Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Marcel Dettling

Institute für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

ETH Zürich, 18. April 2018

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Was ist Statistik?

Statistische Datenanalyse ist dazu da, um unter der Präsenz von Unsicherheit und Variation korrekte Fakten zu gewinnen und intelligente Aussagen abzuleiten, die nicht von den Launen des Zufalls beeinflusst sind. Es geht darum, Messwerte und Beobachtungen in systematische Effekte und zufällige Variation zu separieren.

Anwendungen:

- Business
- Lesen von Fachliteratur
- Studentenarbeiten & Forschung

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Beispiel 1: Wirksamkeit von Schlafmitteln

Untersucht werden soll die Wirksamkeit von zwei Schlafmitteln A und B. Dazu wurde bei 10 Probanden die durchschnittliche Schlafverlängerung von A vs. B in Stunden gemessen. Die beobachteten Werte sind:

+1.2 +2.4 +1.3 +1.3 +0.0 +1.0 +1.8 +0.8 +4.6 +1.4

Fragestellungen:

- Mittlere Schlafverlängerung (Schätzen)
- Genauigkeitsangabe für mittlere Schlafverlängerung (VI)
- Ist A signifikant besser als B? Sicherheit der Aussage? (Test)

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

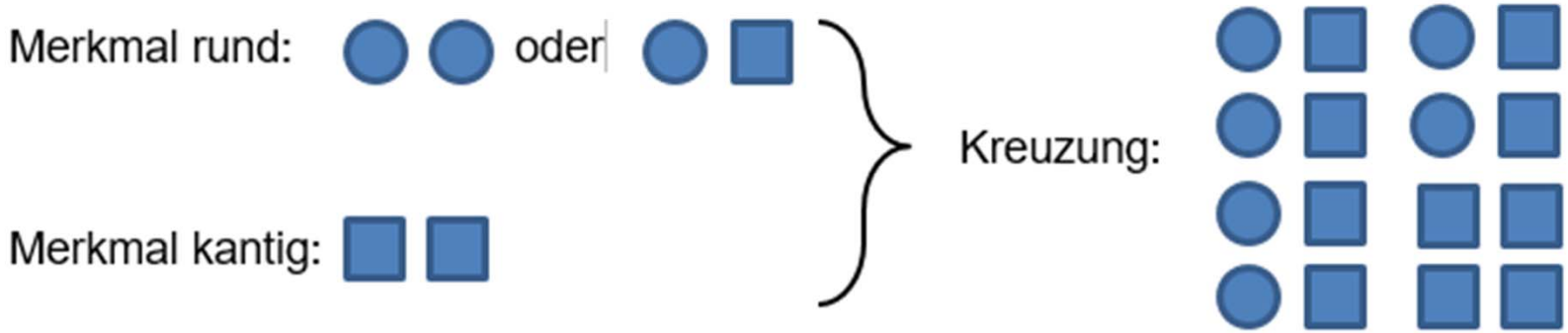
Beispiel 2: Mendels Vererbungsgesetze

Gregor Mendel publizierte im Jahr 1866 eine Studie über den Vererbungsvorgang bei Merkmalen, deren Ausprägung nur von einem einzelnen Gen bestimmt wird. Als Beispiel studierte er 2 Erbsensorten mit runden bzw. kantigen Samen. Dabei werden runde Samen dominant vererbt.

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Beispiel 2: Mendels Vererbungsgesetze

Wenn man runde und kantige Erbsen kreuzt und Mendels Modell gilt, dann sollte man bei den Nachkommen ein Verhältnis von 3:1 zugunsten der runden Samen beobachten.



Experimentelles Resultat:

7’324 Samen, wovon 5’474 rund und 1’850 kantig (2.96:1).

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Beispiel 2: Mendels Vererbungsgesetze

Experimentelles Resultat:

7’324 Samen, wovon 5’474 rund und 1’850 kantig (2.96:1).

Fragestellung:

Gilt das Vererbungsmodell oder gibt es einen Widerspruch?

Zu Bedenken:

- Wie stark kann/darf das Verhältnis schwanken?
- Mit einem Test/Vertrauensintervall erhält man eine Antwort
- Wie viele Samen muss man zählen, um «sicher» zu sein?

Beispiel 3: Korrosion nach Eisengehalt

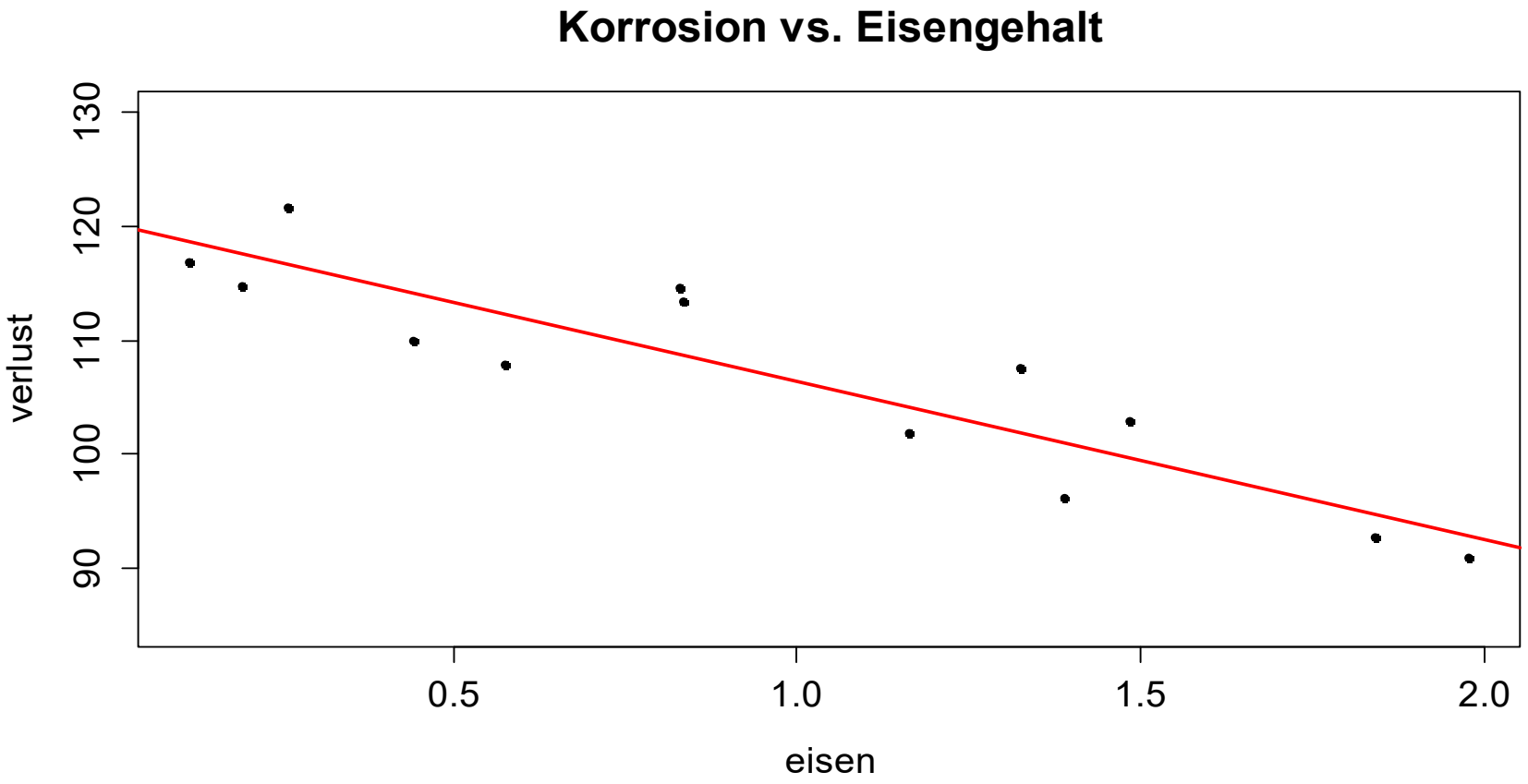
In einem Experiment soll der Zusammenhang zwischen der Korrosion einer Kupfer-Nickel-Legierung in Abhängigkeit von deren Eisengehalt studiert werden. Dazu wurden 13 Räder hergestellt und während 60 Tagen gedreht. Gemessen wurde der korrosionsbedingte Gewichtsverlust in Milligramm.

Fragestellungen:

- Abnahme, wenn Eisengehalt um eine Einheit grösser?
- Ist dieser Effekt statistisch gesichert oder bloss Zufall?
- Wie genau kann man den Gewichtsverlust vorhersagen?

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Beispiel 3: Korrosion nach Eisengehalt



Gibt es Zufall?

Ein Zufallsexperiment ist ein Versuch, bzw. eine Situation, wo das Ergebnis nicht deterministisch vorbestimmt ist.

Um zu entscheiden, ob eine Situation ein Zufallsexperiment ist, stellt man sich am besten die Frage, ob bei einer Wiederholung exakt dasselbe Resultat erneut auftreten würde.

- *Würfel- oder Münzenwurf*
- *Anzahl Studenten im Hörsaal*
- *Regenmenge innerhalb 24h am Züriberg*

Gibt es Zufall?

Ob Zufall wirklich existiert oder nicht, ist eine philosophische Frage, welche in der Statistik bzw. Mathematik nicht im Zentrum steht. Hier wird Zufall als ein Konzept benutzt, das uns davon entbindet, alle möglichen Einflussfaktoren zu verstehen und zu kontrollieren.

Immer dann, wenn man das Ergebnis eines Versuchs oder einer Beobachtung nicht mit Sicherheit voraussagen kann, benützt man die Konzepte von Zufall und Wahrscheinlichkeit. Anstatt das Ergebnis exakt vorherzusagen, begnügt man sich mit der Angabe und dem Studium der Wahrscheinlichkeiten.

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Grundkonzepte der W'keitsrechnung

- Ereignis- bzw. Wahrscheinlichkeitsraum Ω
- Ereignisse A als Teilmengen des W'keitsraums
- Die Wahrscheinlichkeit $P[A]$ als Funktion für Ereignisse.
- Zufallsvariablen und Wahrscheinlichkeitsverteilung.

→ **Siehe Wandtafel…**

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Diskrete Wahrscheinlichkeitsverteilung

Es sei X eine beliebige diskrete Zufallsvariable. Wir bezeichnen die Werte, die X annehmen kann mit $x_1, x_2, ..., x_n$. Die zugehörigen Wahrscheinlichkeiten notieren wir mit $p(x_1), p(x_2), p(x_3), ...$

Es ist also: $p(x_i) = P(X = x_i)$

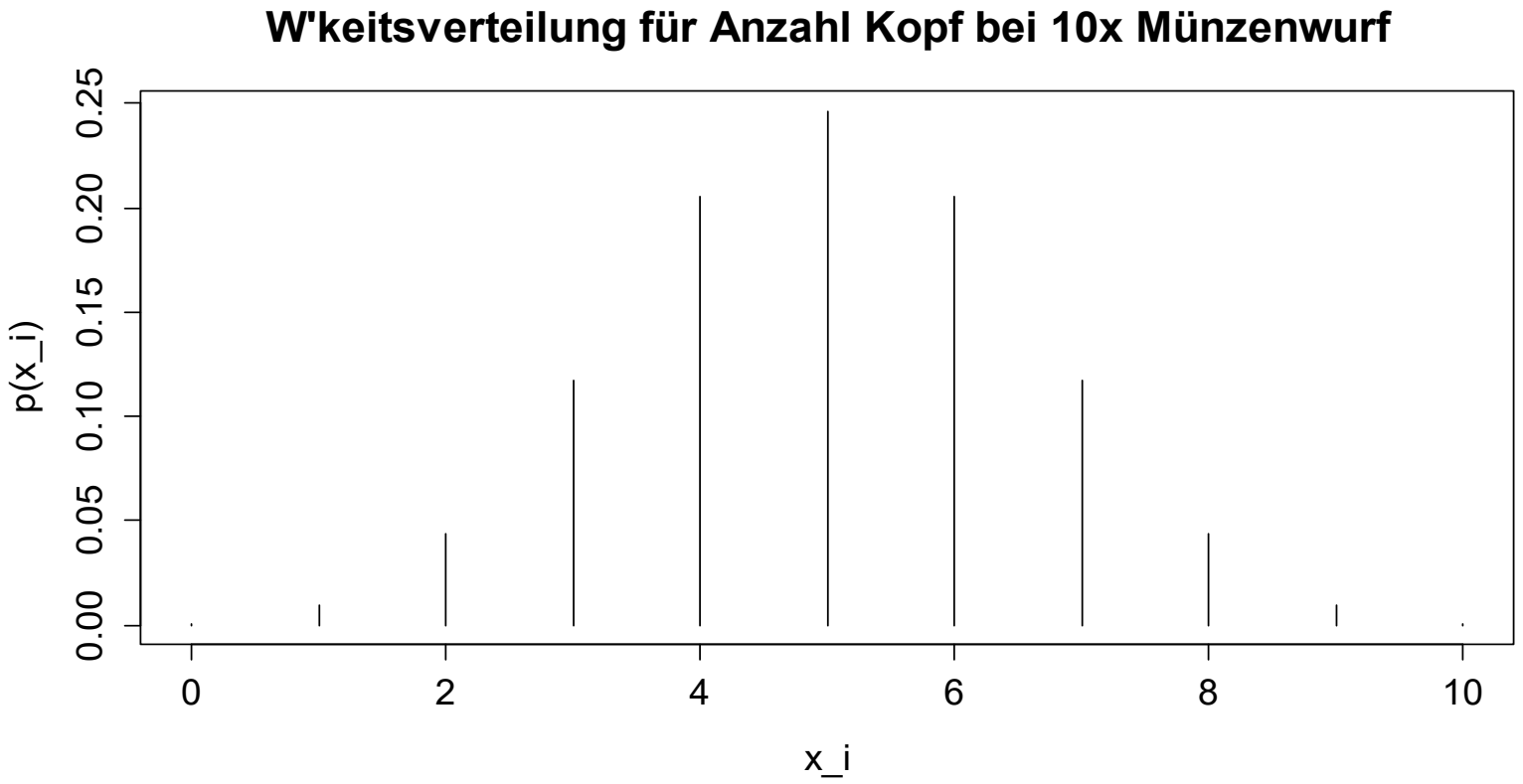
Weil $P(\Omega) = 1$ sein muss, gilt auch $\sum_i p(x_i) = 1$. Man kann die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariable mit einer Tabelle darstellen:

X	x_1	x_2	x_3	…	x_k
$p(\cdot)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	…	$p(x_k)$

Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Diskrete Wahrscheinlichkeitsfunktion

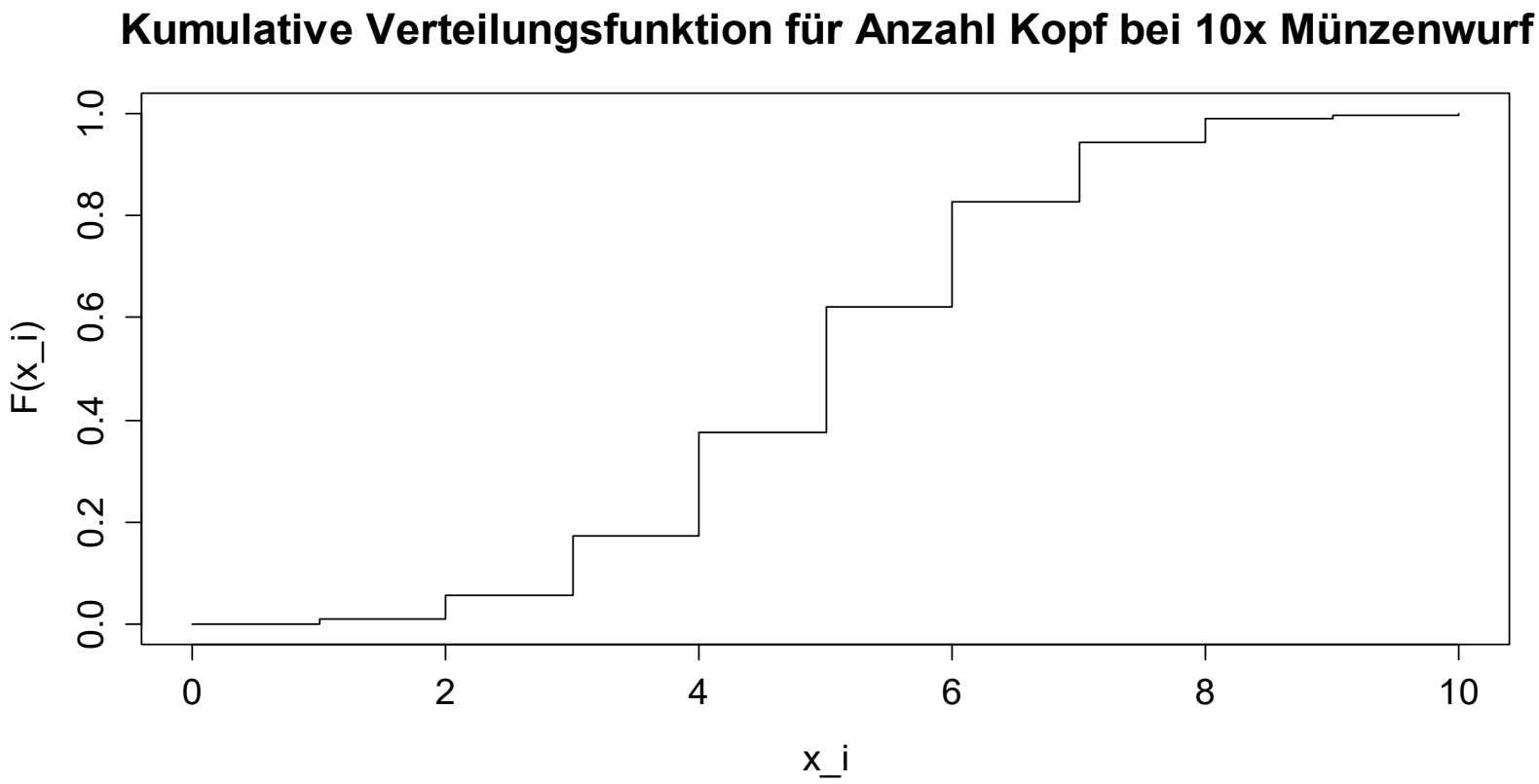
Y = «Anzahl Kopf bei 10x Münzenwurf»



Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Kumulative Verteilungsfunktion

$F(x_i) = P(X \leq x_i)$, wobei $F(-\infty) = 0$ und $F(+\infty) = 1$.



Grundlagen der Mathematik II
Lineare Algebra und Statistik
FS 2018 – Woche 08 & 09

Diskrete Wahrscheinlichkeitsverteilung

Grundsätzlich hat jede (diskrete) Zufallsvariable ihre eigene, individuelle Verteilung. In gewissen Situationen ist diese einfacher zu bestimmen, in anderen wird das schon schwieriger.

→ Siehe Beispiel an der Wandtafel mit dem Punktewert einer zufällig aus einem gut gemischten Stapel gezogenen Jasskarte.

Wie wir aber in der Folge sehen werden, tauchen einige prototypische Verteilungsfamilien bzw. Wahrscheinlichkeitsmodelle immer wieder auf. Wir führen diese nun der Reihe nach ein.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Marcel Dettling

Institute für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

ETH Zürich, 2. Mai 2018

Diskrete Wahrscheinlichkeitsverteilung

Es sei X eine beliebige diskrete Zufallsvariable. Wir bezeichnen die Werte, die X annehmen kann mit $x_1, x_2, ..., x_n$. Die zugehörigen Wahrscheinlichkeiten notieren wir mit $p(x_1), p(x_2), p(x_3), ...$

Es ist also: $p(x_i) = P(X = x_i)$

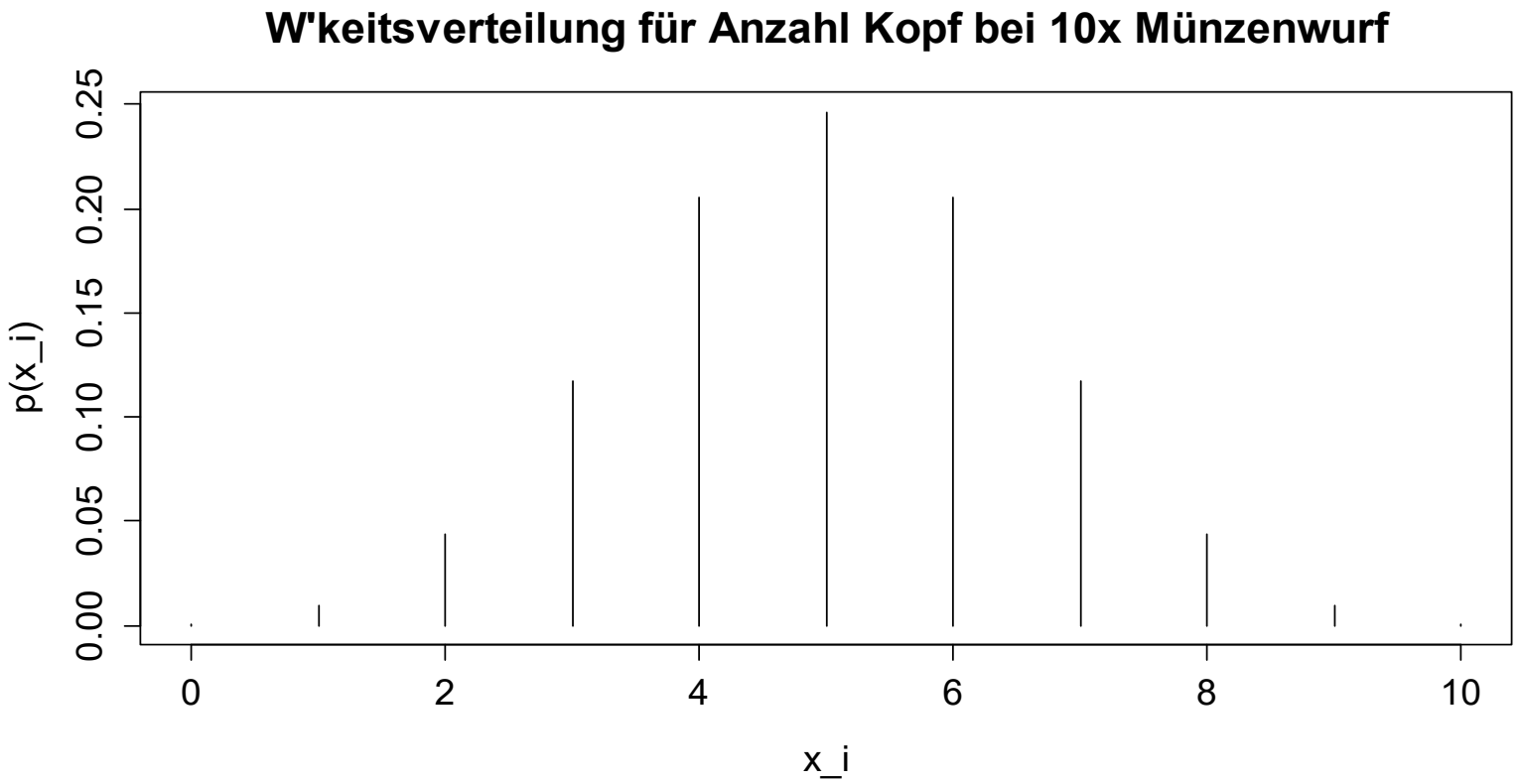
Weil $P(\Omega) = 1$ sein muss, gilt auch $\sum_i p(x_i) = 1$. Man kann die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariable mit einer Tabelle darstellen:

X	x_1	x_2	x_3	…	x_k
$p(\cdot)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	…	$p(x_k)$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

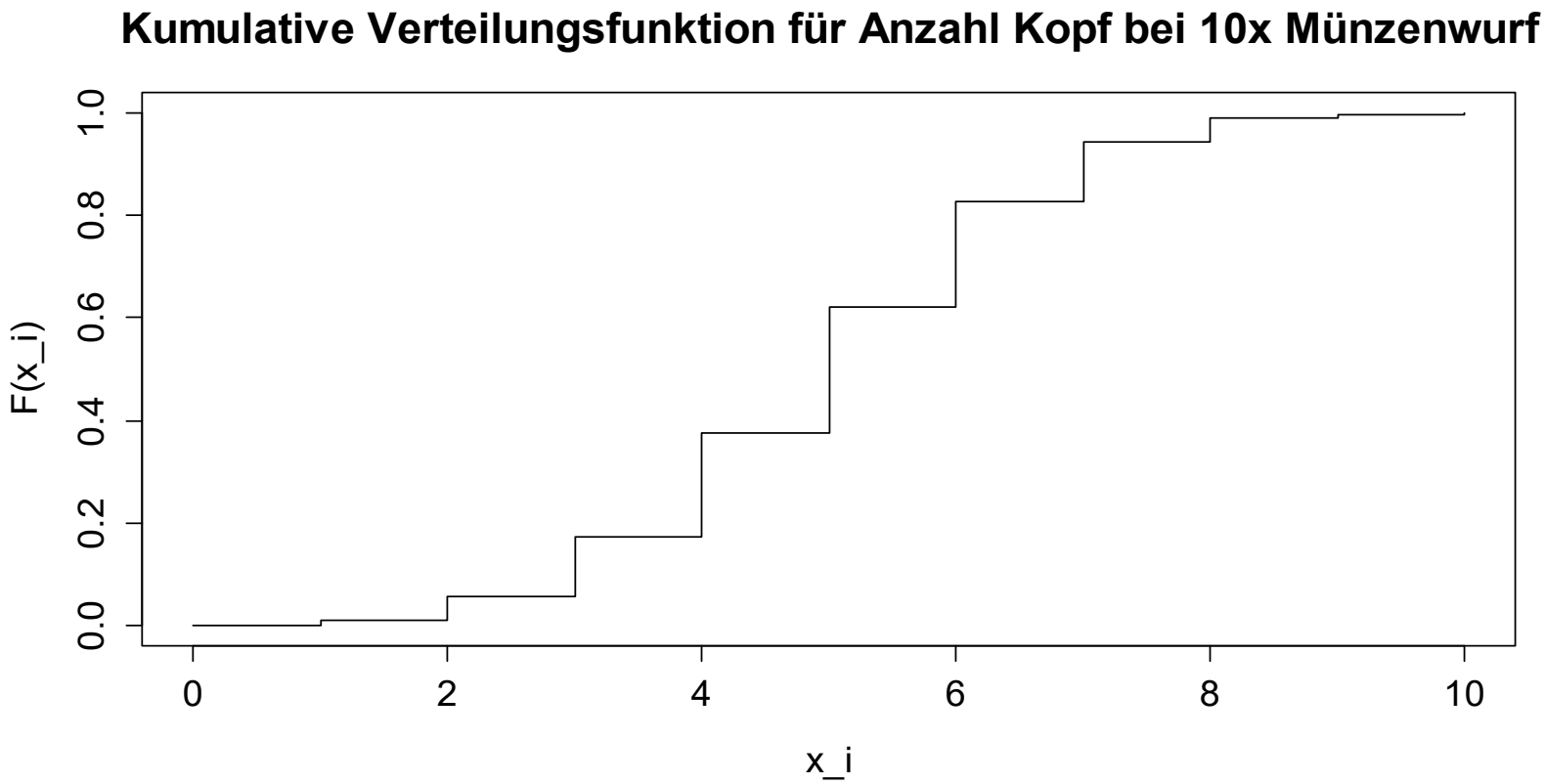
Diskrete Wahrscheinlichkeitsfunktion

Y = «Anzahl Kopf bei 10x Münzenwurf»



Kumulative Verteilungsfunktion

$F(x_i) = P(X \leq x_i)$, wobei $F(-\infty) = 0$ und $F(+\infty) = 1$.



Bernoulli-Verteilung

- Vermutlich die einfachste der diskreten Verteilungen
- Kommt bei ***Bernoulli-Experimenten*** zur Anwendung:
  - *es gibt nur 2 Ergebnisse*
  - *„Erfolg“ und „Misserfolg“*
  - *„Ja“ und „Nein“*
  - *Codiert mit 0 und 1*
- Zentral ist die ***Erfolgswahrscheinlichkeit*** p:

$p = P(X = 1)$, daraus erhält man auch:

$P(X = 0) = 1 - p$

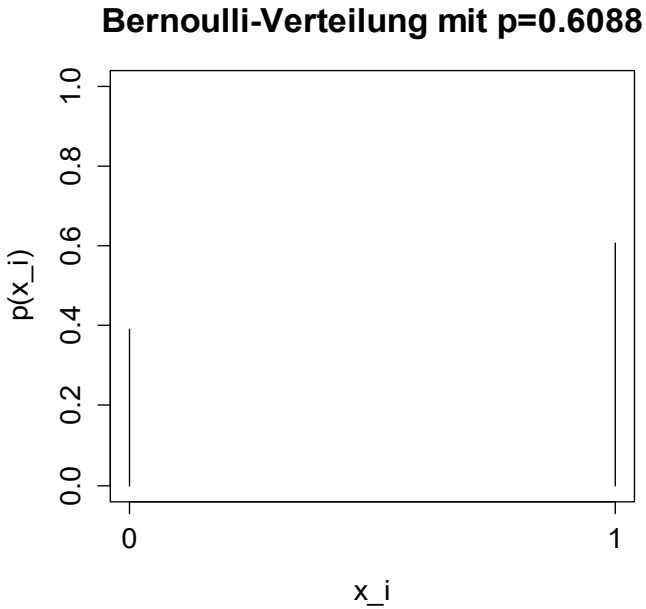
Beispiel zur Bernoulli-Verteilung

$X =$ "Ein Kandidat besteht die Fahrprüfung" $\sim Bernoulli(p)$

Den Parameter p können wir aus vergangenen Daten *schätzen*.
In ZH bestanden im Jahr 2011 total 15'100 von 24'801 Personen.

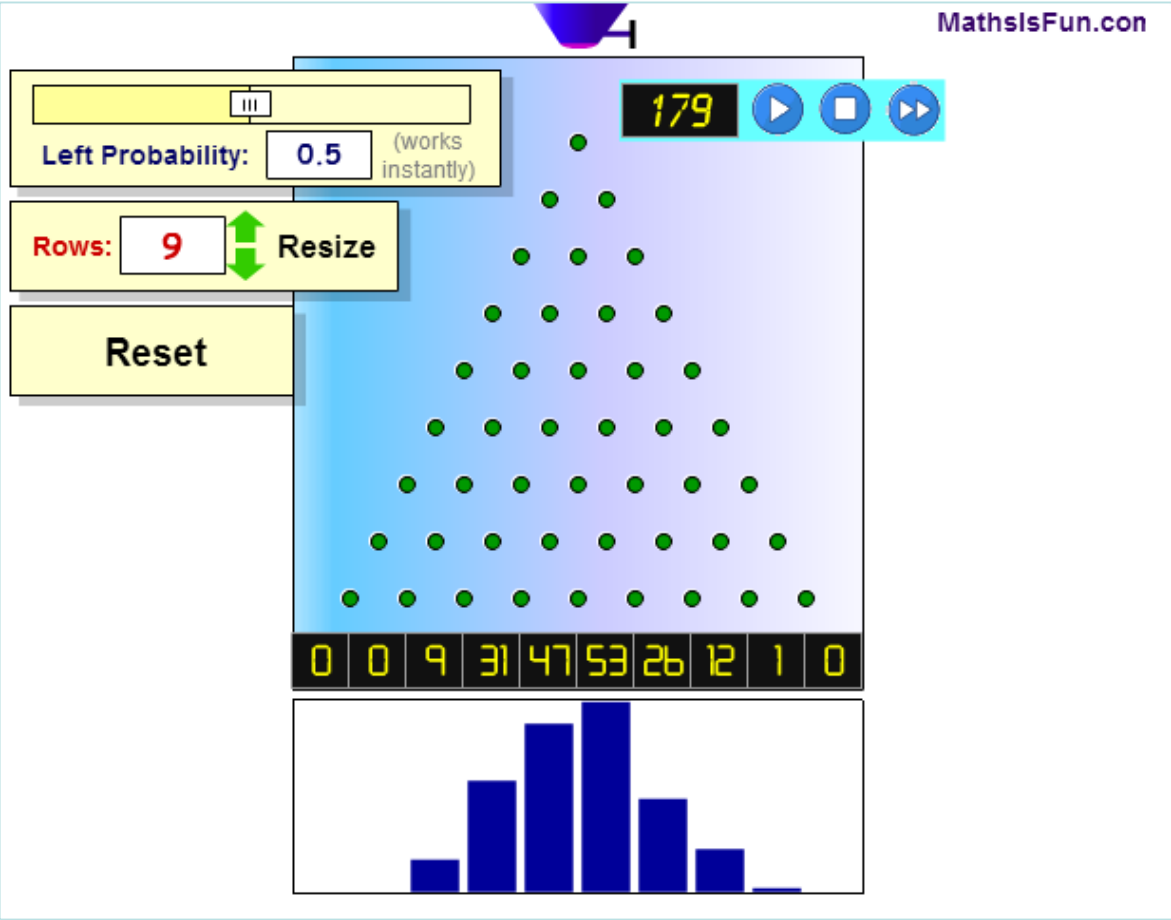
$$\hat{p} = \frac{15'100}{24'801} = 0.6088 = 60.88\%$$

Wir können die Verteilung grafisch als Stabdiagramm darstellen, siehe rechts.



GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Mehrstufige Bernoulli-Experimente



Wir haben:

9 Stufen

10 Urnen

$$p_{links} = p_{rechts} = 0.5$$

Mehrstufige Bernoulli-Experimente

Berechnung der Anzahl Möglichkeiten:

Es gibt offenbar die Tendenz, dass die Bälle bevorzugt in die Urnen in der Mitte fallen als in jene am Rand. Warum?

Wie viele Möglichkeiten, d.h. wie viele verschiedene Wege gibt es im 9-stufigen Brett, um in eine bestimmte Urne zu kommen?

1) 0x nach rechts und 9x nach links
2) 1x nach rechts und 8x nach links
3) 2x nach rechts und 9x nach links
4) …

Mehrstufige Bernoulli-Experimente

Berechnung der W'keit für jeden Pfad:

Wir gehen davon aus, dass die W'keiten $p_{links} = p_{rechts} = 0.5$.
Wie gross ist die Wahrscheinlichkeit für jeden einzelnen Pfad, wo die Kugel genau:

1) 0x nach rechts und 9x nach links gesprungen ist.
2) 1x nach rechts und 8x nach links gesprungen ist.
3) 2x nach rechts und 7x nach links gesprungen ist.

…

10) 9x nach rechts und 0x nach links gesprungen ist.

Mehrstufige Bernoulli-Experimente

Asymmetrische Situation: $p_{Misserfolg} = 0.75$, $p_{Erfolg} = 0.25$

Das Nagelbrett ist nun nicht mehr die beste Illustration. Wir denken an ein Geschicklichkeitsexperiment mit Schülern, wo die Erfolgsw'keit 25% beträgt.

Wir überlegen uns erneut Anzahl Möglichkeiten, um in die einzelnen Urnen zu gelangen, sowie die W'keiten von jedem Pfad, welcher in dieselbe Urne führt…

1) 0x Erfolg und 9x Misserfolg.
2) 1x Erfolg und 8x Misserfolg.
3) …

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Binomialverteilung

- Es werden n unabhängige Zufallsexperimente gemacht
- Jedes Experiment hat genau 2 Ausgänge: „Erfolg/Misserfolg“
- Die Erfolgswahrscheinlichkeit ist konstant und gleich p

Die Anzahl Erfolge X hat dann eine Binomialverteilung, wo die Wahrscheinlichkeit für k Erfolge gegeben ist durch:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Es handelt sich um eine diskrete W‘keitsverteilung, wo die Zahlen 0,1,2,...,n positive Masse haben. Notation:

$$X \sim Bin(n, p)$$

Beispiel zur Binomialverteilung

X = „Anzahl Fahrschüler von 7, welche die Prüfung bestehen“

→ Jeder Prüfling stellt für sich ein Bernoulli-Experiment mit $p = 0.6088$ dar. Natürlich können alle bestehen, oder alle durchfallen. Der Wertebereich ist $X \in \{0,1,2,3,4,5,6,7\}$.

→ Achtung, nicht jedes der Resultate ist gleich wahrscheinlich! Zur Bestimmung der Wahrscheinlichkeiten verwenden wir:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \quad \text{für } k = 0,1,2,...,n$$

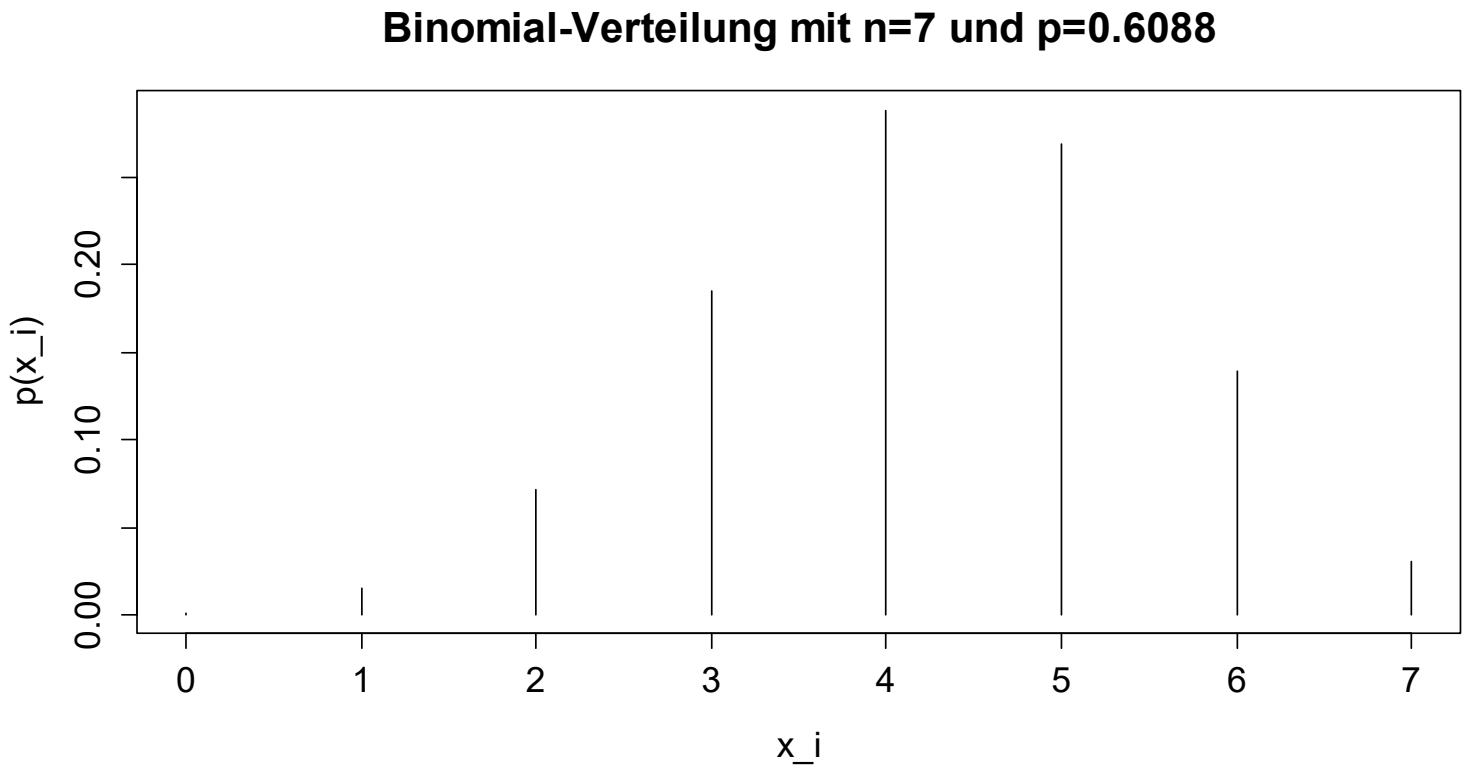
→ In R kann man solche W’keiten einfach berechnen, z.B.:

```
> dbinom(5, size=7, prob=0.6088)
[1] 0.2687758
```

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

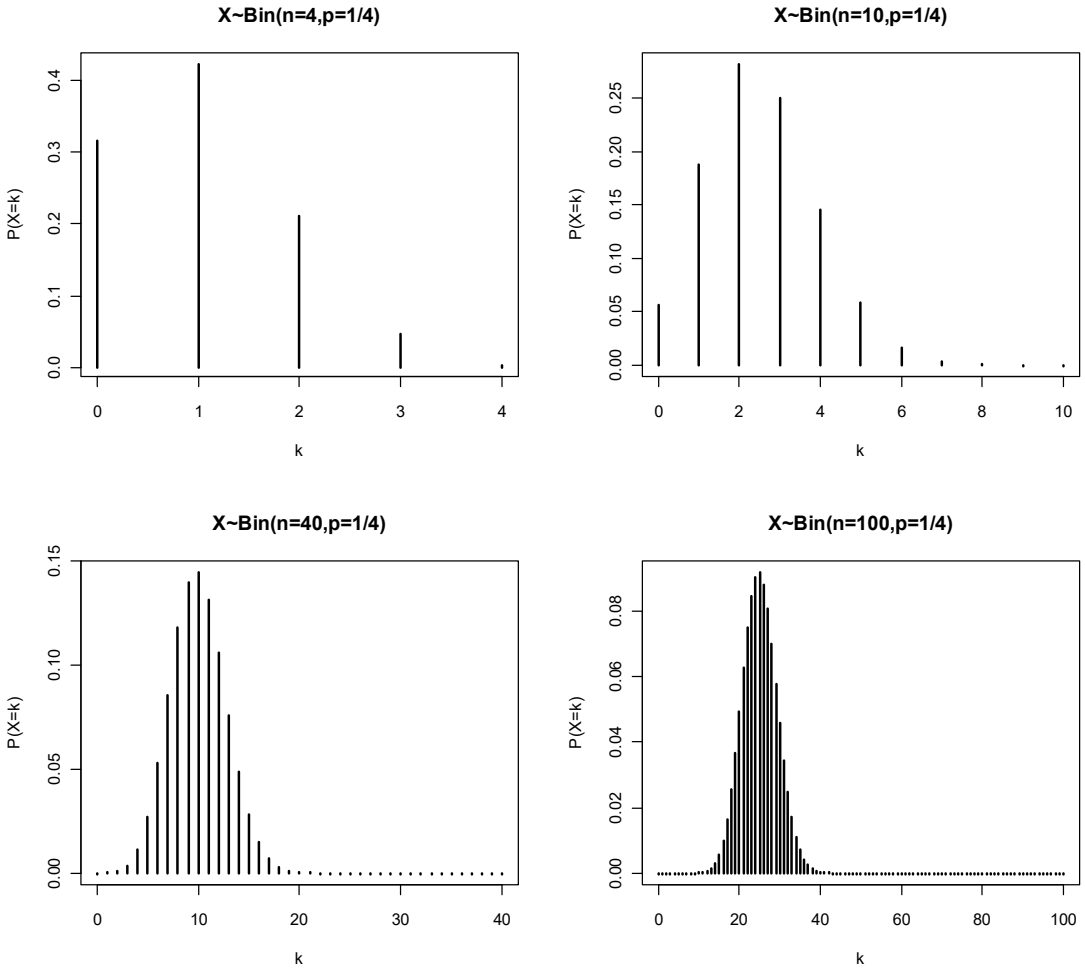
Grafische Darstellung der Verteilung in R

```
> xx <- 0:7
> yy <- dbinom(xx, size=7, prob=0.6088)
> plot(xx, yy, type="h", xlab=...)
```



GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Aussehen der Binomialverteilung



- Symmetrie für $p = \frac{1}{2}$
- Bei wachsendem n auch für $p \neq 1/2$ immer symmetrischer
- Faustregel: falls $np(1-p) > 10$, gilt eine jede $Bin(n, p)$-Verteilung als symmetrisch

Bestimmung der Parameter

- Die Anzahl Versuche n ist aus dem Kontext meist einfach abzulesen und macht keine Probleme.
- Die Erfolgswahrscheinlichkeit p muss hingegen meist aus den Daten geschätzt werden.

Beispiel: Ein Fussballkenner spielt seit über 10 Jahren fast jede Woche Sport-Toto. Er hat bereits 507 Mal ein Tippkolonne mit total $507 \cdot 13 = 6'591$ Tipps abgegeben. Davon hat er total 2902x richtig getippt. Seine Erfolgsw'keit beträgt:

$$\hat{p} = \frac{2902}{6591} = 0.44$$

Bestimmung der Parameter

Aufgabe:

Der Minigolfprofi E. Inlocher rühmt sich damit, dass er der "Hole-In-One"-Spezialist sei. Bei einer Vorführung trifft er in 28 von 34 Versuchen in einem Schlag.

a) wie gross schätzen sie die Erfolgsw'keit p*?*

b) geben sie die Verteilung der Zufallsvariablen $X = \text{"}Anzahl\ Treffer\ in\ 22\ Versuchen\text{"}$ *an.*

c) wie gross ist die W'keit, dass Inlocher in 22 Versuchen 20 oder mehr "Hole-In-Ones" schafft?

Poissonverteilung

Die Poisson-Verteilung eignet sich für Vorfälle, die im Laufe der Zeit eintreten oder sich an einem bestimmten Ort ereignen.

Man interessiert sich für die *Anzahl Vorkommnisse* in einer bestimmten Zeitspanne, oder einem festgelegten Gebiet.

Beispiele:

1) *Unfälle in einer Fabrik, auf Strassen, oder anderswo*

2) *Defekte in Geräten, an Fahr- oder Flugzeugen*

3) *Das Eintreffen von Klienten an einem Schalter*

Abgrenzung zur Binomialverteilung

- Wir haben es nicht mehr mit einer bekannten Anzahl von n Einzelversuchen zu tun, sondern die Grösse der Population ist unbekannt
- Der Wertebereich der Zufallsvariablen (d.h. die Anzahl Ereignisse, welche auftreten können) hat keine klar definierbare Obergrenze, d.h. $k = 0,1,2,...$
- Wir kennen nicht mehr wie bei der Binomial-Verteilung eine Erfolgsw'keit p für den Einzelversuch, sondern nur noch eine Rate λ, die beschreibt, mit welcher Häufigkeit (pro Zeiteinheit, Fläche, etc.) das Ereignis auftritt.

Poisson-W'keitsverteilung

- Es soll nicht der Anteil, sondern die absolute Häufigkeit eines bestimmten Ereignisses untersucht werden.
- Wenn die Ereignisse unabhängig voneinander mit einer konstanten Rate λ passieren, dann hat X = "Anzahl Ereignisse" eine Poisson-Verteilung.
- Die Wahrscheinlichkeit für *k* Ereignisse ist

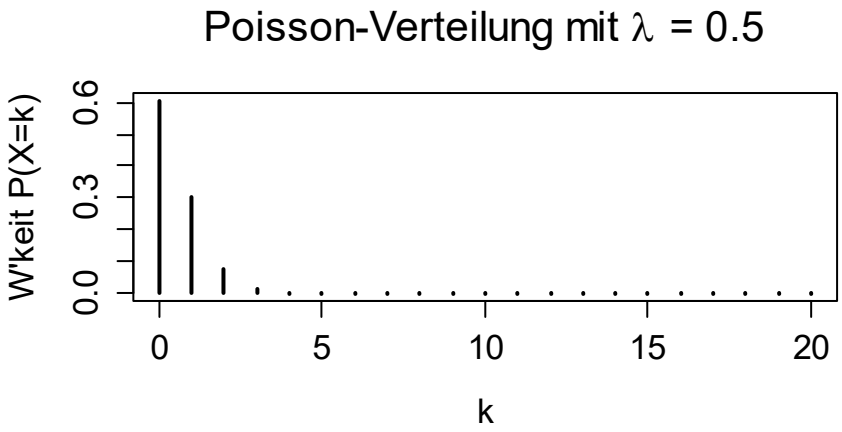
$$P(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

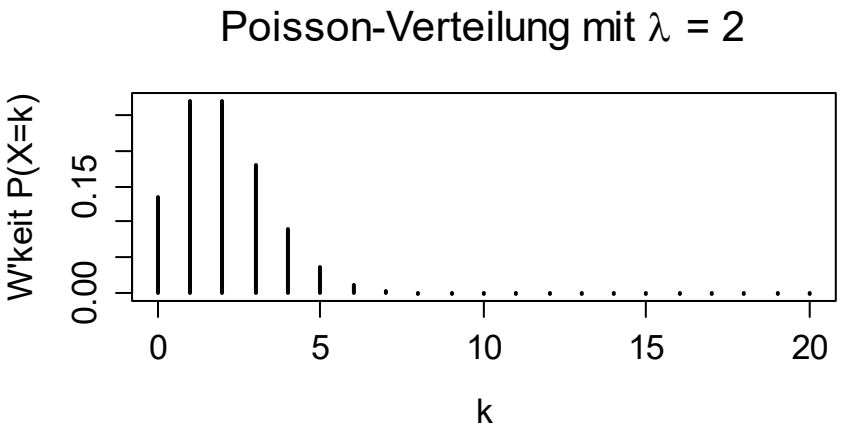
- Kleines λ: stark rechtsschief; grosses λ: symmetrisch

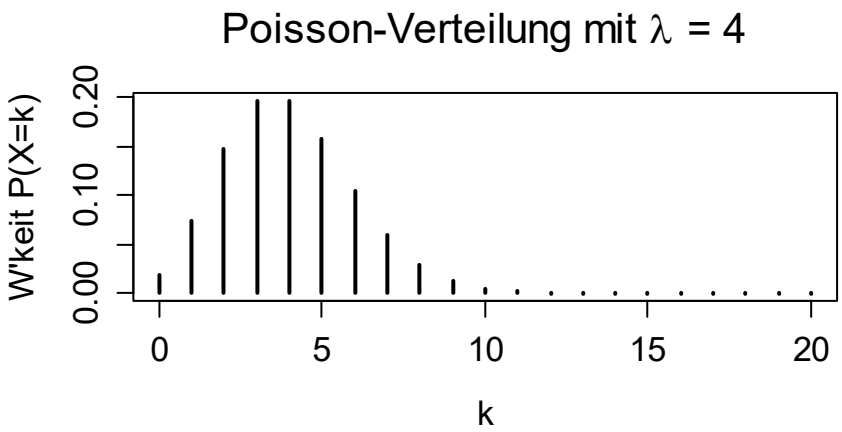
GdM 2: LinAlg & Statistik

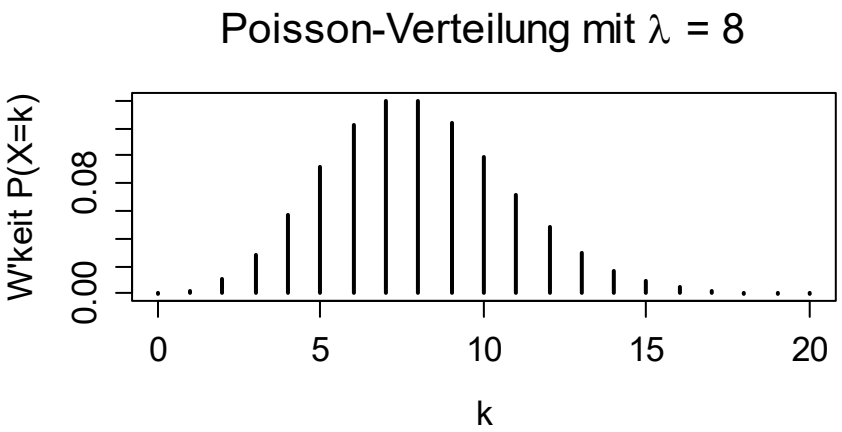
FS 2018 – Woche 10

Beispiele zur Poissonverteilung









Interpretation von λ

- Der Parameter λ der Poisson-Verteilung charakterisiert bereits die Rate, mit welcher die Ereignisse eintreffen. Wir "erwarten" also innerhalb einer Zeit/ Flächen-Einheit gerade λ Ereignisse.

Beispiel:

Wenn X die Anzahl tödliche Verkehrsunfälle pro Jahr in der Schweiz beschreibt, so gilt bei durchschnittlich 350 tödlichen Unfällen:

$$X \sim Pois(350)$$

Faustregel für die Poissonverteilung

Wir können damit eine rasche, grobe Abschätzung treffen, wie viele Ereignisse sich in der nächsten Periode abspielen. Es gilt nämlich: ist $\lambda > 10$, so beobachten wir "normalerweise"

$$\lambda \pm 2 \cdot \sqrt{\lambda} \quad \text{Ereignisse,}$$

bzw. genauer:

$$P[\lambda - 2\sqrt{\lambda} < X < \lambda + 2\sqrt{\lambda}] \approx 95\%$$

Übungsaufgabe:
Rechnen sie die Faustregel für die Verkehrsunfälle nach!

Bestimmung der Parameter: Poisson

Aufgabe:

Von der Rega-Basis Dübendorf werden im Schnitt pro Tag 7 Einsätze geflogen.

a) wie gross schätzen ist die Rate λ?

b) wir betrachten nun eine Schicht von 8 Stunden Dauer. Geben sie die Verteilung von $X = \text{"}Anzahl\ Einsätze\text{"}$ an.

c) wie gross ist die W'keit, dass die Crew in den 8 Stunden i) gar nicht, bzw. ii) mehr als 3 mal ausrücken muss?

Wahl der passenden Verteilung

1) In einer Packung mit 100 Schrauben sind 10 defekte darunter. Grösse von Interesse: Anzahl defekter Schrauben unter 20 zufällig ohne Zurücklegen heraus gegriffenen.

   □ Binomial □ Poisson □ Andere

2) Im Schnitt kommt auf 20 Seiten eines Buches 1 Druckfehler. Grösse von Interesse: Anzahl Druckfehler in einem 250-seitigen Buch. Wie sieht es aus, wenn wir stattdessen die Zufallsvariable „Anzahl Seiten in einem 250-seitigen Buch mit mind. 1 Druckfehler“ betrachten?

   □ Binomial □ Poisson □ Andere

Wahl der passenden Verteilung

1) Die Erfolgsquote beim Elfmeter sei 75%. Grösse von Interesse: Versenkte Elfmeter (von 5) im Penaltyschiessen.

   □ Binomial □ Poisson □ Andere

2) In der preussischen Armee starben in 20 Jahren 122 Soldaten an den Folgen eines Huftritts. Grösse von Interesse: Anzahl Tote im nächsten Jahr

   □ Binomial □ Poisson □ Andere

3) Die Lampen der Pistenbeleuchtung werden alle 2 Wochen, kontrolliert. Grösse von Interesse: Kontrolle, bei welcher die Lampe defekt gefunden wird.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Anwendungsaufgabe Geburten

Appenzell gehen die Frauen aus

SCHREI ⇢ Laut Bundesamt-Statistik werden in Appenzell und Obwalden die meisten Buben geboren.

Genau 80 808 Kinder wurden 2011 in der Schweiz geboren. Und die Mütter waren im Schnitt 31 Jahre alt. Dies die neusten Erhebungen des Bundesamtes für Statistik (BfS). Rund 92 Prozent der Neugeborenen wurden als Termingeburten erfasst.

Interessant ist das mittlere Geburtsgewicht der Kinder: Es pendelte sich laut BfS in den letzten 35 Jahren bei 3296 Gramm ein. **Lediglich 2,2 Prozent der Neugeborenen kamen mit einem Fliegengewicht von weniger als 2000 Gramm zur Welt.**

Witzig ist die Aufschlüsselung der Geburten nach Kantonen und Geschlecht. Seit mehr als einem Jahrhundert werden laut BfS in der Schweiz mehr Knaben als Mädchen geboren.

1898 wurde der niedrigste Wert (103) und 1972 der höchste Wert mit 107 Knaben pro 100 Mädchen verzeichnet.

Aktuell könnte man **Appenzell Ausserrhoden als Bubenkanton bezeichnen.** Dort wurden 2011 41 Prozent mehr Buben geboren als Mädchen. Auch in Obwalden kamen 20 Prozent mehr Knaben zur Welt. Doch die Girls holen auf. Als Frauenkanton bezeichnen könnte man dagegen Schaffhausen und das Wallis. In diesen beiden Kantonen veränderte sich der Bubenüberschuss zur pinken Mädchenwelt.

Das Geschlechterverhältnis schlägt bei den jungen Erwachsenen (25 bis 30 Jahre) und bei den Personen im Rentenalter ins Gegenteil um – dort verzeichnet das BfS dann wahre Frauenüberschüsse. **kmu**

Neuer Erdenbürger
So sieht ein glückliches Durchschnittsbaby aus.

Fotos: Keystone (2), www.plainpicture.com, Screenshot Tele Züri, Webcam SRK, ZVG

Blick am Abend, 16.10.12

- Was ist an den Kantonen Appenzell AR und Obwalden speziell?
- Warum sind die hier gemachten Angaben (41%) schwer interpretierbar?
- Wie müsste man vergleichen, um sinnvolle Aussagen machen zu können?

GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Anwendungsaufgabe Geburten

Datenquelle: Bundesamt für Statistik, www.bfs.admin.ch

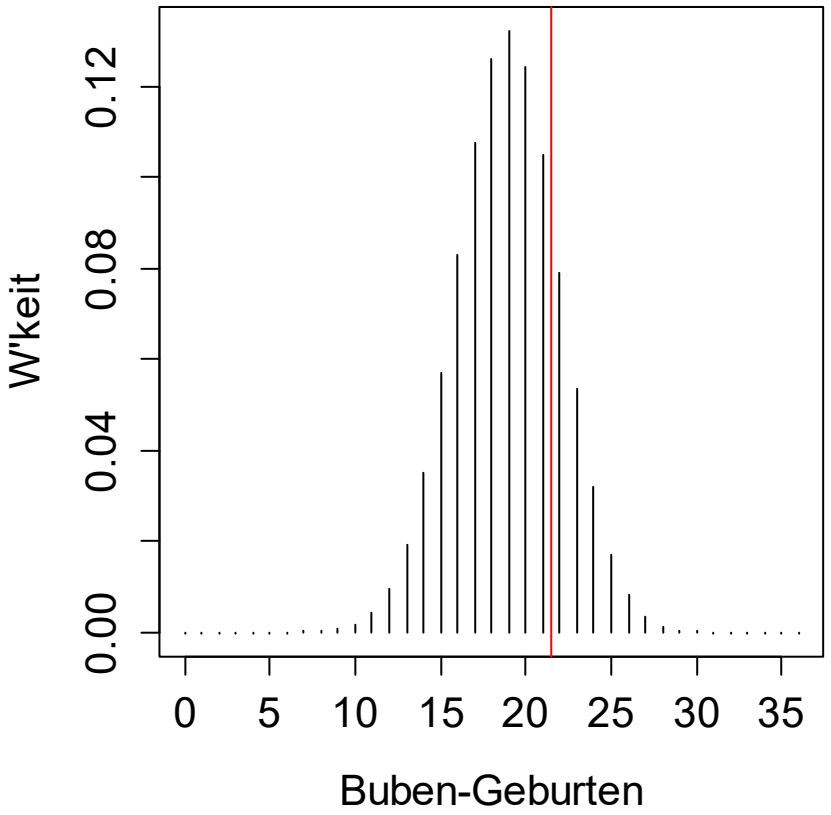
Aug 12 Lebendgeburten nach Kanton

Kantone	Lebendgeburten							
	Total	Nach Geschlecht		Nach Zivilstand der Mutter		Nach Staatsangehörigkeit des Kindes		Anteil Knaben
		Knaben	Mädchen	Verheiratete Mütter	Nicht verheiratete Mütter	Schweiz	Ausland	Total
Schweiz	6 962	3 658	3 304	5 630	1 332	5 287	1 675	0.525
Zürich	1 310	680	630	1 046	264	992	318	0.519
Bern	884	452	432	712	172	772	112	0.511
Luzern	363	192	171	292	71	295	68	0.529
Uri	36	22	14	31	5	29	7	0.611
Schwyz	124	64	60	108	16	99	25	0.516
Obwalden	39	22	17	33	6	36	3	0.564
Nidwalden	39	18	21	30	9	36	3	0.462
Glarus	30	12	18	21	9	23	7	0.400
Zug	115	62	53	97	18	93	22	0.539
Fribourg	261	141	120	197	64	199	62	0.540

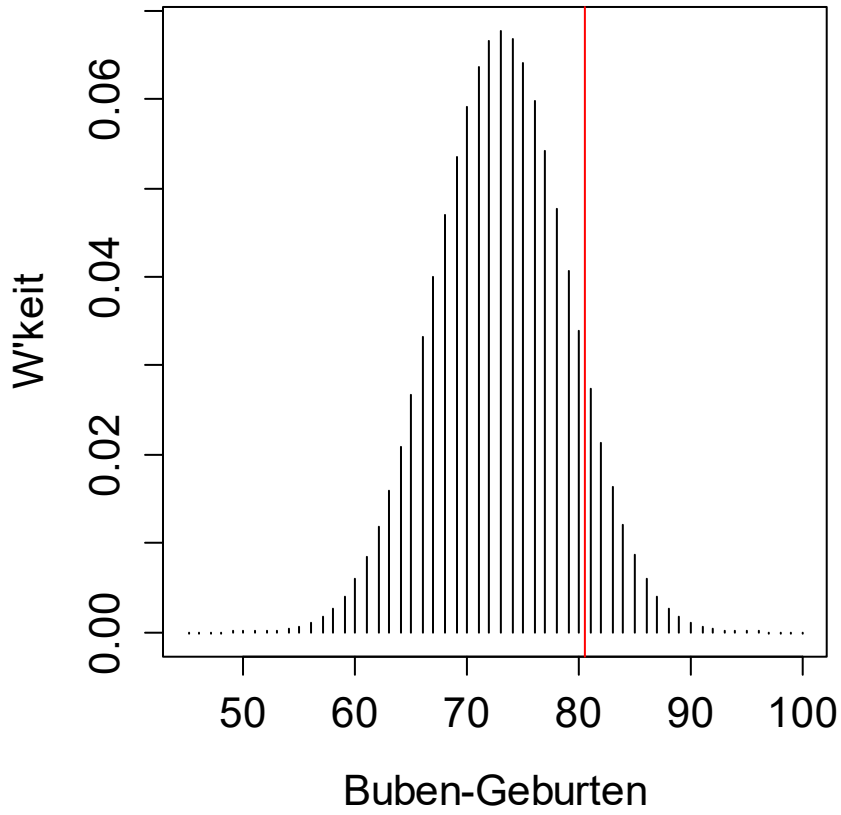
GdM 2: LinAlg & Statistik
FS 2018 – Woche 10

Anwendungsaufgabe Geburten

UR: Bin(n=36, p=0.525), 22 Buben



GR: Bin(n=139, p=0.525), 81 Buben



GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Marcel Dettling

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

Winterthur, 9. Mai 2018

Zufallsvariablen: Diskret oder Stetig?

- *Eine diskrete Zufallsvariable X liegt dann vor, wenn sie endlich oder unendlich abzählbar viele Werte annimmt.*

  Beispiele: - überall, wo gezählt wird
  - Fahrzeuge, Krankheitsfälle, Ausschuss, ...

- *Eine stetige Zufallsvariable X liegt dann vor, wenn sie jeden beliebigen Wert eines Intervalls annehmen kann.*

  Beispiele: - überall, wo gemessen wird
  - Wartezeit, Hämatokrit, Regenmenge, ...

Wahrscheinlichkeitsverteilungen

Eine Wahrscheinlichkeitsverteilung gibt an, welche Werte eine Zufallsvariable mit welcher Wahrscheinlichkeit annimmt.

Je nach Ausprägung der Zufallsvariablen (diskret/stetig) gibt es die entsprechenden Wahrscheinlichkeitsverteilungen

- ***Diskrete Wahrscheinlichkeitsverteilungen***
- ***Stetige Wahrscheinlichkeitsverteilungen***

Die diskrete Wahrscheinlichkeitsfunktion ist bekanntlich das Stabdiagramm. Es stellt sich nun die Frage, wie sich das Konzept auf stetige Zufallsvariablen verallgemeinern lässt.

Wartezeiten am Flughafen

An einem Schalter (z.B. am Flughafen) können pro Minute 6 Personen abgefertigt werden. Es treffen im langfristigen Schnitt nur 4 Personen pro Minute ein.

Bisher: X = „Anzahl eintreffender Personen“ $\sim Poi(\lambda)$

Neu: Y = „Zeitdauer bis nächster Kunde kommt“ $\sim ???$

Während z.B. die Anzahl der eintreffenden Personen nur ganzzahlige Werte annehmen kann, ist dies für die Wartezeit nicht der Fall. Diese ist von kontinuierlichem Charakter.

→ Y ist eine stetige Zufallsvariable mit stetiger Verteilung

Wie könnte diese Verteilung aussehen?

Wir betrachten zu einem beliebigen Zeitpunkt die nächste Sekunde. Die Anzahl Kunden X , welche in der nächsten Sekunde am Schalter eintrifft, hat eine Poisson-Verteilung:

$$X \sim Pois(\lambda = 4/60)$$

Wir bestimmen die W'keit, dass in der nächsten Sekunde kein Kunde eintrifft, die Wartezeit also länger als 1s dauert:

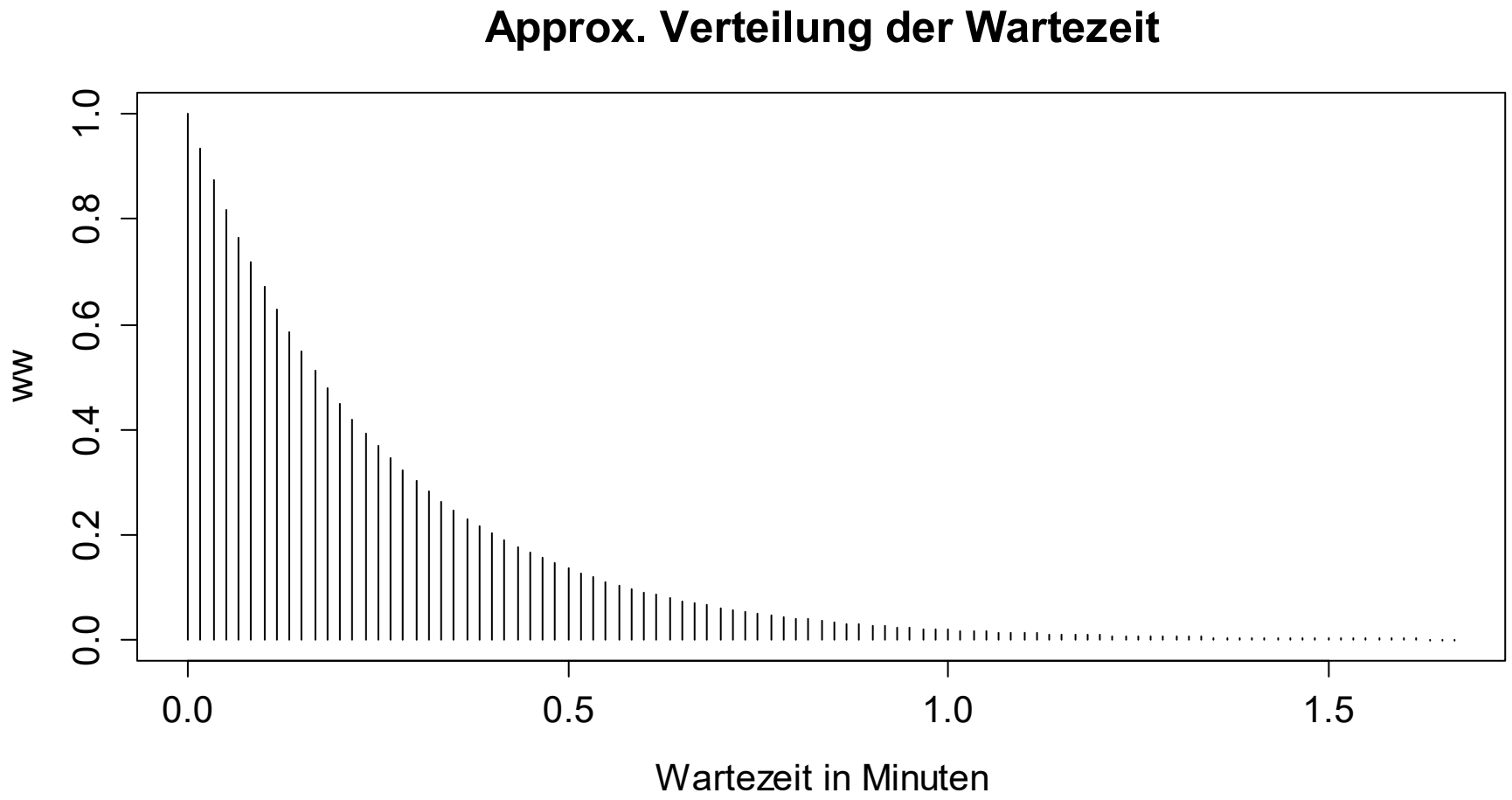
$$P(X = 0) = \exp(-4/60) \approx 0.936$$

Die W'keit, dass nach y Sekunden noch kein Kunde da ist:

$$P(Y > y) \approx P(X = 0)^y \approx 0.936^y$$

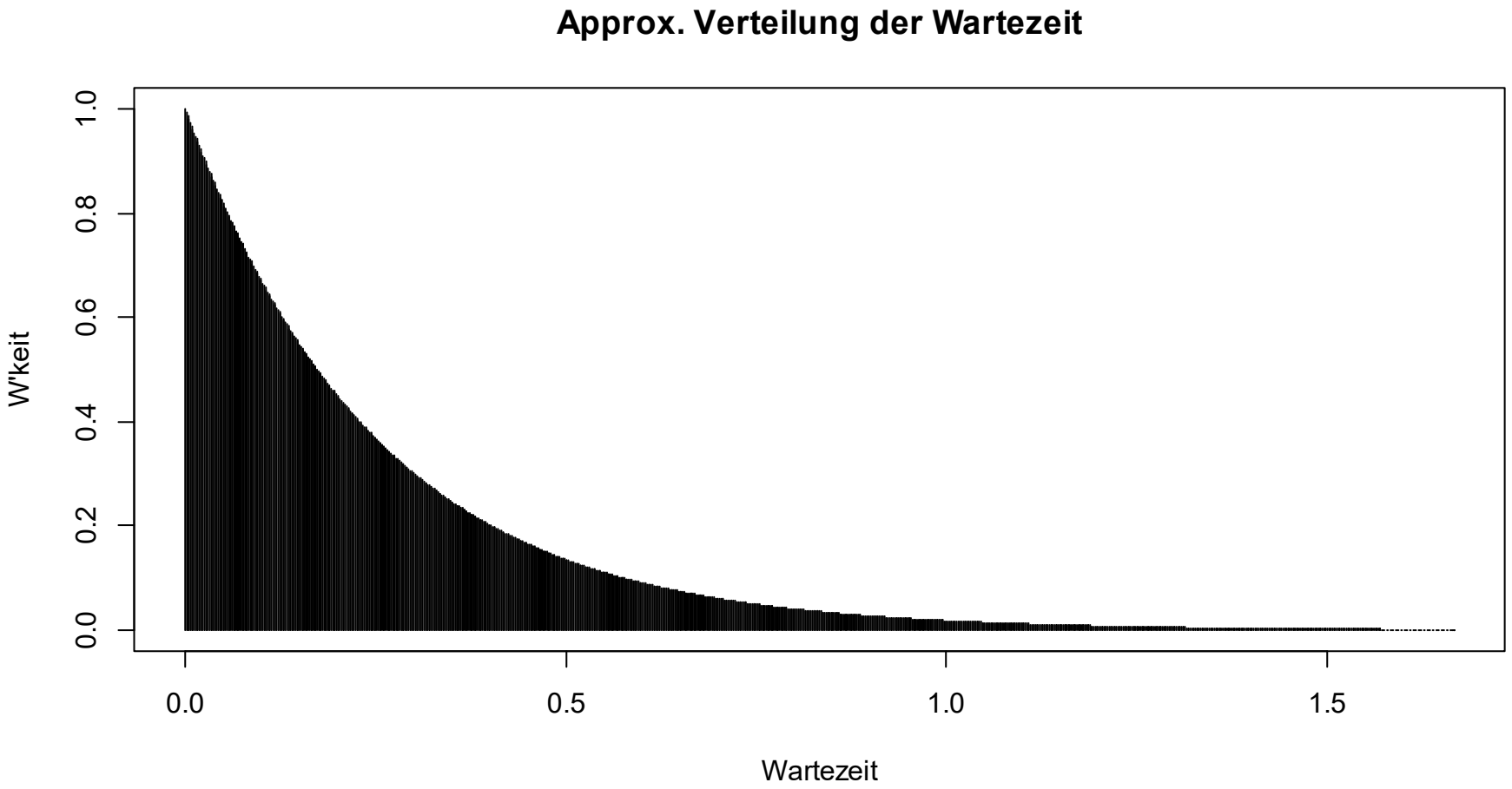
GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Wie könnte diese Verteilung aussehen?



GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Wie könnte diese Verteilung aussehen?



Dichtefunktion

→ Im Grenzübergang erhalten wir eine Dichtefunktion.

Für die Wartezeiten zwischen dem unabhängigen Eintreffen von Personen ist die Exponentialverteilung ein geeignetes Modell. Deren Dichtefunktion ist wie folgt definiert:

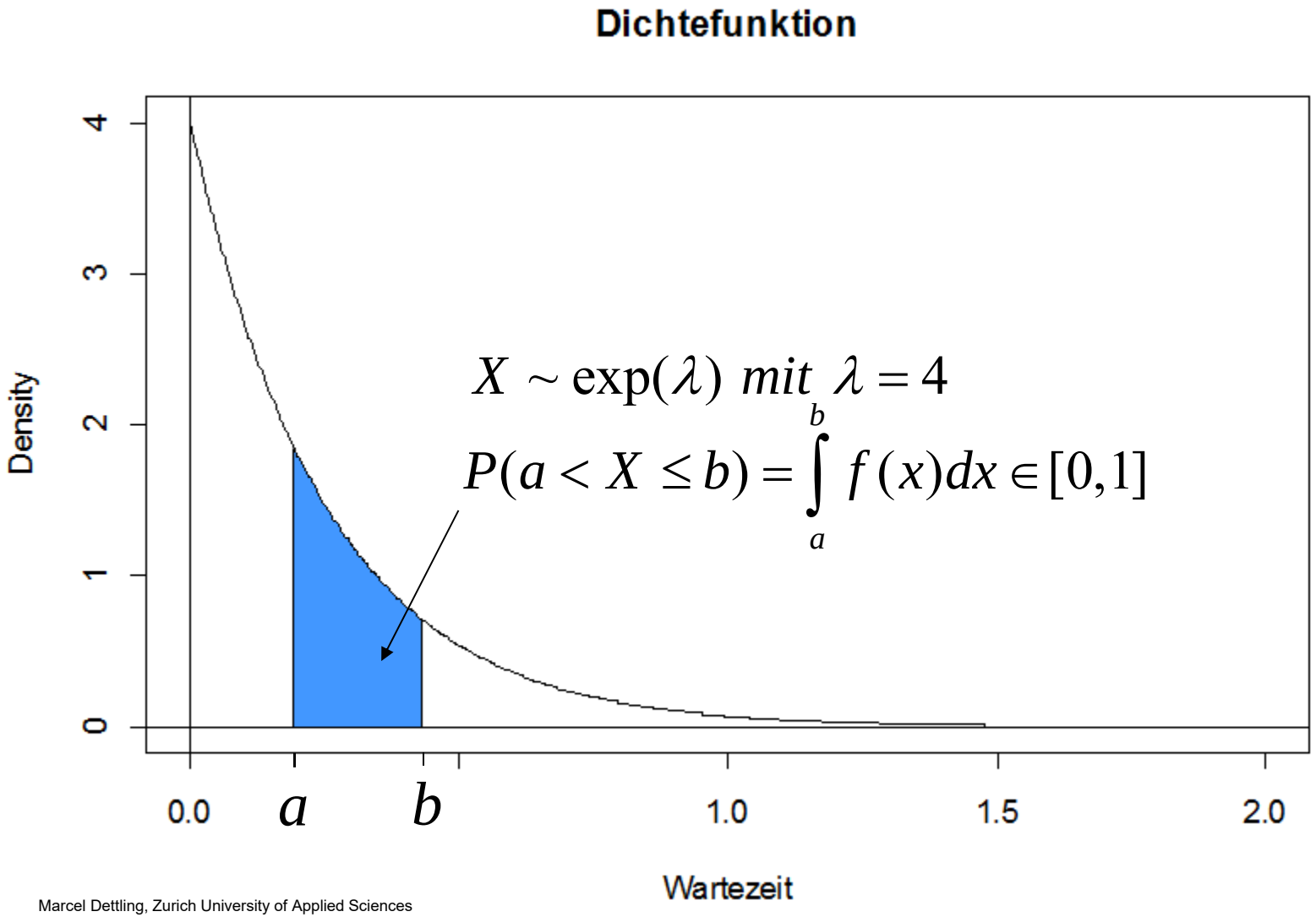
$$f(x) = \lambda \cdot e^{-\lambda x} \quad \textit{für} \quad x \geq 0$$

$$f(x) = 0 \qquad \textit{für} \quad x < 0$$

Die Exponentialverteilung bezeichnet man als *"gedächtnislos"*. Das Eintreffen des letzten Kunden hat keinen Einfluss auf das Eintreffen des nächsten und alle folgenden Kunden.

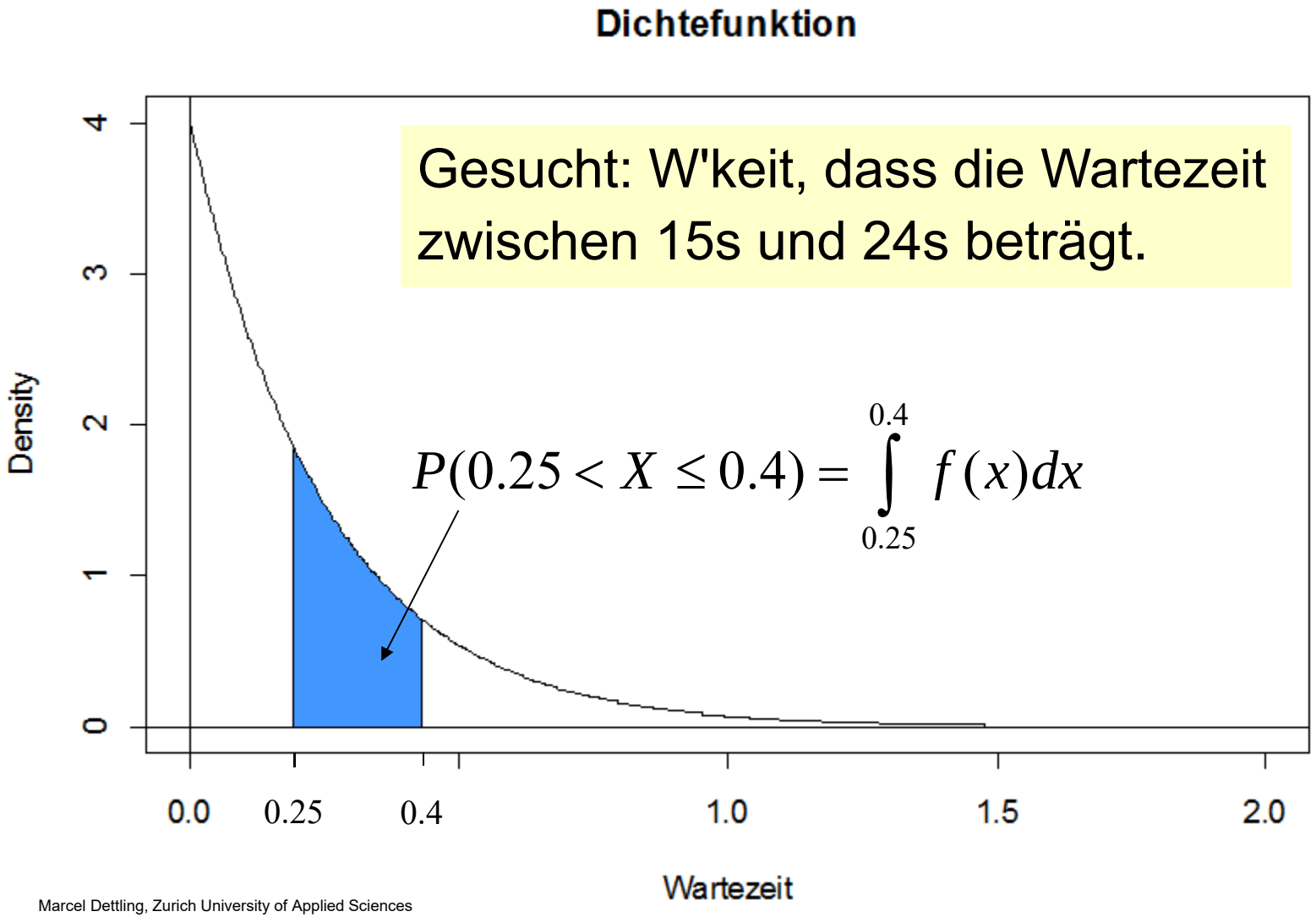
GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Bedeutung der Dichtefunktion



GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Bedeutung der Dichtefunktion



Kumulative Verteilungsfunktion

Die kumulative Verteilungsfunktion $F(\cdot)$ ist definiert durch:

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

Es ist die Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert kleiner als x annimmt. Sie ist bestimmt durch das Integral über die Dichtefunktion vom linken Rand bis zum Wert x.

Konkrete Durchführung für die Exponentialverteilung:

→ **siehe Wandtafel...**

Allgemeine Definition der Dichtefunktion

Die Dichte $f(\cdot)$ ist eine stückweise stetige Funktion mit

$$f(x) \geq 0$$

und

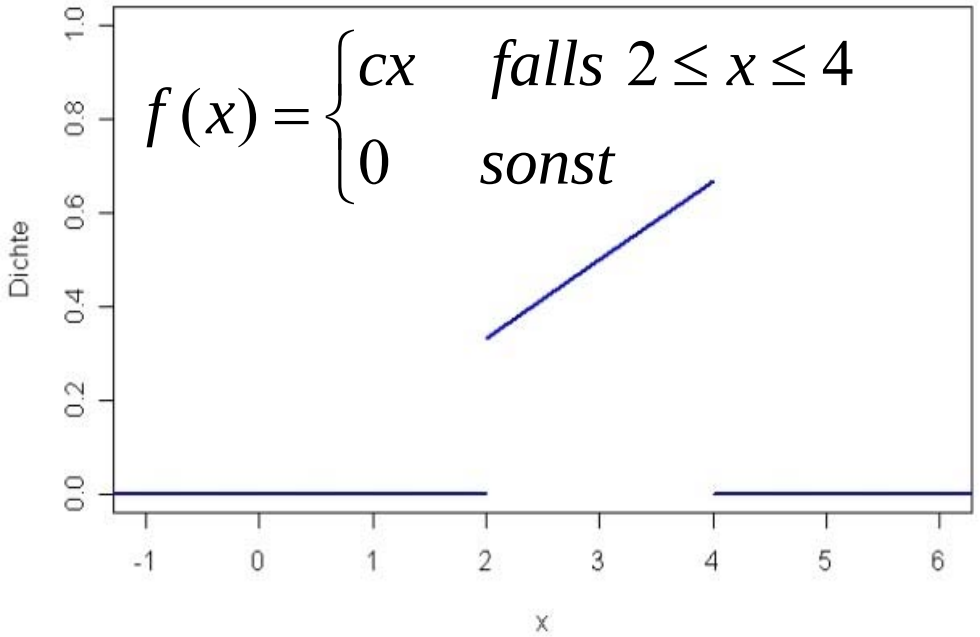
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Oder in Worten: die Dichtefunktion nimmt keine negativen Werte an, und das Integral über den ganzen Bereich der reellen Zahlen ist gleich 1.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Rechnen mit stetigen Verteilungen

In einem Betrieb verdienen Aushilfskräfte zwischen 2000 und 4000 CHF brutto. Es sei X der Lohn, für welchen folgende Dichtefunktion gilt:



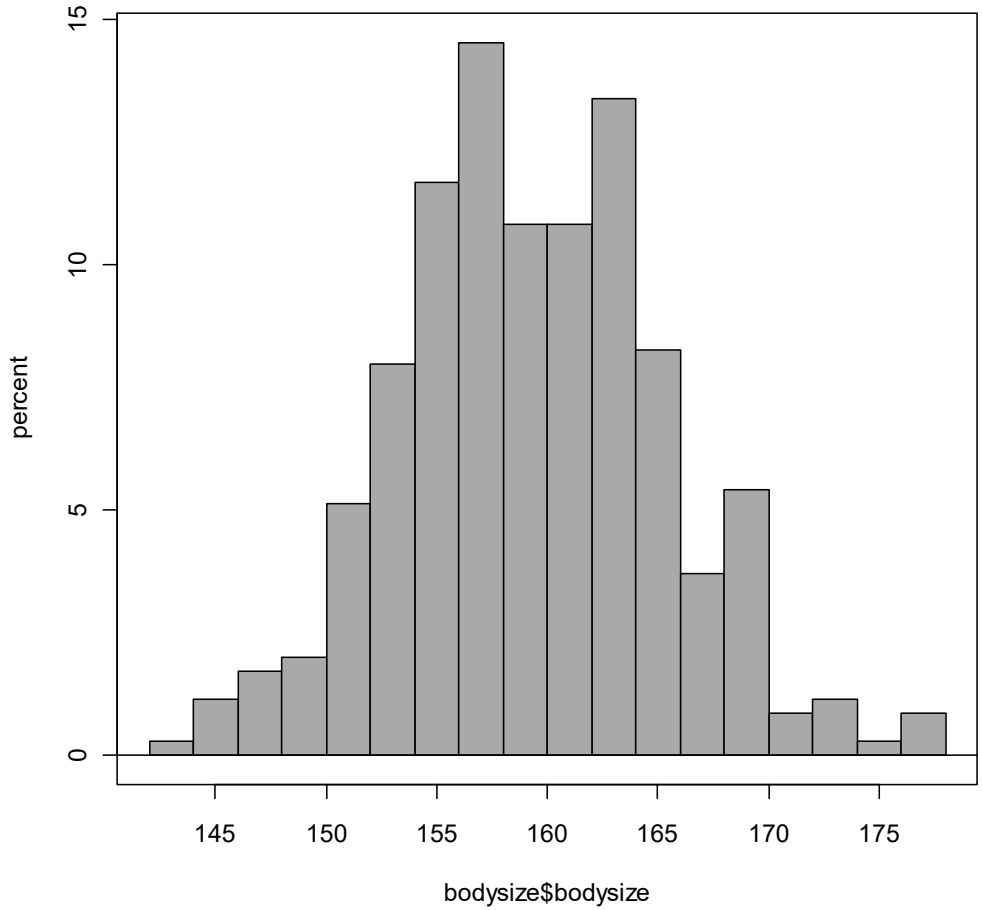
a) Gibt es mehr Leute, die zwischen 3000 und 4000 als unter 3000 verdienen?

b) Wie ist die Konstante c ?

c) Wie gross ist die W'keit für einen Lohn zwischen 2500 und 3000?

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Welche Verteilung haben Messdaten?



Histogramm der Körper-grösse von 350 zufällig ausgewählten Frauen

Normalverteilung

- Die Normalverteilung kommt dann zum Einsatz, wenn stetige Variablen einer symmetrischen, unimodalen Verteilung folgen.
- Dies ist typischerweise der Fall, wenn sich das Resultat eines Zufallsexperiments aus der additiven Überlagerung von vielen kleinen Einflüssen ergibt.
- Die Dichtefunktion der Normalverteilung lautet:

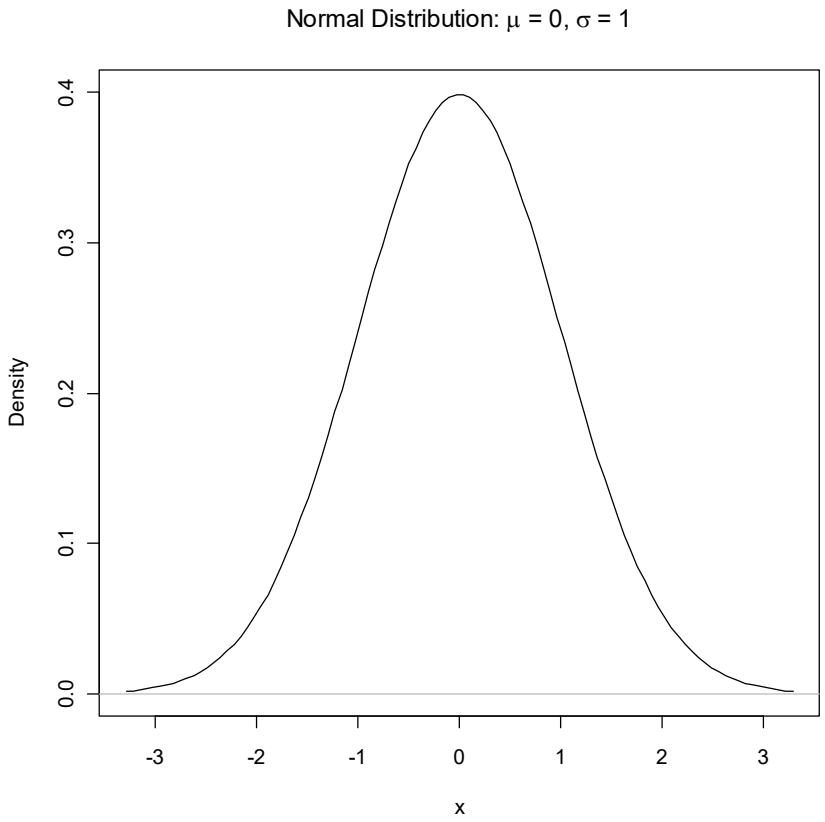
$$f(x) := \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left\{ -\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2} \right\}$$

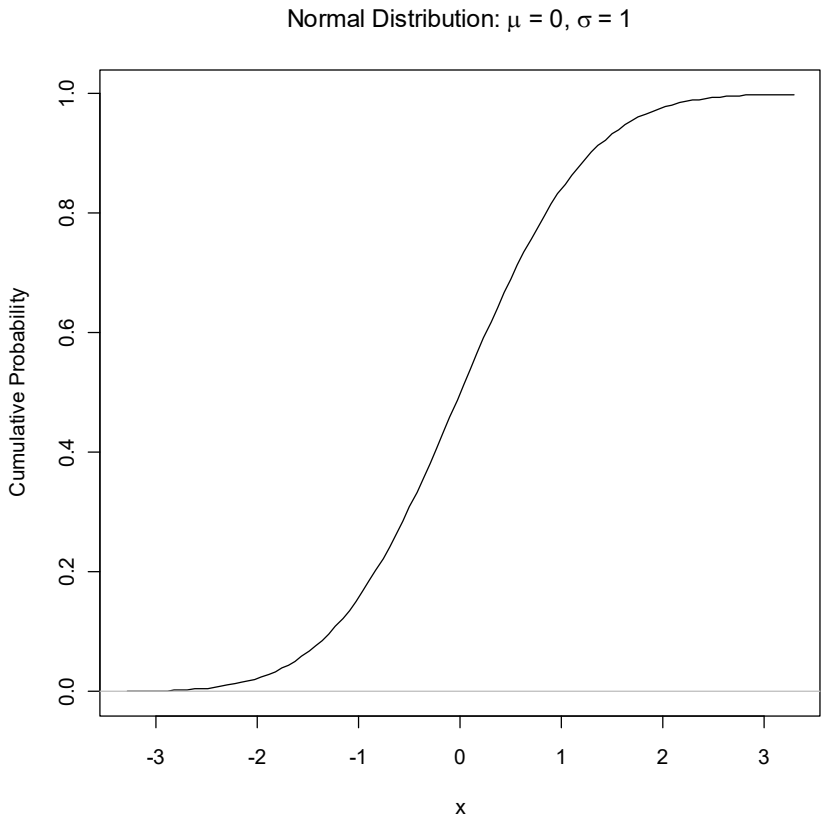
Die Parameter der Verteilung sind μ und σ^2. Sie beschreiben Lage und Streuung. Wir schreiben kurz $X \sim N(\mu, \sigma^2)$.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Dichte und Verteilungsfunktion

Wir betrachten die Zufallsvariable $X \sim N(0,1)$

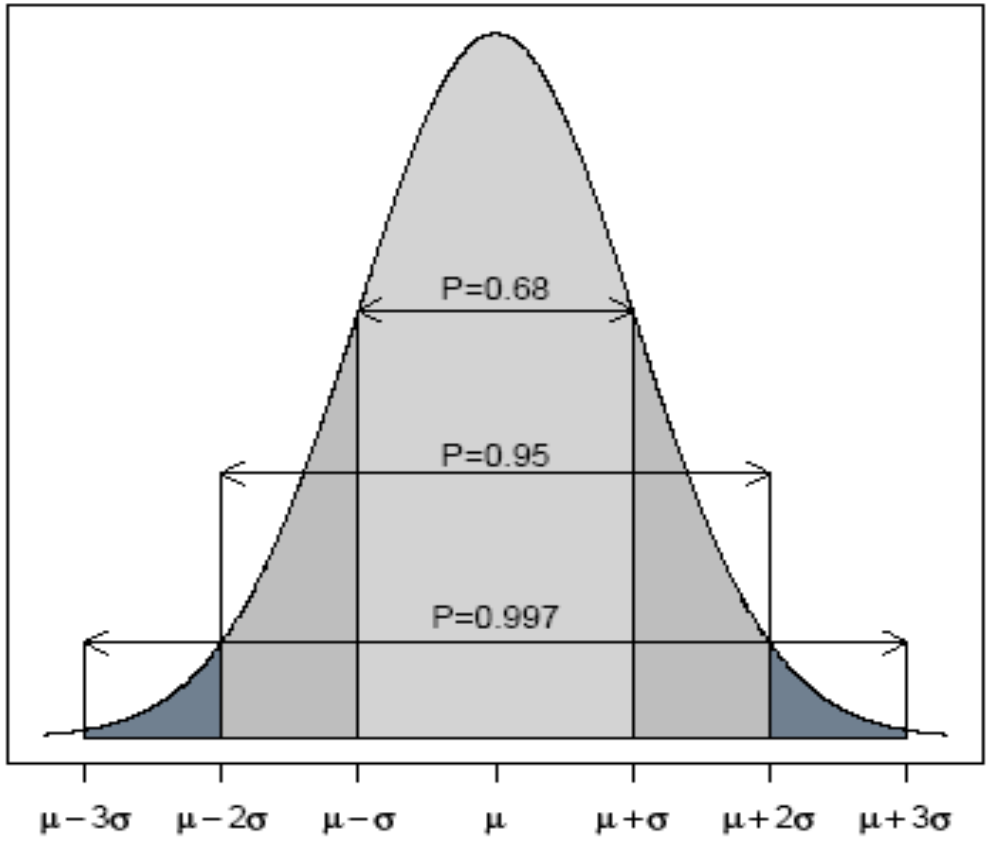




Eigenschaften der Normalverteilung

- Die Normalverteilung mit $\mu = 0$ und $\sigma^2 = 1$ nennt man Standard-Normalverteilung. Wir schreiben $X \sim N(0,1)$.
- Das Integral über die Dichtefunktion existiert, aber es kann nicht in geschlossener Form angegeben werden.
- Es gibt also keine "Formel" für die Verteilungsfunktion. Für die Berechnung von Wahrscheinlichkeiten muss man deshalb auf Tabellen oder Computerprogramme zurückgreifen.
- Es gibt unendlich viele Normalverteilungen, je nach Wahl der beiden Parameter. Die Glockenform haben alle gemeinsam.
- Parameter: μ verschiebt die Glocke, σ^2 streckt/staucht sie.

Schwankungsbereich der Normalverteilung



Diese Grafik liefert uns eine nützliche Faust-regel: das Intervall

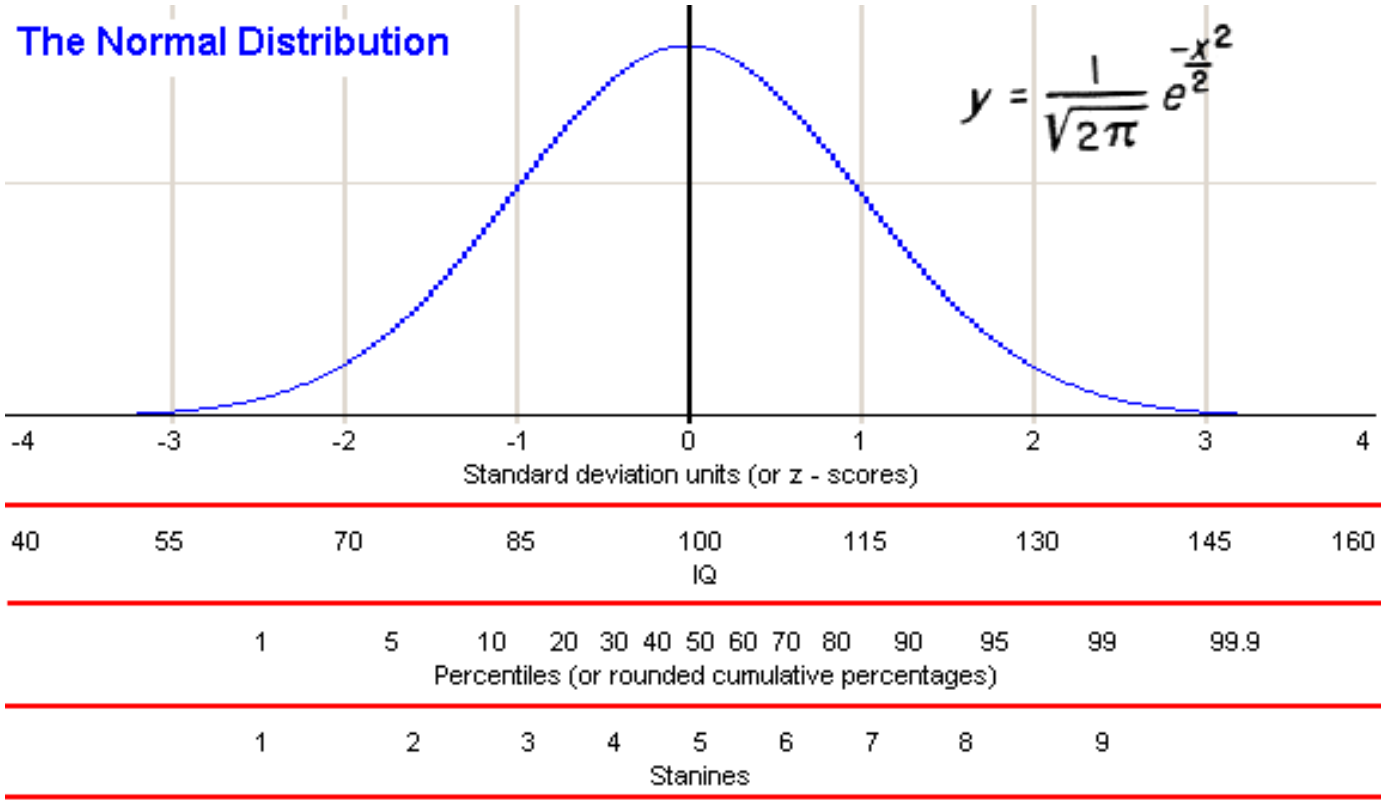
$$[\mu - 2\sigma, \mu + 2\sigma]$$

enthält rund 95% der Wahrscheinlichkeit, d.h.

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

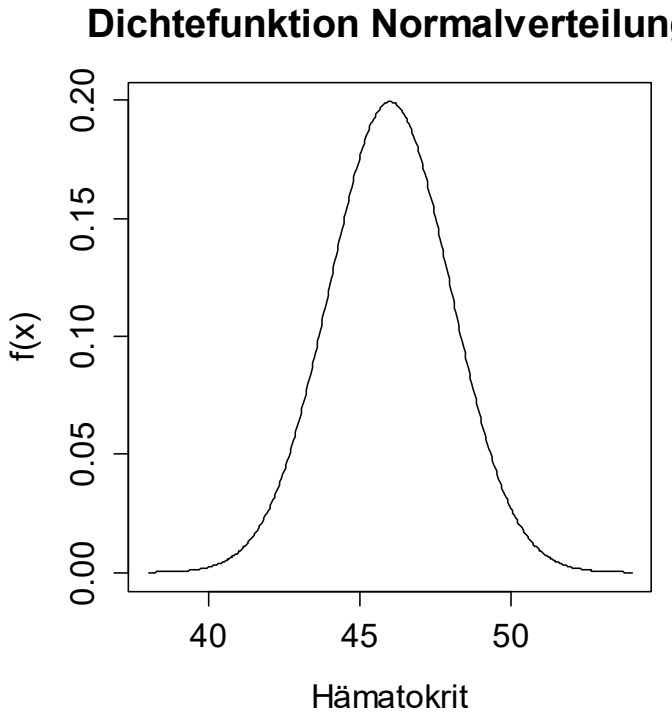
Quantile und Anwendungsbeispiel

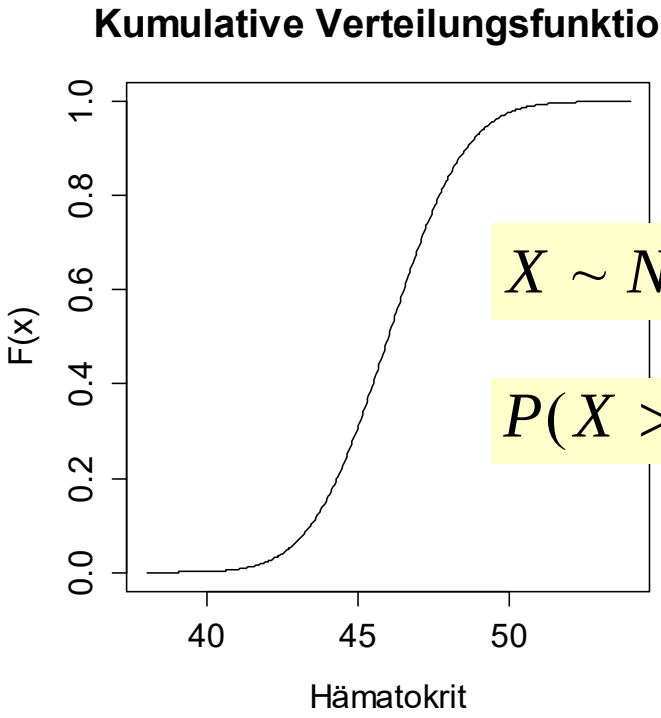


→ Der IQ über alle Personen hat die Verteilung $N(100, 225)$

Beispiel: Hämatokrit

Anteil der Erythrozyten am Volumen des Bluts. Diese stellen rund 99% des Gesamtvolumen der Blutzellen dar. Normale Werte beim Menschen liegen zwischen 42% und 50%.





$$X \sim N(\mu = 46, \sigma^2 = 2^2)$$

$$P(X > 51.5) = 0.3\%$$

Schätzung der Parameter

I.A. sind die Parameter einer Normalverteilung noch nicht bekannt und müssen aufgrund von Daten geschätzt werden.

Wir verwenden dazu den Stichproben-Mittelwert für μ und die Stichproben-Standardabweichung für σ^2.

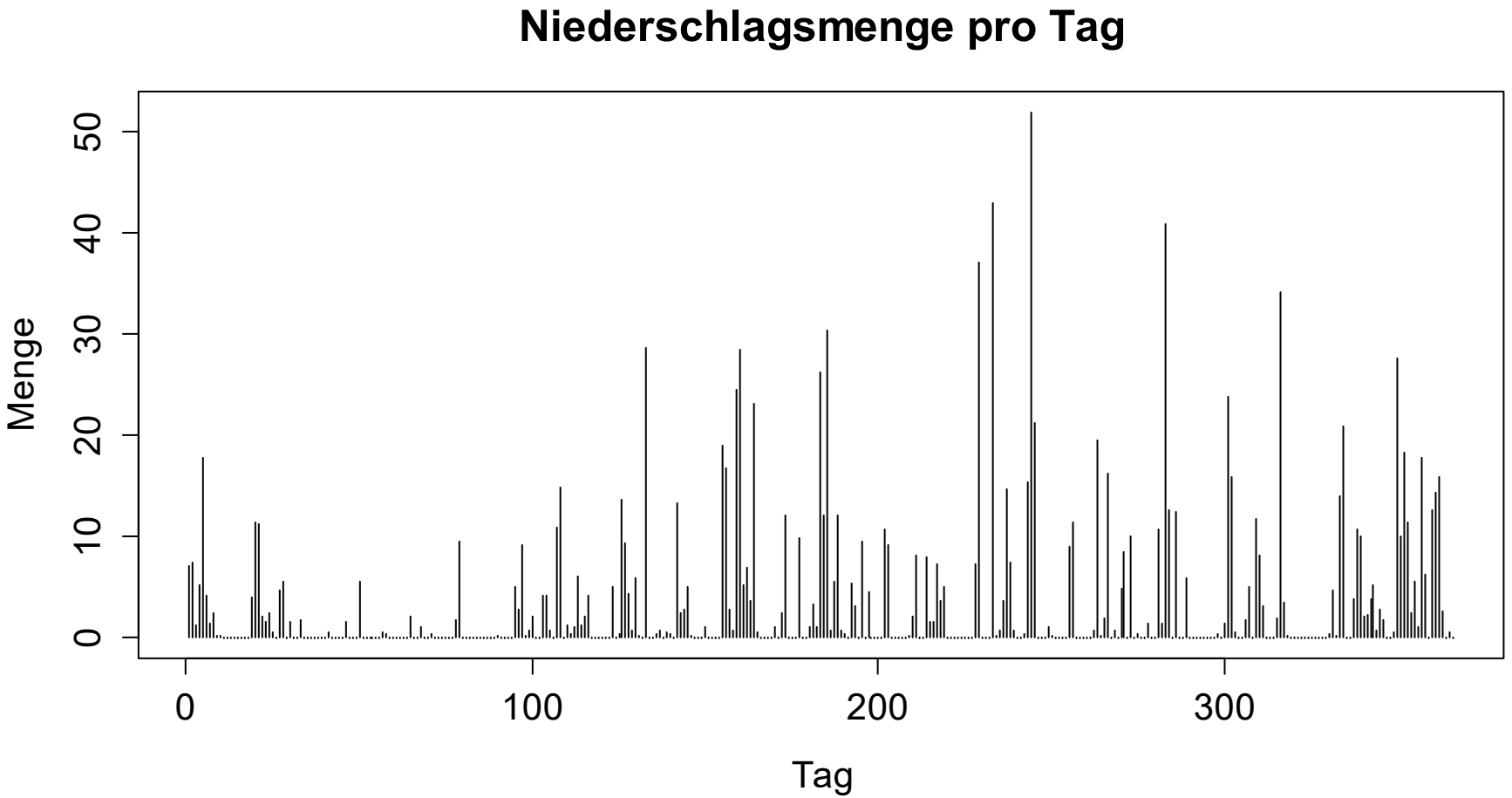
$$\hat{\mu} = \overline{x} = \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Lognormal-Verteilung

$Z =$ **"Regenmenge an einem Tag mit Niederschlag"**



Lognormal-Verteilung

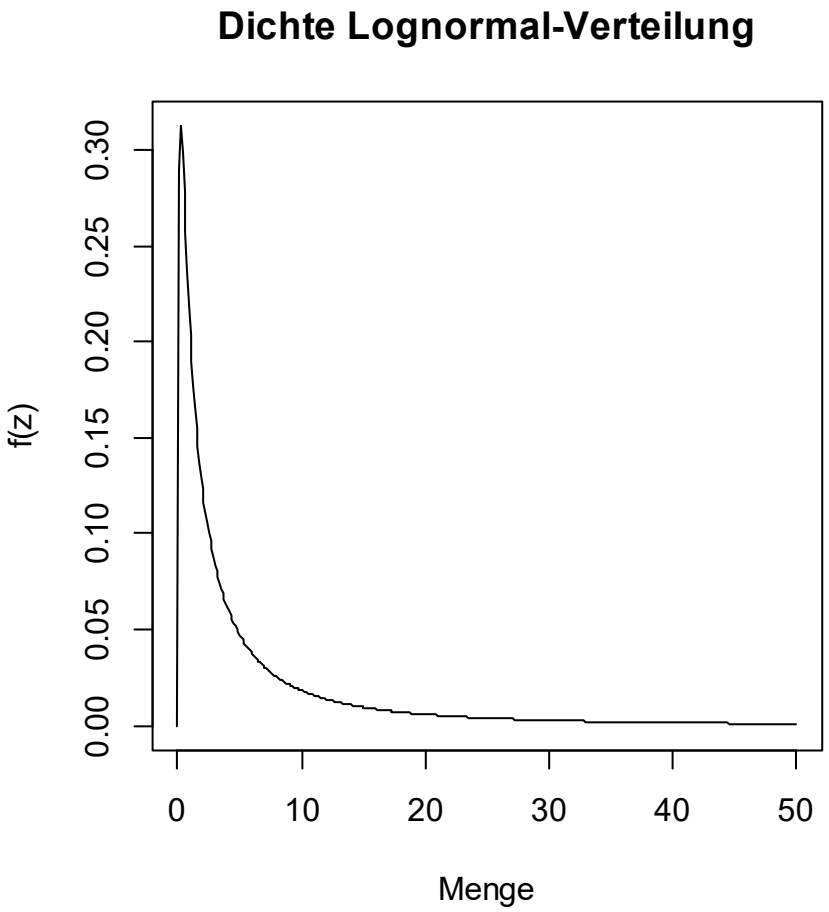
- Die Lognormal-Verteilung kommt für stetige Variablen zum Einsatz, welche nur positive Werte annehmen können und die eine rechtsschiefe Verteilung aufweisen.
- Charakteristisch ist es, dass für solche Grössen Unterschiede besser durch Verhältnisse als durch Differenzen ausgedrückt werden, bzw. sich das Resultat aus einer multiplikativen Überlagen von vielen kleinen Einflüssen ergibt.
- Zur Definition/Rechnung benützen wir folgende Eigenschaft:

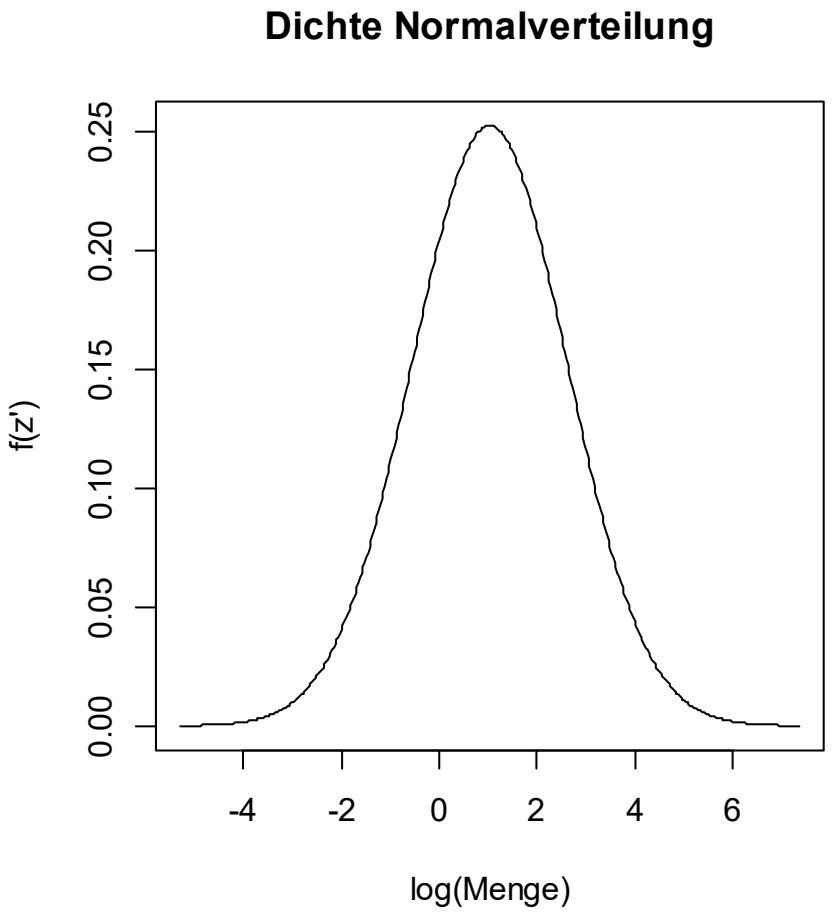
  Es gilt $Z \sim \log N(\mu, \sigma^2)$ falls $\log(Z) \sim N(\mu, \sigma^2)$

  Wir schreiben in diesen Fällen kurz: $Z \sim \log N(\mu, \sigma^2)$.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

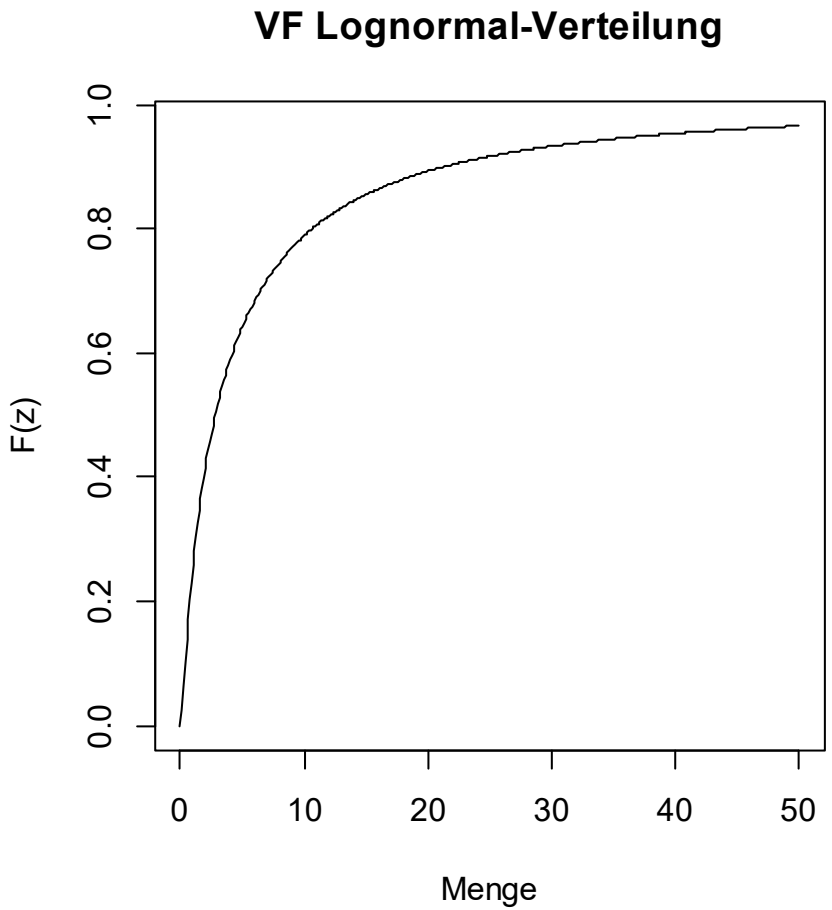
Dichtefunktionen im Beispiel

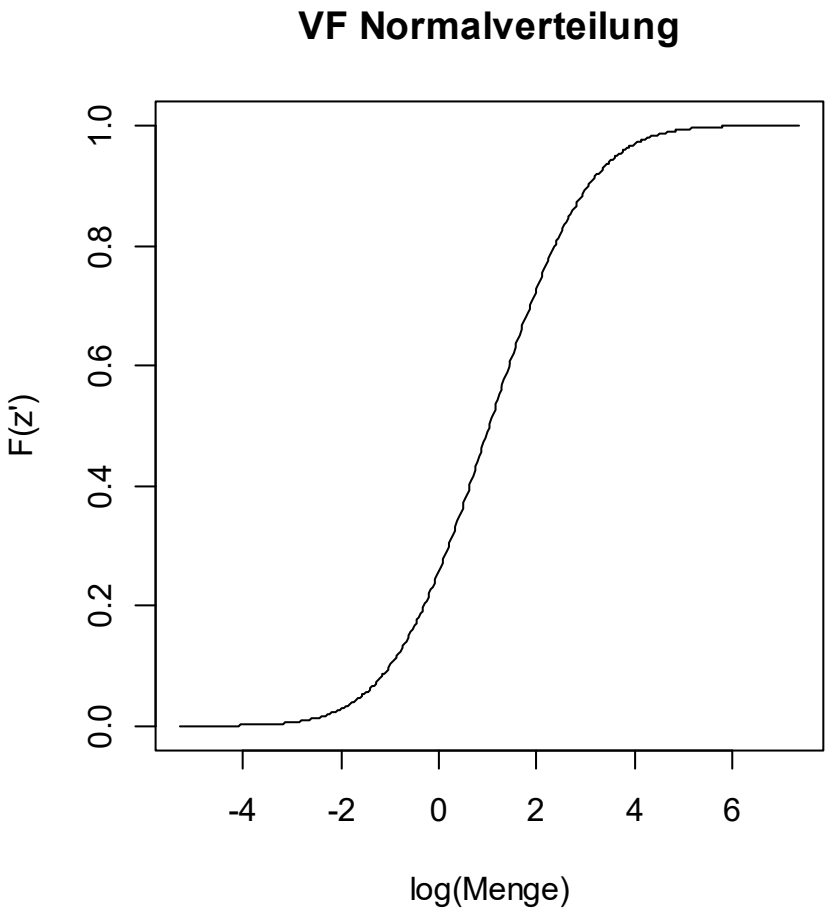




GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Dichtefunktionen im Beispiel





Anpassen und Rechnen

Die Parameter μ, σ^2 einer Lognormal-Verteilung werden durch Bilden von Mittelwert und Stichproben-Varianz auf den mit dem Logarithmus transformierten Daten bestimmt:

```
mw <- mean(log(menge))
sv <- var(log(menge))
```

In R benützt man `dlnorm()`, `plnorm()`, `qlnorm()`:

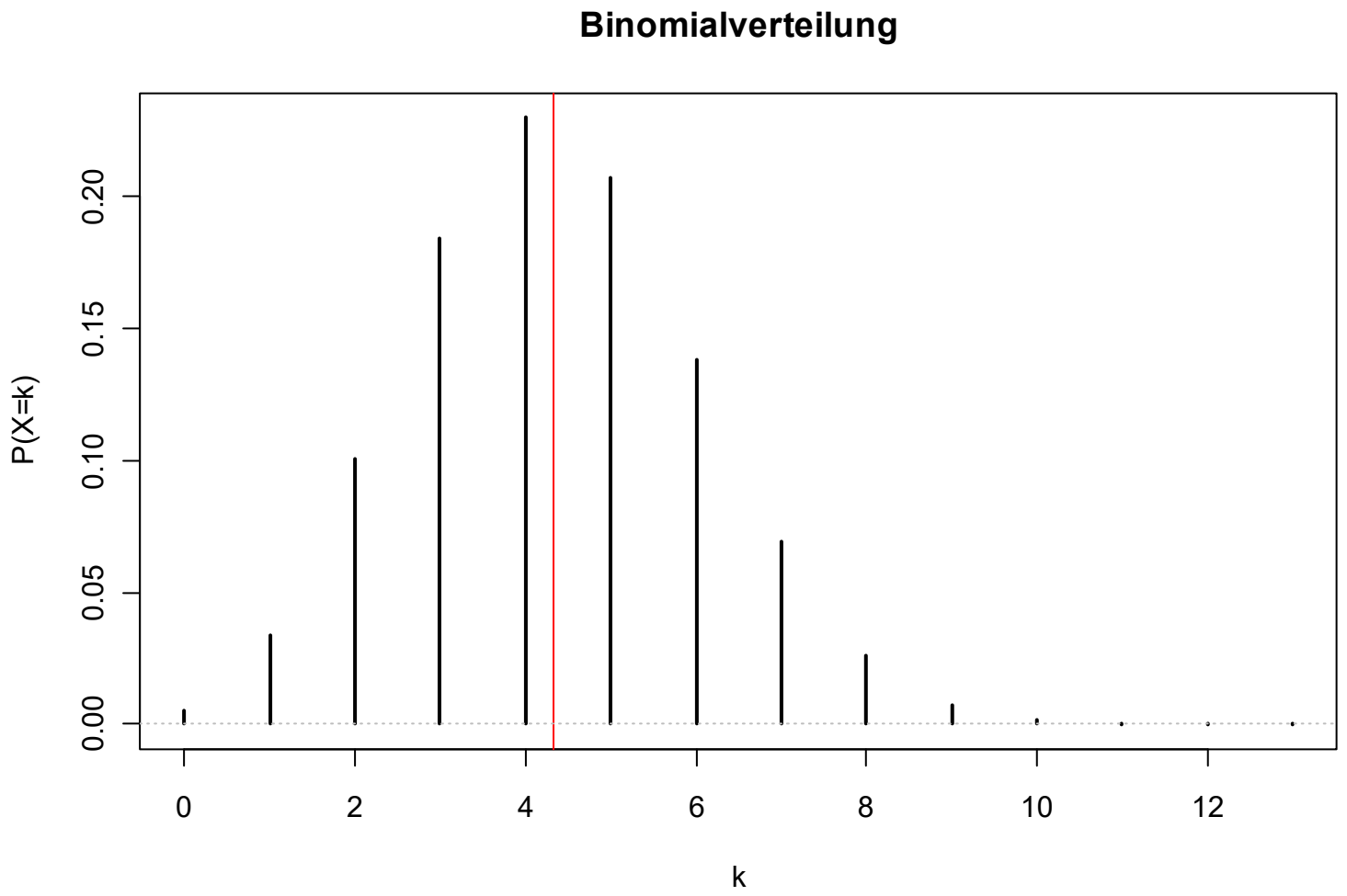
```
> plnorm(10, mw, sqrt(sv))
[1] 0.7829
> pnorm(log(10), mw, sqrt(sv))
[1] 0.7829
```

Erwartungswert

- Der Erwartungswert einer Zufallsvariablen ist das, was man im Schnitt, bei unendlich vielen Realisierungen, erhält.
- Er ist ein Lagemass für eine Zufallsvariable. Bildlich handelt es sich um die Position der Verteilung auf der x-Achse
- Man kann den Erwartungswert geometrisch bestimmen: Es ist der Schwerpunkt (in x-Richtung) der Wahrschein-lichkeits- bzw. Dichtefunktion
- Natürlich kann man den Erwartungswert einer Verteilung auch mittels Formeln rechnerisch bestimmen

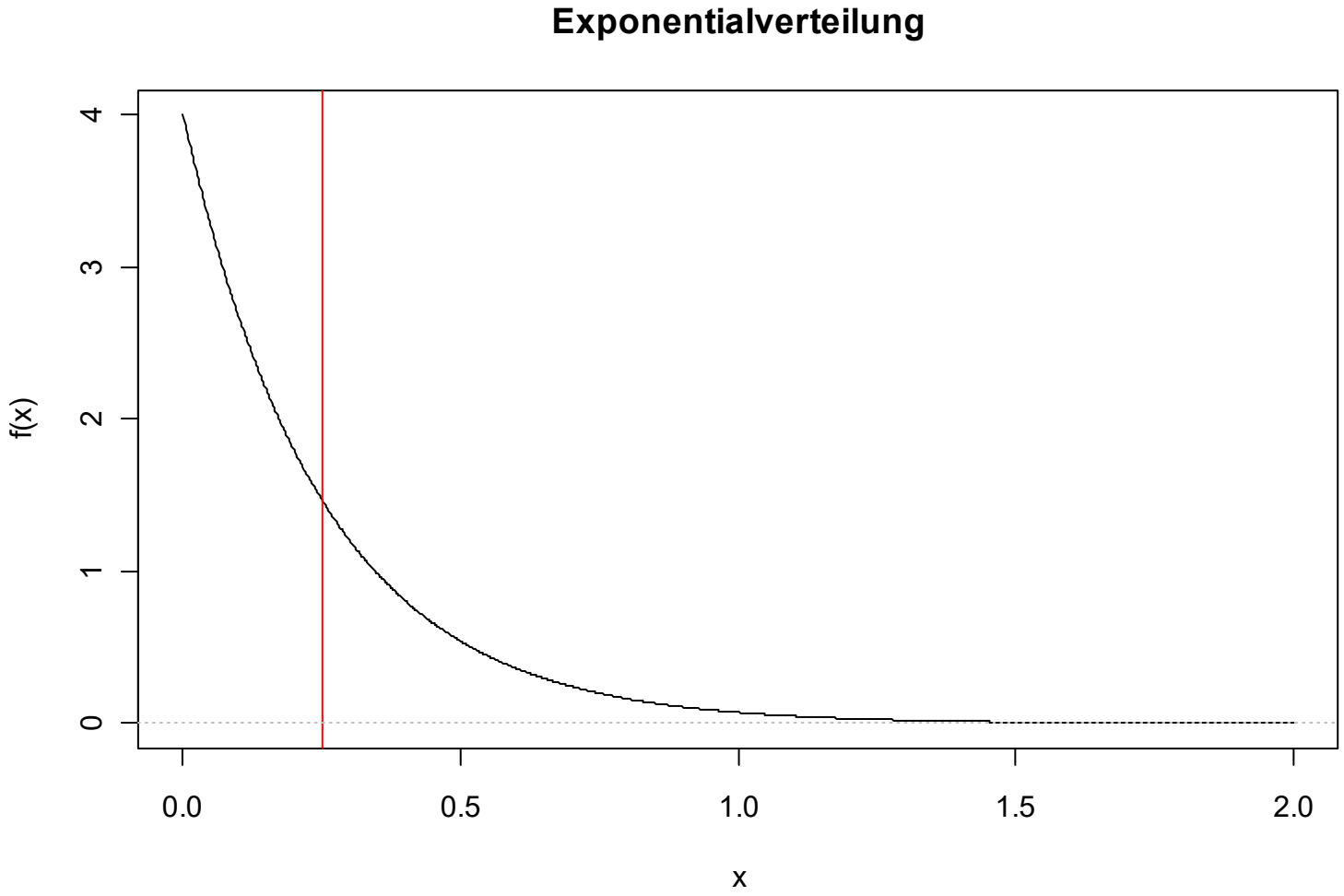
GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Schwerpunkt/EW der Binomialverteilung



GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Schwerpunkt/EW der Exponentialverteilung



Berechnung des Erwartungswerts

1) **Für diskrete Zufallsvariablen** X

Es sei X eine diskrete Zufallsvariable mit W'keitsfunktion $p(\cdot)$. Der Erwartungswert $E[X]$ ist definiert als:

$$E(X) = \sum_i x_i \cdot P(X = i) = \sum_i x_i \cdot p(x_i) = \sum_i x_i \cdot p_i$$

falls diese Summe existiert. Wir schreiben oft auch $\mu = E(X)$.

2) **Für stetige Zufallsvariablen** Z

Sei Z eine stetige Zufallsvariable mit Dichtefunktion $f(\cdot)$. Dann ist, falls das Integral existiert:

$$\mu = E(Z) = \int_{-\infty}^{+\infty} z \cdot f(z)\, dz$$

Varianz und Standardabweichung

Sei X eine Zufallsvariable mit Erwartungswert $E[X]$.
Dann ist die Varianz von X gegeben durch:

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

falls dieser Erwartungswert existiert. Die Standardabweichung ist die Wurzel aus der Varianz. Wir schreiben auch

$$\sigma^2, \ bzw. \ \sigma$$

für die Varianz, bzw. die Standardabweichung. Die Varianz ist ein Streuungsmass der Zufallsvariablen. Sie bestimmt grob gesagt die Breite der Verteilung, bzw. wird durch die Breite der Verteilung bestimmt.

Berechnung der Varianz

1) **Für diskrete Zufallsvariablen** X

Es sei X eine diskrete Zufallsvariable mit W'keitsfunktion $p(\cdot)$ und Erwartungswert $\mu = E[X]$. Die Varianz berechnet sich als:

$$Var(X) = \sum_i (x_i - \mu)^2 \cdot p(x_i)$$

falls diese Summe existiert.

2) **Für stetige Zufallsvariablen** Z

Sei Z eine stetige Zufallsvariable mit Dichtefunktion $f(\cdot)$ und Erwartungswert $\mu = E[Z]$. Dann ist, falls das Integral existiert:

$$Var(Z) = \int_{-\infty}^{+\infty} (z - \mu)^2 \, f(z) dz$$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Erwartungswerte bestimmter Verteilungen

Für alle parametrischen Verteilungsfamilien sind Erwartungswert und Varianz bereits als Funktion der Parameter bekannt:

Verteilung		E[X] bzw. E[Z]	Var(X) bzw. Var(Z)
Bernoulli	$X \sim Bernoulli(p)$	p	$p(1-p)$
Binomial	$X \sim Bin(n,p)$	np	$np(1-p)$
Poisson	$X \sim Pois(\lambda)$	λ	λ
Exponential	$Z \sim Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$
Uniform	$Z \sim Unif(u,o)$	$(u+o)/2$	$(o-u)^2/12$
Normal	$Z \sim N(\mu,\sigma^2)$	μ	σ^2
Lognormal	$Z \sim \log N(\mu,\sigma^2)$	$e^{(\mu+0.5\sigma^2)}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

Rechenregeln für Erwartungswert/Varianz

- Es ist nicht selten der Fall, dass wir an der Summe S von zwei beliebigen Zufallsvariablen X und Y interessiert sind.
- Selbst wenn wir die Verteilung von X, Y kennen, so ist es nur in wenigen Ausnahmefällen (z.B. Poissonverteilung, Normalverteilung) möglich, die Verteilung von S anzugeben.
- Einfacher ist die Sache für die Kenngrössen Erwartungswert und Varianz. Diese lassen sich für lineare Transformationen und für Summen einfach angeben.

→ siehe nächste Folie…

Rechenregeln für Erwartungswert/Varianz

Es seien a, b skalare Grössen und $S = X + Y$ Zufallsvariablen. Dann gelten die folgenden Rechenregeln ***immer***:

1) Für den Erwartungswert

- $E[aX + b] = aE[X] + b$
- $E[S] = E[X] + E[Y]$

2) Für die Varianz

- $Var(aX + b) = a^2 Var(X)$
- $Var(S) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$

Der Term $Cov(X, Y)$ beschreibt den linearen Zusammenhang der beiden Zufallsvariablen. Er ist nur in Ausnahmefällen gleich null, z.B. wenn X, Y unabhängig sind.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 11

Beispielaufgabe

Beispiel: Jasskarten – wir studieren die Zufallsvariable

$$X = "Punktewert\ einer\ zufällig\ herausgegriffenen\ Jasskarte"$$

Wir studieren Bedeutung, Form und Wert:

- der Wahrscheinlichkeitsverteilung
- des Erwartungswerts $E[X]$
- der Varianz $Var(X)$
- der Summe $S = X + X$

→ *Siehe Wandtafel…*

GdM 2: LinAlg & Statistik
FS 2018 – Woche 12

Marcel Dettling

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

Winterthur, 16. Mai 2018

GdM 2: LinAlg & Statistik
FS 2018 – Woche 12

Verteilung einer Summe

Beispiel: Jasskarten – wir studieren die Zufallsvariable

$$X = "Punktewert\ einer\ zufällig\ herausgegriffenen\ Jasskarte"$$

Wir studieren die Verteilung von $S_n = X_1 + ... + X_n$

→ **Siehe Demonstration…**

Schliessende Statistik

- In der schliessenden Statistik möchte man aus einer Stichprobe Rückschlüsse auf eine Population ziehen. Die beiden wichtigsten Bereiche sind ***Schätzer*** und ***Tests***.

Was ist ein Schätzer?

- Ein Schätzer ist ein Verfahren, mit dem man aufgrund von Beobachtungen einer Stichprobe einen passenden Wert für die Kenngrösse einer Verteilung oder einen Parameter bestimmt.
- Ein Schätzer ist eine Funktion der Zufallsstichprobe

Zufallsstichproben

- Damit die Schlussfolgerungen aus einer Datenanalyse zuverlässig sind, muss die Stichprobe repräsentativ sein. Jede Beobachtung ist dann das Ergebnis eines Zufallsexperiments.
- Wir gehen davon aus, dass die Beobachtungen unabhängig und unter identischen Bedingungen gemacht werden.
- Zur i-ten Beobachtung gehört die Zufallsvariable X_i. Alle X_i haben dieselbe Verteilung und die Schar von Zufallsvariablen, $X_1, ..., X_n$ nennen wir eine Zufallsstichprobe vom Umfang n.
- Man spricht dann von i.i.d. (*identically and independently distributed*) Zufallsvariablen.

Verteilung des Mittelwerts

Der populärste ***Schätzer*** ist der Mittelwert:

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Er ist ein Schätzer für den Erwartungswert $E[X_i]$

*Der numerische Wert, den der Schätzer annimmt, wenn eine Stichprobe mit Werten $x_1, ..., x_n$ vorliegt, nennt man **Schätzung**.*

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Wichtig**: während die Schätzung $\bar{x}$ ein fester, auf einer bestimmten Stichprobe angenommer Wert ist, handelt es sich beim Schätzer $\bar{X}$ um eine Zufallsgrösse

Weiteres zu Schätzern

Es gibt weitere Schätzer:

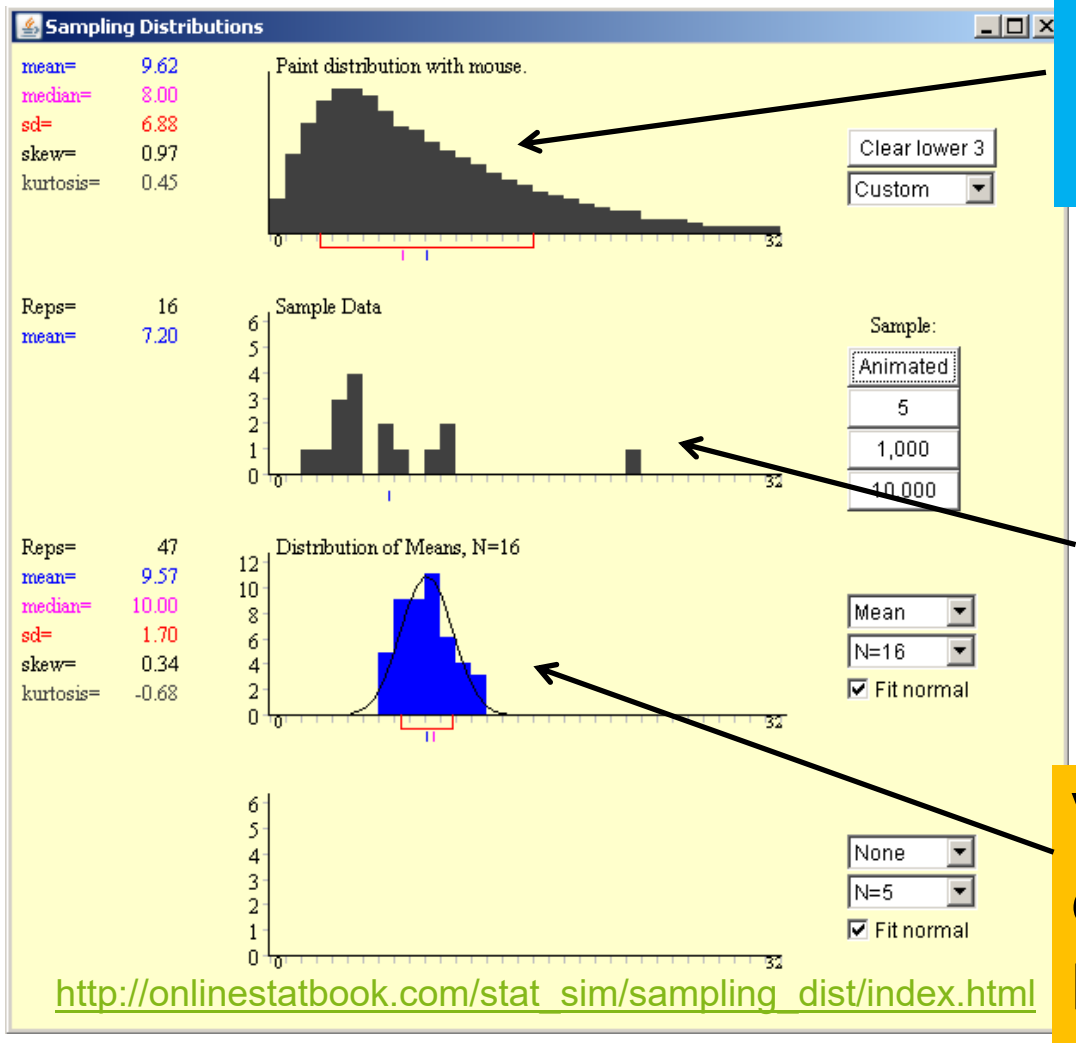
- Der Median, auch für den Erwartungswert $E[X]$
- Die Stichproben-Varianz s^2 für die Varianz $Var(X)$
- ...

Weil Schätzer Zufallsvariablen sind, kann man deren Eigenschaften studieren. Wir interessieren uns für:

a) den Erwartungswert eines Schätzers, z.B. $E[\bar{X}]$
b) die Varianz eines Schätzers, z.B. $Var(\bar{X})$
c) die Verteilung eines Schätzers, z.B. $\bar{X} \sim ...$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 12

Simulationsstudie



Verteilung, von der die Einzelwerte stammen, bzw. simuliert werden

Histogramm einer Stichprobe vom Umfang n=16

Visualisierung der Verteilung der Mittelwerte von Stichproben mit n=16

Resultate

- $E[\bar{X}] = E[X]$, bzw. $\mu_{\bar{X}} = \mu_X$

  Im Mittel liegt unsere Schätzung richtig. Wir sagen auch, sie sei erwartungstreu.

- $Var(\bar{X}) = \frac{1}{n} \cdot Var(X)$, bzw. $\sigma_{\bar{X}} = \sigma_X / \sqrt{n}$

  Die Streuung der Mittelwerte nimmt mit zunehmender Stichprobengrösse n ab. Je grösser die Stichprobe, desto genauer können wir den Erwartungswert $E[X]$ schätzen.

- $\bar{X} \sim N(\mu_X, \sigma_X^2 / n)$

  Zentraler Grenzwertsatz, siehe nächste Folie...

Zentraler Grenzwertsatz

Der **Mittelwert** $\bar{X}$ einer Zufallsstichprobe $X_1, ..., X_n$ vom Umfang n, die aus einer beliebigen Verteilung mit Erwartungswert μ_X und Varianz σ_X^2 entsteht, ist bei unabhängigen X_i **genähert normalverteilt**, d.h. es gilt:

$$\bar{X} \sim N(\mu_X, \sigma_X^2 / n)$$

- In der Regel ist $n > 25$ gross genug für eine gute Näherung
- Es gibt (wenige) Ausnahmen, wo der ZGS nicht gilt!
- Der Beweis erfordert Zeit und Mathematik: wir verzichten
- Einer der Gründe, warum die Normalverteilung wichtig ist

Gütekriterien eines Schätzers

Erwartungstreue

Ein Schätzer streut zwar, aber ein guter Schätzer soll ohne systematischen Fehler um den wahren Wert „herumstreuen“. Der Mittelwert von vielen Schätzungen ist der wahre Wert.

Effizienz

Der Schätzer soll so wenig wie möglich streuen. Mit anderen Worten: seine Varianz soll so klein wie möglich sein

Konsistenz

Je grösser die Stichprobe ist, desto genauer soll der Schätzer werden, d.h. je weniger um den wahren Wert herumstreuen

Statistisches Testen

Nutzen:
Ein statistischer Test beantwortet die Frage, ob die beobachteten Daten mit einer Anfangshypothese (Vorwissen, andere Methode, etc.), der Nullhypothese H_0 vereinbar sind.

→ Wir können mit einem statistischen Test beurteilen, ob der Unterschied zwischen einem Schätzwert auf einer Stichprobe und einem gegebenen Sollwert zufällig oder systematisch ist.

→ Wie sich später zeigt, lässt sich das Vorgehen auch Verallgemeinern, so dass auch zwei Stichproben, bzw. die daraus gewonnenen Schätzwerte miteinander verglichen werden können.

Allgemeines zum Testen

Beispiel:

Bei Kindern ist es vorerst Glückssache, ob sie die Schuhe richtig anziehen. Es gibt aber einen Zeitpunkt, wo ihnen die Unterscheidung zwischen dem linken und dem rechten Schuh gelingt – die Frage ist bloss wann.

Wir testen, indem ein Kind über 12 Schuhanzieh-Vorgänge beobachtet wird und zählen die Anzahl Erfolge.

Wann ist der «Beweis» erbracht?

Grundüberlegungen zum Testen

- Bei Zufallsexperimenten kann man nie mit Sicherheit sagen, dass beobachtete Werte einem stochastischen Modell zwingend widersprechen.
- Subjektiv wird man dies aber annehmen, wenn ein Ereignis beobachtet wird, das im Vergleich zum stochastischen Modell extrem (d.h. ausserhalb des üblichen Bereichs) ist.
- Um dies zu objektivieren, hat man statistische Tests eingeführt. Modelle werden abgelehnt, wenn das Ereignis nicht im Bereich der 95% an «normalen» Ereignissen, sondern ausserhalb davon liegt.

Binomialtest für einen Anteil p_0

Bsp: Mendels Vererbungsgesetze

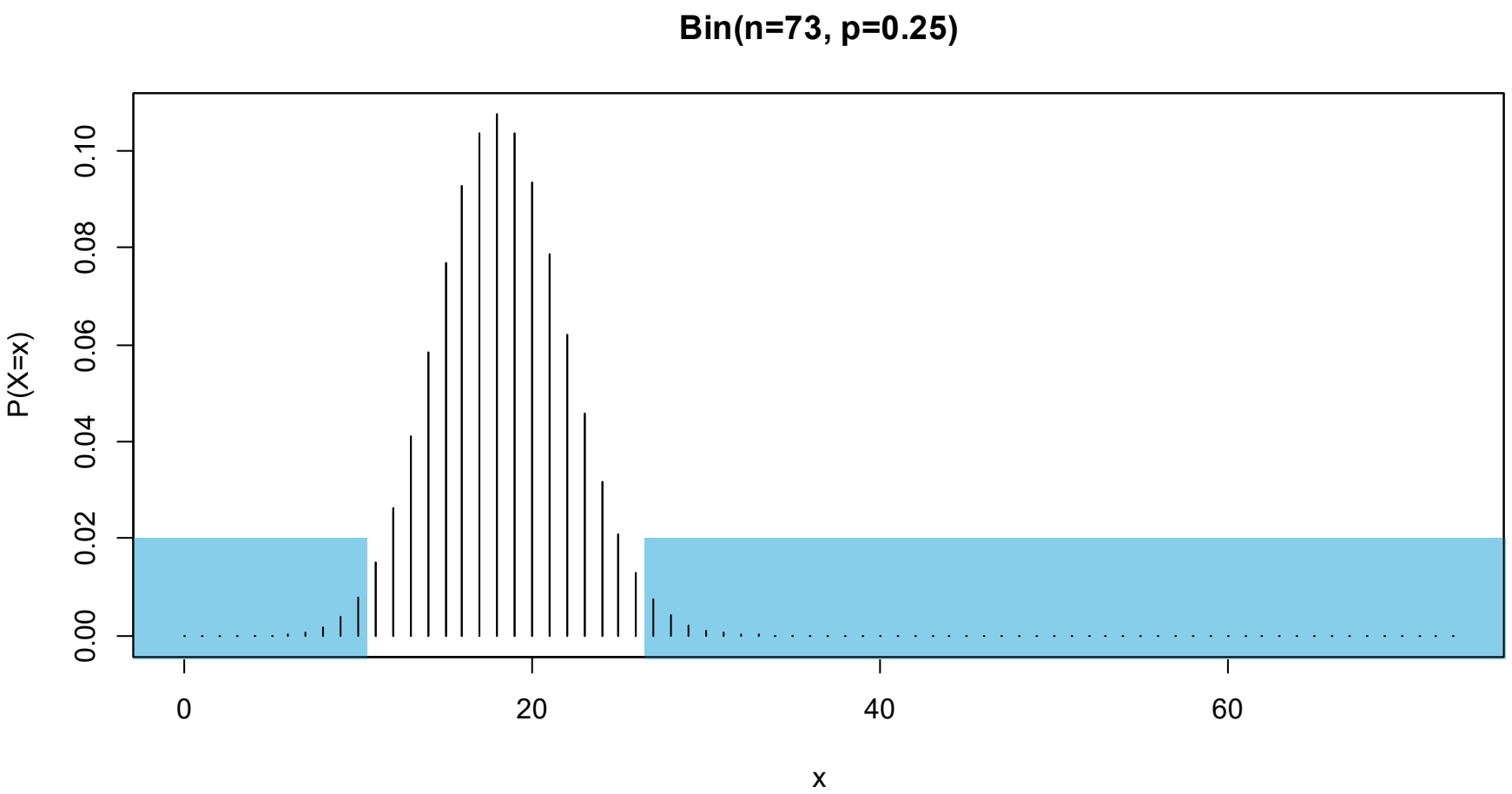
Untersucht wurden 73 Samen, wovon 55 rund und 19 kantig waren. Dies entspricht einem Verhältnis von 2.89:1 und damit nicht exakt den postulierten 3:1.

Ist dies ein Widerspruch zur Theorie?

→ *Siehe Wandtafel für einen Binomialtest…*

GdM 2: LinAlg & Statistik
FS 2018 – Woche 12

Binomialtest für einen Anteil p_0



Grundsätzliches zum Testen

- Die Absicht eines statistischen Tests ist, herauszufinden ob ein bestimmtes Modell "gilt" (bzw. "gelten kann"). Der Test ist eine Regel, die festlegt, wann wir diese Frage mit "ja" und wann mit "nein" beantworten sollen.
- Um diese Regel festzulegen, benützt man eine Art Widerspruchsbeweis: wir gehen davon aus, dass die Testgrösse dem Modell der Nullhypothese entspricht. Ist der beobachtete "unplausibel" so betrachten wir das als Widerspruch zur Nullhypothese, welche dann verworfen wird.
- Man hat dann in den Daten statistische Evidenz gegen die in der Nullhypothese aufgestellte Behauptung gefunden.

Beibehalten der Nullhypothese

- Wenn der beobachtete Wert nicht im Verwerfungsbereich liegt, d.h. wenn der p-Wert grösser als 0.05 ist, so wurde kein Widerspruch zur Nullhypothese gefunden.
- Achtung: dies heisst nicht, dass die Nullhypothese richtig oder gar bewiesen ist – sie bleibt lediglich plausibel – wie einige andere Modelle ebenso.
- Wir sagen in diesem Fall, die Nullhypothese sei beibehalten worden – nicht mehr und nicht weniger.
- Experimente lassen sich in der Wissenschaft besser «verkaufen», falls eine Hypothese verworfen werden konnte.

Fehler beim statistischen Testen

Es gibt 2 grundsätzliche Fehlerquellen beim Testen:

1) Fehler 1. Art

Nullhypothese wird verworfen, obwohl sie korrekt ist.

Die Entscheidungsregel beim statistischen Test ist so eingerichtet, dass dies gerade in 5% aller Fälle passiert. Man nennt $\alpha = 5\%$ auch Irrtumsw'keit oder Niveau. Manchmal benützt man auch kleinere Irrtumsw'keiten von nur $\alpha = 1\%$ oder $\alpha = 0.1\%$.

2) Fehler 2. Art

…

Fehler beim statistischen Testen

Es gibt 2 grundsätzliche Fehlerquellen beim Testen:

1) Fehler 1. Art

…

2) Fehler 2. Art

Nullhypothese wird beibehalten, obwohl sie falsch ist.

Es kann sein, dass die zwar die Alternative richtig ist, die Beobachtung aber trotzdem in den Annahmebereich der Nullhypothese fällt, obwohl diese falsch ist. Die Berechnung der W'keit eines solchen Fehlers 2. Art ist nicht trivial, wir verzichten hier darauf.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Marcel Dettling

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

Winterthur, 23. Mai 2018

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Ausblick

Wir werden noch die folgenden Tests diskutieren:

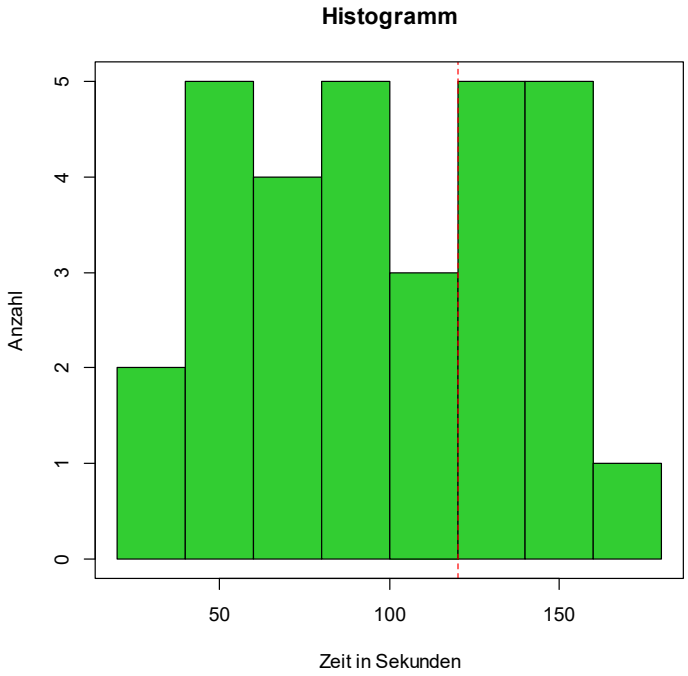
- **t-Test für einen Erwartungswert** μ_0
  Dient dazu, den Mittelwert einer Stichprobe mit einem Sollwert μ_0 zu vergleichen. Die Testgrösse lautet:
  $$T = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\hat{\sigma}_X}$$
  Unter H_0 hat T eine t-Verteilung mit $n-1$ Freiheitsgraden

- **2-Stichproben-Tests**
  Hier wird nicht mehr eine Stichprobe gegen einen Sollwert verglichen, sondern es wird bei zwei Stichproben geprüft, ob Anteil / Anzahl / Mittelwert identisch sind.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

t-Test für einen Erwartungswert μ_0

Ein schmerzstillendes Medikament soll im Schnitt spätestens nach 120 Sekunden Erleichterung bringen. Die tatsächliche Zeit schwankt jedoch von Patient zu Patient.



Datenlage:

$n = 30$ Patienten

Sollwert $\mu_0 = 120s$

Mittelwert $\hat{\mu} = \overline{x} = 100.65s$

Standardabweichung $\hat{\sigma}_X = 40.41s$

t-Verteilung

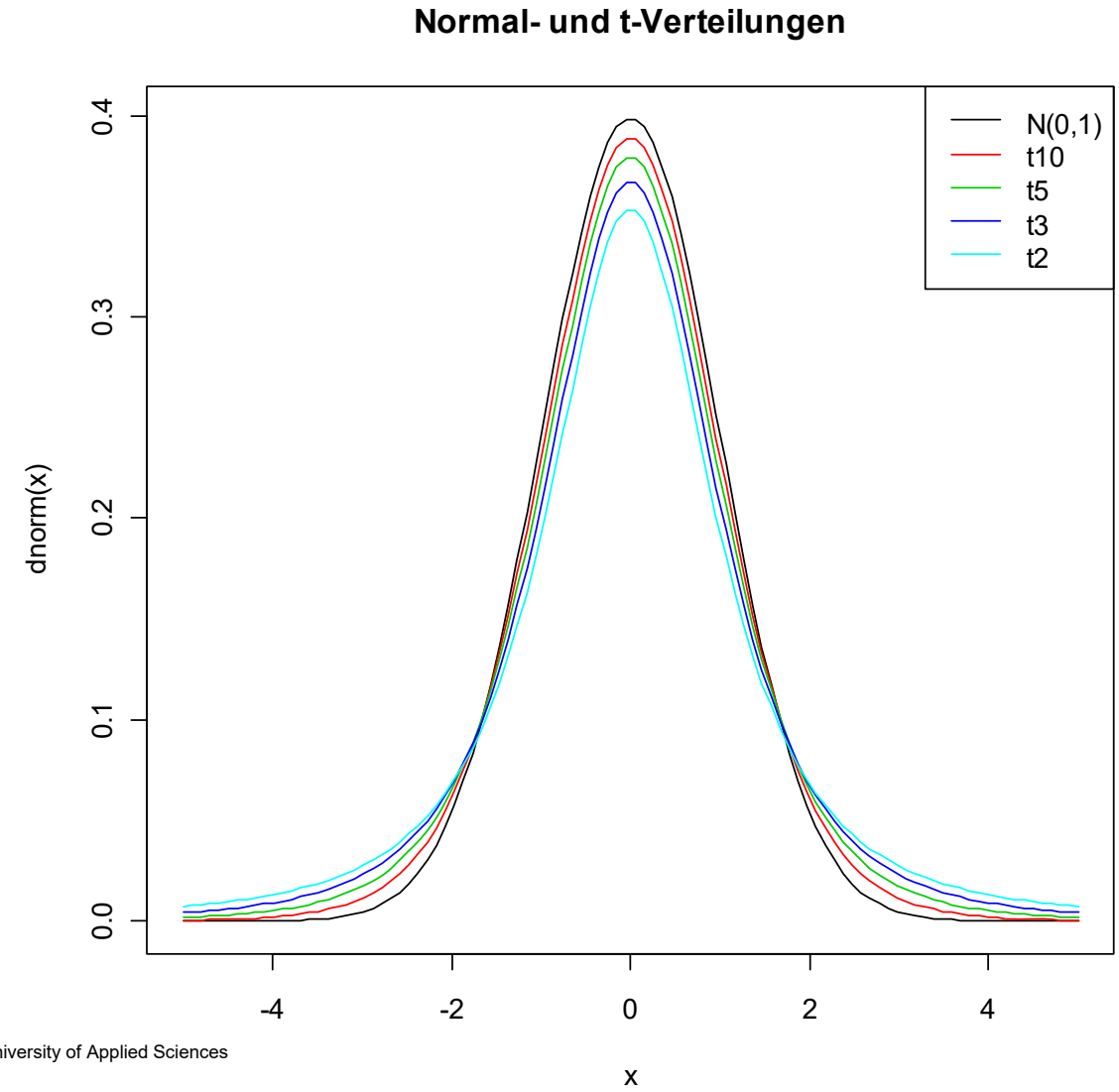
Wenn wir nun nicht mit der "wahren" Standardabweichung σ_X normalisieren können, und uns stattdessen mit ihrer geschätzten, empirischen Variante $\hat{\sigma}_X$ behelfen müssen, so gilt:

$$Z = \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sigma_X} \sim N(0,1) \qquad T = \sqrt{n} \cdot \frac{\bar{X} - \mu}{\hat{\sigma}_X} \sim t_{(n-1)}$$

Durch das Einfügen der Schätzung $\hat{\sigma}_X$ entsteht eine zusätzliche Quelle der Variabilität, weshalb T eine etwas breitere Verteilung als Z hat. Es handelt sich um eine sogenannte t-Verteilung mit $(n-1)$ Freiheitsgraden.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

t-Verteilung



GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Kochrezept für den t-Test

1) Nullhypothese H_0: Problem formulieren
2) Alternativhypothese H_A: welche Abweichung ist gesucht?
3) Testgrösse T wählen und Verteilung unter H_0 bestimmen
4) Verwerfungs-/Annahmebereich bestimmen: $ca.\ [-2, 2]$
5) Realisierten Wert $T = t$ ablesen/bestimmen
6) p-Wert berechnen: $p - Wert = 2 \cdot P(T \leq -|t|)$
7) mit p-Wert oder V/A-Bereich Testentscheidung treffen

→ **siehe Wandtafel…**

Vertrauensintervall für μ: Ziel

- Der Schätzer $\hat{\mu} = \overline{X}$ für die Zeit bis zur Wirkung μ ist eine Zufallsvariable. Grund: wenn neue bzw. weitere Messungen hinzugefügt würden, so ergäbe sich ein anderer Mittelwert $\overline{x}$.

Ziel: (*beide sind äquivalent…*)

- Angeben, welche Werte plausibel sind für den unbekannten Erwartungswert μ. Denn das Resultat von $\overline{x} = 100.65$ hat ja auf einer weiteren Stichprobe nicht bestand.
- Wir möchten die Schätzung $\overline{x} = 100.65$ mit Genauigkeits-angabe versehen. Dieses Intervall ist um $\overline{x}$ zentriert, seine Breite bemisst sich an Variabilität und Stichprobengrösse.

Vertrauensintervall für μ via Dualität

Intuitive Idee:

Das 95% -Vertrauensintervall soll alle Werte μ enthalten, die auf dem 95% -Niveau mit den Daten vereinbar sind.

Dies sind z.B. alle Nullhypothesen H_0, für welche der zugehörige t-Test mit Irrtumsw'keit 5%, gegeben die vorliegenden Beobachtungen, nicht verworfen wird.

- Zur expliziten Bestimmung des Vertrauensintervalls könnte man auf „Ausprobieren" zurückgreifen. Das ist aber mühsam, mit Mathematik geht es viel eleganter...

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Formel & Herleitung

Das 95%-Vertrauensintervall für den Erwartungswert μ lautet:

$$\overline{x} \pm qt_{0.975;(n-1)} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}$$

Herleitung: **siehe Wandtafel...**

Wir betrachten einige Szenarien. Wenn alle anderen Parameter identisch bleiben, aber…

- die Stichprobe grösser wird: → VI wird kürzer!
- die Streuung der Einzelwerte grösser ist: → VI wird länger!
- wir mehr Sicherheit (höheres Niveau) wollen: → VI wird länger!

Vertrauensintervall für p

Beispiel: Meinungsumfrage vor einer Abstimmung

- Das Schätzen eines Anteils p durch $\hat{p} = X / n$, die relative Häufigkeit an Erfolgen/Zustimmung/... ist ein sehr häufig auftretendes Anwendungsproblem.
- Klar ist, dass die Schätzung $\hat{p}$ nicht exakt ist, d.h. von den ausgewählten Personen/Exemplaren/... abhängt.
- Wenn neue und/oder weitere Messungen durchgeführt werden, so wird sich eine andere Schätzung $\hat{p}$ ergeben.

Ziel: Bestimmung eines Vertrauensintervalls für den wahren/tatsächlichen Anteil p in der Grundgesamtheit.

Vertrauensintervall für p: Berechnung

- Das 95%-Vertrauensintervall für p enthält alle Nullhypothesen $H_0 : p = p_0$, für welche der Binomial-Test mit 5% Irrtumsw'keit nicht verworfen wird.
- Die Grösse dieses Bereichs hängt von der Anzahl Beobachtungen n und dem Anteil $\hat{p} = x/n$ ab. Eine exakte, explizite Formel kann man nicht angeben.
- Entweder benützt man die Funktion **`binom.test()`** in R. Oder, falls $n\hat{p}(1-\hat{p}) \geq 10$ auch die Näherungsformel:

$$\hat{p} \pm qnorm_{0.975;(0,1)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Warum ist das VI falsch?

Umfragen: Warum 20 Minuten genauer war als die SRG

ZÜRICH. Zehntausende Teilnehmer und moderne statistische Verfahren: So funktionieren 20-Minuten-Abstimmungsumfragen.

Wissenschaft trifft Online-Befragung: Seit knapp einem Jahr berechnen die Politologen Lucas Leemann und Fabio Wasserfallen für 20 Minuten die Haltung der Stimmbevölkerung. Das müssen Sie über das Projekt wissen:

Wie präzis sind die Ergebnisse der 20-Minuten-Umfragen?

Bei sieben der letzten zwölf eidgenössischen Abstimmungsvorlagen waren die Analysen von 20 Minuten zutreffender als die etablierten GFS-Trendstudien im Auftrag der SRG. Bei vier Vorlagen hatte die SRG die Nase vorn. Einmal lagen beide Umfragen gleichauf.

Die Politologen Fabio Wasserfallen und Lucas Leemann.

Wo liegt der Unterschied zwischen den beiden Umfragen?

An den Online-Umfragen von 20 Minuten nehmen regelmässig über 20 000 Personen teil. Deren Antworten werden nach verschiedenen Variablen gewichtet, so dass die Stichprobe möglichst gut der Struktur der Stimmbevölkerung entspricht. Das GFS führt seine Befragungen dagegen telefonisch durch. Aus der Stimmbevölkerung wird eine repräsentative Stichprobe von rund 1400 Personen interviewt.

Wurde die 20-Minuten-Umfrage schon von Hackern angegriffen oder manipuliert?

Ja. Sowohl Einzelpersonen als auch politische Akteure versuchten schon, die Umfrage durch Mehrfachteilnahmen oder technische Tricks zu manipulieren. «Zur Bekämpfung dieser Angriffe haben wir zusammen mit IT-Experten eine Reihe von Sicherheitselementen eingebaut», sagt Wasserfallen. Eine Manipulation sei so nicht komplett unmöglich – «die finanziellen und technischen Mittel, die dafür nötig wären, sind jedoch beträchtlich». JACQUELINE BÜCHI

Mehr zum Thema gibt es auf 20minuten.ch

«Longchamp ist die grösste Diva»

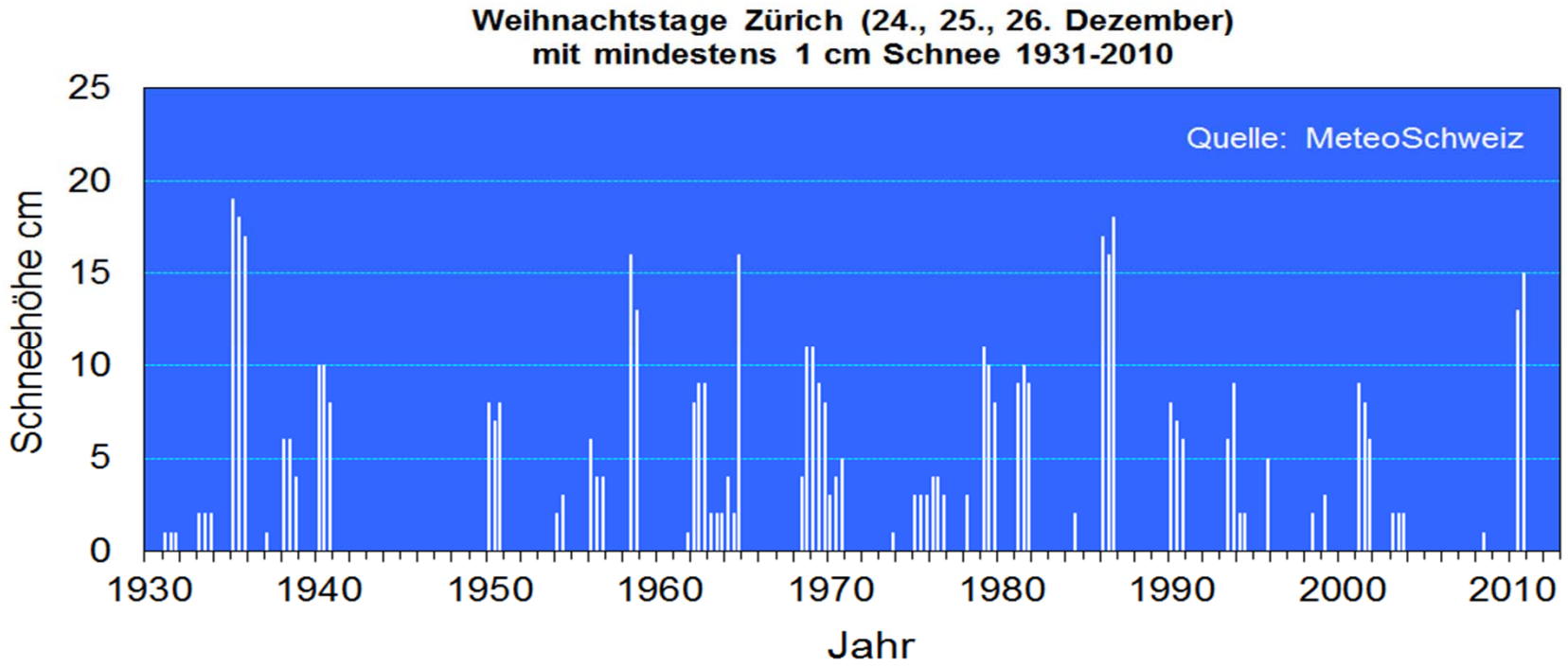
BERN. Nicht nur 20 Minuten führt im Vorfeld der nationalen Abstimmungen jeweils eigene Umfragen durch. Zwei junge Politologen wollen den GFS-Chef Claude Longchamp mit einem neuen Projekt frontal angreifen. «Longchamp ist die grösste Diva im Land. Niemand getraut sich, gegen ihn anzutreten, weil er zu dominant ist», kritisierte Michael Hermann gestern in der NZZ. Wenn die SRG den Umfrageauftrag für die neue Legislatur ausschreibt, wollen er und sein Berufskollege Thomas Milic sich darum bewerben. «Wir versuchen, es besser zu machen», so Milic mit Blick auf die 13-prozentige Abweichung, die Longchamp bei Ecopop verbuchen musste. Longchamp sagt im Artikel, die SRG mache «ein hartes Assessment». Bei der letzten Ausschreibung des Umfrageauftrags setzte er sich gegen mehrere Mitbewerber durch. JBU

Vergleich von 2 Stichproben

- In der Praxis ist es eher selten, dass ein Test gegen einen vorgegebenen Sollwert durchgeführt werden kann. Wenn, dann ist dieser Wert meist aus einer Stichprobe geschätzt.
- In diesem Fall vergleicht man aber zwei Stichproben, und nicht eine Stichprobe gegen einen Sollwert. Weil somit beide Grössen eine Unsicherheit aufweisen, muss man diese beim Testen berücksichtigen.
- Wir besprechend 2-Stichproben-Tests für:
  - zwei Anteile $\hat{p}_1$ und $\hat{p}_2$
  - zwei Mittelwerte $\hat{\mu}_1$ und $\hat{\mu}_2$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Beispiel: Weisse Weihnachten



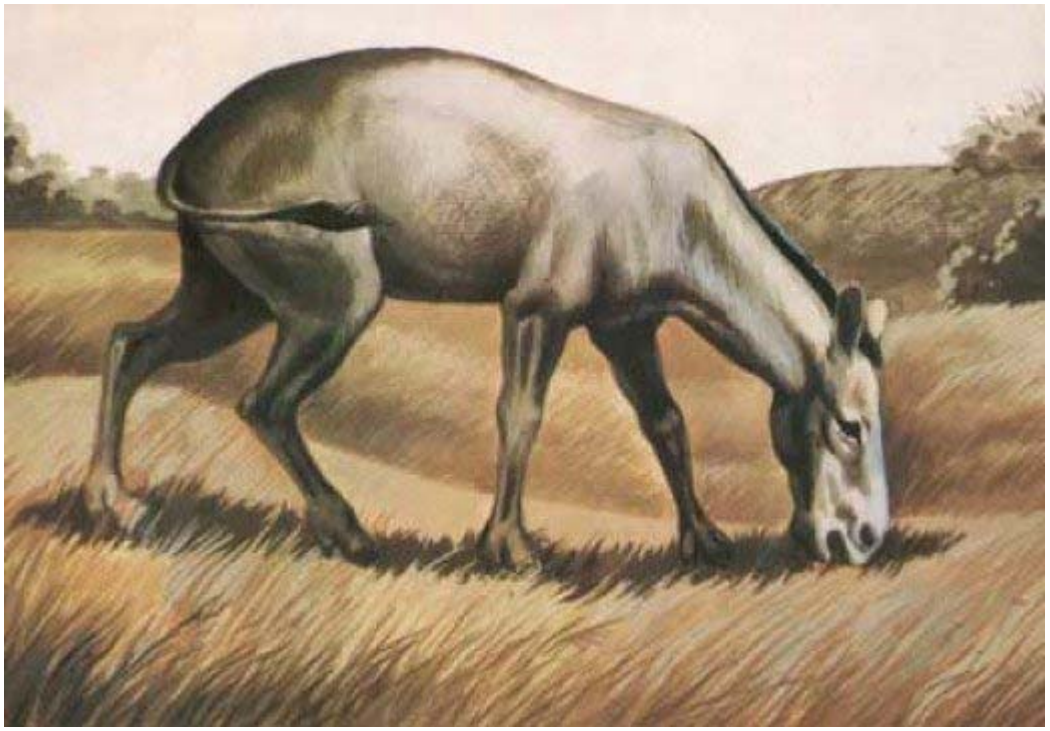
Gab es früher öfter weisse Weihnachten?

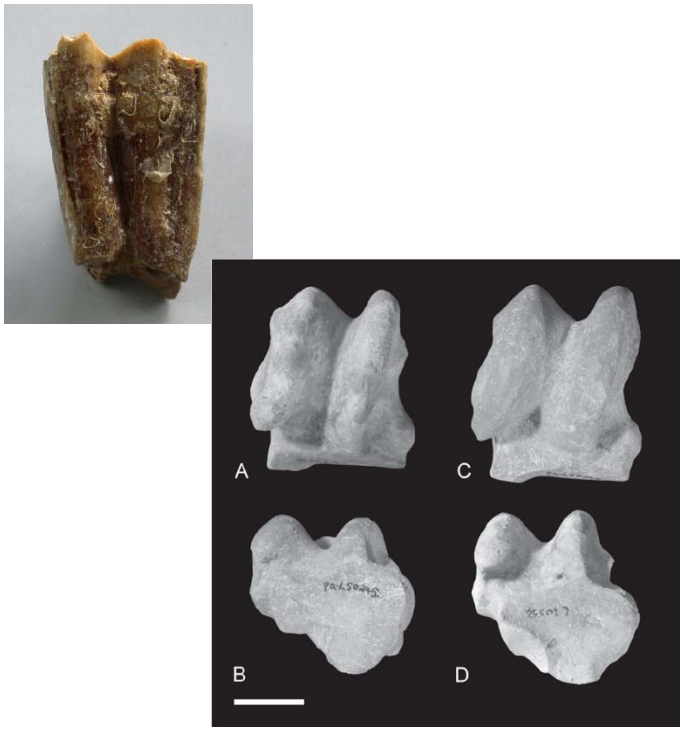
1961-1990: 39 von 90 Tagen mit Schnee

1991-2010: 16 von 60 Tagen mit Schnee

GdM 2: LinAlg & Statistik
FS 2018 – Woche 13

Beispiel: Hipparions





Datenlage

Hipparion Africanum: $n_A = 39,\ \hat{\mu}_A = 25.9mm,\ \hat{\sigma}_A = 2.2mm$

Hipparion Lybicum: $n_L = 38,\ \hat{\mu}_L = 28.4mm,\ \hat{\sigma}_A = 4.3mm$

Ad-Hoc Test via VI-Überlapp

Frage: **Überlappen sich die beiden VIs?**

Testantwort: Falls sich die beiden VIs **nicht überlappen**, so ist der Unterschied zwischen p_1 / μ_1 und p_2 / μ_2 statistisch signifikant.

Falls sich die beiden VIs überlappen, ist hingegen keine Aussage möglich. Es ist immer noch möglich, dass ein genauerer Test die Nullhypothese $p_1 = p_2$ bzw. $\mu_1 = \mu_2$ verwirft.

→ **Durchführen des Tests für die 2 Beispiele…**

2-Stichproben-Binomial-Test

Vergleich von 2 Proportionen (d.h. Anteilen p_1 und p_2):

Wir testen die Nullhypothese $H_0 : p_1 = p_2$, bzw. $p_1 - p_2 = 0$
Falls $n_1 \hat{p}_1 (1 - \hat{p}_1)$ und $n_2 \hat{p}_2 (1 - \hat{p}_2)$ beide ≥ 5, so können wir die folgende Teststatistik benützen:

- $$T_2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sim \chi_1^2$$ → `prop.test()` in R

Es gibt andere, ähnliche Teststatistiken, welche wir hier wegen der fehlenden Implementation in R jedoch nicht besprechen.

2-Stichproben-t-Test

Vergleich von 2 Mittelwerten μ_1 und μ_2:

Wir testen die Nullhypothese $H_0 : \mu_1 = \mu_2$, bzw. $\mu_1 - \mu_2 = 0$ unter der Annahme, dass die Streuung in den beiden Stichproben unbekannt aber identisch ist, d.h. $\sigma_1 = \sigma_2$

Dazu benützen wir die folgende Teststatistik:

- $$T = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$, wobei $$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

  die gepoolte Schätzung der Standardabweichung ist.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Marcel Dettling

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

https://www.zhaw.ch/de/ueber-uns/person/dtli

ETH Zürich, 30. Mai 2018

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Regression

Beispiel:

In Indien behindern basische Böden Pflanzen beim Wachstum. Es werden daher Baumarten gesucht, die eine hohe Toleranz gegen solche Umweltbedingungen haben. In einem Freilandversuch wurden auf einem Feld mit grossen lokalen Schwankungen des pH-Wert 120 Bäume einer bestimmten Art gepflanzt. Nach 3 Jahren wurde von jedem Baum die Höhe gemessen. Gleichzeitig war auch der pH-Wert des Bodens an der entsprechenden Stelle bekannt. Die Daten können in einem Scatterplot dargestellt werden.

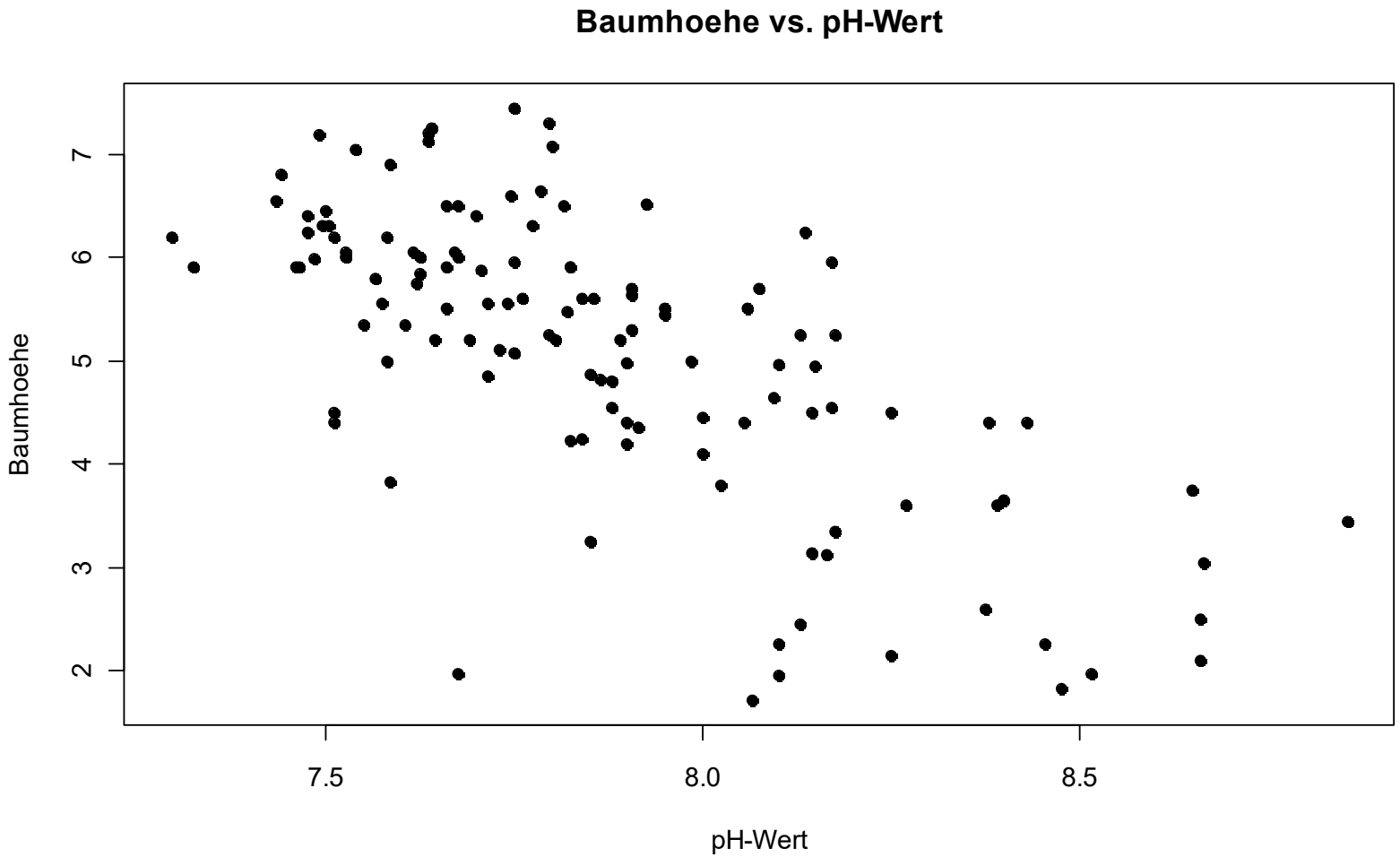
GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Ausschnitt aus der Daten-Tabelle

Baum	Höhe	pH	SAR
1	5.91	7.325	0.0969
2	5.20	7.690	0.4393
3	4.40	7.900	1.0000
4	4.50	8.145	1.3160
5	6.05	7.615	0.0607
6	6.00	7.525	0.2041
7	…	…	…

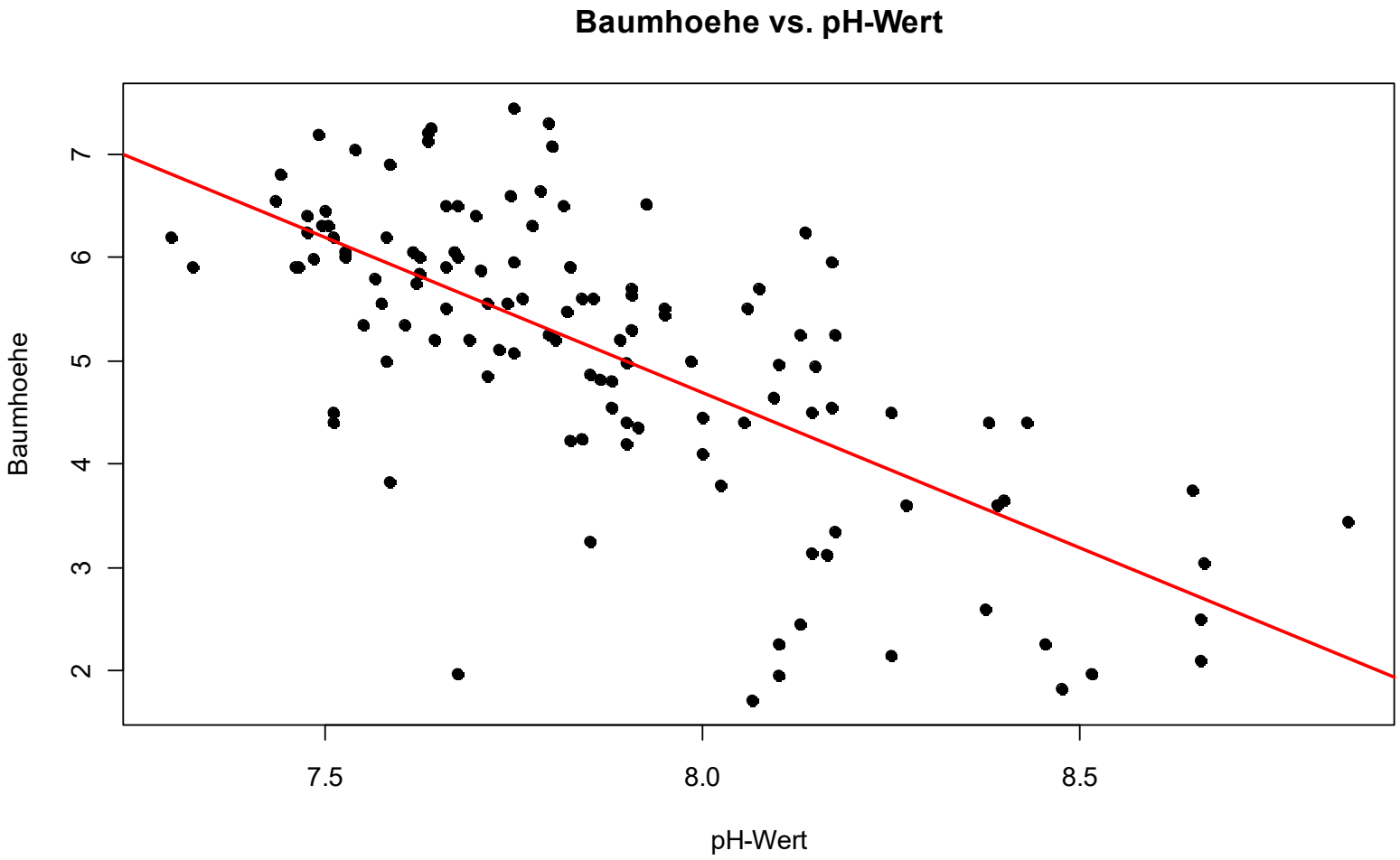
GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Scatterplot Baumhöhe vs. pH-Wert



GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Regressionsgerade



Einfache lineare Regression

Mit zunehmendem pH-Wert nimmt die Baumhöhe tendenziell ab. Der Zusammenhang scheint linear. Es bietet sich also an, eine Gerade zur Beschreibung zu verwenden:

$$f(x) = \beta_0 + \beta_1 x, \text{ bzw. } Höhe = \beta_0 + \beta_1 \cdot pH$$

Name/Bedeutung der Grössen in der Geradengleichung:

$\beta_0 =$ "Intercept"

$\beta_1 =$ "Slope"

Die Anpassung einer Geraden in einen 2-dimensionalen Scatterplot heisst **einfache lineare Regression,** weil:

- es nur eine erklärende Grösse gibt ("*einfach*").
- wir ein linearen Zusammenhang haben ("*linear*").

Modell & Zufallsfehler

Nun bringen wir die Daten ins Spiel. Die Gerade führt nicht durch jeden Datenpunkt, d.h. es gibt (zufällige) Abweichungen:

$$y_i = \beta_0 + \beta_1 x_i + E_i, \quad \text{für alle } i = 1, \ldots, n$$

Bedeutung der Grössen:

y_i ist die Zielvariable (Baumhöhe) der i-ten Beobachtung.

x_i ist die erklärende Grösse (pH) der i-ten Beobachtung.

β_0, β_1 sind die Regressionskoeffizienten, welche erst noch aus den Daten bestimmt/geschätzt werden müssen.

E_i ist der zufällige Rest oder Fehler, d.h. die zufällige Abweichung zwischen Beobachtung und Gerade.

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Anpassung der Geraden

Einfache lineare Regression - was ist die Aufgabe???

Gesucht ist eine Gerade, die möglichst gut "zu den Daten passt". Wir müssen also β_0, β_1 so festlegen, dass die Abweichungen zwischen Datenpunkten und Gerade klein sind...

Intuitives Hineinlegen in den Scatterplot

→ **Welche Kriterien benützt man dabei...**

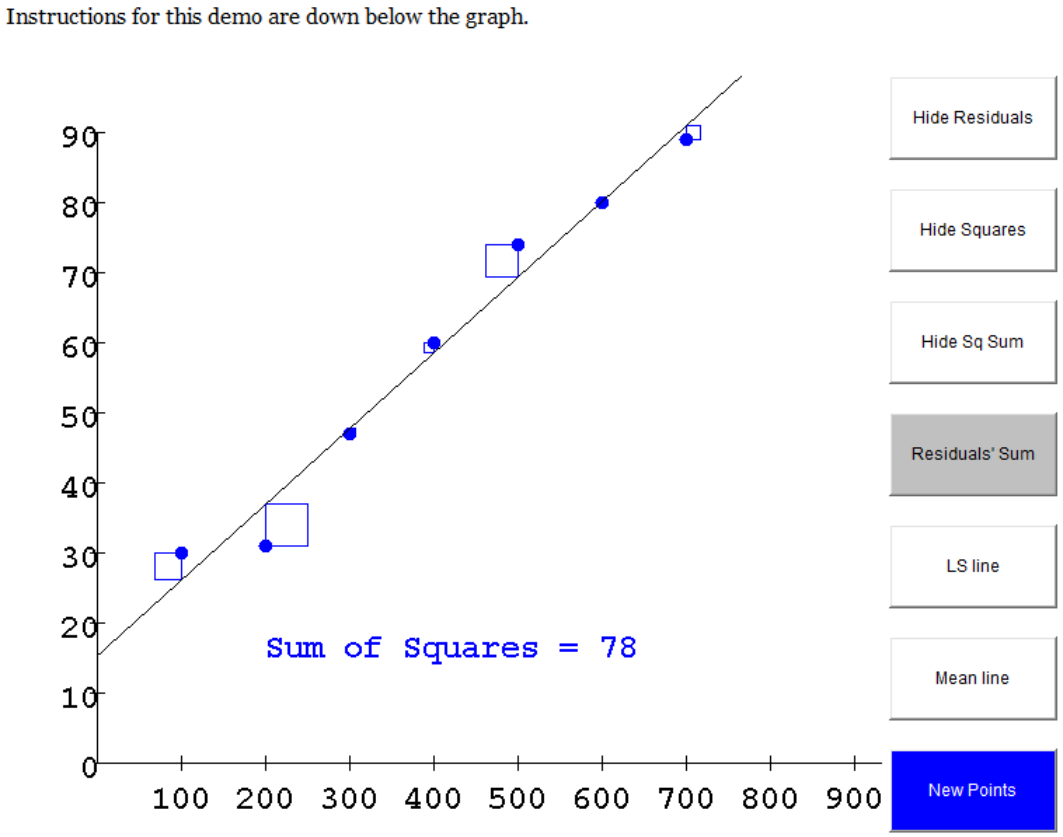
Diskussion und Erklärung

→ **Plenum / Wandtafel ...**

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Kleinste Quadrate: Applet

→ http://demonstrations.wolfram.com/LeastSquaresCriteriaForTheLeastSquaresRegressionLine/



Wir müssen eine Gerade durch die Punkte legen.

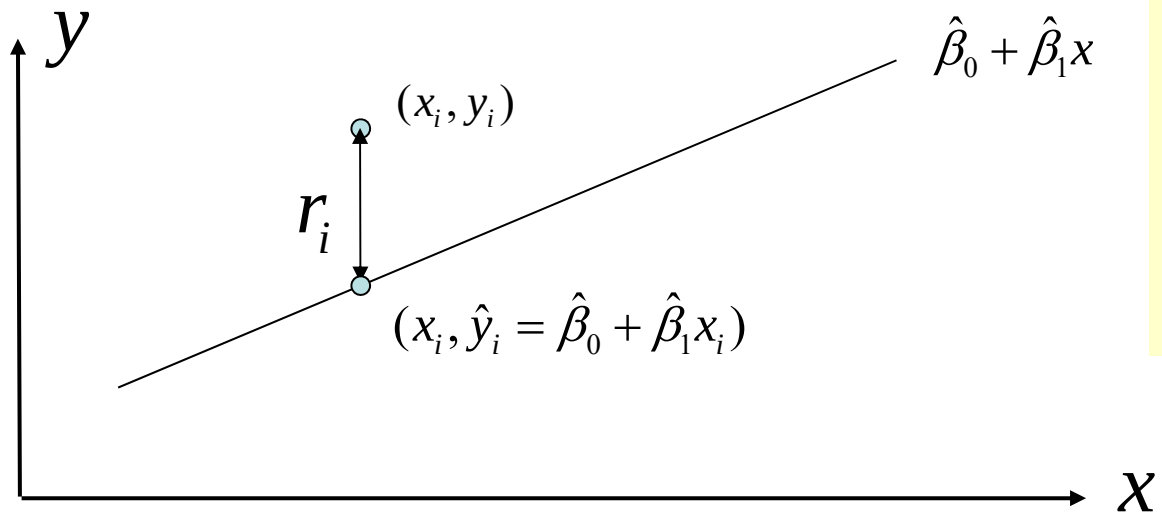
Es gibt viele Lösungen. Einige sind "gut", andere sind weniger geeignet.

Unser Paradigma: die Summe der Fehlerquadrate soll minimal sein!

Residuen vs. Fehler

Das Residuum $r_i = y_i - \hat{y}_i$ ist die Differenz zwischen dem beobachteten und dem angepassten y-Wert für den i-ten Datenpunkt. Achtung: der Fehler E_i ist ein Konzept und eine Zufallsvariable, das Residuum r_i ist ein numerischer Wert.

Illustration der Residuen



Wir bestimmen die Gerade so, dass die Quadratsumme der Residuen möglichst klein ist: $\sum_{i=1}^{n} r_i^2$

Kleinste Quadrate: Mathematisch

In Worten / Mathematisch...

Durch eine Punktwolke von Daten $(x_i, y_i)_{i=1,\ldots,n}$ soll die Gerade so gelegt werden, dass die Summe der quadrierten Abstände r_i zwischen dem Beobachtungswert y_i und dem zugehörigen Punkt auf der Geraden $\hat{y}_i$ minimal ist. Die Funktion

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \min!$$

misst, wie gut die durch (β_0, β_1) definierte Gerade zu den Daten passt. Sie soll einen möglichst kleinen Wert annehmen.

Lösung: → **siehe nächste Folie...**

Lösungsidee: Partielle Ableitungen

- Wir leiten die Funktion $Q(\beta_0, \beta_1)$ partiell nach den beiden Argumenten β_0 und β_1 ab, und setzen die Ableitungen gleich null:

$$\frac{\partial Q}{\partial \beta_0} = 0 \quad \text{und} \quad \frac{\partial Q}{\partial \beta_1} = 0$$

- Es entsteht ein lineares Gleichungssystem, mit (hier) zwei Unbekannten β_0, β_1 und zwei Gleichungen. Diese Gleichungen heissen *Normalengleichungen.*

- Man kann die Lösung für β_0, β_1 *explizit* als Funktion der Datenpaare $(x_i, y_i)_{i=1,\ldots,n}$ aufschreiben, siehe nächste Folie...

Kleinste Quadrate: Lösung

Die gemäss Kleinsten Quadraten optimale Lösung für die Gerade, d.h. die aus den Daten geschätzten Regressionskoeffizienten sind:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{und} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Für eine gegebene Punktwolke $(x_i, y_i)_{i=1,\ldots,n}$ können wir also die Gerade (mit dem TR, besser mit R) bestimmen.

- **Ergebnis für unser Beispiel "Baumhöhe":**

  $\hat{\beta}_0 = 28.7,\ \hat{\beta}_1 = -3.0$ mit Hilfe von Softwarepaket gerechnet!

→ **Probieren sie es selber aus (in den Übungen)...**

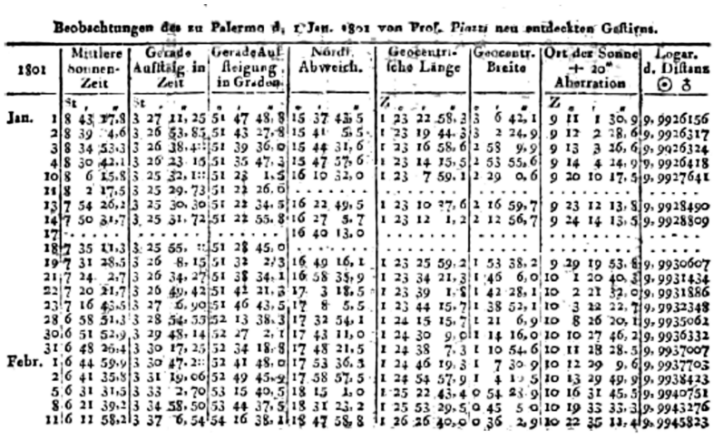
GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Anhang: warum Kleinste Quadrate?

Historisches...

Die Methode wurde innert weniger Jahre (1801, 1805) zweimal unabhängig voneinander zum Lösen von Problemen in der Astronomie entwickelt...

Quelle: → http://de.wikipedia.org/wiki/Methode_der_kleinsten_Quadrate



Beobachtungen des zu Palermo d. 1. Jan. 1801 von Prof. Piazzi neu entdeckten Gestirns.

1801		Mittlere Sonnen-Zeit	Gerade Aufsteig. in Zeit	Gerade Aufsteigung in Graden	Nördl. Abweich.	Geocentrische Länge	Geocentr. Breite	Ort der Sonne + 20" Aberration	Logar. d. Distanz ☉ δ
		h ′ ″	h ′ ″	° ′ ″	° ′ ″	♈ ° ′ ″	° ′ ″	♈ ° ′ ″	
Jan.	1	8 43 37,8	3 27 21, 25	51 47 48, 8	15 37 42, 5	1 23 22 58, 2	3 6 42, 1	9 11 1 30, 9	9, 9926156
	2	8 39 4,6	3 26 53, 85	51 43 27, 8	15 41 5, 5	1 23 19 44, 2	3 2 24, 9	9 12 2 28, 6	9, 9926317
	3	8 34 32,3	3 26 38, 4	51 39 36, 0	15 44 31, 6	1 23 16 58, 6	2 58 9, 9	9 13 3 26, 6	9, 9926324
	4	8 30 0,1	3 26 23, 15	51 35 47, 2	15 47 57, 6	1 23 14 15, 5	2 53 55, 6	9 14 4 24, 9	9, 9926418
	10	8 6 15,8	3 25 32, 1	51 23 1, 5	16 10 32, 0	1 23 7 59, 1	2 29 0, 6	9 20 10 17, 5	9, 9927641
	11	8 2 17,5	3 25 29, 73	51 22 26, 0	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
	13	7 54 26,2	3 25 30, 30	51 22 34, 5	16 22 49, 5	1 23 10 27, 6	2 16 59, 7	9 23 12 13, 8	9, 9928490
	14	7 50 31,7	3 25 31, 72	51 22 55, 8	16 27 5, 7	1 23 12 1, 2	2 12 56, 7	9 24 14 13, 5	9, 9928809
	17	. . . . .	. . . . .	. . . . .	16 40 13, 0	. . . . .	. . . . .	. . . . .	. . . . .
	18	7 35 11,3	3 25 55, 1	51 28 45, 0	. . . . .	. . . . .	. . . . .	. . . . .	. . . . .
	19	7 31 28,5	3 26 8, 15	51 31 2, 3	16 49 16, 1	1 23 25 59, 2	1 53 38, 2	9 29 19 53, 8	9, 9930607
	21	7 24 2,7	3 26 34, 27	51 38 34, 1	16 58 35, 9	1 23 34 21, 3	1 45 6, 0	10 1 20 40, 3	9, 9931434
	22	7 20 21,7	3 26 49, 42	51 42 21, 3	17 3 18, 5	1 23 39 1, 8	1 42 28, 1	10 2 21 32, 0	9, 9931886
	23	7 16 45,6	3 27 6, 90	51 46 43, 5	17 8 5, 5	1 23 44 15, 7	1 38 52, 1	10 3 22 22, 7	9, 9932348
	28	6 58 51,3	3 28 54, 53	52 13 38, 3	17 32 54, 1	1 24 15 15, 7	1 21 6, 9	10 8 26 20, 1	9, 9935062
	30	6 51 52,9	3 29 48, 14	52 27 2, 1	17 43 11, 0	1 24 30 9, 0	1 14 16, 0	10 10 27 36, 2	9, 9936332
	31	6 48 26,4	3 30 17, 25	52 34 18, 8	17 48 21, 5	1 24 38 7, 3	1 10 54, 6	10 11 28 28, 5	9, 9937007
Febr.	1	6 44 59,9	3 30 47, 2	52 41 48, 0	17 53 36, 2	1 24 46 19, 3	1 7 30, 9	10 12 29 9, 6	9, 9937703
	2	6 41 35,8	3 31 19, 66	52 49 45, 9	17 58 57, 5	1 24 54 57, 9	1 4 1, 5	10 13 29 49, 9	9, 9938423
	5	6 31 31,5	3 33 2, 70	53 15 40, 5	18 15 1, 0	1 25 22 43, 4	0 54 23, 9	10 16 31 45, 5	9, 9940751
	8	6 21 39,2	3 34 58, 50	53 44 37, 5	18 31 23, 2	1 25 53 29, 5	0 45 5, 0	10 19 33 33, 3	9, 9943276
	11	6 11 58,2	3 37 6, 54	54 16 38, 1	18 47 58, 8	1 26 26 40, 0	0 36 2, 9	10 22 35 12, 4	9, 9945823

Carl Friedrich Gauss

Adrien-Marie Legendre

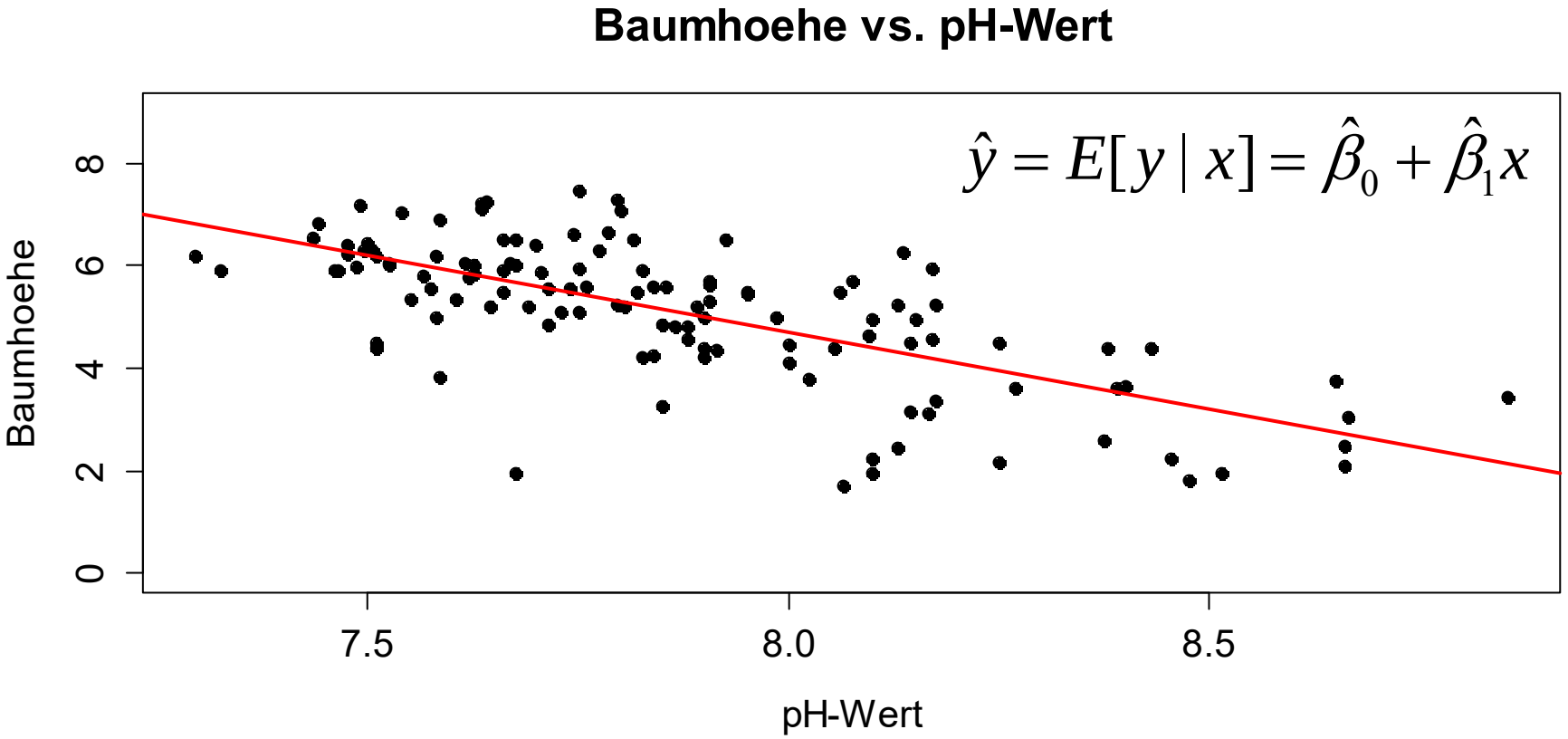
Anhang: warum Kleinste Quadrate?

Mathematisches...

- Das Verfahren ist einfach in dem Sinne, dass die Lösung explizit als Funktion von $(x_i, y_i)_{i=1,\dots,n}$ bekannt ist.
- Die Gerade geht durch den Daten-Schwerpunkt $(\overline{x}, \overline{y})$
- Die Summe der Residuen addiert sich zu null: $\sum_{i=1}^{n} r_i = 0$
- Tiefer gehende, mathematische Optimalität lässt sich beweisen, indem man die Schätzeigenschaften von $\hat{\beta}_0, \hat{\beta}_1$ untersucht, speziell bei normalverteilten Fehlern E_i.

Fitted Values und Regressionsgerade

Die geschätzten Parameter $\hat{\beta}_0, \hat{\beta}_1$ können wir nun benützen, um die *angepassten Werte* $\hat{y}$ (engl. *Fitted Values*) anzugeben. Es handelt sich um einen bedingten Erwartungswert:



GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Haben wir ein gutes Modell für die Baumhöhen-Vorhersage gefunden?

a) Ausserhalb der Punktwolke

Unklar, ziemlich sicher nein...

b) Innerhalb der Punktwolke?

Ja, unter den folgenden, zu prüfenden Bedingungen

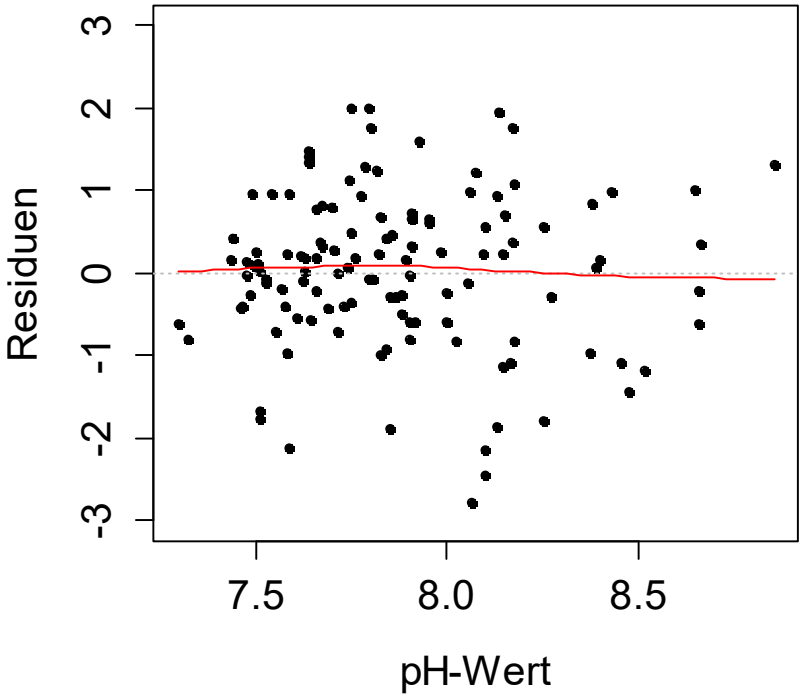
- der Zusammenhang ist eine Gerade ist, d.h. $E[E_i] = 0$
- die Streuung der Fehler konstant ist, d.h. $Var(E_i) = \sigma^2$
- die Fehler unkorreliert sind (repräsentative Stichprobe!)
- die Fehler (approximativ) normalverteilt sind

→ **Gedankenfutter: Schattige Ecke auf dem Feld?**

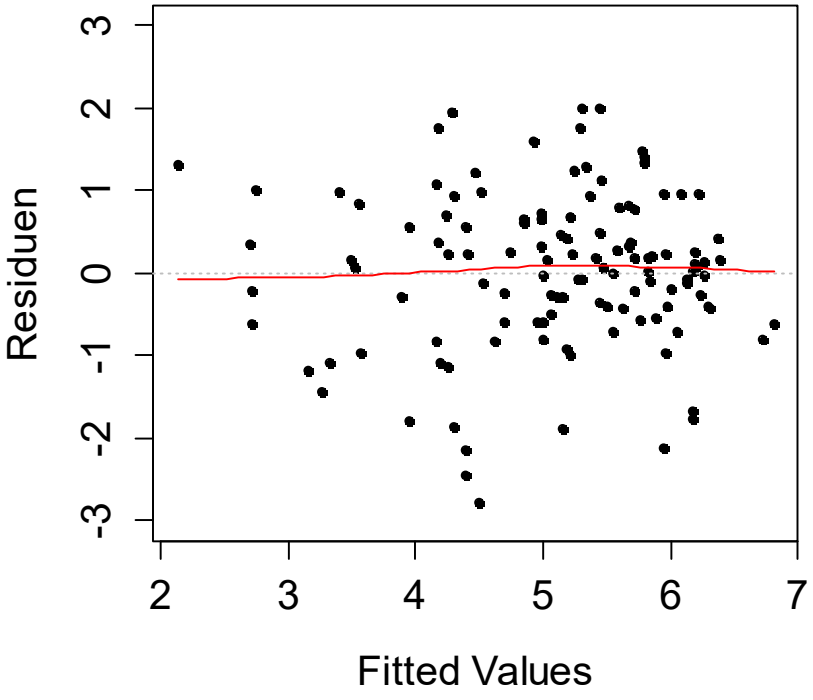
Modelldiagnostik

Um die Vertrauenswürdigkeit der Regressionsgerade zu evaluieren, müssen die getroffenen Annahmen überprüft werden. Für $E[E_i] = 0$ und $Var(E_i) = \sigma^2$ betrachten wir:

Residuen vs. Prädiktor



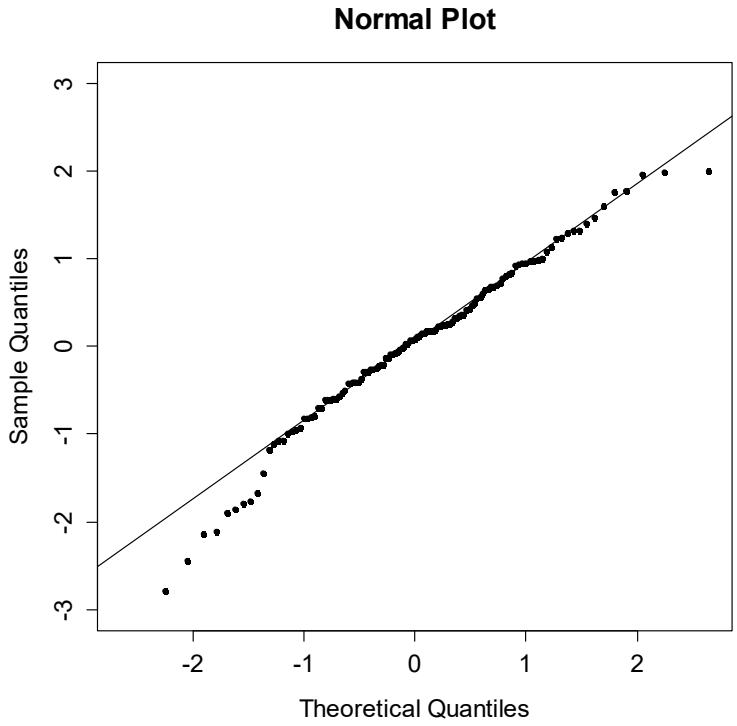
Tukey-Anscombe-Plot



GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Modelldiagnostik

Um die Vertrauenswürdigkeit der Regressionsgerade zu evaluieren, müssen die getroffenen Annahmen überprüft werden. Für die Normalverteilung betrachten wir:



Es gibt auch noch weitere, verfeinerte Diagnoseplots.

Diese bespricht man in der Regel erst bei der multiplen linearen Regression.

Eigenschaften der Schätzer

Die KQ-Schätzer sind erwartungstreu, d.h.

$$E[\hat{\beta}_0] = \beta_0 \text{ und } E[\hat{\beta}_1] = \beta_1$$

Die Varianzen der Schätzer sind wie folgt:

$$Var(\hat{\beta}_0) = \sigma_E^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \text{ und } Var(\hat{\beta}_1) = \frac{\sigma_E^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Präzise Schätzungen erhält man durch:

- eine grosse Anzahl Beobachtungen n
- eine ausreichende Streuung der x_i
- einen informativen Prädiktor, so dass σ_E^2 klein ist

Schätzen der Fehlervarianz σ_E^2

Neben den Regressionskoeffizienten ist auch eine Schätzung der Fehlervarianz σ_E^2 von Interesse. Sie ist ein wichtiger Input für alle Tests und Konfidenzintervalle, die wir besprechen:

Die Schätzung basiert auf der *residual sum of squares* (RSS):

$$\hat{\sigma}_E^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} r_i^2$$

In unserem Beispiel ergibt sich als `Residual standard error`:

```
> summary(fit)
...
Residual standard error: 1.008 on 121 degrees of freedom
```

Nutzen von Regression

1) Untersuchen der Beziehung zwischen y und x

Die Absicht ist, genau zu verstehen, wie und wie stark die Zielvariable vom Prädiktor abhängt. Es gibt diverse Kenngrössen und statistische Tests, welche sich dieser Frage widmen.

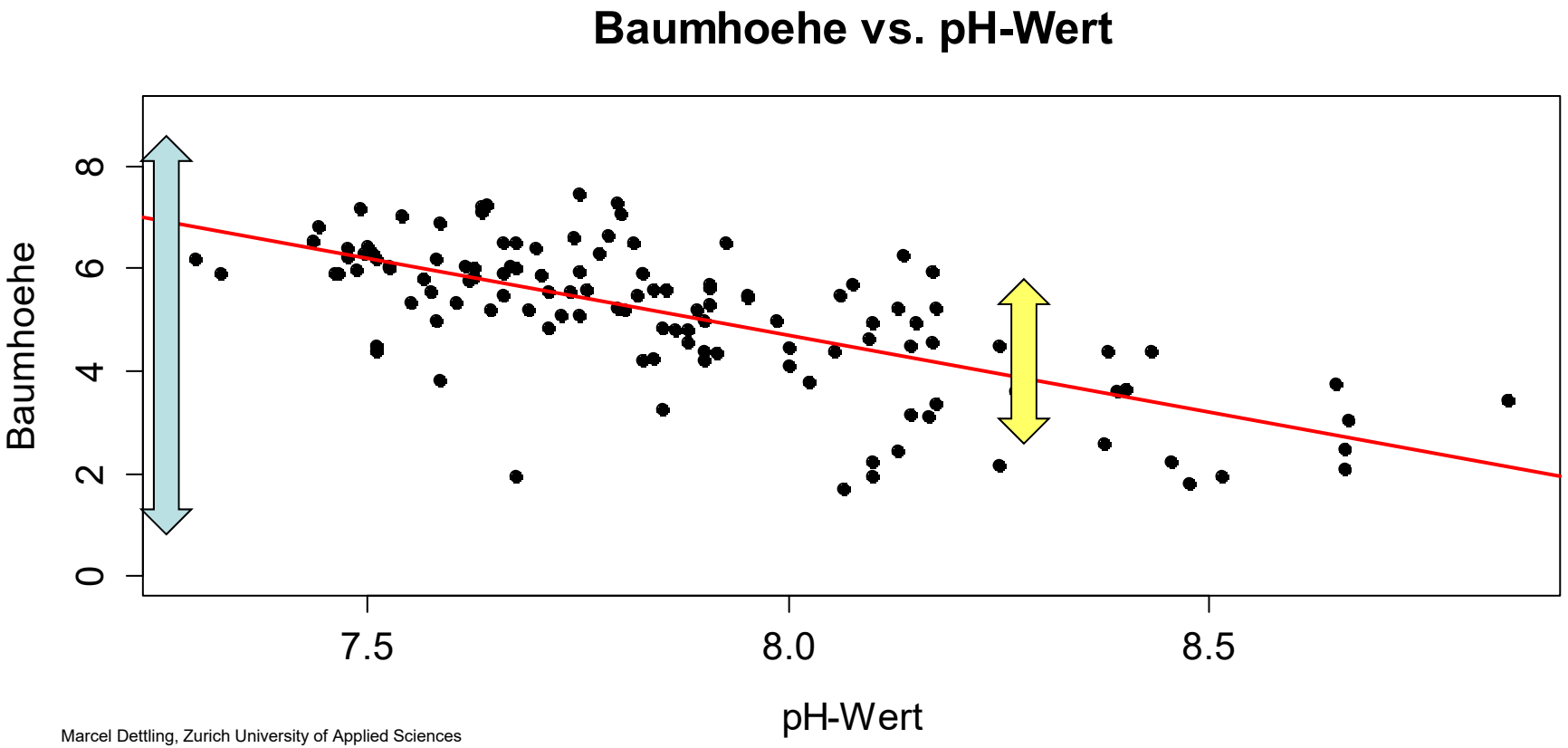
2) Vorhersage

Wir können die Regressionsgleichung, bzw. –gerade benützen, um für einen beliebigen ph-Wert die Baumhöhe anzugeben.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

R^2: Erklärungsgehalt der Regressionsgerade

Intuitiv: je grösser der blaue Pfeil im Vergleich zum gelben ist, desto grösser ist der Erklärungsgehalt der Regressionsgerade



Das Bestimmtheitsmass R^2

Der Erklärungsgehalt der Regressionsgeraden wird mit R^2 gemessen. Man nennt R^2 das *Bestimmtheitsmass*, bzw. englisch *Coefficient of Determination*. Es handelt sich um das Verhältnis zwischen dem gelben und dem blauen Pfeil. Es ist der Anteil der Gesamtstreuung, welche durch die Gerade erklärt wird.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \in [0,1]$$

Je grösser R^2, desto enger streuen die Punkte um die Gerade. Es gibt aber kein formelles Kriterium, wie gross R^2 sein muss.

Vertrauensintervall für die Steigung β_1

Das 95%-VI für die Steigung β_1 ist um die Punktschätzung $\hat{\beta}_1$ zentriert und enthält alle Werte die ebenfalls plausibel sind. Die Unsicherheit ist durch die Streuung der Datenpunkte bedingt..

95%-VI für β_1 : $\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \hat{\sigma}_{\hat{\beta}_1}$, bzw.

$$\hat{\beta}_1 \pm qt_{0.975;n-2} \cdot \sqrt{\hat{\sigma}_E^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Salopp: $-3 \pm 2 \cdot 0.28 = [-3.56; -2.44]$

Exakt: $[-3.566353; -2.440355]$ aus Statistikpaket

Test für die Steigung β_1

Es gibt einen statistischen Test, mit welchem man feststellen kann, ob die Steigung der Regressionsgerade signifikant von null oder einem beliebigen anderen Wert b verschieden ist:

$$H_0 : \beta_1 = 0, \text{ bzw. } H_0 : \beta_1 = b$$

Man testet zweiseitig auf dem 95%-Niveau. Die Alternative ist:

$$H_A : \beta_1 \neq 0, \text{ bzw. } H_A : \beta_1 \neq b$$

Als Teststatistik verwenden wir:

$$T_{H_0:\beta_1=0} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ bzw. } T_{H_0:\beta_1=b} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}, \text{ beide haben } t_{n-2}\text{-Verteilung.}$$

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Lesen von Output

```
> summary(fit)

Call: lm(formula = height ~ ph, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.7227     2.2395   12.82   <2e-16 ***
ph           -3.0034     0.2844  -10.56   <2e-16 ***
---
Residual standard error: 1.008 on 121 degrees of freedom
Multiple R-squared:  0.4797, Adjusted R-squared:  0.4754
F-statistic: 111.5 on 1 and 121 DF,   p-value: < 2.2e-16
```

→ **Man beachte die Bedeutung der Grössen!**

Test für die Steigung β_1

Praxisbeispiel:

Man nehme die Baumhöhen-Daten und überprüfe mit einem Test die Hypothese $H_0 : \beta_1 = -2$. Die Informationen von Folie 28 dürfen verwendet werden. Man beantworte auch:

a) *Formulieren sie umgangssprachlich, was sie eben getestet haben und was einem dies in der Praxis nützt.*

b) *Worin liegt der Zusammehang zwischen dem Testresultat und dem 95%-VI von Folie 25? Hätten wir das Testresultat schon vom VI vorhersehen können?*

→ **Siehe Wandtafel…**

Test für den Achsenabschnitt β_0

Für den Achsenabschnitt gibt es analoge Tests.

- Egal wie das Testresultat ausfällt, den Achsenabschnitt soll man stets im Regressionsmodell belassen!
- Der Achsenabschnitt bietet Schutz vor Nichtlinearität und Kalibrationsfehlern. Wird er weggelassen, so sind die Resultate für die Praxis meist weniger brauchbar.
- Falls einem (physikalische) Theorie diktiert, dass es keinen Achsenabschnitt geben darf, er aber trotzdem signifikant ist, dann heisst das, dass die lineare Beziehung nicht bis zum Punkt $x = 0$ extrapoliert werden darf.

Vorhersage

Mit der Regressionsgerade können wir den y-Wert an beliebiger x-Stelle vorhersagen. Wir benützen die Gleichung:

$$E[y \mid x] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \text{ a.k.a } \textbf{\textit{“fitted value”}}$$

Beispiel: Für einen pH-Wert von 8.0 erwarten wir:

$$28.7 - 3.0 \cdot 8.0 = 4.7 \text{ Meter Baumhöhe}$$

Aber Achtung:

Interpolation im Bereich der beobachteten x-Werte ist i.d.R. problemlos. Extrapolation (z.B. für pH-Werte von 1 oder 10) funktioniert in der Regel nicht und ist gefährlich!

Vertrauensintervall für $E[y \mid x]$

Wir haben gelernt, wie man den Fitted Value $\hat{\beta}_0 + \hat{\beta}_1 x$ bestimmt, d.h. die erwartete Baumhöhe für gegebenen pH-Wert. Achtung, es handelt sich um eine Schätzung mit Unsicherheit.

Ein 95%-VI für den angepassten Wert an der Stelle x ist:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

Die Werte können mit Statistik-Software berechnet werden.

```
Fitted    Lower     Upper
4.695861 4.501321 4.8904
```

Prognoseintervall für y

Das 95%-VI für $E[y \mid x]$ zeigt die Variabilität des Fitted Value. Es beinhaltet jedoch nicht die Streuung der Daten um die Gerade und definiert darum nicht die Region, in der ein zukünftiger Datenpunkt zu liegen kommt. Ein 95%-Prognoseintervall an der Stelle x ist gegeben durch:

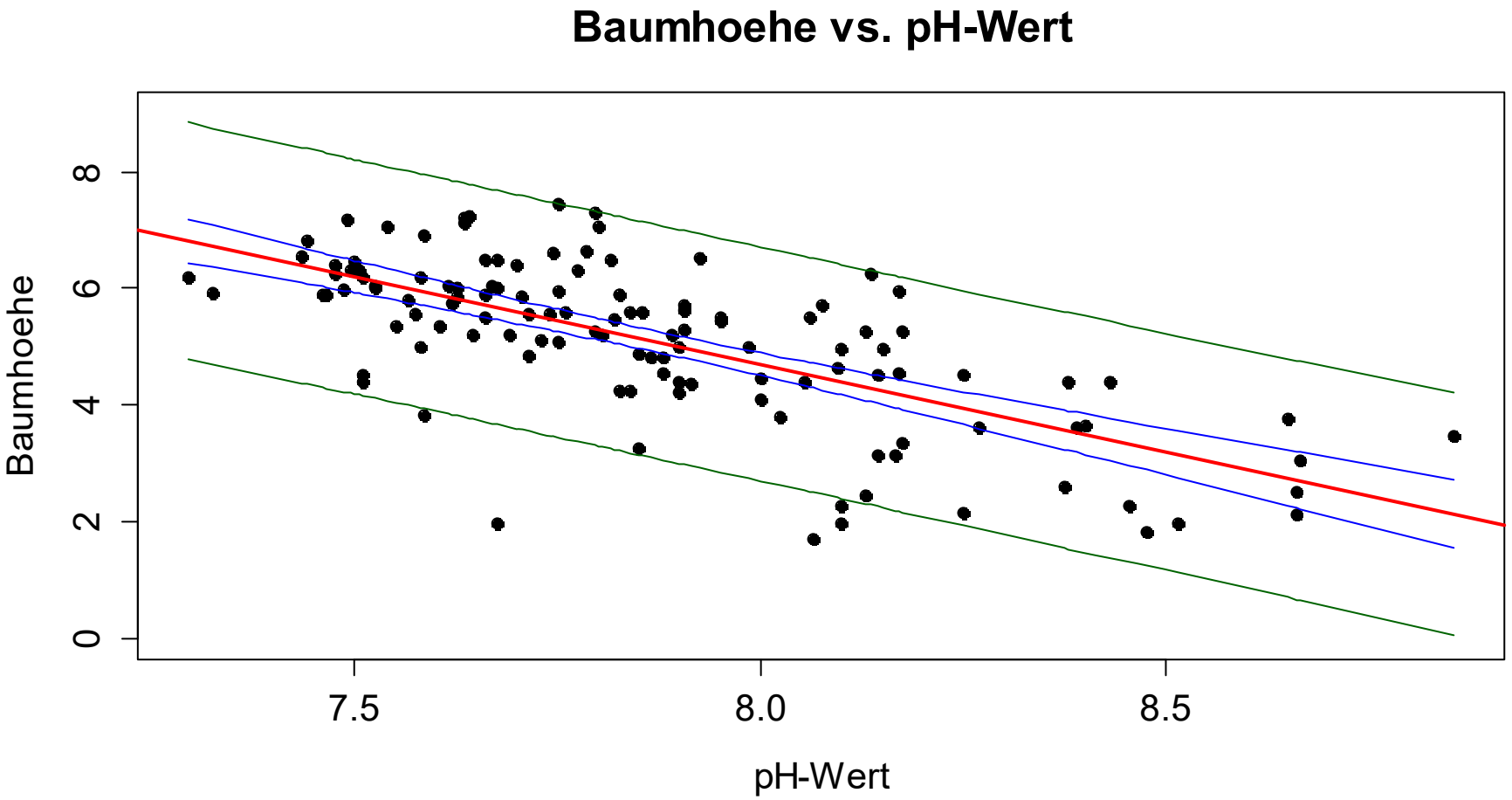
$$\hat{\beta}_0 + \hat{\beta}_1 x \pm qt_{0.975;n-2} \cdot \hat{\sigma}_E \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Konkret:

```
Fitted    Lower     Upper
4.695861 2.690581 6.70114
```

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Vertrauens- und Vorhersagebereich



Ausblick: Multiple Regression

In der realen Welt wird die Zielgrösse meist von mehreren Prädiktoren gleichzeitig beeinflusst. Es lohnt sich also, den multiplen Zusammenhang zu studieren. Das Modell lautet:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + E_i$$

Aufgabe ist wiederum, die Koeffizienten $\beta_0, \beta_1, \ldots, \beta_p$ aus den zur Verfügung stehenden Daten zu schätzen. Hinweis: das Resultat fällt dabei im Allg. anders aus, wie wenn man mehrere einfache Regression von der Zielgrösse gegen jeden Prädiktor separat ausführt!

→ Zur Schätzung benützt man weiterhin die KQ-Methode.

Ausblick: Multiple Regression

Output aus Statistikpaket:

```
Call: lm(height ~ ph + l.sar, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.9466     2.7445   9.818  < 2e-16 ***
ph           -2.7558     0.3603  -7.649  5.6e-12 ***
l.sar        -0.2519     0.2255  -1.117    0.266
---
Residual standard error: 1.007 on 120 degrees of freedom
Multiple R-squared:  0.485,  Adjusted R-squared:  0.4764
F-statistic: 56.51 on 2 and 120 DF,   p-value: < 2.2e-16
```

GdM 2: LinAlg & Statistik
FS 2018 – Woche 14

Informationen zur Prüfung

- Dauer: 90min, je ~2 Aufgaben aus LinAlg und Statistik
- Stoff: alles, was in Vorlesung und Übungen vorkam
- Aufgaben: Transferleistung nötig, Verständnis wichtig
- Es sind beliebige schriftliche Hilfsmittel erlaubt
- Taschenrechner sind verboten!
- Ferienpräsenz: siehe Webpage

→ **Schöne Ferien und viel Erfolg bei der Vorbereitung!**