Normalverteilung

Peter Büchel

HSLU I

ASTAT: Block 07

Kontinuierliche Messdaten

- In vielen Anwendungen: Keine diskreten Daten, sondern *Messdaten*
- Messdaten können jeden Wert in einem bestimmten Bereich annehmen
- Bsp: Gemessene Körpergrössen (in cm) können *jeden* Wert im Intervall $[0, 500]$ annehmen
- Also auch:

$$145.325\,986\,54\ldots$$

- Voraussetzung: Beliebig genaue Messung möglich

Definitionen

- *Wertebereich* W_X einer Zufallsvariable $\rightarrow$ Menge aller Werte, die X annehmen kann

- Zufallsvariable X *stetig*: Wertebereich W_X kontinuierlich

- Kontinuierliche Menge: Hier Ausschnitt aus der Zahlengeraden

- Kontinuierlich: „Zusammenhängend" und nicht „löchrig", wie Menge $\{1, 2, 3\}$

- Wichtige kontinuierliche Wertebereich:

$$W_X = \mathbb{R},\ \mathbb{R}^+ \quad \text{oder} \quad [0, 1]$$

- Letzter Fall: Zahlen 0 und 1 *und* alle Zahlen dazwischen

Intervalle

- Intervall, wo die Grenzen innerhalb oder ausserhalb des Intervalls sein sollen $\quad\rightarrow\quad$ eckige und runde Klammern
  - Runde Klammer: Wert ausserhalb des Intervalls
  - Eckige Klammer: Wert innerhalb des Intervalls
- Intervall $(a, b]$: Alle Punkte x mit $x > a$ und $x \leq b$

Beispiel

- Intervall

$$(1.2, 2.5]$$

  - Enthält die Zahl 1.2 nicht, die Zahl 2.5 schon

- Unterschied zum Intervall

$$[1.2, 2.5]$$

- Es enthält nur den einen Punkt 1.2 der Zahlengeraden mehr

- Statistik: Spielt keine Rolle, ob 1. oder 2. Intervall verwendet wird

Punktwahrscheinlichkeit 0

- W'keitsverteilung einer *diskreten* Zufallsvariablen: „Punkt"-W'keiten $P(X = x)$ für alle möglichen x im Wertebereich

- Vorher: X Körpergrösse auf cm gerundet $\quad \rightarrow \quad P(X = 174)$ nicht 0

- Stetige Zufallsvariable X: Für alle $x \in W_X$ gilt:

$$P(X = x) = 0$$

- Folgerung: W'keitsverteilung von X kann *nicht* mittels „Punkt"-W'keiten beschrieben werden

Beispiel: Körpergrösse

- Messen Körpergrösse von Personen
- W'keit *genau* eine Körpergrösse von $182.254\,680\,895\,434\ldots$ cm zu messen ist gleich 0:

$$P(X = 182.254\,680\,895\,434\ldots) = 0$$

- Verwendung der W'keit einen exakten Messwert zu messen, bringt nichts für W'keitsverteilung der Körpergrösse
- Summe aller W'keiten müsste 1 ergeben $\quad\rightarrow\quad$ Ist nicht der Fall
- Aber möglich: W'keit, dass ein Messwert in einem bestimmten *Bereich* liegt

- Beispiel: zwischen 174 und 175 cm:

$$P(174 < X \leq 175)$$

- Diese W'keit ist dann nicht mehr 0

- Da

$$P(X = 174) = P(X = 175) = 0$$

  gilt

$$P(174 < X \leq 175) = P(174 \leq X \leq 175) = P(174 < X < 175)$$

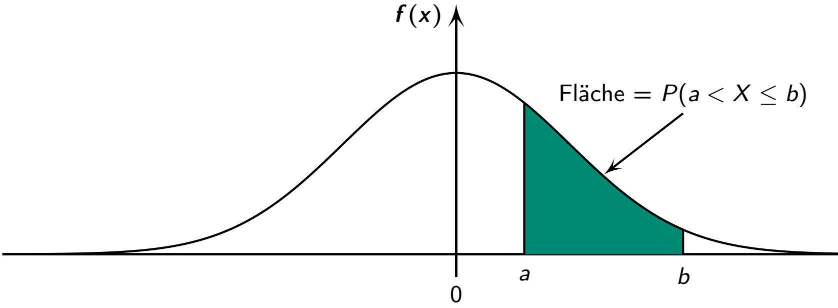
- Neuer Begriff: W'keitsdichte

Eigenschaften Wahrscheinlichkeitsdichte

Für eine W'keitsdichte $f(x)$ gelten folgende Eigenschaften:

- Es gilt:
  $$f(x) \geq 0$$
  Das heisst, die Kurve liegt überhalb der x-Achse
- W'keit
  $$P(a < X \leq b)$$
  entspricht der Fläche zwischen a und b unter $f(x)$
- Die gesamte Fläche unter der Kurve ist 1:
  - Dies ist die W'keit, dass *irgendein* Wert gemessen wird

- Skizze:



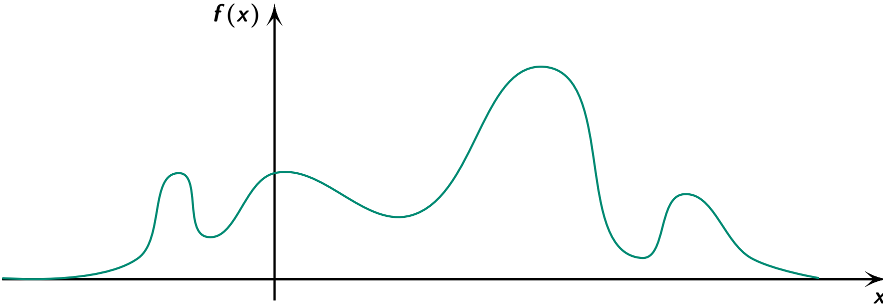
- Wichtig: Zusammenhang zwischen W'keit und Flächen:

Merkregel

Für stetige W'keitsverteilungen entsprechen W'keiten Flächen unter der Dichtefunktion.

Beispiel: Wahrscheinlichkeitsdichtefunktion

- Skizze:

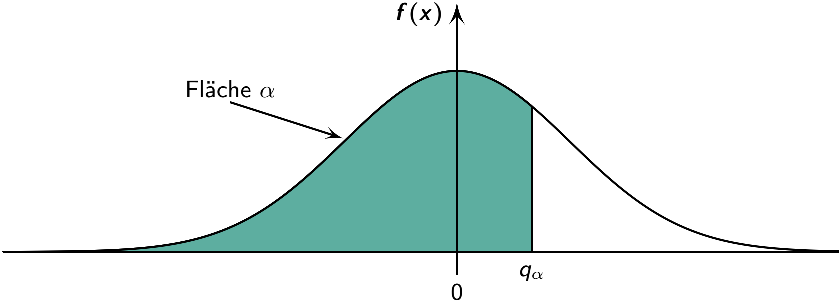


- W'keitsdichtefunktionen müssen keine „schöne" Form haben

- Normalerweise aber „schöne" Form vorhanden

Quantile

- Stetige Verteilungen: α-Quantil q_α derjenige Wert, wo die Fläche (W'keit) unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht
- 50 %-Quantil: *Median*
- Skizze:



Beispiel: Körpergrösse

- Messen wieder die Körpergrösse
- Beispiel: Für $\alpha = 0.75$ ist das zugehörige Quantil

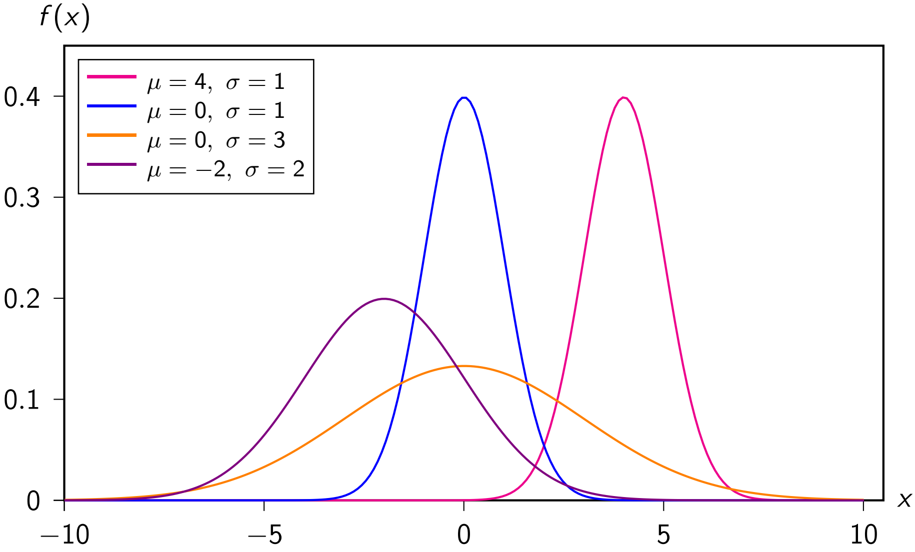
$$q_\alpha = 182.5$$

- D.h.: 75 % der gemessenen Personen kleiner oder gleich 182.5 cm

Normalverteilung (Gaussverteilung): $X \sim \mathcal{N}(\mu, \sigma^2)$

- Definition muss man einmal gesehen haben
- *Wertebereich*
$$W = (-\infty, \infty)$$
- *Dichte*
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$
- *Erwartungswert*
$$\mathrm{E}[X] = \mu$$
- *Varianz*
$$\mathrm{Var}(X) = \sigma^2$$

Normalverteilung: Illustration Dichten



Eigenschaften der Normalverteilung

- Dichtefunktionen „glockenförmig"
- Durch Parameter μ Verschiebung der Kurve:
  - Nach rechts, falls μ positiv
  - Nach links, falls μ negativ
- Durch Parameter σ wird die Kurve
  - schmal und hoch um μ, falls σ klein (nahe bei 0)
  - weit und tief um μ, falls σ gross

Beispiel mit R: Verteilung von IQ

- Anwendung: Häufigste Verteilung für Messwerte
- Beispiel: IQ Tests folgen einer Normalverteilung mit Mittelwert 100 und Standardabweichung 15
- X misst den IQ einer zufällig ausgewählten Person
- X normalverteilt mit $\mu = 100$ und $\sigma = 15$
- Notation:

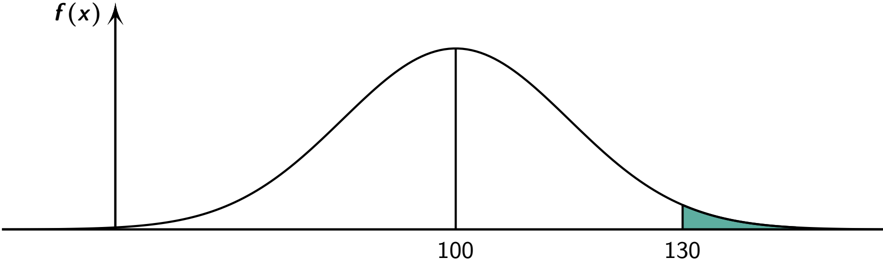
$$X \sim \mathcal{N}(100, 15^2)$$

Beispiel

- Wie gross die W'keit ist, dass jemand einen IQ von mehr als 130 hat, also als hochbegabt gilt?

- $P(X > 130)$, wobei $X \sim \mathcal{N}(100, 15^2)$

- Skizze:



- Berechnung von $P(X > 130)$ mit R-Befehl `pnorm(...)`

- Dieser berechnet die W'keit:

$$P(X \leq 130)$$

- Beachte: Richtung des Ungleichheitszeichens!

- Berechnung:

```
pnorm(q = 130, mean = 100, sd = 15)
## [1] 0.9772499
```

- Befehl `pnorm(...)` berechnet Fläche (W'keit) von $-\infty$ bis `q = 130` unter der Normalverteilungskurve mit $\mu = 100$ und $\sigma = 15$

- Dies ist aber *nicht* gesuchte W'keit $P(X > 130)$

- Aber: Gesamtfläche unter der Kurve 1 $\quad \rightarrow \quad$ Gesuchte W'keit wie folgt schreiben:

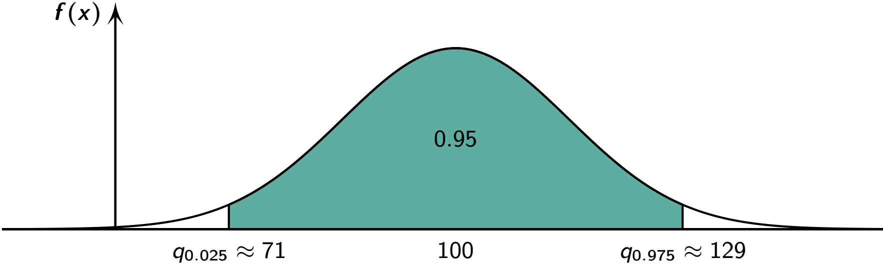
$$P(X > 130) = 1 - P(X \leq 130)$$

- Berechnung mit R:

```
1 - pnorm(q = 130, mean = 100, sd = 15)
## [1] 0.02275013
```

- Also rund 2 % der Bevölkerung ist hochbegabt

- Welches Intervall enthält 95 % der IQ's um den Mittelwert $\mu = 100$?
- W'keit als Fläche:



- Grüne Fläche: 95 % der Gesamtfläche
- Kleine weissen Flächen links und rechts: Jeweils 0.025.

- W'keiten gegeben $\rightarrow$ Suchen die zugehörigen Werte
- Bestimmung der Quartile $q_{0.025}$ und $q_{0.975}$
- R:

```r
qnorm(p = 0.025, mean = 100, sd = 15)
## [1] 70.60054
qnorm(p = 0.975, mean = 100, sd = 15)
## [1] 129.3995
```

- Oder kürzer:

```r
qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)
## [1]  70.60054 129.39946
```

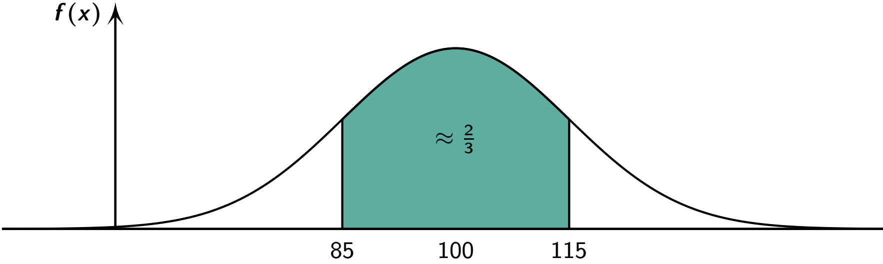
- 95 % der Menschen haben einen IQ zwischen ungefähr 70 und 130
- Entspricht Abstand von etwa 2 Standardabweichungen vom Mittelwert $\mu = 100$.

- Wieviel Prozent der Bevölkerung liegen innerhalb einer Standardabweichung vom Mittelwert liegen?

- Gesucht W'keit:

$$P(85 \leq X \leq 115)$$

- W'keit als Fläche:



- Mit R:

```
pnorm(q = 115, mean = 100, sd = 15) - pnorm(85, 100, 15)
## [1] 0.6826895
```

- D.h.: Etwa $\frac{2}{3}$ der Bevölkerung hat einen IQ zwischen 85 und 115

Normalverteilung: Eigenschaften

- Letzte Resultat aus Beispiel gilt für alle Normalverteilung $\mathcal{N}(\mu, \sigma^2)$
- Die W'keit, dass eine Beobachtung eine höchstens Standardabweichung vom Erwartungswert abweicht, ist etwa $\frac{2}{3}$:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \frac{2}{3}$$

- Normalverteilung: Konkrete Aussage für die Streuung als „mittlere" Abweichung vom Erwartungswert
- W'keit, dass eine Beobachtung höchstens zwei Standardeinheiten vom Erwartungswert abweicht:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

Normalverteilung: Eigenschaften