Hypothesentest
z-Test
t-Test

Peter Büchel

HSLU I

Stat: Block 06

Hypothesentest: 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* 500 ml
  - Mal einen Tropfen mehr, mal einer weniger
- Brauerei interessiert, dass die Abfüllmaschine möglichst genau abfüllt:
  - Füllt die Maschine zuviel ab, so ist dies schlecht für die Brauerei, da sie zuviel Bier für denselben Preis verkauft
  - Füllt sie zuwenig ab, sind die Kunden und der Konsumentenschutz unzufrieden, da sie für den entsprechenden Preis zuwenig Bier bekommen

- Herstellerfirma behauptet, dass Maschine die Büchsen normalverteilt mit $\mu = 500\,\text{ml}$ und $\sigma = 1\,\text{ml}$ abfüllt
- 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, wo 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 kommende Verfahren liefert *niemals einen Beweis*, dass beispielsweise eine Grösse nicht zu einer Messreihe passt

- Wir 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!

Problemstellung

- Vorgehen beim Hypothesentest wird durch folgendes Beispiel erklärt
- *Datensatz*: Methode zur Bestimmung der Schmelzwärme von Eis
- Wiederholte Messungen der freigesetzten Wärme beim Übergang von Eis bei $-0.7\,^\circ$C zu Wasser bei $0\,^\circ$C ergaben die Werte (in cal/g):

Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04
Methode A	79.97	80.05	80.03	80.02	80.00	80.02	

- Messungen als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen X_i betrachten
- Beispiel: 2. Messwert $x_2 = 80.04$ Realisierung der Zufallsvariable X_2

Allgemein

- Messdaten $x_1, \dots, x_n$ als Realisierungen von

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

- Zwei Kennzahlen der Zufallsvariablen X_i sind:

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

- Normalfall: Kennzahlen *unbekannt*
- Rückschlüsse darüber aus Daten wie in Beispiel Methode A machen
- Nähern $\mathrm{E}(X_i)$ und σ_X^2 (wahre, aber unbekannte Werte) durch Mittelwert und Varianz der gegebenen Daten an
- Wir sprechen von einer *Schätzung* $\widehat{\mu}$ von $\mathrm{E}(X_i)$

- Analog: Schätzung $\widehat{\sigma}_X^2$ von σ_X^2

- Notation: Dach $\widehat{}$ bezeichnet Schätzung einer Grösse

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

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

- Schätzer hier als Funktionen der Zufallsvariablen $X_1, \ldots, X_n$

- $\widehat{\mu}$ und $\widehat{\sigma}_X^2$ selbst wieder Zufallsvariablen mit Verteilungseigenschaften von $\widehat{\mu}$

- Hier: Schätzungen für das Beispiel der Methode A

Beispiel: Methode A

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

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

- Berechnung mit **Python**:

```
from pandas import Series
import numpy as np

methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

methodeA.mean()
methodeA.std()
## 80.02076923076923
## 0.023965787580611863
```

- Problem: *Andere* Messreihen haben andere Schätzwerte

Simulation von neuen Messreihen

- Simulieren neue Messreihen, mit „ähnlichen" Werten wie Methode A
- *Annahme*: Messwerte in Methode A normalverteilt mit

$$\mu = 80 \qquad \text{und} \qquad \sigma_X^2 = 0.02^2$$

- Generierung von Zufallszahlen mit `norm.rvs()` aus `Scipy Stats`, die dieser Verteilung folgen
- Hier: Messreihen der Länge 6
- Rundung der meisten Resultate meist auf zwei Nachkommastellen: `np.round(...,2)`
- Mit `np.random.seed(...)` werden Zufallszahlen festgelegt
  - Es werden immer die gleichen Zufallszahlen erzeugt

- Python-Code

```
from scipy.stats import norm
np.random.seed(1)

methodeA_sim1 = Series(np.round(norm.rvs(size=6, loc=80, scale=
0.02),2))

methodeA_sim1

methodeA_sim1.mean()
methodeA_sim1.std()
```

```
## 0    80.03
## 1    79.99
## 2    79.99
## 3    79.98
## 4    80.02
## 5    79.95
## dtype: float64
## 79.99333333333333
## 0.028751811537128993
```

- Geschätzten Werte $\widehat{\mu}$ und $\widehat{\sigma}^2$ (leicht) anders als vorher

- Fünf Simulationen mit Mittelwerten nahe bei $\mu = 80$

```
np.random.seed(17)
for i in range(5):
    methodeA_sim1 = Series(np.round(norm.rvs(size=6, loc=80, scale=
0.02),2))
    print("Mittelwert:", np.round(methodeA_sim1.mean(), 3))
    print("Standardabw.:", np.round(methodeA_sim1.std(), 3))
    print()
```

```
## Mittelwert: 80.01
## Standardabw.: 0.027
##
## Mittelwert: 80.007
## Standardabw.: 0.02
##
## Mittelwert: 79.992
## Standardabw.: 0.028
##
## Mittelwert: 79.995
## Standardabw.: 0.016
##
## Mittelwert: 79.992
## Standardabw.: 0.013
```

- Mittelwert kann auch „weit" von $\mu = 80$ entfernt liegen:

```
np.random.seed(463137)

methodeA_sim2 = Series(np.round(norm.rvs(size=6, loc=80, scale=
0.02),2))

methodeA_sim2

methodeA_sim2.mean()
methodeA_sim2.std()
```

```
## 0    80.07
## 1    80.06
## 2    80.03
## 3    80.03
## 4    80.02
## 5    80.03
## dtype: float64
## 80.04
## 0.019999999999998862
```

- Werte hier müssen „weit" über 80 liegen: Nicht sehr wahrscheinlich, aber möglich

- Aber was heisst hier „nicht sehr wahrscheinlich”?
- Verteilung des Mittelwertes dieser Messreihe:

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

- W'keit, einen Mittelwert von *80.04 oder höher* zu erhalten ist $5 \cdot 10^{-7}$:

$$P(\overline{X}_6) \geq 80.04$$

```
1 - norm.cdf(x=80.04, loc=80, scale=0.02/np.sqrt(6))
## 4.816785043049165e-07
```

- Punktw'keit $P(\overline{X}_6 = 80.04)$ ist 0 und bringt hier nichts
- Obwohl Daten zufällig aus der Verteilung $\mathcal{N}(80, 0.02^2)$ stammen, ist diese Wahrscheinlichkeit sehr klein
- Wir können daran zweifeln, ob der wahre Mittelwert wirklich 80 ist

- Ein weiteres Beispiel:

```
np.random.seed(647)
methodeA_sim3 = Series(np.round(np.random.normal(size=6, loc=80,
scale= 0.02),2))

methodeA_sim3

methodeA_sim3.mean()
methodeA_sim3.std()
```

```
## 0    79.98
## 1    79.99
## 2    80.00
## 3    79.93
## 4    80.00
## 5    79.98
## dtype: float64
## 79.98
## 0.026076809620810759
```

- Mittelwert *unter* 80

- W'keit, dass ein Durchschnit 79.98 *oder kleiner* ist:

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(6))
## 0.007152939217724509
```

- Nicht so krass wie vorher

- Möglich, aber doch wohl eher unwahrscheinlich, wenn wir $\mu = 80$ annehmen

- Frage: Welche Mittelwerte sind noch tolerierbar oder bei welchen Werten beginnen wir am wahren (unbekannten) Wert $\mu = 80$ zu zweifeln

Fragestellungen:

- 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?
- Frage: Was heisst „nahe" oder „weit"?
- Beachte: Der wahre Mittelwert ist grundsätzlich *nicht* bekannt
- Verwenden Hypothesentest

Vorgehen Hypothesentest

- Annahme: Daten normalverteilt sind mit $\mu = 80.00$ und $\sigma = 0.02$
- Wie können wir überprüfen, ob der Mittelwert $\mu = 80$ auch stimmt?
- Grundidee: Mit einer Messreihe überprüfen, ob unter dieser Annahme $\mu = 80$, die Messreihe wahrscheinlich ist oder nicht
- Wählen dazu eine Messreihe der Länge 6 aus mit Modell

Modell

Die 6 Messwerte sind Realisierungen der ZV $X_1, X_2, \ldots, X_6$, 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 überprüfen, ob *Annahme* $\mu = 80$ auch gerechfertigt ist
- Dazu führen wir 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:

```
## 0    79.98
## 1    79.99
## 2    80.00
## 3    79.93
## 4    80.00
## 5    79.98
## dtype: float64
## Mittelwert: 79.98
```

- (Geschätzter) Mittelwert ist $\widehat{\mu} = 79.98$

- Konkret: Was heisst es, dass dieser Mittelwert (un)wahrscheinlich ist?

- W'keit

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

  bringt uns hier nicht weiter, da diese 0 ist

- Da $\widehat{\mu} < 80$ ist, können wir folgende W'keit betrachten:

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

- Unter Annahmen $\mu = 80$ und $\sigma = 0.02$ ist $\overline{X}_6$ verteilt wie:

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

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

Teststatistik

Verteilung der Teststatistik T unter der Nullhypothese H_0:

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

- W'keit:

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(6))
## 0.007152939217724509
```

- Diese W'keit ist klein, 0.7 %

- Ist sie aber *zu* klein?

- *Abmachung*: Es hat sich als praktisch erwiesen, diese Grenze, was zu klein ist und was nicht bei 2.5 % festzulegen

- Gemäss dieser Abmachung ist

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

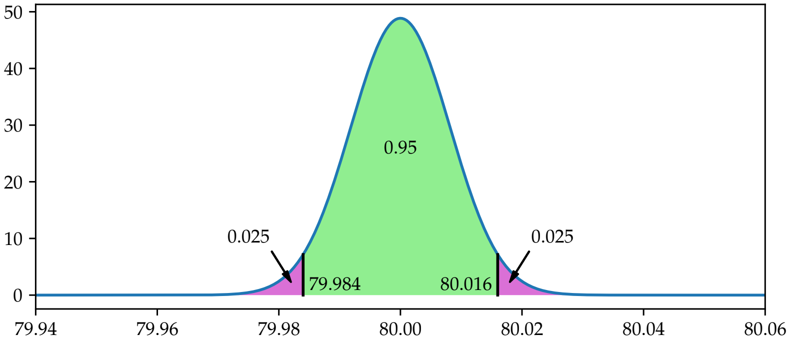
- Betrachten diesen geschätzen Mittelwert $\widehat{\mu} = 79.98$ als *zu unwahrscheinlich*, als dieser zum Wert $\mu = 80$ passen könnte

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

- Wir sagen: „Wir verwerfen die Nullhypothese und nehmen die Alternativhypothese an!"

Graphische Darstellung

- Normalverteilungskurve in drei Teile aufteilen:
  - Symmetrischer Teil um Mittelwert $\mu = 80$ soll 0.95, also 95 % betragen
  - Die beiden Teilen links und rechts müssen zusammen 0.05 ergeben
  - Also ergibt sich für jeden Teil 0.025
  - Abbildung:



- Begriff:

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: α-Wert von 0.05 bzw. 0.01
- Verwenden hier

$$\alpha = 0.05$$

- Signifikanzniveau legt roten Bereich in Abbildung vorher fest
- Der rote Bereich in der Abbildung heisst *Verwerfungsbereich*

- Grenzen Verwerfungsbereich; 0.025- und 0.975-Quantilen:

```
norm.ppf(q=0.025, loc=80, scale=0.02/np.sqrt(6))
norm.ppf(q=0.975, loc=80, scale=0.02/np.sqrt(6))
## 79.98399696107882
## 80.01600303892118
```

oder

```
norm.ppf(q=[0.025, 0.975], loc=80, scale=0.02/np.sqrt(6))
## [79.98399696 80.01600304]
```

- Liegt der gemessene Mittelwert im roten Bereich in Abbildung, so verwerfen wir die Nullhypothese, dass $\mu = 80$

- Dieser Bereich heisst

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 werden muss

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

- Machen den sogenannten

Testentscheid

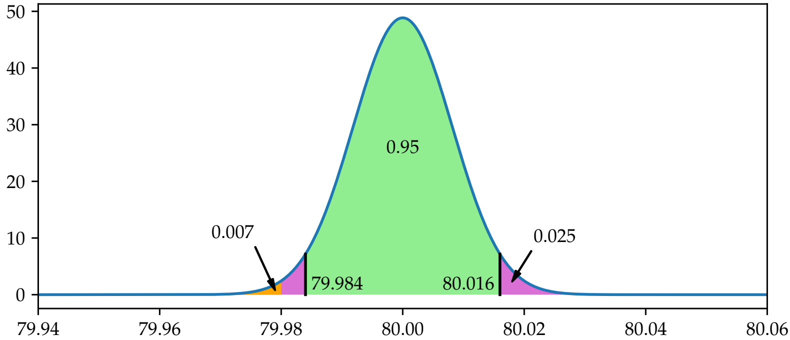
In unserem Beispiel hatten wir

$$\overline{X}_6 = 79.98 \in K$$

- Dieser Wert liegt im Verwerfungsbereich

- Wir verwerfen die Nullhypothese und nehmen Alternativhypothese an

- Graphisch:


50
40
30
20
10
0
0.95
0.007
0.025
79.984
80.016
79.94
79.96
79.98
80.00
80.02
80.04
80.06

Bemerkungen

- Warum haben wir hier den Verwerfungsbereich nach oben und nach unten aufgeteilt, wenn wir schon wissen, dass der gemessene Mittelwert kleiner als $\mu = 80$ ist?
- Nun, das wussten wir *vor* der Messung nicht
- Der gemessene Mittelwert hätte also auch grösser als $\mu = 80$ sein können (siehe Beispiel nachher)
- Wir sprechen in diesem Fall von einem *zweiseitigen Test*
- Es gibt auch *einseitige Tests* (siehe Beispiel später)

- Wir haben hier eine *Annahme* gemacht, dass der gesamte Verwerfungsbereich 5 % (Signifikanzniveau 5 %) betragen soll

- Diese Annahme hat sich als praktisch erwiesen, aber wir hätten auch 1 % wählen können, was auch ab und zu gemacht wird

- In Beispiel oben folgt die zufällige Messreihe der Normalverteilung $\mathcal{N}(80, 0.02^2/6)$ und doch wird der Parameter $\mu = 80$ hier als unwahrscheinlich verworfen

- Dies heisst, wir haben hier einen *Fehler* gemacht

- Auf diese Problematik gehen wir später ein

- Wählen als *andere* Messreihe mit Modell, Nullhypothese, Alternativhypothese, Teststatistik, Signifikanzniveau und Verwerfungsbereich gleich wie im Beispiel vorher

- Müssen also nur noch den Testentscheid durchführen:

```
## 0    80.07
## 1    80.06
## 2    80.03
## 3    80.03
## 4    80.02
## 5    80.03
## dtype: float64
## Mittelwert: 80.04
```

- Geschätzte Mittelwert ist im Verwerfungsbereich und somit wird auch hier die Nullhypothese verworfen.

- W'keit:

$$P(\overline{X}_6 > 80.04)$$

```
1 - norm.cdf(x=80.04, loc=80, scale=0.02/np.sqrt(6))
## 4.816785043049165e-07
```

- Bei weitem kleiner als 0.025 und damit so unwahrscheinlich, dass wir auch auf diese Weise $\mu = 80$ als nicht richtig annehmen (müssen)
- Wir *verwerfen* die Nullhypothese
- Verwerfungsbereich: Nur Entscheidung möglich, ob der geschätzte Mittelwert im Verwerfungsbereich liegt oder nicht
- Wert von $P(\overline{X}_6 > 80.04)$ macht noch Aussage über die Sicherheit des Verwerfen
- Hier: $5 \cdot 10^{-7}$ *sehr viel kleiner* als 0.025 und damit können wir mit grosser Sicherheit davon ausgehen, dass $\mu = 80$ *nicht* gilt

Grössere Messreihen

- Beispiel vorher: Grösse $\mu = 80$ verworfen, da die Messreihe einen zu tiefen und dann einen (viel) zu grossen Mittelwert lieferte
- Wie sieht es nun aber aus, wenn wir eine neue Messreihe bilden, die aus *beiden* Messreihen besteht?
- Oder anders gefragt: Welchen Einfluss hat die Anzahl der Messungen auf den Verwerfungsbereich?
- Wählen Messreihen verschiedener Länge n, die alle den geschätzten Mittelwert $\widehat{\mu} = 79.78$ haben

- Dann bestimmen wir für alle Messreihen den Wert

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

mit

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

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

- Für $n = 2$ erhalten wir folgenden Wert für

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(2))
## 0.07864960352518385
```

- Nullhypothese wird also auf Signifikanzniveau 5 % nicht verworfen

- Für $n = 4$ erhalten wir

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(4))
## 0.022750131948200674
```

- Hier wird die Nullhypothese (knapp) verworfen

- Für $n = 6$ erhalten wir

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(6))
## 0.007152939217724509
```

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

- Und schlussendlich noch für $n = 8$:

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

```
norm.cdf(x=79.98, loc=80, scale=0.02/np.sqrt(8))
## 0.0023388674905277422
```

- Die Nullhypothese wird noch klarer verworfen, als bei $n = 8$

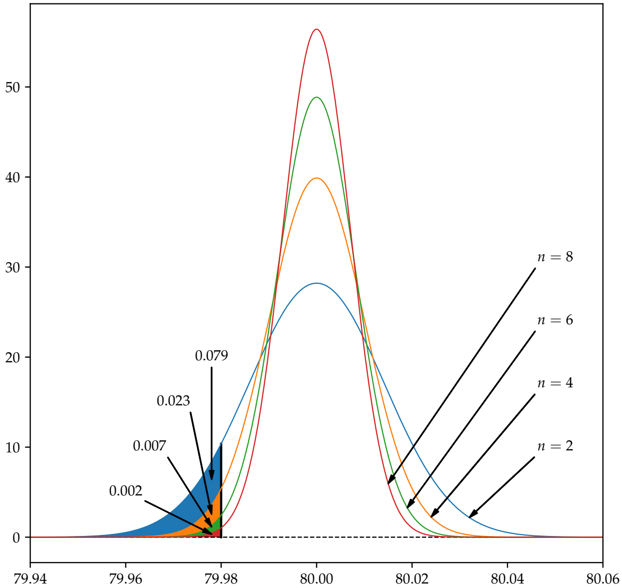
- Mit zunehmendem n wird der Wert

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

  immer kleiner

- Dies liegt daran, dass die Standardabweichung mit grösser werdendem n kleiner wird und damit werden die Normalverteilungskurven schmaler (Abbildung nächste Folie).

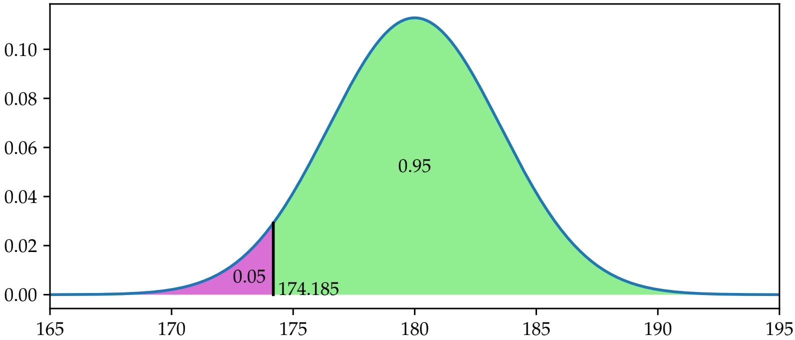
- Abbildung:


0.079
0.023
0.007
0.002
$n = 8$
$n = 6$
$n = 4$
$n = 2$
50
40
30
20
10
0
79.94
79.96
79.98
80.00
80.02
80.04
80.06

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 Mittelwert ist zu gross ist
- Hier macht ein zweiseitiger Test wenig Sinn, da wir „wissen", dass dieser Mittelwert zu gross ist
- Das heisst, der wahre Wert liegt wohl eher tiefer
- Überlegung ähnlich vorher, aber Verwerfungsbereich nicht auf beide Seiten verteilen, sondern nur nach unten, da wir erwarten, dass der wahre Mittelwert tiefer als $\mu = 180$ ist (Abbildung nächste Folie)
- Machen einen *einseitigen* Test

- Abbildung:


0.10
0.08
0.06
0.04
0.02
0.00
0.95
0.05
174.185
165
170
175
180
185
190
195

- *Modell*:

$$X_1, \ldots, X_n \text{ i.i.d.} \qquad X_i \sim \mathcal{N}(180, 10^2)$$

- Annahme: Der wahre Mittelwert ist 180 cm

- *Nullhypothese:*

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

- *Alternativhypothese*:

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

- Untersuchung unter n Personen und testen ob jetzt der Wert

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

  ist oder nicht

- Der Verwerfungsbereich ist hier also einseitig nach unten

- In Abbildung vorher ist der Verwerfungsbereich für $n = 8$ eingezeichnet pink eingezeichnet

- *Teststatistik unter der Nullhypothese H_0:*

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

- *Signifikanzniveau:*

$$\alpha = 0.05$$

- Grenze des Verwerfungsbereichs:

```
norm.ppf(q=0.05, loc=180, scale=10/np.sqrt(8))
## 174.18456423161663
```

- *Verwerfungsbereich* (siehe Abbildung vorher):

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

- Dieser Verwerfungsbereich 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

- Wählen nun zufällig acht erwachsene Frauen aus, messen deren Körpergrösse und bestimmen den Mittelwert, der bei 172 cm liegt

- *Testentscheid:*
  So ist Wert im Verwerfungsbereich und somit *verwerfen* wir die Nullhypothese, dass das wahre $\mu = 180$ gilt

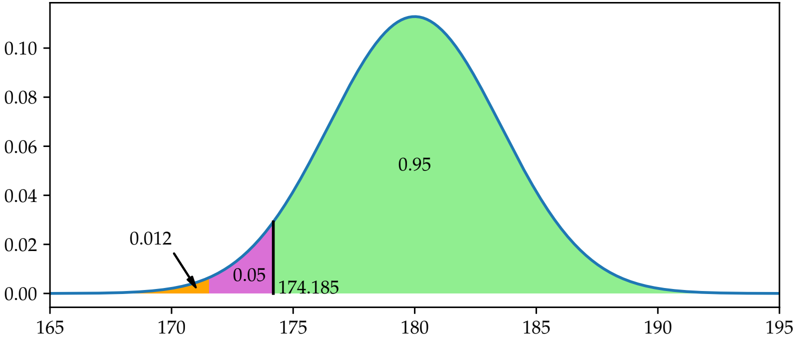
- Dieser Mittelwert der zufällig ausgewählten acht Frauen erscheint immer noch relativ hoch, aber er reicht schon, damit wir an der Annahme $\mu = 180$ zweifeln müssen

- Wert für $P(\overline{X}_6 < 172)$ ist (siehe Abbildung unten).

$$P(\overline{X}_6 < 172) = 0.012$$

```
norm.cdf(x=172, loc=180, scale=10/np.sqrt(8))
## 0.011825808327677984
```

- Abbildung:



- Dieser Wert heisst p-Wert und gibt die Sicherheit mit der wir den Testentscheid treffen

- 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 $\alpha = 0.05$) ist

p-Wert

- p-Wert ist ein Wert zwischen 0 und 1
- Gibt an, wie gut *Nullhypothese* und *Daten* zusammenpassen:
  - 0: passt gar nicht
  - 1: passt sehr gut
- Genauer: p-Wert ist die W'keit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein *extremeres* zu erhalten

- Mit 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

p-Wert ist W'keit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten wie das aktuell beobachtete

- Testentscheid mit Hilfe des p-Wertes durchführen

p-Wert und Statistischer Test

- Man kann anhand des p-Werts direkt den Testentscheid ablesen: Wenn der p-Wert kleiner als das Niveau ist, so verwirft man H_0, ansonsten nicht
- Verglichen mit dem reinen Testentscheid enthält der p-Wert aber mehr Information, da man direkt sieht, „wie stark" die Nullhypothese verworfen wird
- 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 indirekt, indem der p-Wert angegeben wird
- Zusätzlich zu dieser Entscheidungsregel quantifiziert der p-Wert, *wie signifikant* eine Alternative ist (d.h. wie gross die Evidenz ist für das Verwerfen von H_0)
- Manchmal werden sprachliche Formeln oder Symbole anstelle der p-Werte angegeben:

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

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

Beispiel: Statistischer Test für Durchschnittsgrösse

- Durchschnittsgrösse der Schweizer Frauen beträgt 1.64 m (Bundesamt für Statistik)
- Stimmt diese Aussage?
- Wie kann man sie überprüfen?
- *Lösung*: In Stadt Luzern zum Beispiel Grösse von 150 Frauen messen
- Gemessene Körpergrössen $x_1, \ldots x_{150}$ als Realisierungen auffassen von

$$X_1, \ldots, X_{150} \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2)$$

Beispiel: Statistischer Test für Durchschnittsgrösse

- Bestimmen das arithmetische Mittel $\overline{x}_{150}$ der 150 Messpunkte
- *Frage:* Wenn jeden Tag eine solche Messreihe erheben würden, was wird man feststellen?
- *Antwort:* Arithmetischen Mittel $\overline{x}_{150}$ würden variieren
- Das gemessene arithmetische Mittel $\overline{x}_{150}$ wird als Realisierung der Zufallsvariablen $\overline{X}_{150}$ aufgefasst
- *Frage:* Welche Verteilung hat die Durchschnittsgrösse von 150 Frauen?

Beispiel: Statistischer Test für Durchschnittsgrösse

- *Antwort:* Verteilung vom arithmetischen Mittel $\overline{x}_{150}$:

$$\overline{X}_{150} \sim \mathcal{N}(\mu, \sigma^2_{\overline{X}_{150}})$$

- Wie entscheidet man, ob die gemessenen Durchschnittsgrössen zu dem vom Statistikamt angegebenen Wert von 1.64 m passt?

- *Lösung:* Berechnen W'keit, dass gemessener Wert $\overline{x}_{150}$ oder einen extremeren Wert beobachtet wird, unter der Annahme, dass

$$\overline{X}_{150} \sim \mathcal{N}\left(\mu = 1.64, \frac{\sigma_X^2}{150}\right)$$

Beispiel: Statistischer Test für Durchschnittsgrösse

- *p*-Wert bei einseitiger nach oben gerichteter Alternativhypothese:

$$P\left(\overline{x}_{150} < \overline{X}_{150}\right)$$

- *p*-Wert bei einseitiger nach unten gerichteter Alternativhypothese:

$$P\left(\overline{X}_{150} < \overline{x}_{150}\right)$$

- *p*-Wert bei zweiseitiger Alternativhypothese:

$$P\left(\overline{x}_{150} < |\overline{X}_{150}|\right)$$

Beispiel: Statistischer Test für Durchschnittsgrösse

- Unterscheidung von zwei Fällen:
  - σ_X ist bekannt (aus langjähriger Erfahrung) $\quad \rightarrow \quad$ *z-Test*
  - σ_X wurde aus den Daten geschätzt, d.h., $\widehat{\sigma}_X$ bekannt (was in der Praxis normalerweise der Fall ist) $\quad \rightarrow \quad$ *t-Test*

z-Test: σ_X bekannt

1. *Modell*: X_i ist eine kontinuierliche Messgrösse:

$$X_1, \ldots, X_n \text{ iid } \mathcal{N}(\mu, \sigma_X^2), \ \sigma_X \text{ bekannt}$$

2. *Nullhypothese*:

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

Alternative:

$$H_A: \ \mu \neq \mu_0 \quad (\text{oder "`<"' oder "`>"'})$$

3. *Teststatistik*:

$$T = \overline{X}_n$$

Verteilung der Teststatistik unter H_0

$$T \sim \mathcal{N}\left(\mu_0, \frac{\sigma_X^2}{n}\right)$$

z-Test: σ_X bekannt

4. *Signifikanzniveau*:

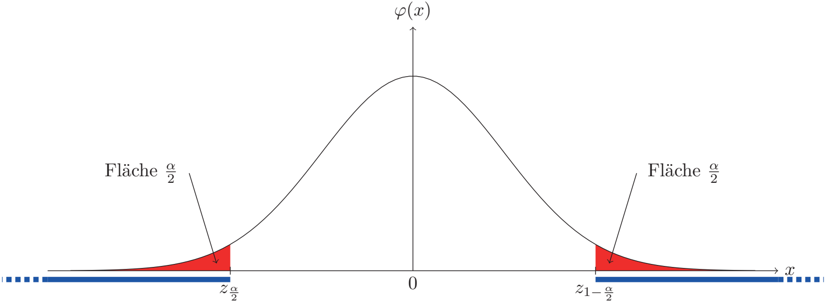
$$\alpha$$

5. *Verwerfungsbereich für die Teststatistik*:

$$K = (-\infty, w_{\frac{\alpha}{2}}] \cup [w_{1-\frac{\alpha}{2}}, \infty) \quad \text{bei } H_A: \ \mu \neq \mu_0$$
$$K = (-\infty, w_{\alpha}] \quad \text{bei } H_A: \ \mu < \mu_0$$
$$K = [w_{1-\alpha}, \infty) \quad \text{bei } H_A: \ \mu > \mu_0$$

6. *Testentscheid*: Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt

Zweiseitiger Verwerfungsbereich beim z-Test



Beispiel: z-Test

- Messreihe für die Körpergrösse von 150 Frauen in Luzern ergab:

$$\overline{x}_{150} = 168\,\text{cm}$$

- Vermutung: Durchschnittsgrösse der Schweizerinnen ist grösser als 164 cm (einseitiger nach oben gerichteter Test)

- σ_X sei bekannt und beträgt $8\,\text{cm}$

Beispiel: z-Test

- Berechnung vom p-Wert (mit **Python**)

$$P[\overline{X}_{150} > 168] = 1 - P[\overline{X}_{150} \leq 168]$$

```
from scipy.stats import norm
import numpy as np

1 - norm.cdf(x=168, loc=164, scale=8/np.sqrt(150))
## 4.5706494145036913e-10
```

- Nullhypothese auf dem Signifikanzniveau $\alpha = 0.05$ verworfen werden

- Angabe vom Bundesamt stimmt nicht

Problem in Praxis: σ_X ist unbekannt!

- Falls σ_X unbekannt $\quad\rightarrow\quad$ Varianz aus den Daten schätzen:

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

- Neue Teststatistik (standardisiert):

$$T = \frac{\overline{X}_n - \mu_0}{\frac{\widehat{\sigma}_X}{\sqrt{n}}}$$

- Historisch wurde standardisiert, aber mit `Python` nicht mehr notwendig

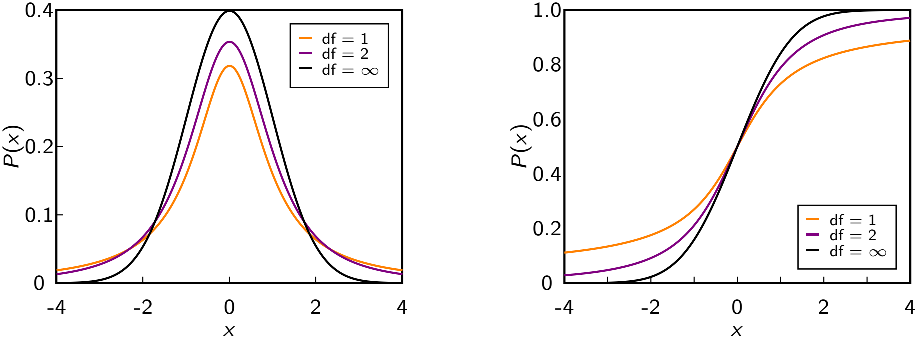
- Verteilung von T, falls H_0 stimmt:

$$T \sim t_{n-1}$$

- t_{n-1} ist die sogenannte *t-Verteilung mit $n-1$ Freiheitsgraden*

„Student's" t-Verteilung

- Gefunden von William Sealy Gosset (1876–1937), Chefbrauer der Guiness-Brauerei
- Werte mit Computer ermittelbar
- Falls $n = \infty$: $t_n = \mathcal{N}(0, 1)$



t-Verteilung: Eigenschaften

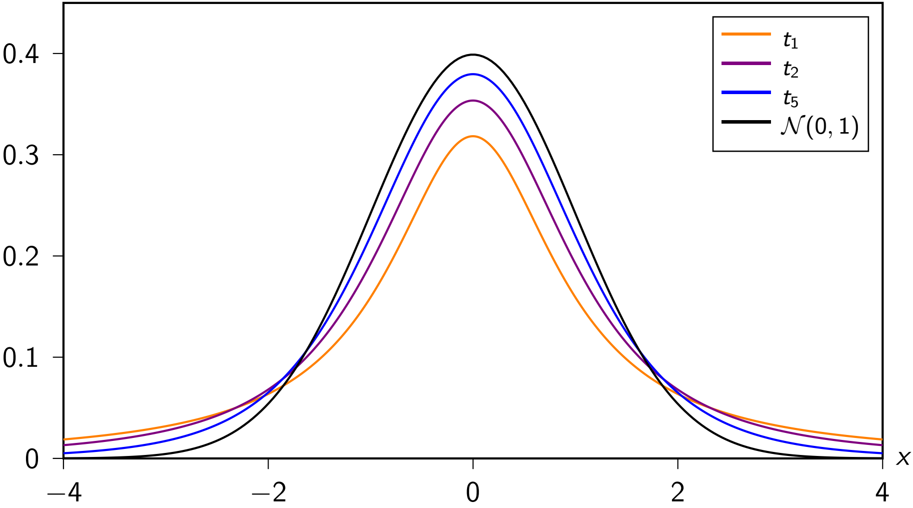
- Wie Standardnormalverteilung: t-Verteilung symmetrisch um 0
- Allerdings: t-Verteilung "‘langschwänziger"’
- Peak in der Mitte ist weniger hoch und "‘weit aussen"’ ist die Dichte grösser (insbesondere falls die Anzahl Freiheitsgrade n klein ist)
- Verglichen mit Standardnormalverteilung liefert sie *eher grosse Werte*
- Es gilt:

$$t_n \rightarrow \mathcal{N}(0, 1) \quad \text{für} \quad n \rightarrow \infty$$

(siehe auch Plot mit Dichten)

t-Verteilung: Dichten

Dichte $f(x)$



t-Test: σ_X unbekannt

- Für Berechnungen: `t.cdf(...)` anstatt `norm.cdf(...)` verwenden
- Achtung: Freiheitsgrad `df=...` nicht vergessen
- Freiheitsgrad 1 kleiner als die Anzahl Beobachtungen

Beispiel: t-Test

- Messreihe für Körpergrösse von 150 Frauen in Luzern:

$$\overline{x}_{150} = 168\,\mathrm{cm}$$

- Vermutung: Durchschnittsgrösse der Schweizerinnen entweder grösser oder kleiner als 164 cm ist (zweiseitiger Test)

- σ_X wurde nun geschätzt, und beträgt $\widehat{\sigma}_X = 10\,\mathrm{cm}$

- Berechnen (mit **Python**) p-Wert für eine einseitige Alternativhypothese

$$P[\overline{X}_{150} \geq 168] = 1 - P[X \leq 168]$$

Beispiel: t-Test

- *Achtung*: Bei der $\overline{X}_{150}$ *muss* Freiheitsgrad 149 gewählt werden!

- **Python**

```
from scipy.stats import t
1 - t.cdf(x=168, df=149, loc=164, scale=10/np.sqrt(150))
## 1.241988350275669e-06
```

oder standardisiert:

```
1 - t.cdf(x=(168-164)/(10/np.sqrt(150)), df=149)
## 1.241988350275669e-06
```

- p-Wert: $2 \cdot 1.241988 \cdot 10^{-6}$

- Nullhypothese auf dem Signifikanzniveau $\alpha = 0.05$ verwerfen

t-Test

- t-Test wichtiger als z-Test, da wahre Standardabweichung nie bekannt
- Falls Daten bekannt: Verfahren in Python implementiert
- Befehl:

```
from scipy.stats import ttest_1samp
from pandas import Series

methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

ttest_1samp(a=methodeA, popmean=80)
## Ttest_1sampResult(statistic=3.1246428367325474, pvalue=0.0087787788
```

Befehlserklärung

- `ttest_1samp`: t-Test für eine Stichprobe (one sample)
- `a=...`: Daten
- `popmean=...`: Nullhypothese (population mean)
- Output:
  - `statistics=...`: Der sogenannte t-Wert, für uns nicht interessant
  - `pvalue=...`: p-Wert, entscheidet ob Nullhypothese verworfen wird oder nicht
  - Achtung: p-Wert ist immer für zweiseitigen Test
  - Für einseitigen Test muss Wert halbiert werden