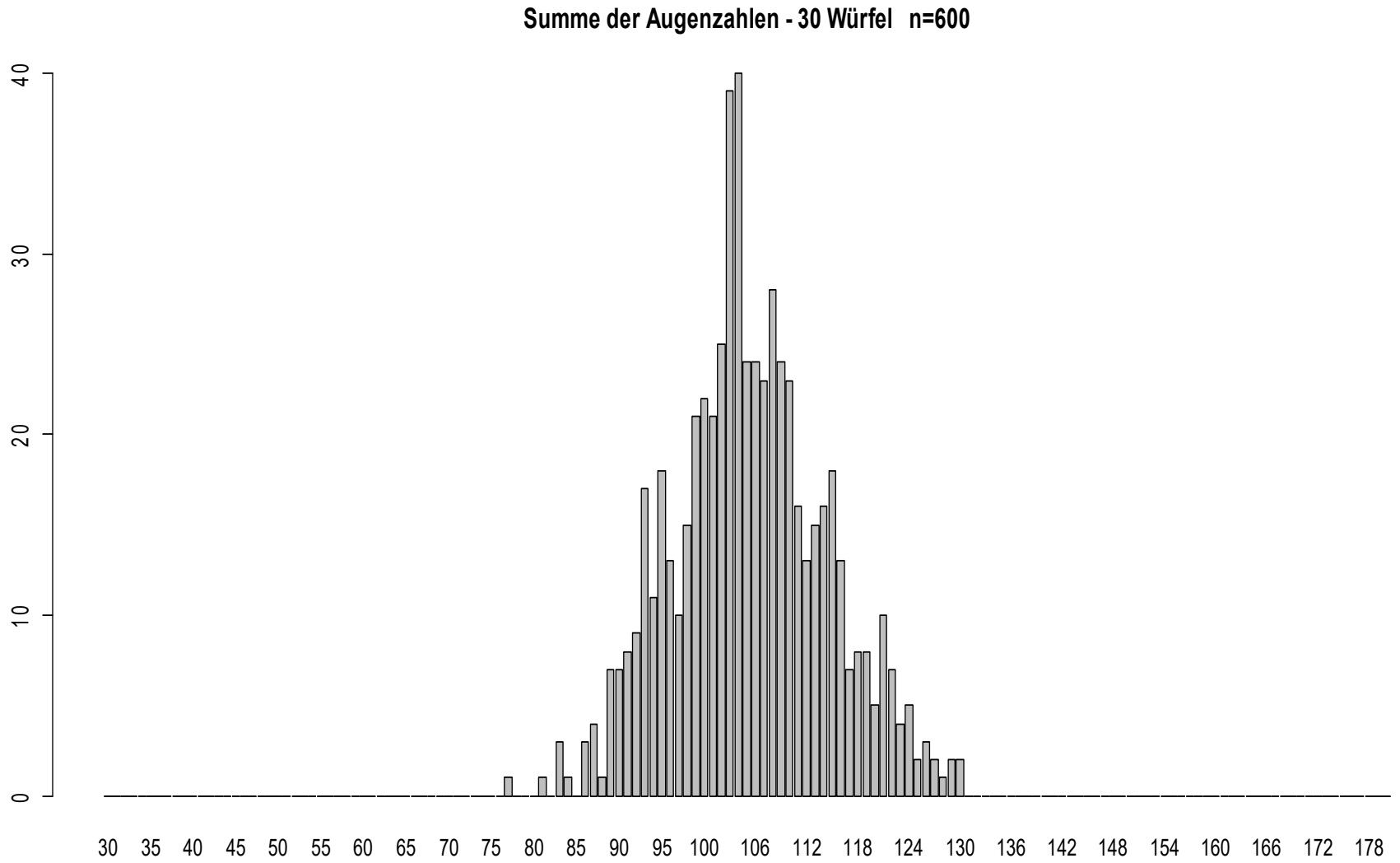

Augenzahl - 1 Würfel n=600
0
20
40
60
80
100
1
2
3
4
5
6


Augenzahl - 2 Würfel n=600
0
20
40
60
80
2
3
4
5
6
7
8
9
10
11
12


Summe der Augenzahlen - 3 Würfel n=600
0
20
40
60
80
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18


Summe der Augenzahlen - 5 Würfel n=600
0
10
20
30
40
50
60
5
7
9
11
13
15
17
19
21
23
25
27
29


Summe der Augenzahlen - 10 Würfel n=600
0
10
20
30
40
10
14
18
22
26
30
34
38
42
46
50
54
58


Summe der Augenzahlen - 30 Würfel n=600
0
10
20
30
40
30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 106 112 118 124 130 136 142 148 154 160 166 172 178

Theoretisches Basiswissen

Seien $X_1, X_2, ..., X_n$ identisch verteilte, unabhängige Zufallsvariablen mit

$$E(X_i) = \mu \qquad \text{und} \qquad V(X_i) = \sigma^2 > 0$$

Dann gilt für die Verteilung Summe

$$S_n = X_1 + X_2 + ... + X_n$$

Erwartungswert $E(S_n) = n\mu$

und Varianz $V(S_n) = n\sigma^2$.

Hinweis:

Nur bei Unabhängigkeit gilt Varianz der Summe ist gleich die Summe der Varianzen!

Zentraler Grenzwertsatz

Seien $X_1, X_2, ..., X_n$ identisch verteilte, unabhängige Zufallsvariablen mit

$$E(X_i) = \mu \qquad \text{und} \qquad V(X_i) = \sigma^2 > 0$$

Dann konvergiert die Verteilung der standardisierten Summe

$$Z_n = \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$$

mit wachsendem n gegen eine Normalverteilung mit Erwartungswert $E(Z_n) = 0$ und Varianz $V(Z_n) = 1$.

$$Z_n \sim N(0, 1^2)$$

Simulation.xls

Theoretische Verteilung:

X	Prob(X=x)	Prob(X≤x)
0	0,4	0,4
1	0,3	0,7
2	0,2	0,9
3	0,1	1

Empirischische Verteilung:

X	Anzahl	Rel. Häuf.
0	34	0,34
1	29	0,29
2	19	0,19
3	18	0,18
	100	
Summe:	121	

Eine Simulation

Index	Zufallszahl	Nachfrage
1	0,63239343	1
2	0,18339057	0
3	0,63104985	1
4	0,59476274	1
5	0,27825077	0
6	0,99244714	3
7	0,61025841	1
8	0,07108634	0
9	0,01420279	0
10	0,8477959	2
11	0,3833616	0
12	0,75876094	2
13	0,95497708	3
14	0,90482995	3
15	0,93719852	3
16	0,46728341	1
17	0,80837425	2
18	0,13575523	0
19	0,57372722	1
20	0,59950231	1
38	0,98817026	3

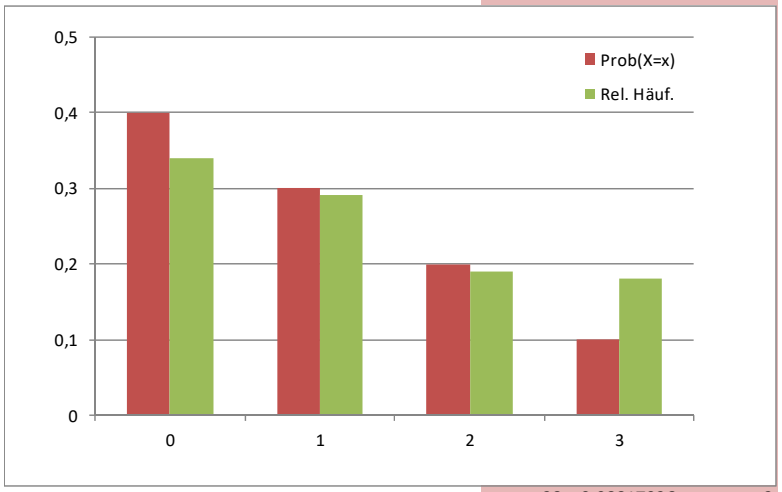
Wiederholte Simulationen

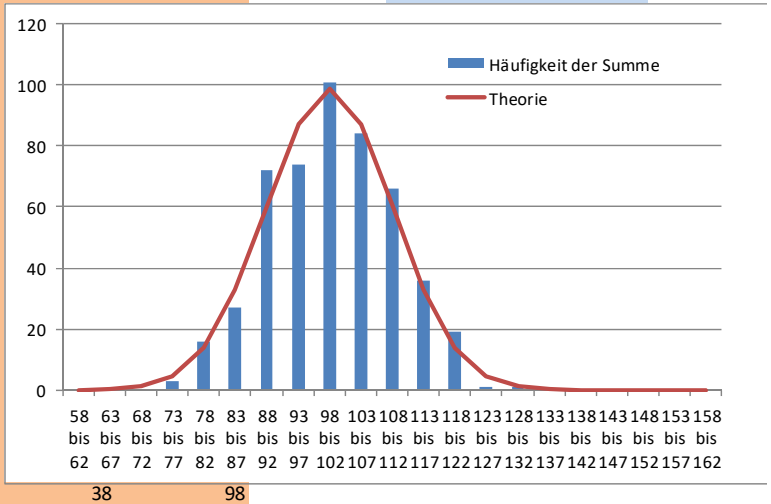
Index	Summe
1	121
2	99
3	116
4	102
5	92
6	102
7	93
8	86
9	100
10	84
11	109
12	114
13	95
14	94
15	103
16	110
17	114
18	129
19	100
20	107
38	98

58	62
63	67
68	72
73	77
78	82
83	87
88	92
93	97
98	102
103	107
108	112
113	117
118	122
123	127
128	132
133	137
138	142
143	147
148	152
153	157

Verteilung der Summe

Bereich	Häufigkeit	Theorie
58 bis 62	0	0,0
63 bis 67	0	0,2
68 bis 72	0	1,2
73 bis 77	3	4,6
78 bis 82	16	13,9
83 bis 87	27	32,8
88 bis 92	72	60,5
93 bis 97	74	87,3
98 bis 102	101	98,7
103 bis 107	84	87,3
108 bis 112	66	60,5
113 bis 117	36	32,8
118 bis 122	19	13,9
123 bis 127	1	4,6
128 bis 132	1	1,2
133 bis 137	0	0,2
138 bis 142	0	0,0
143 bis 147	0	0,0
148 bis 152	0	0,0
153 bis 157	0	0,0





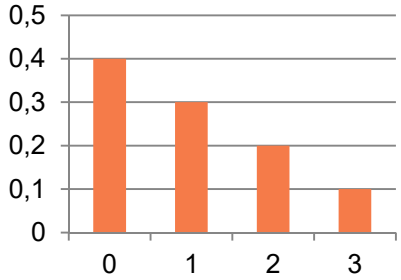
Erkenntnis

- Wir haben ein Merkmal, das eindeutig nicht normalverteilt ist.
- Wenn wir viele Stichproben ziehen und uns dabei von jeder Stichprobe die Merkmalsumme merken, beobachten wir, dass die Verteilung der Merkmalsumme (bzw. auch der Mittelwerte) sich sehr gut an eine Normalverteilung annähert.

Beispiel

- Wahrscheinlichkeitsfunktion für die Anzahl der Verkäufe pro Tag eines bestimmten Produkts sei bekannt

X	0	1	2	3
Prob	0,4	0,3	0,2	0,1



- Wie ist die Anzahl der Verkäufe pro 100 Tage (X100) verteilt, wenn die einzelnen Verkaufstage als unabhängig angesehen werden können?
- Wie groß ist die Wahrscheinlichkeit, dass X100 > 120 ist?
  $X100=X_1+X_2+...X_{100}$

Beispiel (Fortsetzung)

X	0	1	2	3	
Prob	0,4	0,3	0,2	0,1	
X*Prob	0	0,3	0,4	0,3	==> E(X)=1
X^2*Prob	0	0,3	0,8	0,9	==> $E(X^2)=2$

- $V(X) = 2 - 1^2 = 1$
- E(X100)=100
- V(X100)=100
- X100~N(100, 100)
- $P(X100>120) = 1-F_N((120-100)/10) = 1-F_N(2)=0,023$

Seien $X_1, X_2, ..., X_n$ identisch verteilte, unabhängige Zufallsvariablen mit

$$E(X_i) = \mu \qquad \text{und} \qquad V(X_i) = \sigma^2 > 0$$

Dann gilt für die Verteilung des arithmetischen Mittels

$$\overline{x}_n = 1/n(X_1 + X_2 + ... + X_n)$$

Erwartungswert $E(\overline{x}_n) = \mu$

und Varianz $V(\overline{x}_n) = \sigma^2/n.$

i) Auch das arithmetische Mittel der Stichprobe ist eine Zufallsvariable

ii) Die Standardabweichung des arithmetischen Mittels wird auch Standardfehler bezeichnet

Anwendung des zentralen Grenzwertsatzes auf Mittelwert

Seien $X_1, X_2, ..., X_n$ identisch verteilte, unabhängige Zufallsvariablen mit

$$E(X_i) = \mu \qquad \text{und} \qquad V(X_i) = \sigma^2 > 0$$

Dann konvergiert die Verteilung des standardisierten Mittelwertes

$$Z_n = \frac{\frac{1}{n}\sum X_i - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

mit wachsendem n gegen eine Normalverteilung mit Erwartungswert $E(Z_n) = 0$ und Varianz $V(Z_n) = 1$.

$Z_n \sim N(0, 1^2)$

Standardfehler

- Die Varianz bzw. die Standardabweichung des arithmetischen Mittels ergibt sich also durch:

$$\sigma_{\bar{x}}^2 = \sigma^2 / n$$

$$\sigma_{\bar{x}} = \sqrt{\sigma^2 / n} = \sigma / \sqrt{n}$$

- Der Mittelwert schwankt weniger stark als die Einzelwerte
- Die Standardabweichung des Mittelwertes wird auch als Standardfehler (standard error) bezeichnet.
- Wurzel-n Gesetz: Doppelte Genauigkeit benötigt vier-fachen Stichprobenumfang!

Grenzwertsatz von De Moivre und Laplace

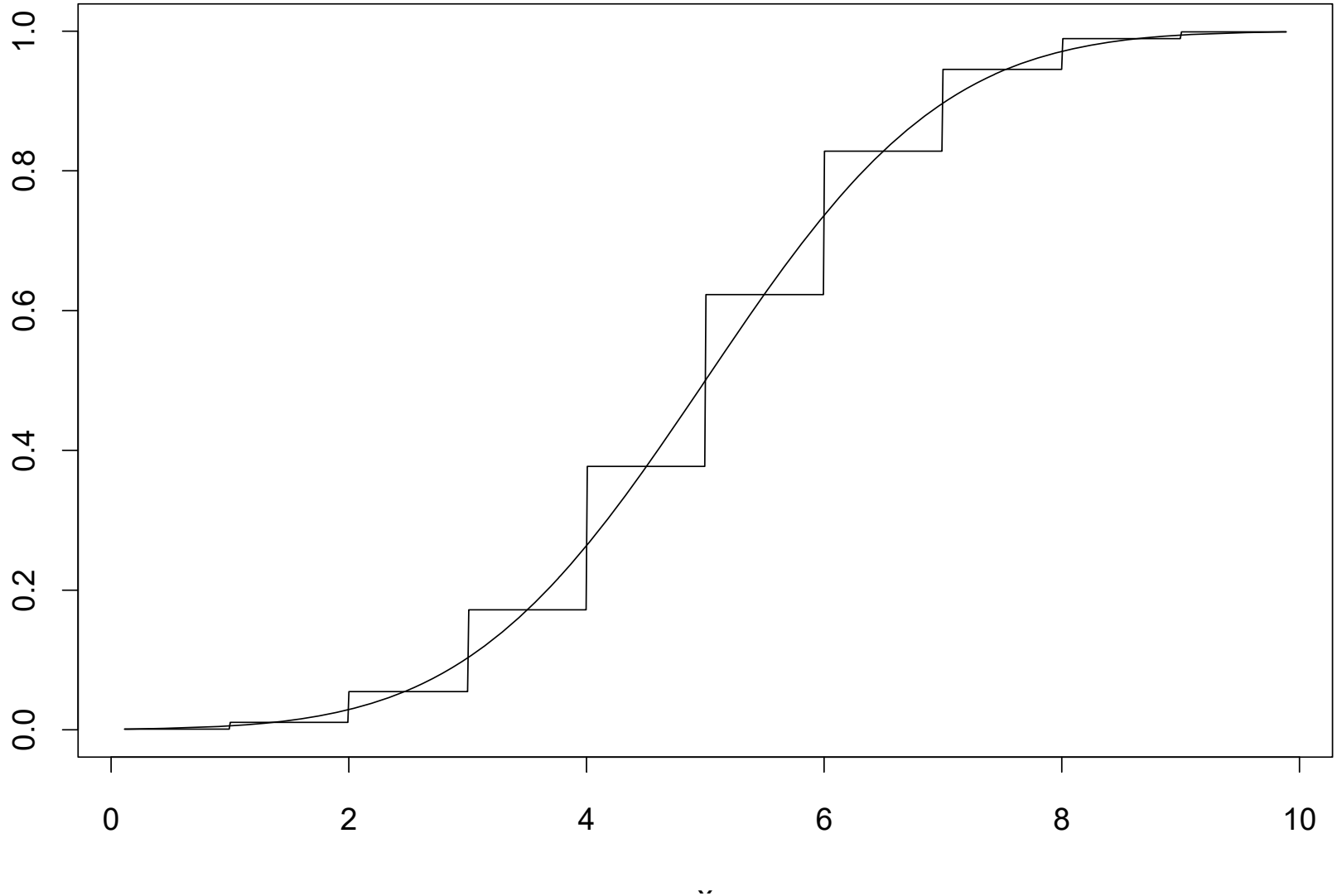
- Falls X binomialverteilt ist mit den Parametern n und p [es sei also X~Bi(n, p)] so gilt:

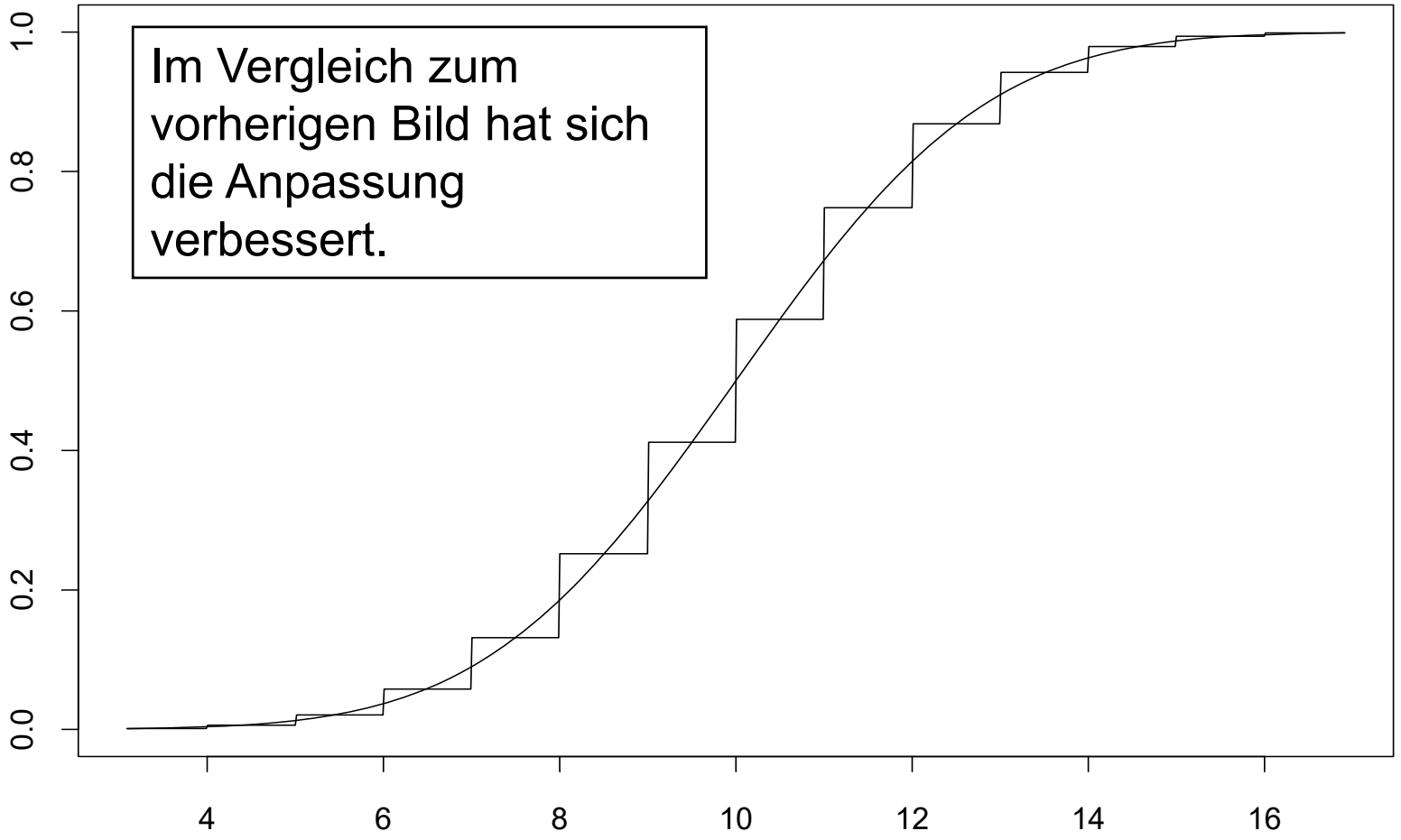
$$\frac{X - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} \approx N(0,1)$$

Die Normalverteilung kann zur Approximation der Binomial-verteilung verwendet werden !

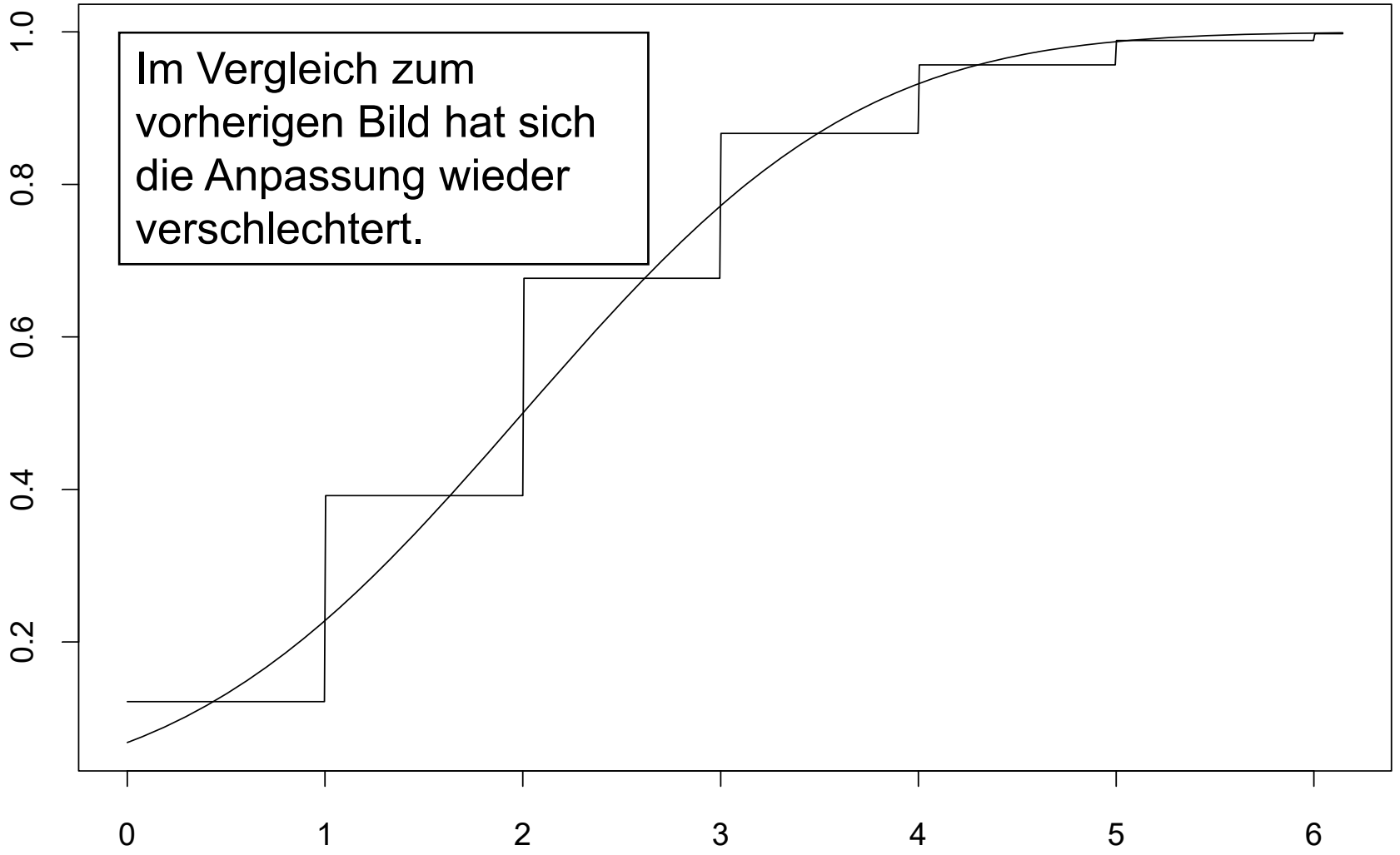
Beachte $E(X) = n \cdot p$ und $V(X) = n \cdot p \cdot (1-p)$

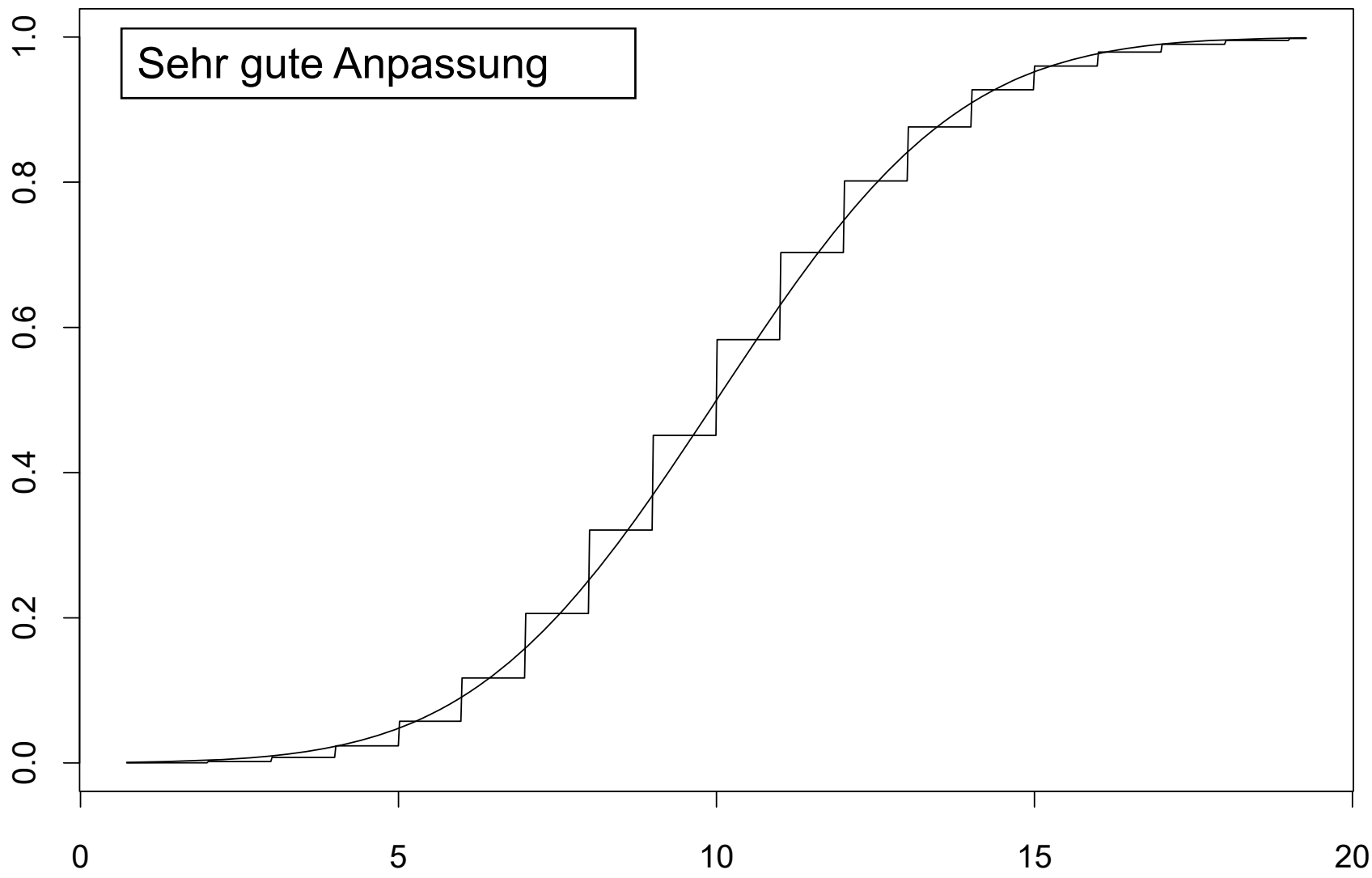
- Die Güte der Anpassung hängt dabei von n und p ab. (Wenn p nahe 1/2 und n möglichst groß ist, so steigt die Güte)
- Faustregel: $np > 10$ und $n(1-p) > 10$


n= 10 p= 0.5
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10


n= 20 p= 0.5
Im Vergleich zum vorherigen Bild hat sich die Anpassung verbessert.
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
16

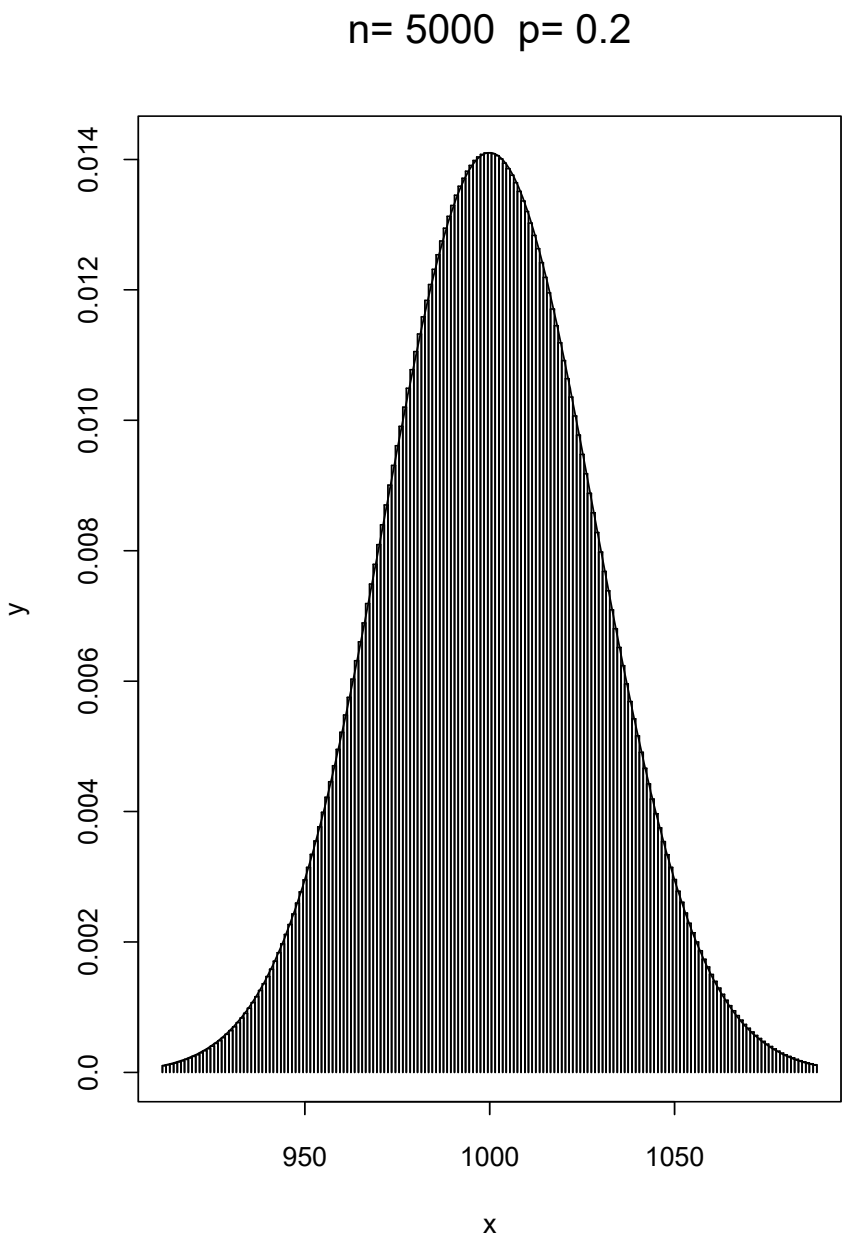
n= 20 p= 0.1

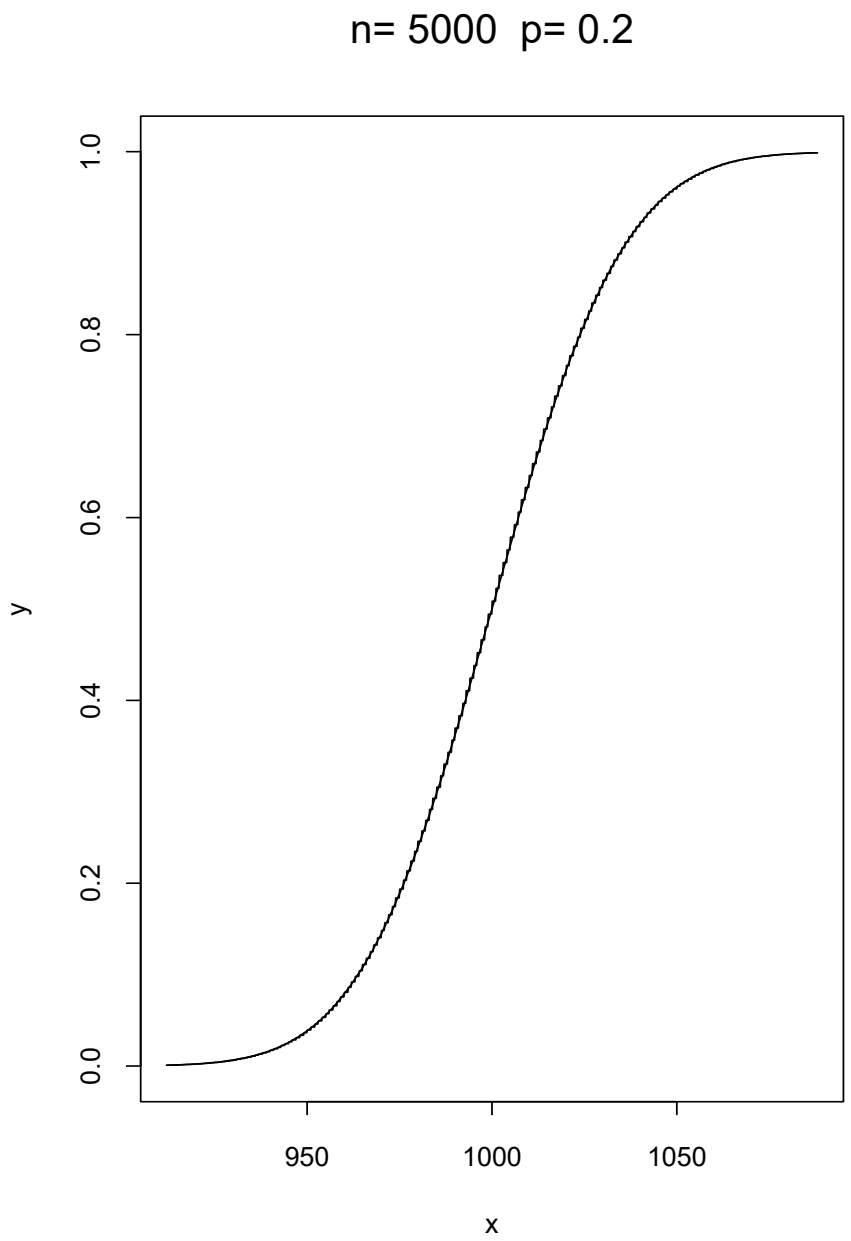

Im Vergleich zum vorherigen Bild hat sich die Anpassung wieder verschlechtert.
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6

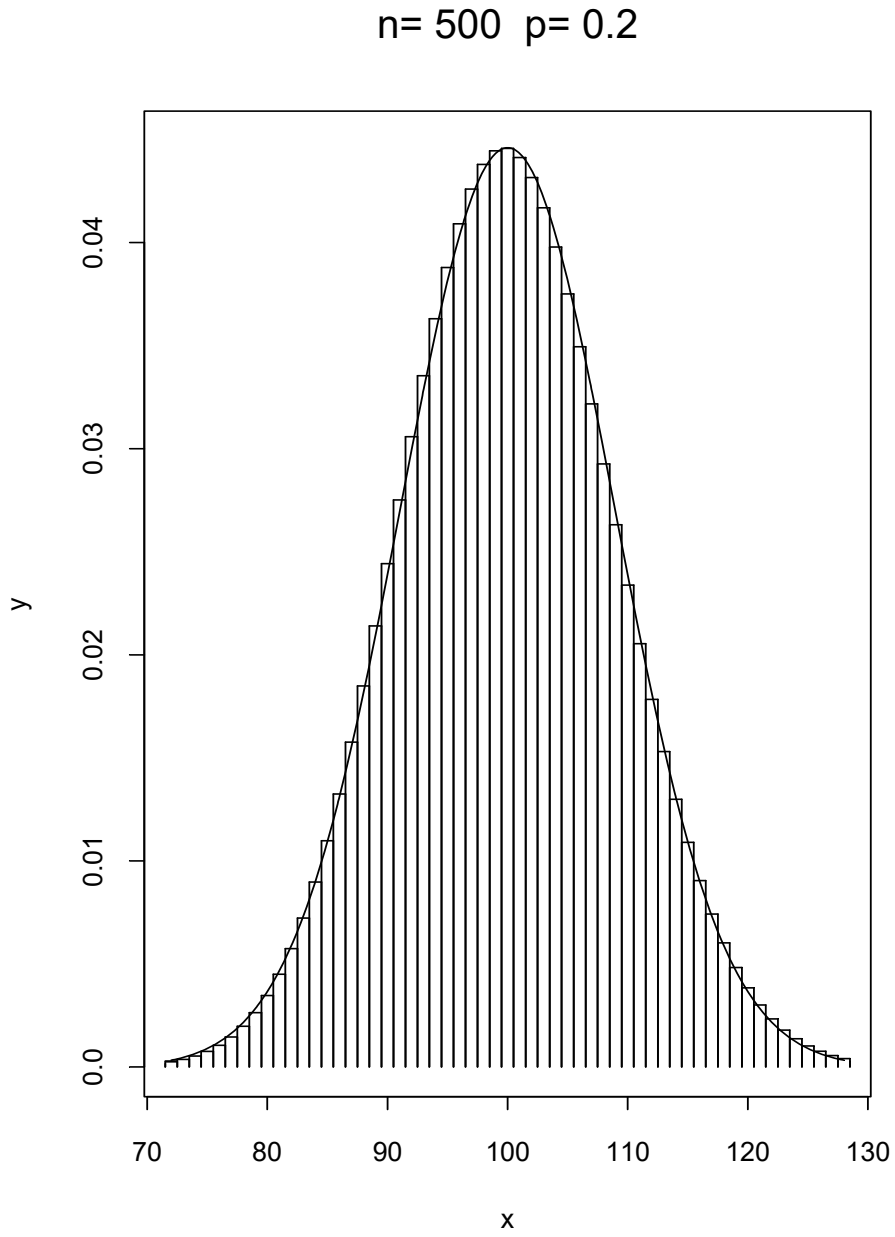

n= 100 p= 0.1
Sehr gute Anpassung
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

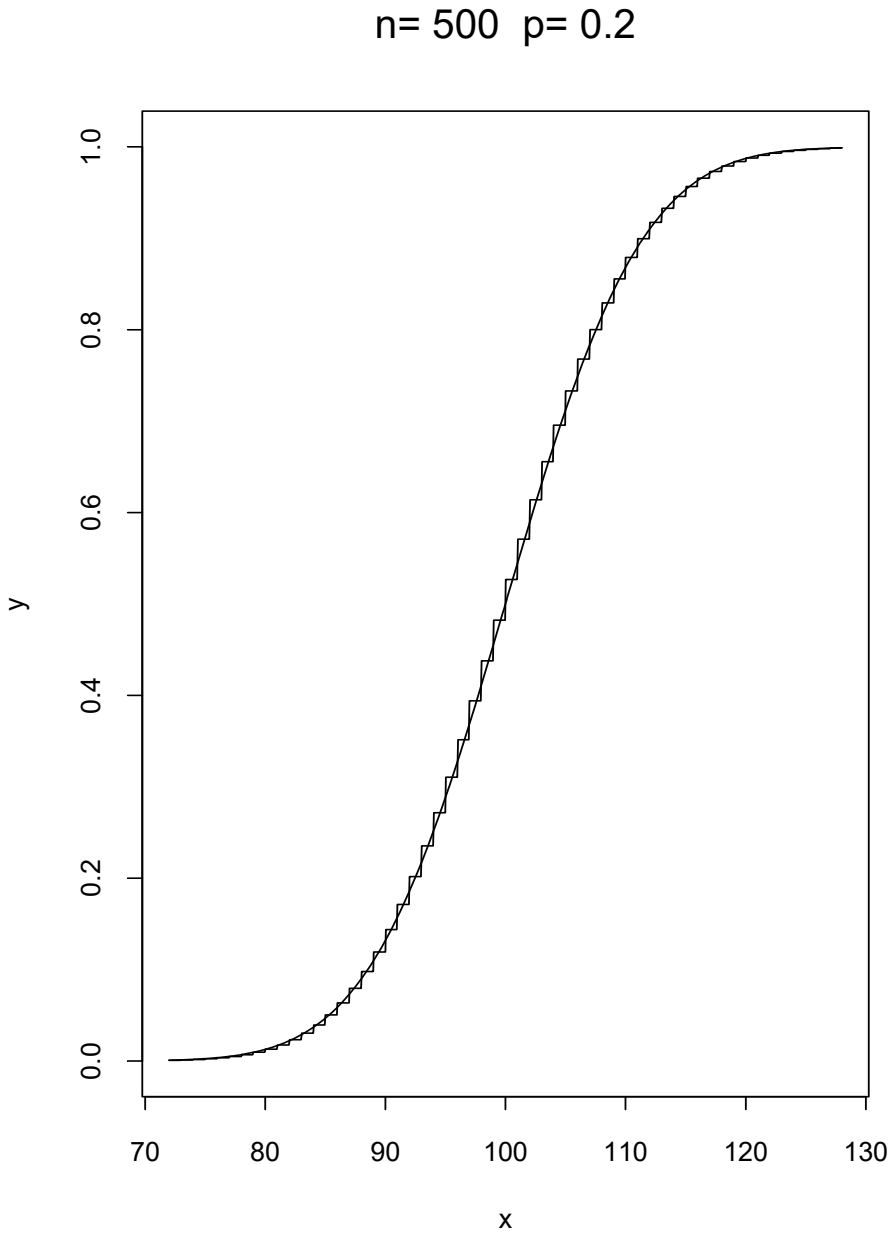
Beispiel: Prognose des Rücklaufs

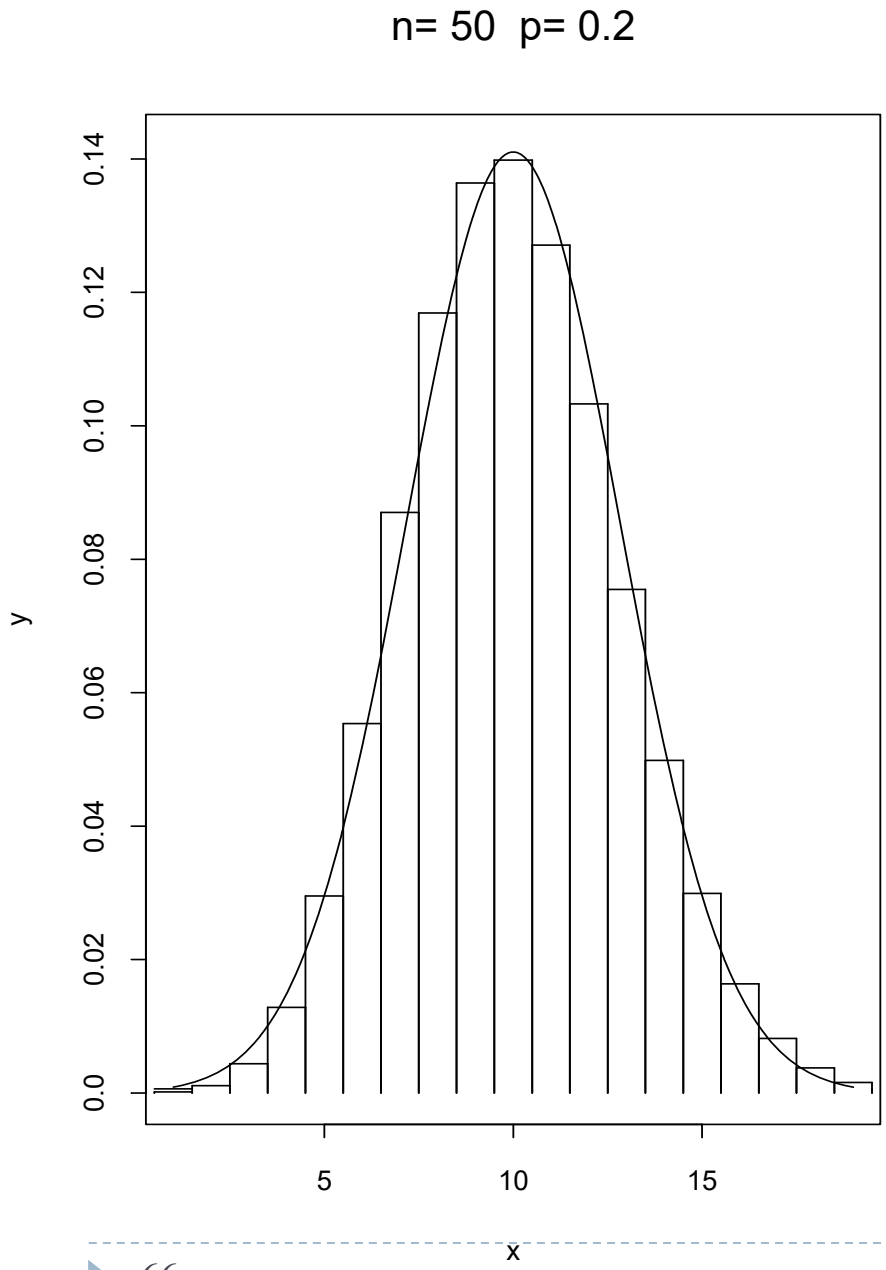
- Bei einer bestimmten schriftlichen Befragung weiß man aus Erfahrung, dass etwa 20% der Befragten tatsächlich antworten.
- Es werden n=5.000 Fragebogen versandt.
- X sei die Anzahl der Antworter
- E(X) = 5.000*0,2 = 1.000 Var(X)=5.000*0,2*0,8 = 800
- X~N(1.000, 800) Std.Abw. =28
- Mehr als 1.000 Antworten: P(X>1.000) = 0,5
- Mehr als 1.200 Antworten: P(X>1.200) = 0,0
- 95% Intervall für die Anzahl der zu erwartenden Antworten:
- P(1000-1,96*28<X<1000+1,96*28) = 0,95
- P(945<X<1055) = 0,95

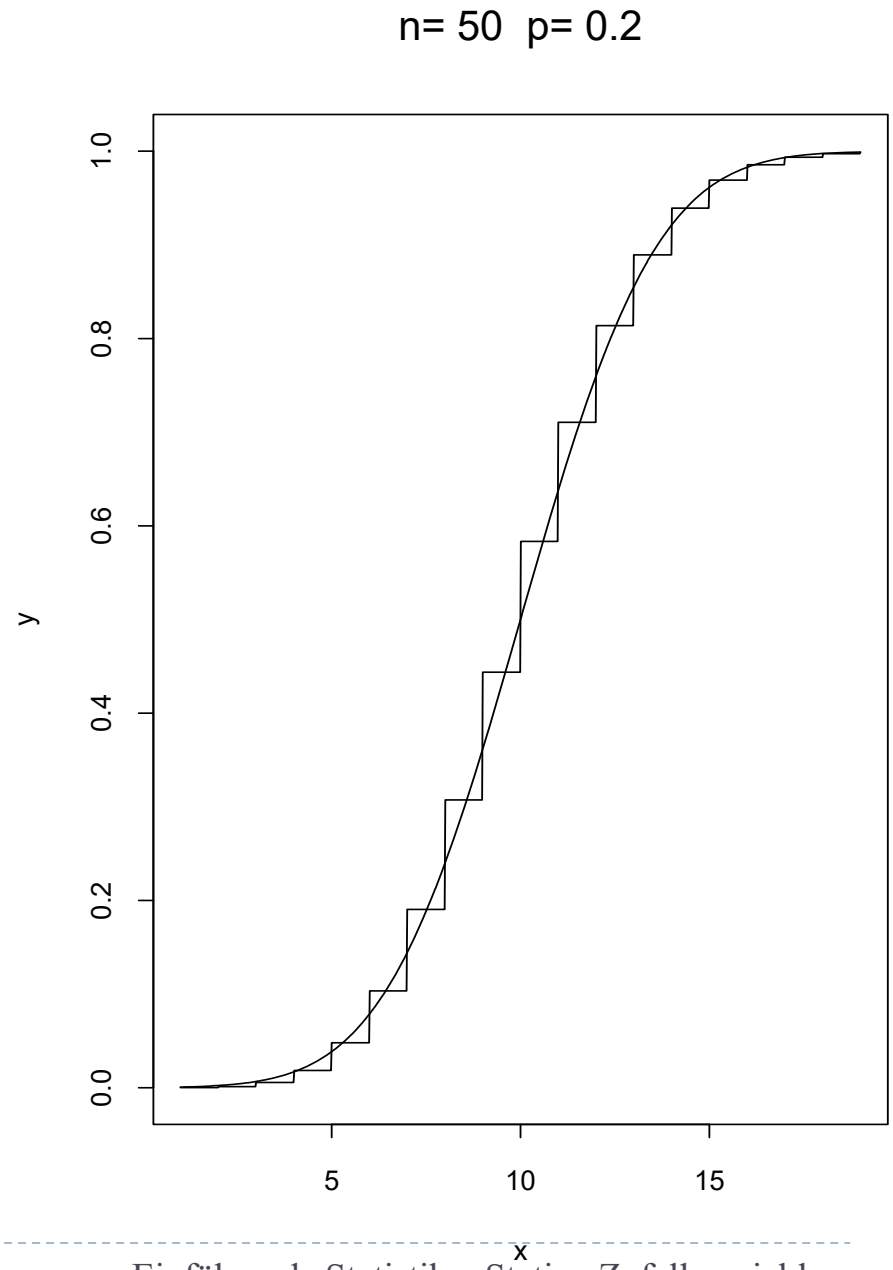

n= 5000 p= 0.2
y
0.0
0.002
0.004
0.006
0.008
0.010
0.012
0.014
950
1000
1050
x

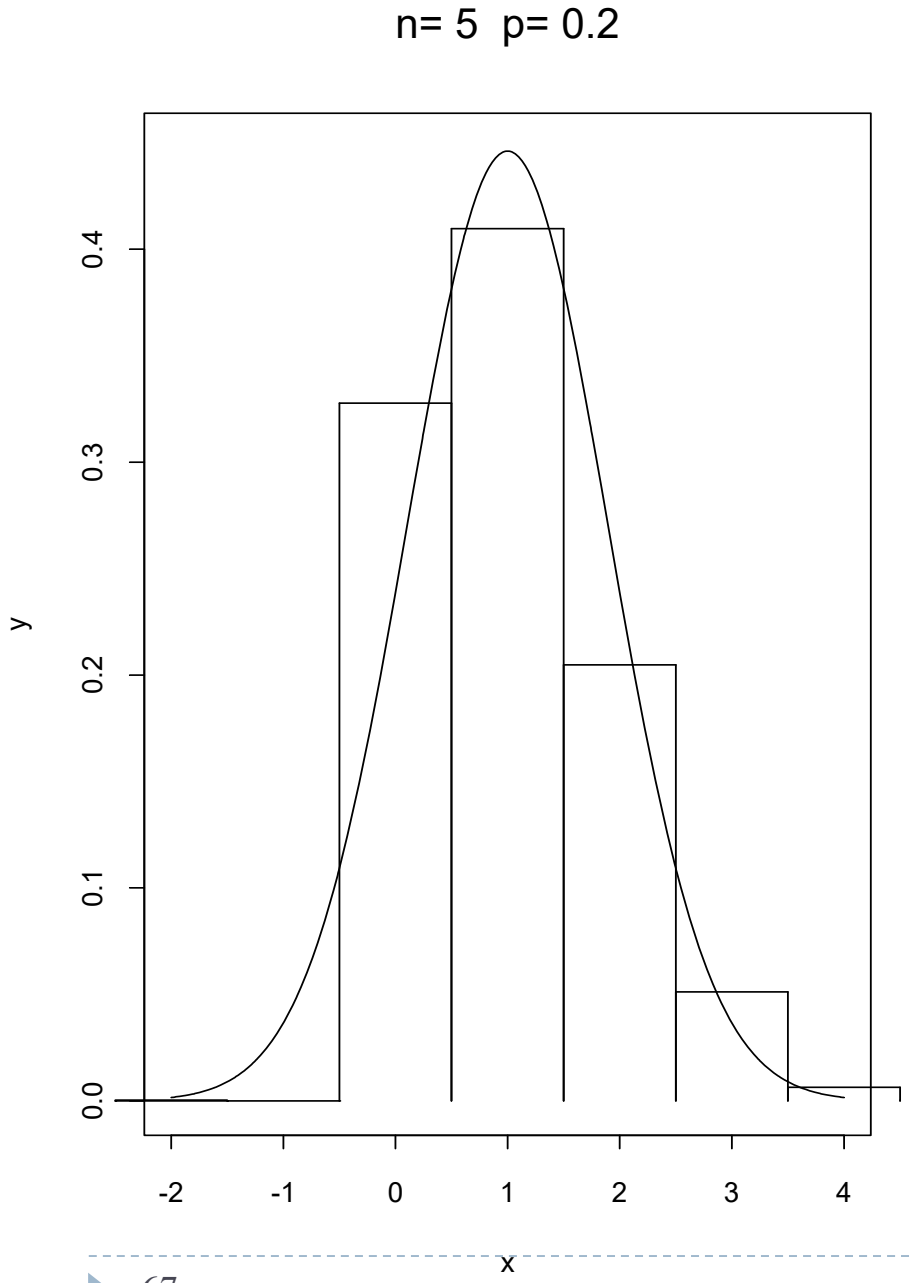

n= 5000 p= 0.2
y
0.0
0.2
0.4
0.6
0.8
1.0
950
1000
1050
x

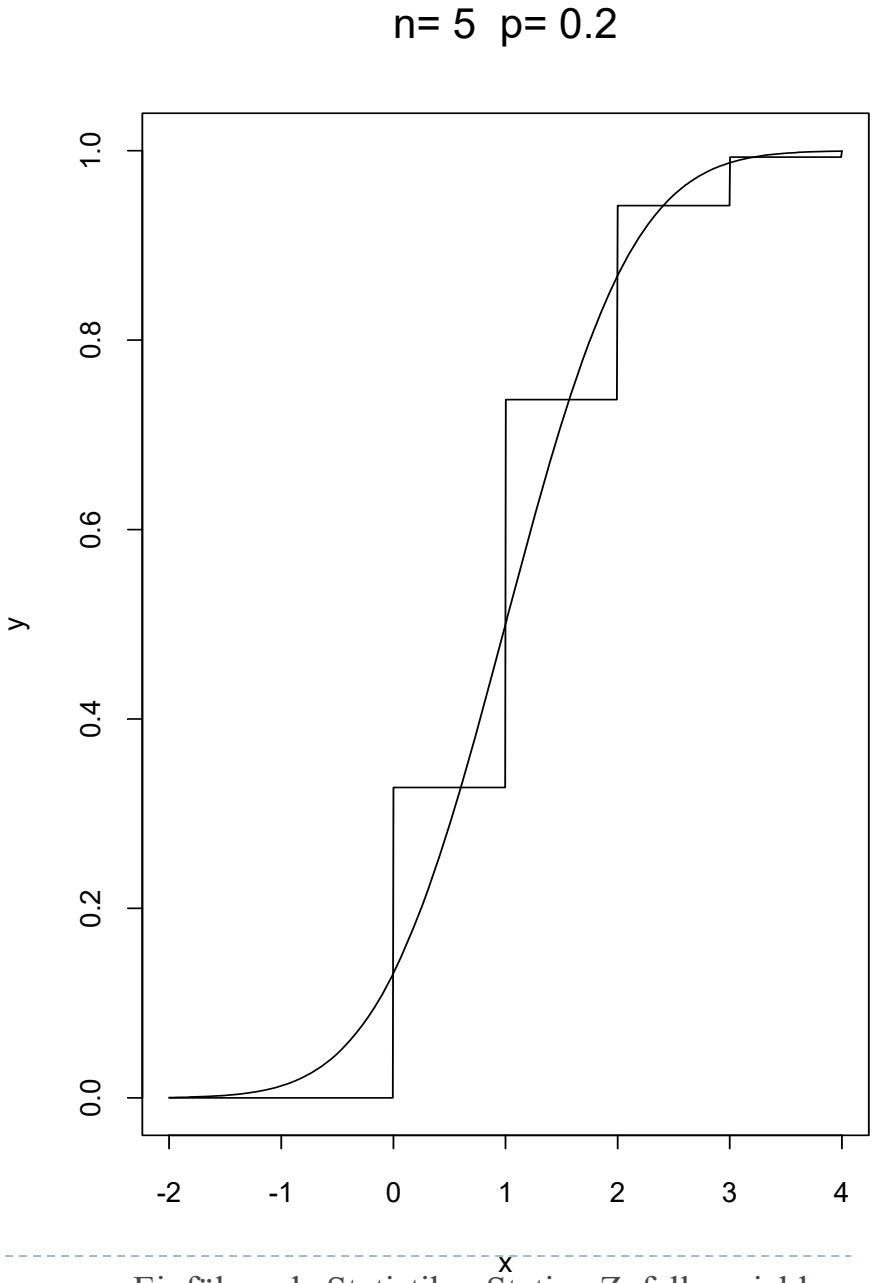

n= 500 p= 0.2
y
0.0
0.01
0.02
0.03
0.04
70
80
90
100
110
120
130
x


n= 500 p= 0.2
y
0.0
0.2
0.4
0.6
0.8
1.0
70
80
90
100
110
120
130
x


n= 50 p= 0.2
y
0.0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
5
10
15
x


n= 50 p= 0.2
y
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
x


n= 5 p= 0.2
y
0.0
0.1
0.2
0.3
0.4
-2
-1
0
1
2
3
4
x


n= 5 p= 0.2
y
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
x

Stetigkeitskorrektur

Bei der Approximation der Binomialverteilung (diskrete ZV) durch die Normalverteilung (stetige ZV) ist eine Stetigkeitskorrektur (Kontinuitätskorrektur) zu berücksichtigen.

- Die diskrete P(X=x) entspricht im stetigen Fall P(X<x+0,5) - P(X<x-0,5)

$$P(X = x) \approx \Phi\left(\frac{x+0{,}5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x-0{,}5-np}{\sqrt{np(1-p)}}\right)$$

bzw.

$$P(X \leq x) \approx \Phi\left(\frac{x+0{,}5-np}{\sqrt{np(1-p)}}\right)$$

Beispiel:

- In einer Bevölkerung sind 60% der Bürger für die Einführung eines neuen Gesetzes. Wie wahrscheinlich ist es, genau 50 Befürworter in einer Stichprobe vom Umfang n=100 zu haben ?
- Binomialverteilung

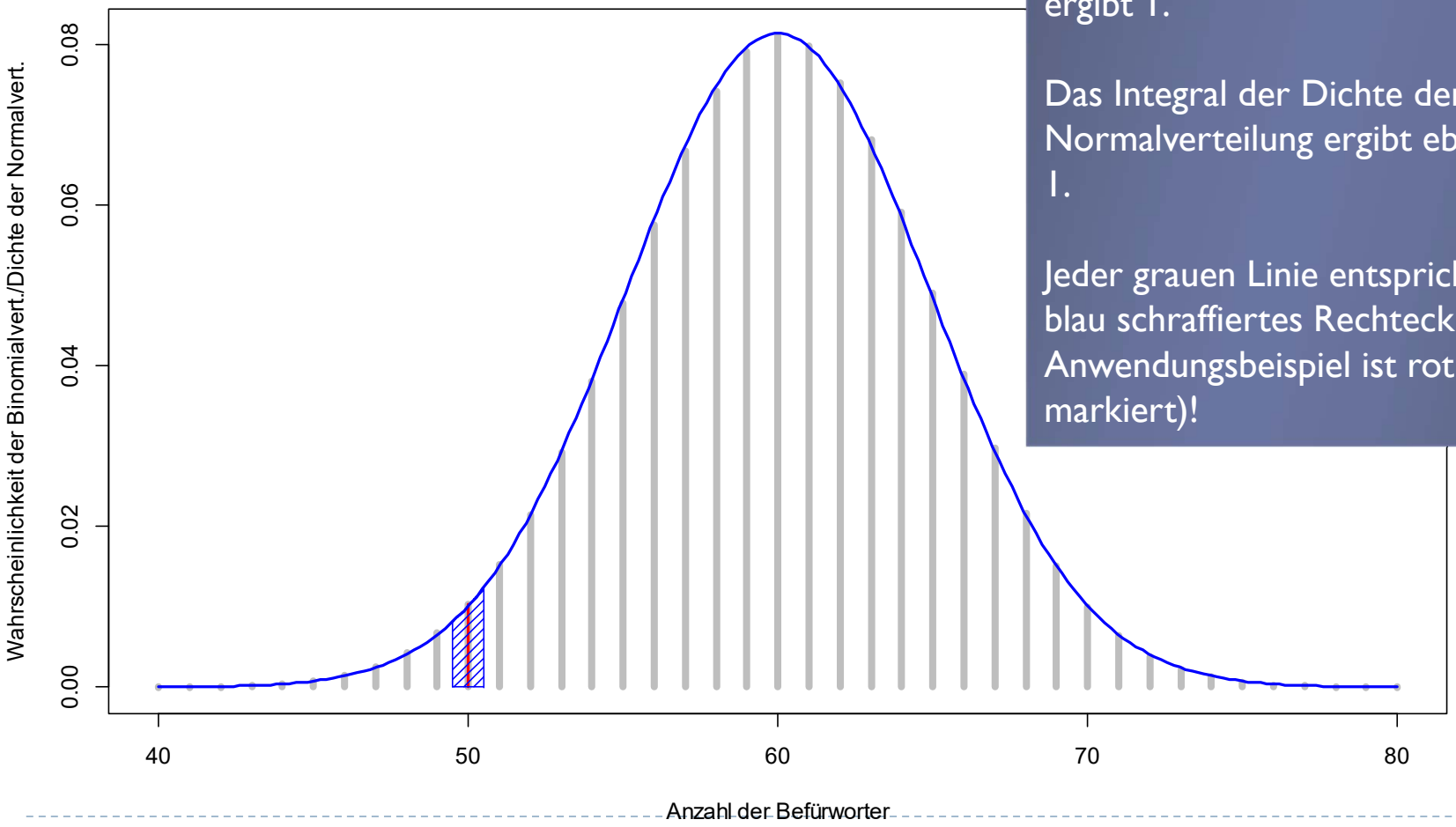
$$P(X = 50) = \binom{100}{50} * 0{,}6^{50} * 0{,}4^{50} = 0{,}0103$$

- Normalverteilung

$$P(X = 50) = \Phi\left(\frac{50 + 0.5 - 60}{\sqrt{24}}\right) - \Phi\left(\frac{50 - 0.5 - 60}{\sqrt{24}}\right) =$$

$$\Phi(-1{,}939) - \Phi(-2{,}143) = 0{,}0262 - 0{,}0160 = 0{,}0102$$

Visualisierung der Stetigkeitskorrektur



Die Summe der Wahrscheinlichkeiten der Binomialverteilung (graue Linien) ergibt 1.

Das Integral der Dichte der Normalverteilung ergibt ebenfalls 1.

Jeder grauen Linie entspricht ein blau schraffiertes Rechteck (das Anwendungsbeispiel ist rot markiert)!

Beispiel:

- In einer Bevölkerung sind 60% der Bürger für die Einführung eines neuen Gesetzes.
- Wie groß ist die Wahrscheinlichkeit, dass sich in einer Stichprobe von 10 (100) Personen, weniger als 5 (50) Befürworter des Gesetzes finden ?
- a) Binomialverteilung mit n=10 und p=0.6
  $P(X<5)=P(X=0) + P(X=1) + ... + P(X=4)=$

  0.000+ 0.002 +0.011+ 0.042+ 0.111=0.166

  (Exaktes Ergebnis durch Einsetzen in die Formel der Binomialverteilung)

Beispiel:

- b) Bei einer Stichprobe von n=100 gibt es 2 Lösungswege:
- b1) Einsetzen in die Formel der Binomialverteilung mit n=100 und p=0.6
  P(X<50)=P(X=0) + P(X=1) + ... + P(X=49)= 0.0168
- b2) Approximation durch Normalverteilung

  X~N(60; 24)

  n.p=100*0,6=60 n.p.(1-p)=60*0,4=24

  Wurzel(n.p.(1-p))=4,899

  $P(X \leq 49) =$

  $F_N((49+0,5-60)/4,899)=$

  $F_N(-2,14)=0,0160$

```
> sum(dbinom(0:49, size=100, prob=0.6))
[1] 0.01676169
> pnorm(49.5, 60, sqrt(24))
[1] 0.01604437
>
```

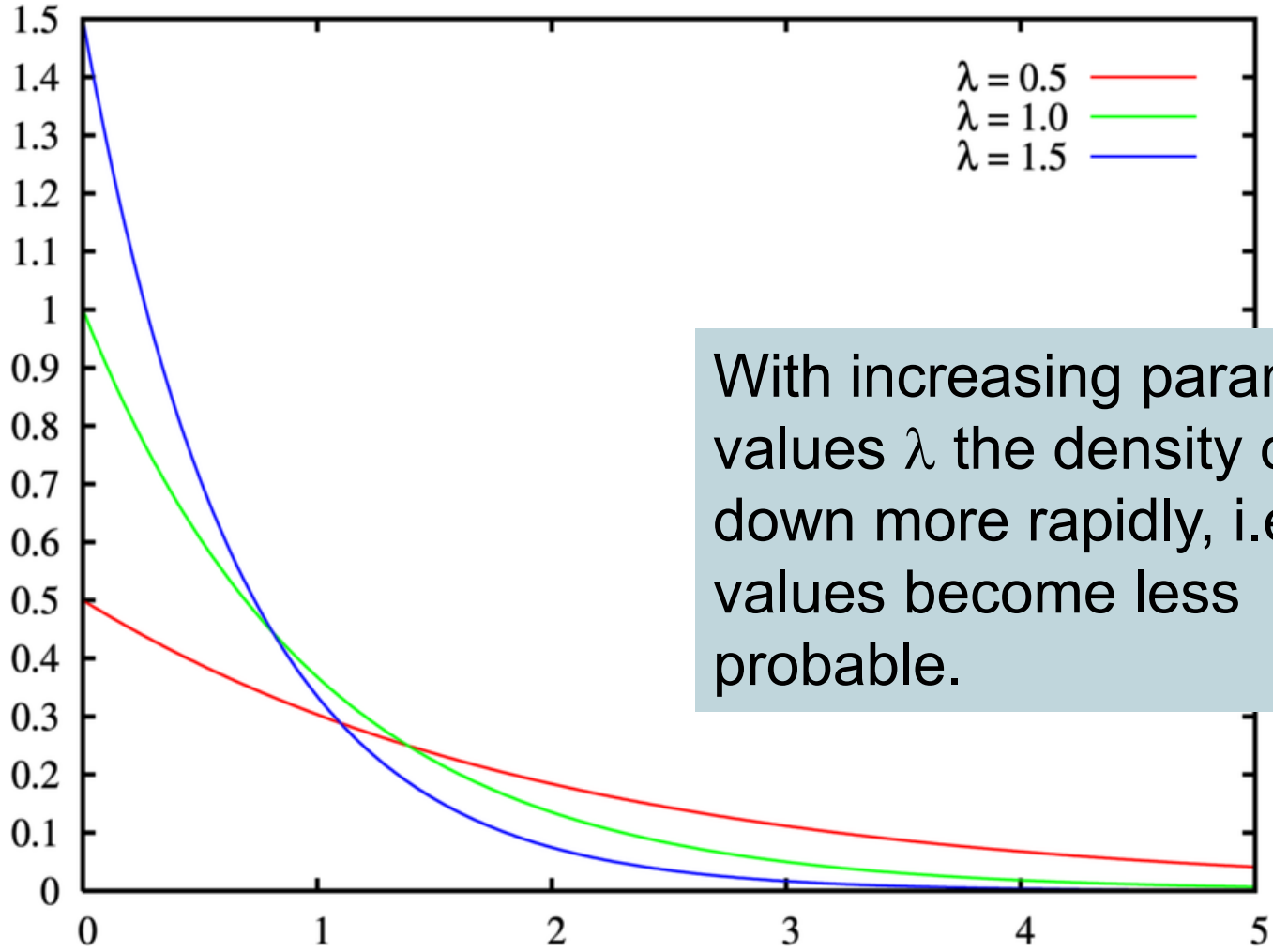
Exponential Distribution

- The exponential distribution represents a family of continuous probability distributions that are often used to model the time between independent events that happen at a constant average rate.
- X is said to be an exponentially distributed random variable with parameter λ [in short X ~ EX(λ)], if the probability density function takes the following form

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & , \ x \geq 0, \\ 0 & , \ x < 0. \end{cases}$$

- A particular member of the family is identified by one single parameter.

3 Exponential Densities with varying λ



With increasing parameter values λ the density drops down more rapidly, i.e. larger values become less probable.

Cumulative Distribution Function

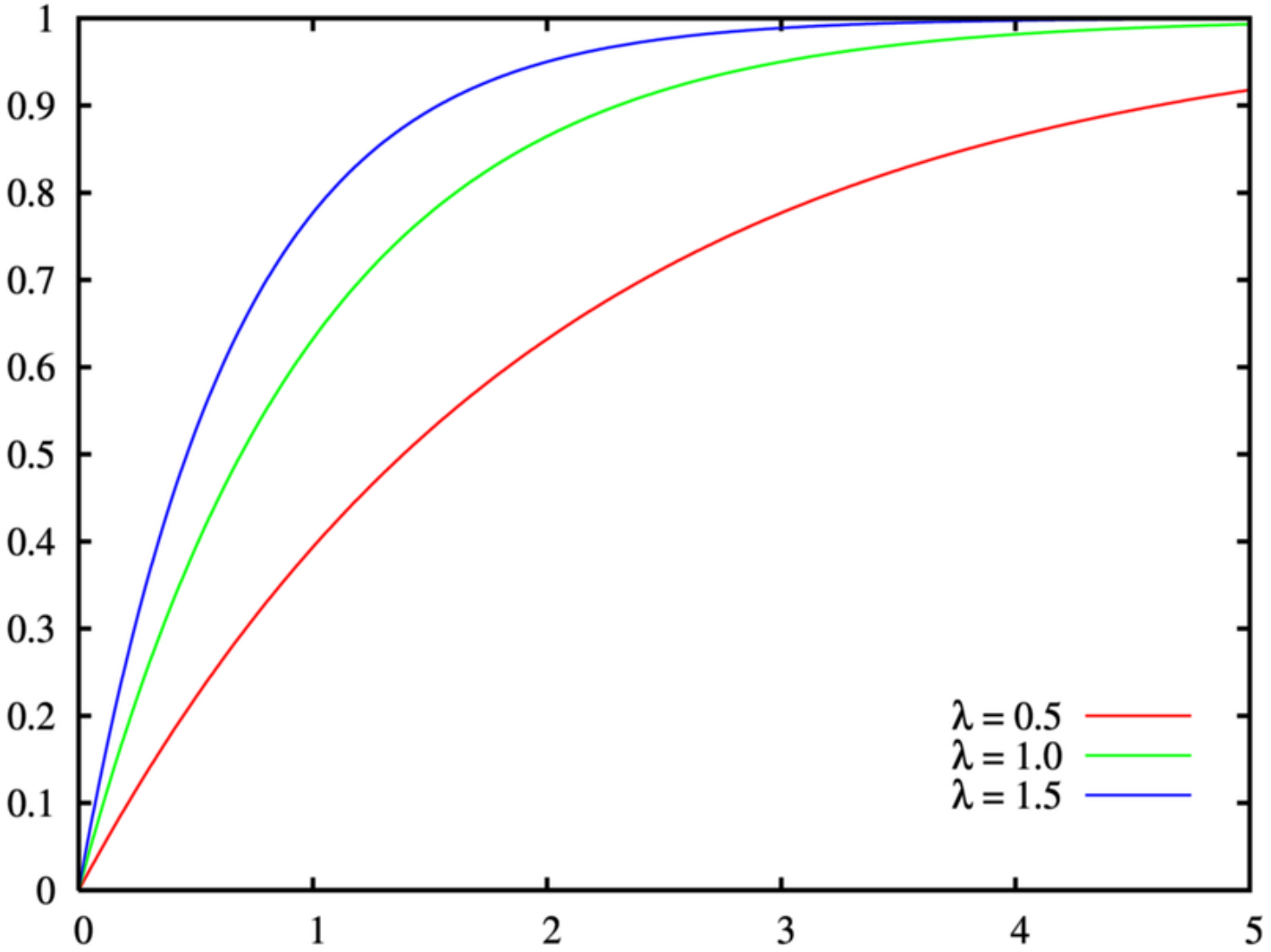
- Integrating the exponential density

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & , \ x \geq 0, \\ 0 & , \ x < 0. \end{cases}$$

- gives the cumulative distribution function

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & , \ x \geq 0, \\ 0 & , \ x < 0. \end{cases}$$

Cumulative Distribution Function



Important Measures

- X~EX(λ),
  - E(X) = 1/λ
  - V(X) = 1/λ² SD(X) = 1/λ
  - Note: In case of an exponential distributed random variable the expectation equals the standard deviation
- The inverse cumulative distribution function (quantile function) of EX(λ) is for $0 \leq p < 1$.

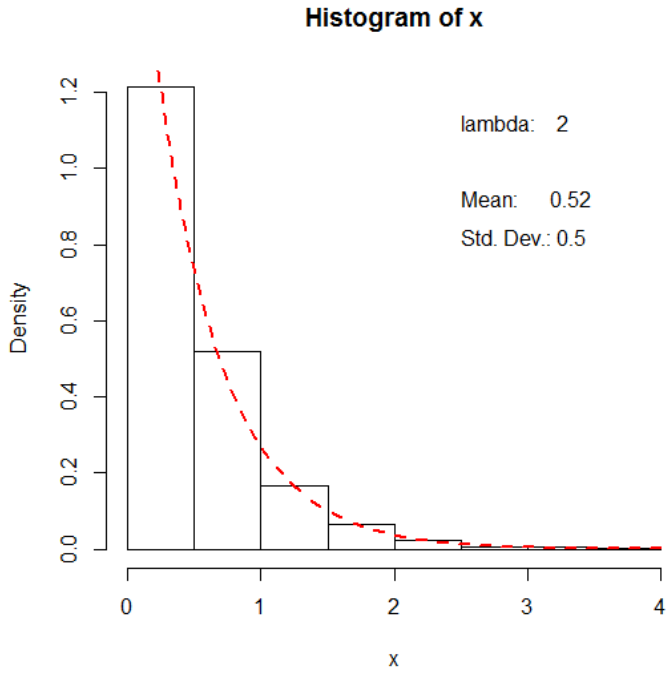
$$F^{-1}(p;\lambda) = \frac{-\ln(1-p)}{\lambda},$$

  - Median (2nd-Quartile) = ln(2)/ λ
  - 1st-Quartile = ln(4/3)/ λ; 3rd- Quartile = ln(4)/ λ

Simulating Exponential RVs

R Console

```
> # Exponential Density
>
> lambda <- 2
> x <- rexp(1000, lambda)
> hist(x, prob=T)
> text(2.5, 1.10, paste("lambda:    ", round(lambda,2), sep=""), adj = c(0,0))
> text(2.5, 0.9, paste("Mean:      ", round(mean(x),2), sep=""), adj = c(0,0))
> text(2.5, 0.80, paste("Std. Dev.: ", round(sd(x),2), sep=""), adj = c(0,0))
> curve(dexp(x, lambda), col = 2, lty = 2, lwd = 2, add = TRUE)
>
```



No Memory property

- The Exponential Distribution has a so-called „no memory“ or „no ageing“ property:

$$P(T > s + t \mid T > s) = P(T > t) \text{ for all } s, t \geq 0.$$

- h(t) = P(t ≤ T < t+dt | T>t) = f(t)dt / (1-F(t))

  h(t) ... hazard-function or instantaneous risk of mortality

  i.e. conditional probability of dying at time-point t given survival up to time-point t
- If T~EX(λ)
- h(t) = λ
- „no ageing“ ~ constant risk of mortality

Relationship to Geometric Distribution

- The constant probability of an event within a fixed time interval is given by p
- The probability of the first occurrence of the event (foe) in interval k+1 is given by the geometric distribution

$$P(\text{foe} = k+1) = p(1-p)^k$$

- The probability of the first occurrence of the event at interval k+1 or later is given by the cumulative distribution function $P(\text{foe} \geq k+1) = (1-p)^k$
- Considering smaller time intervals implies smaller probabilities for each interval (to keep constancy)
- 1/2 interval ➔ p/2 1/3 interval ➔ p/3 etc.

Infinitesimal small time intervals

- $P(X \geq k+1)=(1-p)^k$
- $P(X \geq k+1)=(1-p/2)^{2k}$ with 1/2 interval
- $P(X \geq k+1)=(1-p/3)^{3k}$ with 1/3 interval
- ...
- $P(X \geq k+1)=(1-p/n)^{nk}$ with 1/n interval
- For $n \to \infty$:

  $P(X \geq k+1)= \exp(-pk)$
- Results in the c.d.f. of the exponential distribution

  $P(X \leq x) = 1 - \exp(-\lambda x)$
- Exponential distribution is the continuous analogon of the geometric distribution, which is a discrete distribution

Applications

- The exponential distribution is used to model Poisson processes, which are situations in which an object initially in state A can change to state B with constant probability per unit time λ. The time at which the state actually changes is described by an exponential random variable with parameter λ.
- Exponential variables can also be used to model situations where certain events occur with a constant probability per unit *distance*
- In real-world scenarios, the assumption of a constant rate (or probability per unit time) is rarely satisfied that means more flexible models are often needed

Real life examples

- which might yield approximately exponentially distributed variables:
  - the time until a radioactive particle decays
  - the time between beeps of a Geiger-counter
  - the number of dice rolls needed until you roll a six 12 times in a row
  - the time between two telephone calls in your call-center
  - the time-interval between two claims of a policy holder
  - the distance between mutations on a DNA strand
  - the distance between mortal accidents on a given street