Hypothesentest
z-Test
t-Test

Peter Büchel

HSLU I

ASTAT: Block 09

Hypothesentest: Problemstellung, Beispiele

- Hypothesentests: Wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewisse Grösse „passt"

- Brauerei bestellt neue Abfüllmaschine für 500 ml Büchsen

- Abfüllmaschine füllt *nie genau* 500 ml ab, sondern nur *ungefähr*
  - Mal einen Tropfen mehr, mal einer weniger

- Für Brauerei wichtig, dass Abfüllmaschine möglichst genau abfüllt:
  - Füllt Maschine zuviel ab, so ist dies schlecht für Brauerei, da sie zuviel Bier für den angegebenen Preis verkauft
  - Füllt sie zuwenig ab, sind Kunden unzufrieden, da sie für den angegebenen Preis zuwenig Bier bekommen

- Herstellerfirma behauptet: Maschine füllt Büchsen normalverteilt mit $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ ab
- Brauerei macht 100 Stichproben
- Mittelwert dieser Stichproben ist 499.57 ml
- Weniger als 500 ml, aber liegt dies noch innerhalb der Angaben $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ des Herstellers der Abfüllanlage?
- Wie können wir dies überprüfen?
- Wäre Mittelwert 421.54 ml, so würden wir reklamieren
- Wo ist die Grenze zwischen „ok" und „nicht ok"?

Beispiele

- Allgemeiner: Sie stellen eine Maschine her und müssen sich auf die Angaben der Spezifikationen der Hersteller für die Bestandteile verlassen können

- Wie können wir feststellen, dass die Bestandteile die Spezifikationen auch erfüllen?

- (Fiktive) Anfrage beim Bundesamt für Statistik: Durchschnittliche Körpergrösse der erwachsenen Frauen liegt in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm

- Angabe ist gefühlsmässig wohl falsch, da viel zu hoch

- Wie können wir dies aber mathematisch überprüfen und begründen, ohne uns auf unser Gefühl zu verlassen?

Ziel

- Ziel: Standardisiertes, reproduzierbares Verfahren einzuführen, mit dem wir entscheiden können, ob der Mittelwert einer Messreihe zu einem bestimmten „wahren" Mittelwert μ passt oder nicht

- *Achtung*: Das folgende Verfahren liefert *niemals einen Beweis*, dass beispielsweise eine Grösse nicht zu einer Messreihe passt

- Können mit statistischen Mitteln nur zeigen, dass diese Grösse *mit grosser Wahrscheinlichkeit* nicht zu dieser Messreihe passt

- Lesen Sie in der Zeitung „...mit Statistik bewiesen...", ist das ein Blödsinn!

Beispiel

- Waagebeispiel von früher:

Waage A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05
Waage A	80.03	80.02	80.00	80.02					
Waage B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97	

- Messungen als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen X_i betrachten

- Zweite Messwert $x_2 = 80.04$ der Waage A eine Realisierung der Zufallsvariable X_2

Allgemein

- Betrachten Messdaten $x_1, \ldots, x_n$ als Realisierungen von

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

- Zwei Kennzahlen der Zufallsvariablen X_i sind:

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

- Typischerweise sind diese (und andere) Kennzahlen unbekannt

- Ziel: Rückschlüsse darüber aus den Daten

Schätzungen

- (Punkt-) Schätzungen für den Erwartungswert und die Varianz sind:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\widehat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

Beispiel: Waage A

- Schätzungen für den Mittelwert μ und die Varianz σ_X^2:

$$\widehat{\mu} = 80.02 \qquad \text{und} \qquad \widehat{\sigma}_X^2 = 0.024^2$$

- R:

```r
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
    79.97, 80.05, 80.03, 80.02, 80, 80.02)

mean(waageA)
## [1] 80.02077
sd(waageA)
## [1] 0.02396579
```

- Problem: für *andere* Messreihen lauten diese Schätzwerte praktisch immer anders

- Jetzt: Messreihen simulieren, die „ähnlich aussehen”, wie die Werte in Waage A
- *Annahme*: Messwerte in Waage A normalverteilt mit *wahren* Parametern:
$$\mu = 80 \qquad \text{und} \qquad \sigma_X^2 = 0.02^2$$
- Generieren mit `rnorm` Zufallszahlen, die dieser Verteilung folgen
- Wegen Übersichtlichkeit: Messreihen der Länge 6
- Runden meist auf zwei Nachkommastellen (`round(..., 2)`)
- `set.seed(...)`: bringt immer dieselben Zufallszahlen

- Code:

```r
set.seed(1)
waageA.sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02), 2)

waageA.sim1
## [1] 79.99 80.00 79.98 80.03 80.01 79.98
mean(waageA.sim1)
## [1] 79.99833
sd(waageA.sim1)
## [1] 0.0194079
```

- Geschätzte Werte $\widehat{\mu}$ und $\widehat{\sigma}^2$: (Leicht) anders, als in Beispiel vorher

- Führen dies fünfmal durch:

```
set.seed(10)
for (i in 1:5) {
    waageA.sim1 <- round(rnorm(n = 6, mean = 80, sd = 0.02),
        4)
    cat(round(mean(waageA.sim1), 2), round(sd(waageA.sim1), 4),
        "\n")
}
## 80 0.0131
## 79.99 0.0213
## 80 0.0142
## 79.99 0.0257
## 79.99 0.0087
```

- Mittelwerte sind hier alle nahe bei 80, was auch zu erwarten war
- Keine Zweifel, dass der wahre Mittelwert nicht $\mu = 80$ sein könnte
- Abweichungen sind durchaus zu erwarten

- Beispiel vorher: geschätzte Mittelwerte alle sehr nahe bei $\mu = 80$

- Allerdings sind auch folgende Fälle möglich:

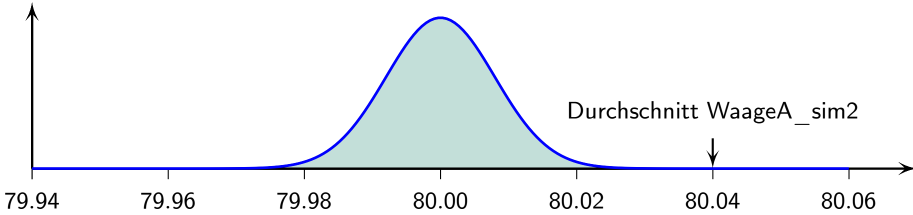
```
set.seed(1450070)
waageA.sim2 <- rnorm(n = 6, mean = 80, sd = 0.02)

waageA.sim2
## [1] 80.05403 80.03896 80.03671 80.06336 80.01052 80.04372
mean(waageA.sim2)
## [1] 80.04122
sd(waageA.sim2)
## [1] 0.01804572
```

- Mittelwert dieser Messreihe verteilt wie (ZGWS):

$$\overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right) = \mathcal{N}\left(80, 0.0082^2\right)$$

- Mittelwert Messreihe fast 5 Standardabweichungen grösser als 80
- Möglich, aber nicht sehr wahrscheinlich
- Abbildung:



- Aber was heisst hier „nicht sehr wahrscheinlich”?
- Erwarten, dass der Mittelwert in der Nähe von $\mu = 80$ liegt, sofern der wahre Mittelwert tatsächlich $\mu = 80$ ist

- Ein weiteres Beispiel:

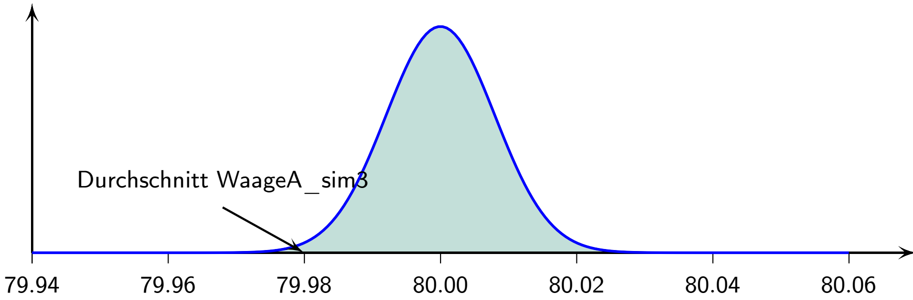
```
set.seed(384)
waageA.sim3 <- rnorm(n = 6, mean = 80, sd = 0.02)

waageA.sim3
## [1] 80.00420 79.95783 79.96086 79.95553 79.97645 79.99413
mean(waageA.sim3)
## [1] 79.97483
sd(waageA.sim3)
## [1] 0.02046691
```

- Mittelwert etwa 3 Standardabweichungen unter 80

- Dies ist zwar immer noch weit weg, aber nicht so stark wie im vorher

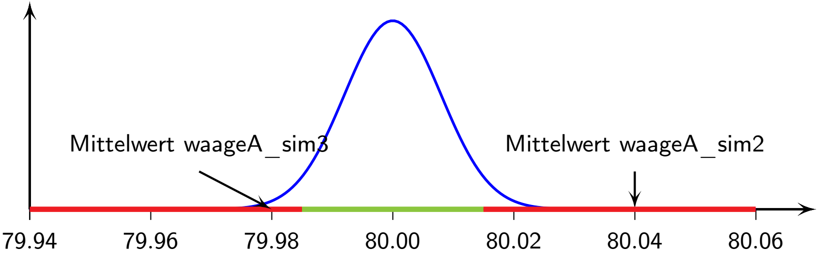
- Abbildung:



- Erwarten allerdings, dass der Durchschnitt der Messreihe in der Nähe vom wahren $\mu = 80$ liegt

- Liegt der Durchschnitt weit weg $\quad \rightarrow \quad$ beginnen zu zweifeln, ob der wahre Mittelwert tatsächlich 80 ist

- Idee: Legen Bereich fest, was „nahe bei" oder „weit entfernt von" μ ist
- Skizze:



Fragestellung

- Ist eine Messreihe mit der Annahme $\mu = 80$ noch kompatibel oder müssen an dieser Annahme zweifeln?
- Das heisst: Liegt der Mittelwert der Messreihe in der „Nähe" des wahren Mittelwertes $\mu = 80$ oder liegt er so „weit" entfernt, dass wir an der Angabe des wahren $\mu = 80$ zweifeln müssen?
- Hier stellt sich natürlich die Frage, was „nahe" heisst (gleich)
- Der wahre Mittelwert ist grundsätzlich *nicht* bekannt

Beispiel: Vorgehen Hypothesentest

- Annahme: Daten normalverteilt sind mit $\mu = 80.00$ und $\sigma = 0.02$
- Wie kann man überprüfen, ob der Mittelwert $\mu = 80$ auch realistisch ist?
- Grundidee: Mit Messreihe überprüfen, ob *unter dieser Annahme* $\mu = 80$, der Mittelwert dieser Messreihe w'lich ist oder nicht
- Wählen dazu eine Messreihe der Länge 6 aus und gehen von folgendem Modell aus:

Modell

6 Messwerte sind Realisierungen der Zufallsvariablen $X_1, X_2, \ldots X_6$, wobei X_i eine kontinuierliche Messgrösse ist. Es soll gelten:

$$X_1, \ldots, X_6 \text{ i.i.d. } \sim \mathcal{N}(80, 0.02^2)$$

- Wollen nun überprüfen, ob die *Annahme* $\mu = 80$ auch gerechfertigt ist
- Führen folgende Begriffe ein:

Nullhypothese

$$H_0 : \quad \mu = \mu_0 = 80$$

Alternativhypothese

$$H_A : \quad \mu \neq \mu_0 = 80 \quad \text{oder „}<\text{" oder „}>\text{"}$$

- Wählen Messreihe `waageA.sim3`:

```
## [1] 80.00 79.96 79.96 79.96 79.98 79.99
## Mittelwert: 79.975
## Standardabweichung: 0.01760682
```

- Geschätzter Mittelwert: $\widehat{\mu} = 79.98$

- Konkretisieren, was es heisst, dass dieser Mittelwert (un)wahrscheinlich ist

- Folgende W'keit bringt uns hier nicht weiter, da diese 0 ist:

$$P(\overline{X}_6 = 79.98) = 0$$

- Da $\widehat{\mu} < 80$ ist, betrachten folgende Wahrscheinlichkeit:

$$P(\overline{X}_6 \leq 79.98)$$

- Verteilung von $\overline{X}_6$ unter unseren Annahmen $\mu = 80$ und $\sigma = 0.02$:

$$\overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

- Testen mit dieser Verteilung, ob die Annahme $\mu = 80$ gerechtfertig ist

Teststatistik

Verteilung der Teststatistik T unter der Nullhypothese H_0:

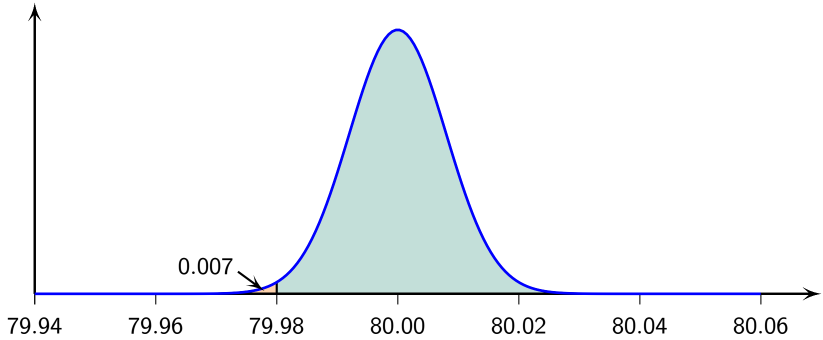
$$T = \overline{X}_6 \sim \mathcal{N}\left(80, \frac{0.02^2}{6}\right)$$

- Erhalten für die W'keit

$$P(\overline{X}_6 \leq 79.98) = 0.007$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))
## [1] 0.007152939
```

- Skizze:



- Diese W'keit ist klein: 0.7 %
- Ist sie aber *zu* klein?
- Nun kommt eine *Abmachung*: Es hat sich als praktisch erwiesen, diese Grenze, was zu klein ist und was nicht bei 2.5 % festzulegen
- Warum dies 2.5 % sind, kommt gleich

- Gemäss dieser Abmachung ist:

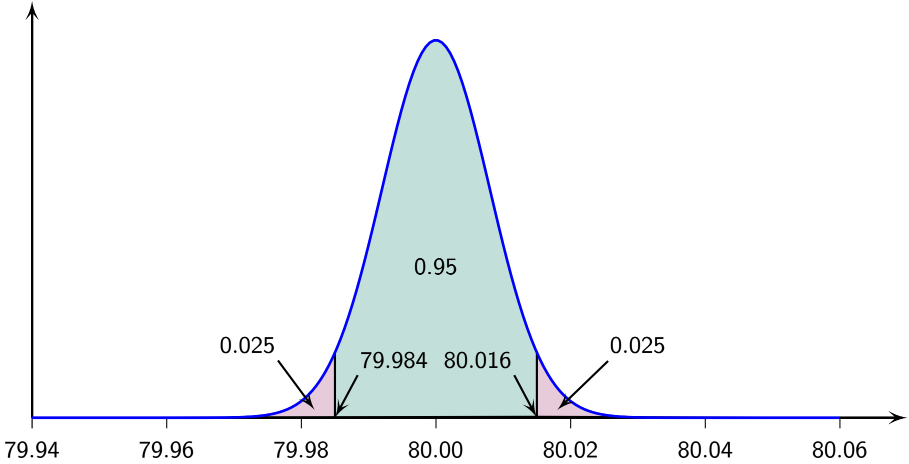
$$P(\overline{X}_6 \leq 79.98) < 0.025$$

- Geschätzer Mittelwert $\widehat{\mu} = 79.98$ zu *zu unwahrscheinlich*, als dieser zum Wert $\mu = 80$ passen könnte

- *Gehen also davon aus, dass der angegebene Mittelwert von* $\mu = 80$ *nicht stimmen kann!*

Graphische Darstellung

- Normalverteilungskurve in drei Teile auf:



- Symmetrischer Teil um Mittelwert $\mu = 80$ soll 0.95 (95 %) betragen
- Beide Teilen links und rechts müssen zusammen 0.05 ergeben
- Also ergibt sich für jeden Teil 0.025
- Grenzen entsprechen den 0.025- und 0.975-Quantilen

```
qnorm(p = c(0.025, 0.975), mean = 80, sd = 0.02/sqrt(6))
## [1] 79.984 80.016
```

- Fläche 0.05 des gesamten roten Bereiches heisst *Signifikanzniveau*

Signifikanzniveau α

Signifikanzniveau α, gibt an, wie hoch das Risiko ist, das man bereit ist einzugehen, eine falsche Entscheidung zu treffen
Für die meisten Tests wird ein α-Wert von 0.05 bzw. 0.01 verwendet. Hier

$$\alpha = 0.05$$

- Liegt der gemessene Mittelwert im roten Bereich in Abbildung, so zweifelt man an der Nullhypothese

$$H_0: \quad \mu = 80$$

- Wir sagen, wir *verwerfen* die Nullhypothese $\mu = 80$

- Bereich, wo die Nullhypothese verworfen wird, heisst deshalb

Verwerfungsbereich

$$K = (-\infty, 79.984] \cup [80.016, \infty)$$

- Gehen davon aus, dass ein Mittelwert einer Messreihe im Verwerfungsbereich so unwahrscheinlich ist, dass an der Richtigkeit von $\mu = 80$ gezweifelt wird

- Müssen annehmen, dass das wahre μ nicht 80 ist

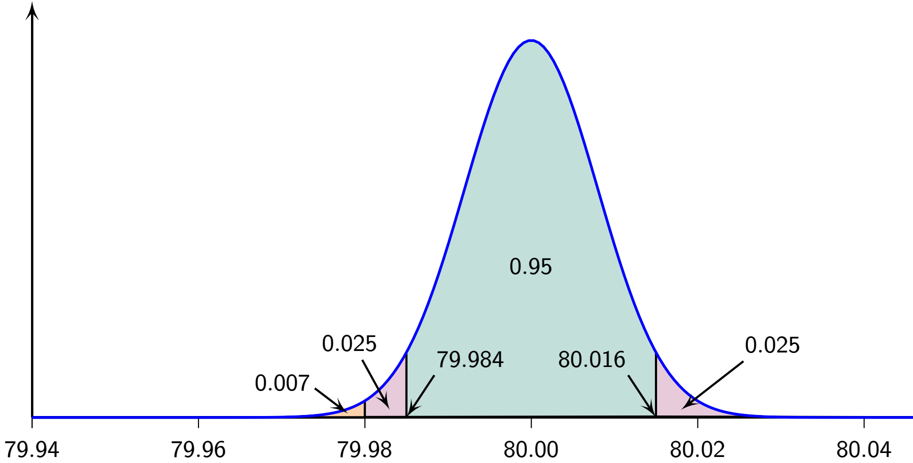
- Mit Messreihe überprüfen, ob deren Mittelwert im Verwerfungsbereich liegt oder nicht

- Machen den sogenannten

Testentscheid

- In Beispiel oben
$$\overline{X}_6 = 79.98 \in K$$
- Dieser Wert liegt im Verwerfungsbereich
- Gehen nicht vom wahren $\mu = 80$ aus, da der Mittelwert der Messreihe nicht zu diesem Parameter passt
- D.h.: Dieser Wert ist zu unwahrscheinlich, als dass $\mu = 80$ plausibel ist
- Nullhypothese wird verworfen und Alternativhypothese angenommen:
$$\mu \neq 80$$

- Abbildung:



- Wählen *andere* Messreihe: Beispiel früher

```
## [1] 80.05403 80.03896 80.03671 80.06336 80.01052 80.04372
## Mittelwert: 80.04122
```

- Modell, Nullhypothese, Alternativhypothese, Teststatistik, Signifikanzniveau und Verwerfungsbereich gleich vorher
- Nur noch den Testentscheid durchführen
- Geschätzter Mittelwert ist im Verwerfungsbereich
- Somit wird auch hier die Nullhypothese verworfen

- Es gilt für die W'keit

$$P(\overline{X}_6 > 80.04) \approx 5 \cdot 10^{-7}$$

```
1 - pnorm(q = 80.04, mean = 80, sd = 0.02/sqrt(6))
## [1] 4.816785e-07
```

- Bei weitem kleiner als 0.025
- Damit so unwahrscheinlich, dass man auch auf diese Weise $\mu = 80$ als nicht richtig annehmen (muss)
- Nullhypothese wird *verworfen*

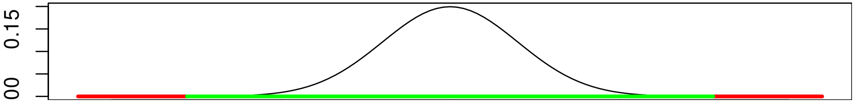
- Verwerfungsbereich: Nur Entscheidung fällen, ob der geschätzte Mittelwert im Verwerfungsbereich liegt oder nicht
- Wert von $P(\overline{X}_6 > 80.04)$ noch eine Aussage über die Sicherheit des Verwerfen
- In diesem Fall ist $5 \cdot 10^{-7}$ *sehr viel kleiner* als 0.025 und damit können wir mit grosser Sicherheit davon ausgehen, dass $\mu = 80$ *nicht* gilt
- Siehe *p*-Wert
- Aber nochmals: Messreihe stammt von der wirklichen Verteilung $\mathcal{N}(80.00, 0.02^2)$
- Allerdings ist sie so unwahrscheinlich, dass an der Annahme $\mu = 80$ gezweifelt werden muss

Bemerkungen

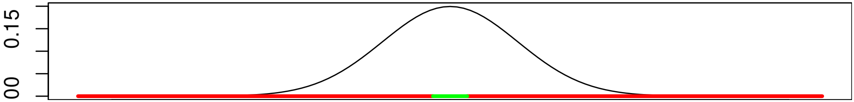
- Warum wurde Verwerfungsbereich nach oben und nach unten aufgeteilt, wenn man schon weiss, dass der gemessene Mittelwert kleiner als $\mu = 80$ ist?
- *Vor* der Messung war dies nicht bekannt
- Der gemessene Mittelwert hätte also durchaus auch grösser als $\mu = 80$ sein können
- Man spricht in diesem Fall von einem *zweiseitigen Test*
- Es gibt auch *einseitige Tests* (Beispiel gleich)
- *Annahme* hier: Gesamter Verwerfungsbereich 5 %
- Annahme hat sich als praktisch erwiesen, aber auch 1 % oft gewählt

Wahl von Signifikanzniveau α

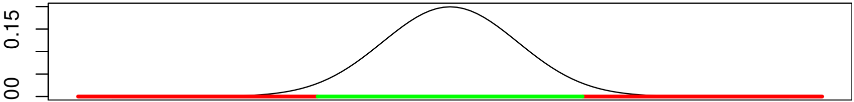
- Graphik: $\alpha = 0.0001$ (nahe bei 0)



- Graphik: $\alpha = 0.8$ (gross)



- Graphik: $\alpha = 0.05$



- Ist α sehr nahe bei null, so Bereich wo *nicht* verworfen wird (grüner Bereich) sehr gross
- D.h.: Es braucht ein sehr Ereignis bis verworfen wird
- Es wird viel zu wenig verworfen
- Im Extremfall $\alpha = 0$: Es wird gar nicht verworfen
- Für α gross: Grüner Bereich sehr klein
- D.h.: Es braucht ein sehr Ereignis bis verworfen wird
- Es wird viel zu wenig verworfen
- Im Extremfall $\alpha = 1$: Es wird immer verworfen
- $\alpha = 0.05$: Kompromiss zwischen den beiden Extremen

Beispiel: Abfüllanlage

- Testen, ob die Angabe in Beispiel Abfüllanlage mit der Testreihe konform ist
- Herstellerfirma behauptet, dass die Maschine die Büchsen normalverteilt mit $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ abfüllt
- Brauerei macht 100 Stichproben
- Mittelwert dieser Stichproben ist $499.84\,\text{ml}$
- Annahme: Messungen sind normalverteilt mit bekanntem $\sigma = 1$

- *Modell*
  X_i: Inhalt der i-ten Büchse

$$X_1, \ldots, X_{100} \text{ i.i.d. } \sim \mathcal{N}(\mu, 1^2)$$

- *Nullhypothese*

$$H_0 : \qquad \mu_0 = 500$$

- *Alternativhypothese*

$$H_A : \qquad \mu \neq \mu_0 = 500$$

- *Teststatistik mit Signifikanzniveau* $\alpha = 0.05$

$$\overline{X}_{100} \sim \mathcal{N}\left(500, \frac{1^2}{100}\right)$$

- *Verwerfungsbereich*
  Grenze des Verwerfungsbereichs:

```r
qnorm(p = c(0.025, 0.975), mean = 500, sd = 1/sqrt(100))
## [1] 499.804 500.196
```

- Also
$$K = (-\infty, 499.804) \cup (500.196, \infty)$$

- *Testentscheid* Es gilt
$$499.84 \notin K$$

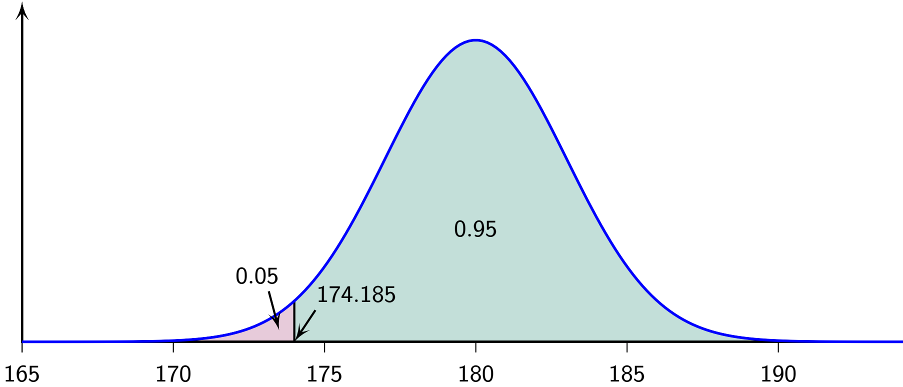
- Nullhypothese wird nicht verworfen

- Vertrauen der Angabe des Hersteller der Abfüllanlage

Beispiel: Körpergrösse Frauen

- Bundesamt für Statistik behauptet, dass die durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz bei 180 cm mit einer Standardabweichung von 10 cm liegt
- Vermutung: dieser Wert ist zu gross
- Zweiseitiger Test macht wenig Sinn, da man „weiss", dass dieser Mittelwert zu gross ist
- D.h.: der wahre Wert liegt wohl eher tiefer
- Überlegung ist an sich dieselbe wie in Beispielen

- Verwerfungsbereich nicht auf beide Seiten verteilen, sondern nur nach unten



- Erwartung, dass der wahre Mittelwert tiefer als $\mu = 180$ ist

- *Einseitiger* Test

- Wählen zufällig 8 erwachsene Frauen aus, deren durchschnittliche Körpergrösse 171.54 cm beträgt (was immer noch sehr gross ist)
- Annahme: Körpergrösse normalverteilt mit $\mathcal{N}(\mu, 10^2)$
- Annahme: Standardabweichung dieselbe, wie vom Bundesamt angegeben
- *Modell:*
  X_i: Körpergrösse der i–ten Frau. Es gilt

$$X_1, \ldots, X_8 \text{ i.i.d. } \sim \mathcal{N}(\mu, 10^2)$$

- Gehen davon aus, dass der wahre Mittelwert wirklich 180 cm ist

- *Nullhypothese*

$$H_0: \qquad \mu_0 = 180$$

- *Alternativhypothese*

$$H_A: \qquad \mu < \mu_0 = 180$$

- Testen ob jetzt der Wert

$$P(\overline{X}_8 < \overline{x}_8) < 0.05$$

  ist oder nicht

- Verwerfungsbereich ist hier also einseitig nach unten
- Abbildung oben: Verwerfungsbereich für $n = 8$ pink eingezeichnet

- *Teststatistik mit Signifikanzniveau* $\alpha = 0.05$

$$\overline{X}_8 \sim \mathcal{N}\left(180, \frac{10^2}{8}\right)$$

- Grenze des Verwerfungsbereichs:

```r
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))
## [1] 174.1846
```

- *Verwerfungsbereich*
  Der Verwerfungsbereich ist also

$$K = (-\infty, 174.185)$$

- Verwerfungsbereich ist natürlich viel zu gross, da wohl kaum Körpergrössen von erwachsenen Frauen unter 50 cm zu erwarten sind

- Arbeiten hier mit einem *Modell*, das eben nur in einem bestimmten Bereich Sinn macht

- *Testentscheid*
  Wert im Verwerfungsbereich und somit wird Nullhypothese *verworfen*, dass das wahre $\mu = 180$ gilt

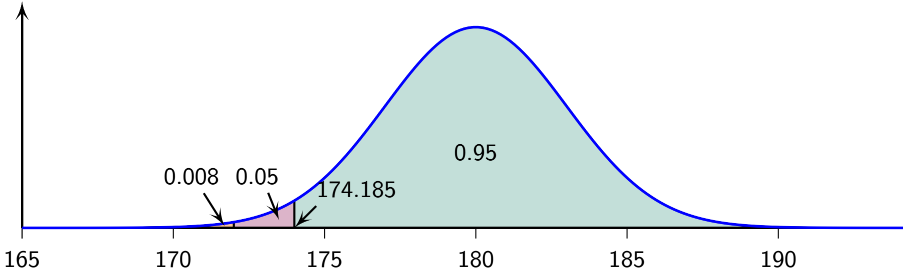
- Mittelwert der zufällig ausgewählten acht Frauen erscheint immer noch relativ hoch, aber er reicht schon, damit an der Annahme $\mu = 180$ gezweifelt wird

- Wert für $P(\overline{X}_6 < 171.54)$:

$$P(\overline{X}_6 < 171.54) = 0.008$$

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))
## [1] 0.008359052
```

- Abbildung:



- Dieser Wert heisst *p*-Wert und gibt die Sicherheit mit der man Testentscheid trifft
- Wird die Nullhypothese verworfen, so deutet ein sehr kleiner *p*-Wert darauf hin, dass die Nullhypothese sicherer verworfen wird, als wenn er in der Nähe des Signifikanzniveaus (hier $\alpha = 0.05$) liegt.

Einfluss der Anzahl Messungen auf Verwerfungsbereich

- Beispiel Waage von früher
- Messreihen verschiedener Länge n, alle mit geschätztem Mittelwert $\widehat{\mu} = 79.78$
- Bestimmen für alle Messreihen den Wert

$$P(\overline{X}_n \leq 79.98) \qquad \text{mit} \qquad \overline{X}_n \sim \mathcal{N}\left(80, \frac{0.02^2}{n}\right)$$

- Ist dieser Wert grösser als 0.025, dann wird die Nullhypothese nicht verworfen, ansonsten schon

- $n = 2$:

$$P(\overline{X}_2 \leq 79.98) = 0.079 > 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(2))
## [1] 0.0786496
```

- Die Nullhypothese wird also nicht verworfen
- Bei 2 Messwerten Abweichung vom wahren Mittelwert als zufällig möglich erachtet

- $n = 4$:

$$P(\overline{X}_4 \leq 79.98) = 0.022 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(4))
## [1] 0.02275013
```

- Hier wird die Nullhypothese (knapp) verworfen

- $n = 6$:

$$P(\overline{X}_6 \leq 79.98) = 0.007 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(6))
## [1] 0.007152939
```

- Nullhypothese wird klarer verworfen als für $n = 4$

- Und schlussendlich noch für $n = 8$:

$$P(\overline{X}_6 \leq 79.98) = 0.002 < 0.025$$

```
pnorm(q = 79.98, mean = 80, sd = 0.02/sqrt(8))
## [1] 0.002338867
```

- Die Nullhypothese wird noch klarer verworfen, als bei $n = 6$

- Mit zunehmendem n wird der Wert

$$P(\overline{X}_n \leq 79.98)$$

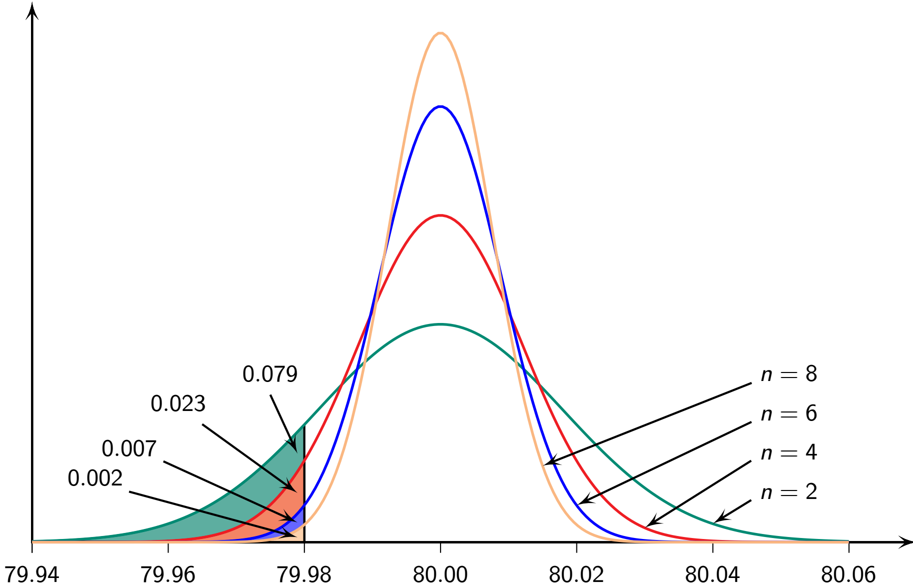
  immer kleiner

- Grund: Standardabweichung mit grösser werdendem n kleiner wird

- Normalverteilungskurven werdem schmaler

- D.h.: je mehr Messungen wir haben, umso gewichtiger ist eine Abweichung von wahren Mittelwert

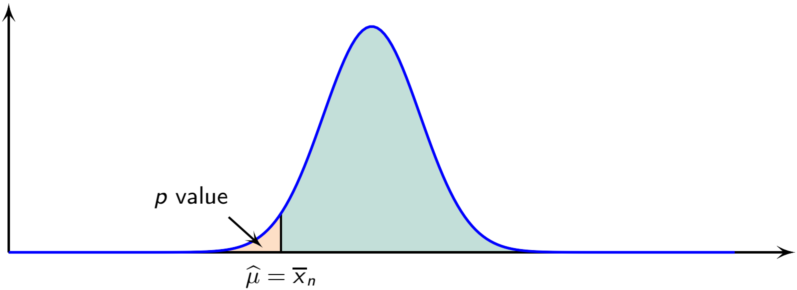
- Abbildung:


0.079
0.023
0.007
0.002
$n = 8$
$n = 6$
$n = 4$
$n = 2$
79.94
79.96
79.98
80.00
80.02
80.04
80.06

p-Wert

- p-Wert ist ein Wert zwischen 0 und 1, der angibt, wie gut *Nullhypothese* und *Daten* zusammenpassen
  - 0: passt gar nicht
  - 1: passt sehr gut

- p-Wert ist die W'keit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein *extremeres* zu erhalten



- Mit dem p-Wert wird also angedeutet, wie extrem das Ergebnis ist
- Je kleiner der p-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese
- Werte kleiner als eine im voraus festgesetzte Grenze, wie 5 %, 1 % oder 0.1 % sind Anlass, die Nullhypothese abzulehnen

p-Wert

Der *P-Wert* ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete.

- Testentscheid auch mit Hilfe des p-Wertes durchführen

p-Wert und Statistischer Test

Bei einem vorgegebenen Signifikanzniveau α (z.B. $\alpha = 0.05$) gilt aufgrund der Definition des p-Werts für einen einseitigen Test:

- Verwerfe H_0 falls p-Wert $\leq \alpha$
- Belasse H_0 falls p-Wert $> \alpha$

- Viele Computer-Pakete liefern den Testentscheid nur mit p-Wert

- *Wie signifikant?*

$$
\begin{aligned}
p\text{-Wert} &\approx 0.05 &&: \text{ schwach signifikant, “.“} \\
p\text{-Wert} &\approx 0.01 &&: \text{ signifikant, “} * \text{“} \\
p\text{-Wert} &\approx 0.001 &&: \text{ stark signifikant, “} ** \text{“} \\
p\text{-Wert} &\leq 10^{-4} &&: \text{ äusserst signifikant, “} *** \text{“}
\end{aligned}
$$

p-Wert für zweiseitigen Test

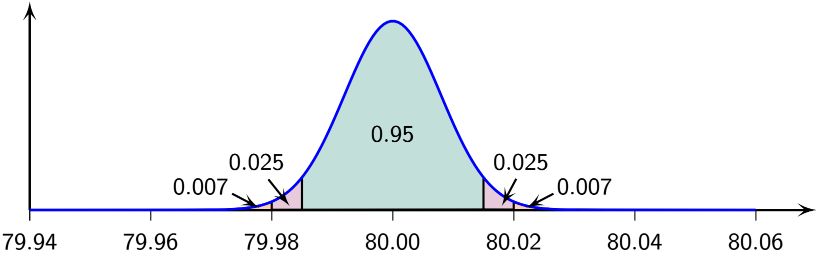
- Haben den p-Wert für einseitige Tests definiert
- Wie sieht nun aber der p-Wert für zweiseitige Tests aus?
- Beispiel von früher:

$$P(\overline{X}_6 \leq 79.98) = 0.007$$

  der kleiner ist als 0.025

- Könnten dies als p-Wert betrachten, tun es aber nicht

- Skizze:



- Da aber das Signifikanzniveau auf $\alpha = 0.05$ liegt, wird die W'keit oben auf 5 % umgerechnet, also verdoppelt:

$$p\text{-Wert} = 2 \cdot P(\overline{X}_6 \leq 79.98) = 0.014$$

- Dieser p-Wert dann mit dem Signifikanzniveau vergleichen

- Computersoftware gibt den p-Wert *immer* auf Signifikanzniveau an

t-Test

- Bisher: Verfahren heisst z-Test
- Stillschweigend vorausgesetzt: Standardabweichung *bekannt*
- Praxis: Praktisch nie der Fall
- Folgender *t-Test*: Setzt keine Standardabweichung voraus
- Darum: t-Test viel wichtiger als z-Test
- Vorgehen sehr ähnlich z-Test $\quad\rightarrow\quad$ Nur andere Verteilung
- Wie vorher Annahme: Daten Realisierungen von

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

- Praxis: Annahme, dass σ_X bekannt ist, meist unrealistisch

- Können aber σ_X aus Daten schätzen $\quad\rightarrow\quad \widehat{\sigma}_X^2$

$$\widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Zusätzliche Unsicherheit $\quad\rightarrow\quad$ Verteilung der Teststatistik ändern

t-Verteilung

Die Verteilung der Teststatistik beim t-Test unter der Nullhypothese

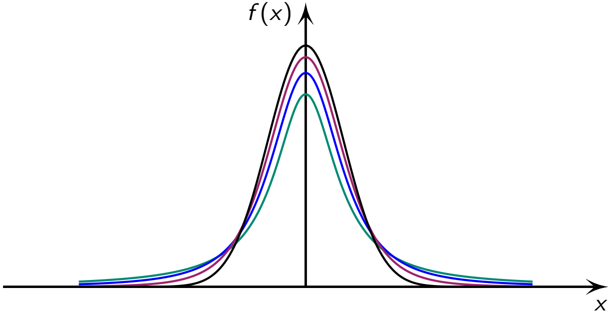
$$H_0: \quad \mu = \mu_0$$

ist gegeben durch

$$T = \overline{X}_n \sim t_{n-1}\left(\mu, \frac{\widehat{\sigma}_X}{\sqrt{n}}\right)$$

wobei t_{n-1} eine t-*Verteilung* mit $n-1$ Freiheitsgraden ist

- Normalverteilung wird also durch eine t-Verteilung ersetzt
- Was aber ist eine t-Verteilung?
- Ähnlich Normalverteilung, aber flacher, wegen grösserer Unsicherheit
- Hängt von der Anzahl Beobachtungen
- Skizze für $\mu = 0$ und $\sigma \approx 1$ (hängt von n ab):



- Grün: $n = 1$, blau: $n = 2$, violet: $n = 5$, schwarz: $\mathcal{N}(0, 1)$
- t_n-Verteilung symmetrische Verteilung um 0, aber langschwänziger ist Standardnormalverteilung $\mathcal{N}(0, 1)$
- Für grosse n ist t_n ähnlich zu $\mathcal{N}(0, 1)$
- t_n strebt für $n \to \infty$ gegen Standardnormalverteilung $\mathcal{N}(0, 1)$
- Wichtig: Für t-Test t_{n-1} verwenden
- t-Verteilung wurde von William Gosset (Chefbrauer Guiness Brauerei) 1908 gefunden

`R`

- Alle Begriffe vom *z*-Test können für den *t*-Test übernehmen
- Verwerfungsbereich mit t `qt(...)` anstatt `qnorm(...)`
- *p*-Wert mit `pt(...)` anstatt `pnorm(...)`
- *t*-Test kommt sehr oft vor: Ganzes Verfahren in `R` implementiert
- Daten in Befehl `t.test(...)` eingeben und `R` übernimmt Arbeit
- Verwerfungsbereich nicht *nicht* ausgegeben
- Aber *p*-Wert wird ausgegeben, reicht für Testentscheid

Beispiel

- Datensatz aus normalverteilten Datenpunkten $x_1, \ldots, x_{20}$

5.9	3.4	6.6	6.3	4.2	2.0	6.0	4.8	4.2	2.1
8.7	4.4	5.1	2.7	8.5	5.8	4.9	5.3	5.5	7.9

- Vermutung: $x_1, x_2, \ldots, x_{20}$ Realisierungen von

$$X_i \sim \mathcal{N}(5, \sigma_X^2)$$

- σ_X unbekannt $\quad \rightarrow \quad$ σ_X also aus Daten schätzen

```
x <- c(5.9, 3.4, 6.6, 6.3, 4.2, 2, 6, 4.8, 4.2, 2.1, 8.7, 4.4,
    5.1, 2.7, 8.5, 5.8, 4.9, 5.3, 5.5, 7.9)

mean(x)
## [1] 5.215
sd(x)
## [1] 1.883802
```

- Nullhypothese lautet in diesem Fall:

$$H_0 : \quad \mu_0 = 5$$

- Test, ob Mittelwert 5.215 zum vermuteten Wert μ_0 passt oder nicht

- Befehl `t.test(...)`:

```
t.test(x, mu = 5)
##
##  One Sample t-test
##
## data:  x
## t = 0.51041, df = 19, p-value = 0.6156
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.333353 6.096647
## sample estimates:
## mean of x
##     5.215
```

Zum R-Output

- `One Sample t-test`
  Es wird ein Einstichprobentest gemacht (Zweistichproben nächste Woche)

- `data: x`
  Datensatz, der verwendet wurde

- `t = 0.51041`
  - *t*-Wert
  - Dieser ist an sich uninteressant
  - „Grosser" *t*-Wert: Nullhypothese wird verworfen
  - *t*-Wert „nahe" bei 0: Nullhypothese wird *nicht* verworfen
  - Entscheidender ist der *P*-Wert weiter unten

- `df = 19`
  Freiheitsgrad (degree of freedom): Auch uninteressant

- `p-value = 0.6156`
  - *p*-Wert
  - Dies ist *der* entscheidende Wert
  - Entscheidet, ob die Nullhypothese verworfen wird oder nicht
  - Hier: Nullhypothese auf Signifikanzniveau 5 % nicht verwerfen, da *p*-Wert grösser als 0.05
- `alternative hypothesis: true mean is not equal to 5`
  Hier wird die Alternativhypothese aufgeführt
- `95 percent confidence interval: 4.33 6.09`
  Vertrauensintervall (wird gleich eingeführt)
- `mean of x 5.215`
  Mittelwert von `x`

Beispiel: Waage A

- Schätzen Standardabweichung σ_X aus den Daten
- Behauptung: Wahres $\mu = 80$
- t-Test auf dem 5 % Signifikanzniveau
- t-Test

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
    79.97, 80.05, 80.03, 80.02, 80, 80.02)

t.test(waageA, mu = 80)
##
##  One Sample t-test
##
## data:  waageA
## t = 3.1246, df = 12, p-value = 0.008779
## alternative hypothesis: true mean is not equal to 80
## 95 percent confidence interval:
##  80.00629 80.03525
## sample estimates:
## mean of x
##  80.02077
```

- p-Wert: 0.009
- Dieser Wert kleiner als Signifikanzniveau 0.05
- Nullhypothese H_0 wird verworfen
- Müssen davon ausgehen, dass der wahre Mittelwert statistisch signifikant *nicht* 80 ist

Beispiel: Körpergrösse Frauen

- Bundesamtes für Statistik: Durchschnittliche Körpergrösse der erwachsenen Frauen in der Schweiz ist 180 cm
- Vermutung: Wert zu gross
- Auf einem Signifikanzniveau von 5 % untersuchen
- Wählen zufällig 10 Frauen aus und messen deren Körpergrösse (in cm)
- Gemessene Grössen:

  165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9, 170.4, 163.4, 152.5

- Vermutung: Durchschnittliche Körpergrösse *kleiner* als 180 cm

- *t*-Test nach unten:

```
groesse <- c(165.7, 156.7, 171.7, 180.3, 163.2, 166.7, 149.9,
    170.4, 163.4, 152.5)

t.test(groesse, mu = 180, alternative = "less")
##
##  One Sample t-test
##
## data:  groesse
## t = -5.4836, df = 9, p-value = 0.0001942
## alternative hypothesis: true mean is less than 180
## 95 percent confidence interval:
##      -Inf 169.382
## sample estimates:
## mean of x
##    164.05
```

- p-Wert: 0.0002, also weit unter dem Signifikanzniveau von 0.05
- Nullhypothese
$$H_0 : \quad \mu_0 = 180$$
verwerfen
- Alternativhypothese
$$H_A : \quad \mu_0 < 180$$
annehmen
- Aussage des Bundesamtes für Statistik stimmt also statistisch signifikant (sehr wahrscheinlich) *nicht*