STAT

Lageparameter

Arithmetisches Mittel ("Durchschnitt") / .mean(): $\bar{x} = \frac{x_1 + x_2 + x_3 \dots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ↗ mean sagt nichts über die Streuung der Daten aus

↳ Mittelwert

↳ μ_x

Median ("Mitte") = Daten sortieren
- → Bei gerader Anzahl: mean() von den mittleren zwei
- → Bei ungerader Anzahl: der Wert in der Mitte

pandas: median()

↳ Vorteil: Robustheit bei extremen Beobachtungen

Trick: $k = np.round(0.5 \cdot n + 0.5) - 1$

Quantile
- Unteres Quartil: Wert, wo 25% kleiner oder gleich und 75% grösser oder gleich sind
- Oberes Quartil: Wert, wo 75% kleiner oder gleich und 25% grösser oder gleich sind

xxx.quantile(q=.25 / .75, interpolation="midpoint") = q^{75}, q^{25} = xxx.quartile(q=[.75, .25], interpolation="midpoint")

Streuungsparameter:

Empirische Varianz / Empirische Standardabweichung

→ Pandas ⇒ .var() → Pandas ⇒ .std()

↗ Durch $\sqrt{\ }$ kommen die Daten wieder in dieselbe Einheit.

$$Var(x) = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$S_x = \sqrt{Var(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

↳ σ

- Var(x) & S_x gross ⇒ so ist die Streuung der Messwerte gross. (+ $\bar{x}$)

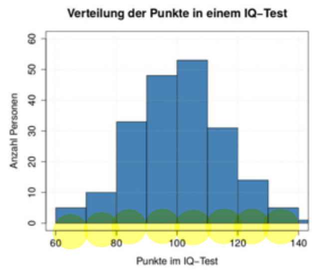
Quartilsdifferenz: oberes Quartil − unteres Quartil

Quartilsdifferenz

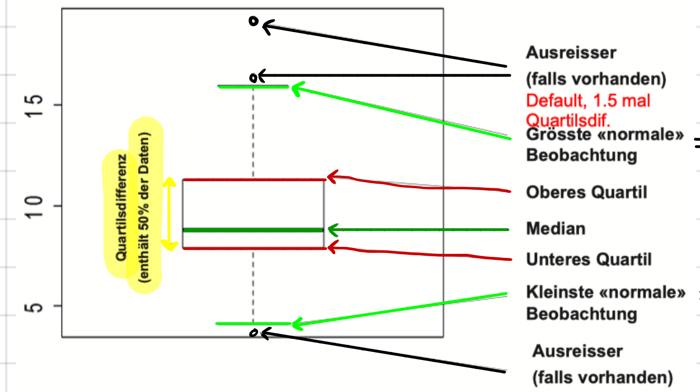
0 25% 50% 75% 100%

Graphische Darstellung

Histogramm: Aufteilung in n-Klassen (Intervall) <
- < 50 Messungen = Klasse 5 bis 7
- > 250 Messungen = Klasse 10 bis 20

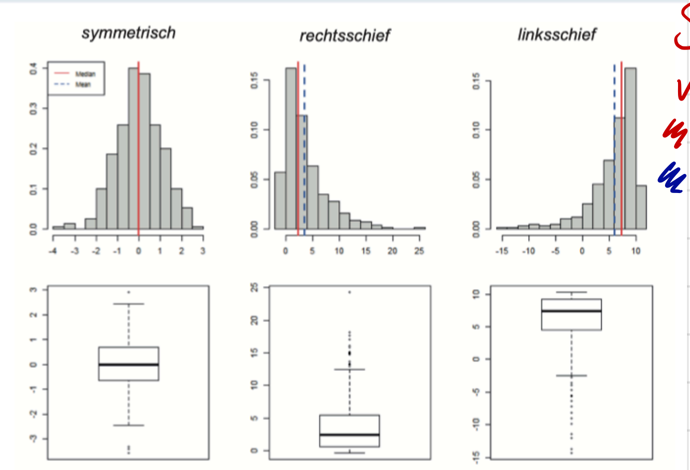


Boxplot



⇒ 1.5 · Quartilsdifferenz

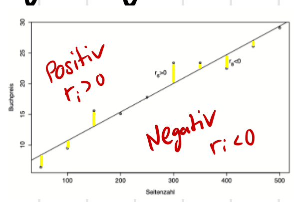
⇒ 1.5 · Quartilsdifferenz



Schiefe von Daten
- = Median
- = Mean

Lineare Regression

Residumm: Abstand der Messpunkte zur Regressionsgerade. $r_i = y_i - a - bx_i$



$$r = \sum_{i=1}^{n}(r_i)^2$$

a+b von linearer Regression

$$b = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Empirische Korrelation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

Numerische Zusammenfassung der lin. Abhängigkeiten von zwei Grössen.

eine Zahl $r \Rightarrow [-1; 1]$

$r = +1 \Rightarrow$ dann liegen Punkte auf steigender Geraden $b > 0$

$r = -1 \Rightarrow$ dann liegen Punkte auf fallender Geraden $b < 0$

$r = 0 \Rightarrow$ x und y unabhängig keinen Zusammenhang

Stetige W'keitsverteilung

Jass bsw: $\omega = As \Rightarrow X(\omega) = 11 \Rightarrow x = 11$

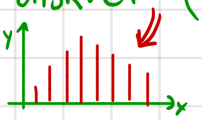
Zufallsvariable X ist eine Funktion (kann auch Y, Z sein)

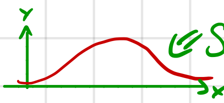
Elementarereignis ω

Zahlenwert / konkreten Wert : x (x eine Realisierung der Zufallsvariable (X))

Grundraum Ω

diskret (endlich abzählbare Menge)





stetig

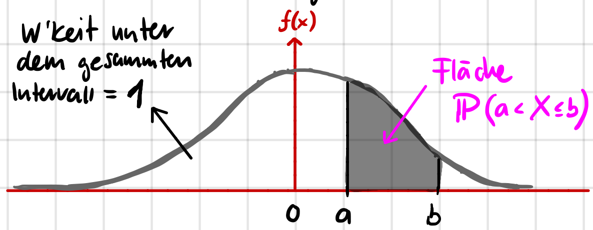
Die Werte einer Zufallsvariablen X (die möglichen Realisationen von X) treten mit gewissen Wahrscheinlichkeiten auf. Die Wahrscheinlichkeit, dass X den Wert x annimmt, berechnet sich wie folgt:

$$P(X = x) = P(\{\omega \mid X(\omega) = x\}) = \sum_{\omega; X(\omega) = x} P(\omega)$$

W'keit von allen Realisierungen $\Rightarrow$ W'keitsverteilung $\sum_{\text{alle möglichen } x} P(X = x) \Rightarrow 1$

Eigenschaften Wahrscheinlichkeitsdichte

1. $f(x) \geq 0$ für alle x (da $F(\cdot)$ monoton wachsend ist)
2. $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$ (Fläche zwischen a und b unter $f(x)$)
3. $\int_{-\infty}^{\infty} f(x)dx = 1$ (wegen 2.)



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

Varianz : $Var(X) = \sigma_X^2 = E(X^2) - E(X)^2$

Kumulative Verteilung : $P(X \leq x) \Rightarrow$ cdf

Nullhypothese H_0
Alternativhypothese H_A
Wert $x_1, x_2 \dots x_n$
Realisierung $X_1, X_2, X_3 \dots X_n$
arithmetisches Mittel $\bar{x}$
Realisierung des gem. $\bar{x} \Rightarrow \bar{X}$

P-Wert

- Wie gut Daten und Nullhypothese passen
- Wert zwischen $[0;1]$ $0 \Rightarrow$ schlecht $1 \Rightarrow$ sehr gut
- W'keit
- Je kleiner der P-Wert, desto mehr spricht gegen die Nullhypothese.

Statistischer Test

- einseitig nach oben gerichteter $H_A \Rightarrow P(\bar{x}_{150} < \bar{X}_{150})$
- einseitig nach unten gerichteter $H_A \Rightarrow P(\bar{X}_{150} > \bar{x}_{150})$
- zweiseitiger $H_A \Rightarrow P(\bar{x}_{150} < |\bar{X}_{150}|)$

2 Fälle

σ = Standardabweichung

σ_x ist bekannt $\Longrightarrow$ z-Test
σ_x wurde geschätzt $\Longrightarrow$ t-Test

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

1. Verwerfe H_0 falls *P*-Wert $\leq \alpha$
2. Belasse H_0 falls *P*-Wert $> \alpha$

z-Test

cookbook:

1. Modell : X_i kontinuierliche Messgrösse
   $X_1 \dots X_n \; \mathcal{N}(\mu, \sigma_x^2)$ σ_x ist bekannt.

2. Nullhypothese : $H_0 : \mu = \mu_0$
   Alternative : $H_A : \mu \neq \mu_0$ (oder $>$ oder $<$)

3. Teststatistik : $Z = \dfrac{(\bar{x} - \mu_0)}{\sigma_{\bar{x}_n}} = \dfrac{(\bar{X} - \mu_0)}{\frac{\sigma_x}{\sqrt{n}}}$
   ↳ Verteilung $Z \sim \mathcal{N}(0,1)$

$\Rightarrow \dfrac{(168 - 164)}{\frac{8}{\sqrt{150}}} = \dfrac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}$

4. Signifikanzniveau : α (oft 5%)

5. Verwerfungsbereich
   $K = (-\infty, z_{\frac{\alpha}{2}}] \cup [z_{1-\frac{\alpha}{2}}, \infty)$ bei $H_A : \mu \neq \mu_0$
   $K = (-\infty, z_\alpha]$ bei $H_A : \mu < \mu_0$
   $K = [z_{1-\alpha}, \infty)$ bei $H_A : \mu > \mu_0$

6. Testentscheid

t-Test

cookbook:

1. Modell ist eine kontinuierliche Messgrösse
   $X_1 \dots X_n$ iid $\mathcal{N}(\mu, \sigma_x^2)$ σ_x wird durch $\hat{\sigma}_x$ geschätzt
   . Std

2. Nullhypothese : $H_0 : \mu = \mu_0$
   Alternative : $H_A : \mu \neq \mu_0$ (oder $>$/$<$)

3. Teststatistik : $T = \dfrac{(\bar{X}_n - \mu_0)}{\hat{\sigma}_{\bar{x}_n}} = \dfrac{(\bar{X}_n - \mu_0)}{\hat{\sigma}_x / \sqrt{n}}$
   ↳ Verteilung $H_0 : T \sim t_{n-1}$
   Freiheitsgrad
   $= \dfrac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}$

4. Signifikanzniveau α (oft 5%)

5. Verwerfungsbereich für die Teststatistik
   $K = (-\infty, t_{n-1; \frac{\alpha}{2}}] \cup [t_{n-1; 1-\frac{\alpha}{2}}, \infty) \Rightarrow H_A : \mu \neq \mu_0$
   $K = (-\infty, t_{n-1; \alpha}] \Rightarrow H_A : \mu < \mu_0$
   $K = [t_{n-1; 1-\alpha}, \infty) \Rightarrow H_A : \mu > \mu_0$

6. Testentscheid

Fehler 1. und 2. Art

Entscheidung / Wahrheit	H_0	H_A
H_0	✓	Fehler 1. Art
H_A	Fehler 2. Art	✓ Macht

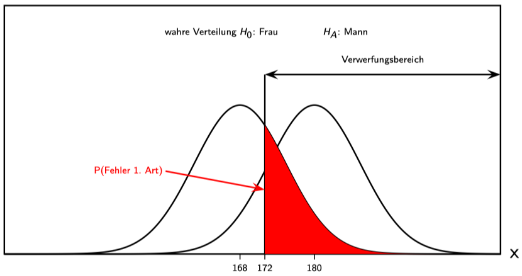
Nullhypothese H_0
Alternativhypothese H_A

- Entscheidung für H_0, aber H_A wäre richtig
  ↳ Fehler 2. Art
- Entscheidung für H_A, aber H_0 wäre richtig
  ↳ Fehler 1. Art

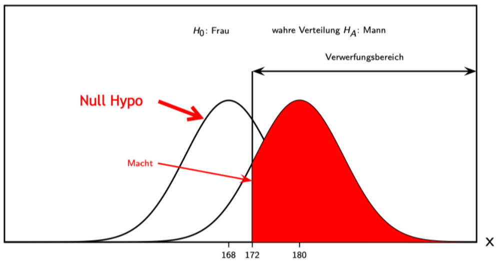
↳ schwerwiegenster Fehler, weil man nimmt H_A obwohl H_0 korrekt ist.

- Signifikanzniveau α möglichst klein. z.b. 0.5%
- Je kleiner α, desto kleiner der Verwerfungsbereich

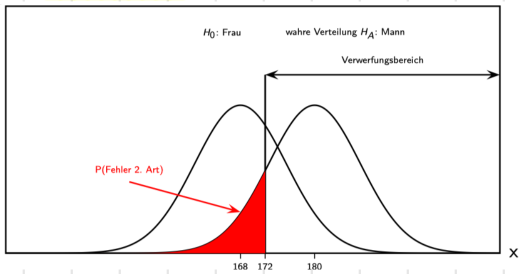
1. Art



Macht



2. Art

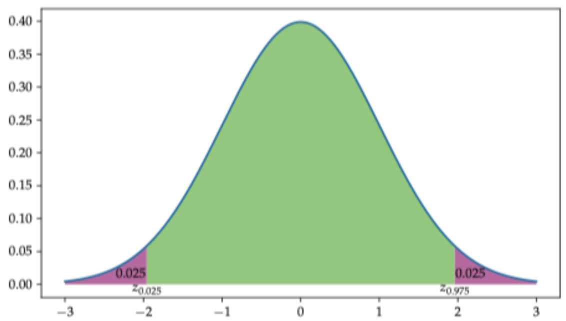


Fehler 2. Art = 1 − Macht

Vertrauensintervall μ

zweiseitigem t-Test:

$$K = (-\infty, t_{n-1;\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty)$$



Zweiseitiges Vertrauensintervall

Dies führt dann auf die folgenden **zweiseitigen Vertrauensintervalle** (die dazugehörigen Tests sind zweiseitig mit Alternative $H_A : \mu \neq \mu_0$) zum Niveau $1 - \alpha$:

$$\left[\overline{x}_n - t_{n-1,1-\alpha/2} \cdot \frac{\widehat{\sigma}_X}{\sqrt{n}}, \overline{x}_n + t_{n-1,1-\alpha/2} \cdot \frac{\widehat{\sigma}_X}{\sqrt{n}}\right]$$

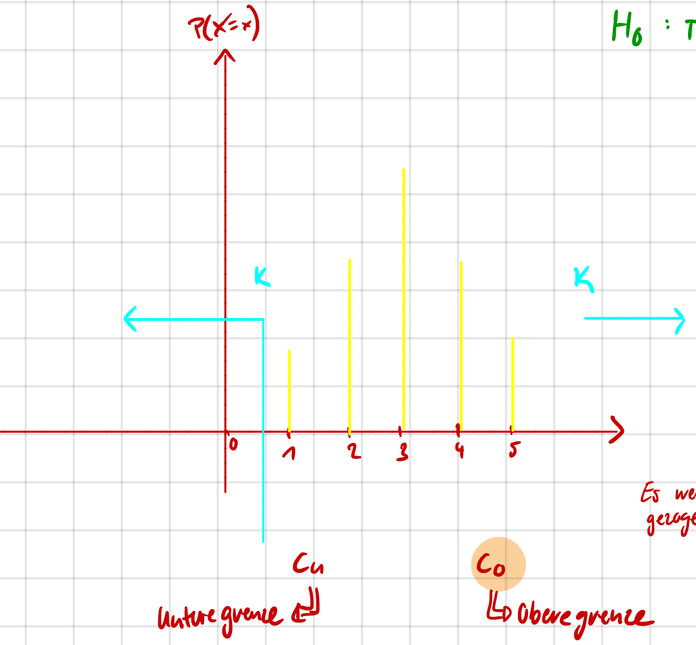
Vorzeichentest

0. QQ-Plot
1. Modell: $x_i \dots x_n$, wobei X_i: beliebige Verteilung
2. Nullhypothese H_0 $\mu = \mu_0$ (μ = ist der Median)
   Alternative H_A $\mu \neq \mu_0$ (einseitig / beidseitig)
3. Teststatistik V: Anzahl X_i's mit ($X_i > \mu_0$)
   ↳ Verteilung H_0: $V \sim \text{Bin}(n, \pi_0)$ $\pi_0 = 0.5$
4. Signifikanzniveau: α (oft 5%)
5. Verwerfungsbereich für Teststatistik
   c_u + c_o mit Binomialverteilung oder Normalapproximation $K = [0; c_u] \cup [c_o, n]$
   untere Grenze obere Grenze
6. Testentscheid

Beispiel: Beobachtungen

- $x_1 = 13$ $x_2 = 9$ $x_3 = 17$
  $x_4 = 8$ $x_5 = 14$
- $H_0 = \mu = \mu_0 = 10$
  $H_A = \mu \neq \mu_0$
- Vorzeichentest $x_i - \mu_0$
  $x_1 = 3\ (+)$ $x_2 = -1\ (-)$ $x_3 = 7\ (+)$
  $x_4 = -2\ (-)$ $x_5 = 4\ (+)$
- Binomialtest
  $H_0 = \pi = 0.5$
  $H_A = \pi \neq 0.5$ $n = 5$ $x = 3$
  st.binom_test (x=3, n=5, p=0.5) Anzahl (+)
  ↳ Resultat: P-Wert

Binomialtest



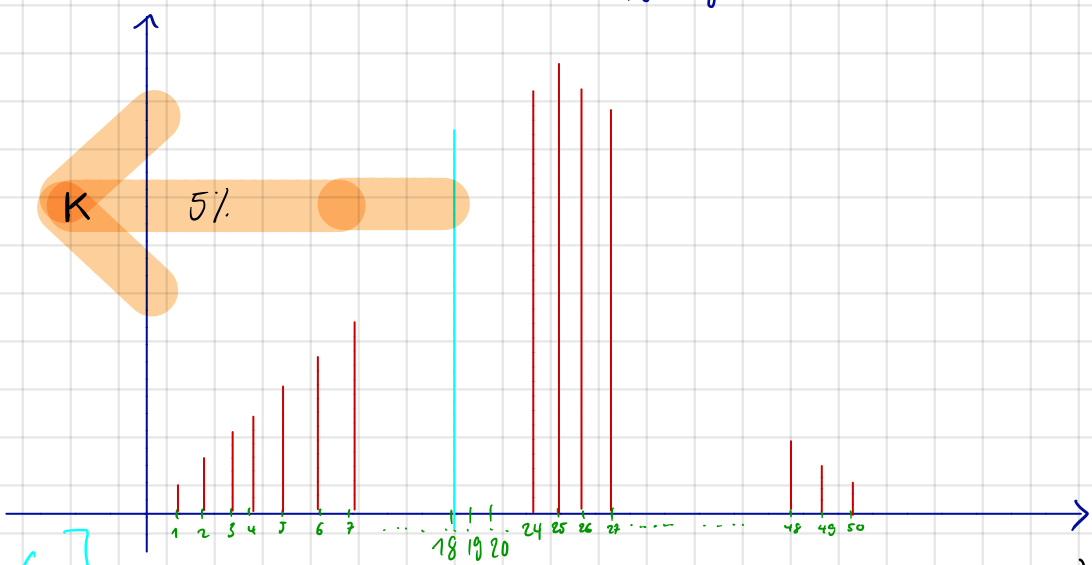
H_0: $\pi_0 = 0.5$
5%

K = Verwerfungsbereich

$K = \ \cup [5]$

Es werden 5 Lösli gezogen.

$n = 50$
X = # Anzahl Gewinne bei n=50 Lösli
$H_0 = \pi_0 = 0.5$ ⇒ P



$C = [0, C_u]$
└ $[0\ 18]$

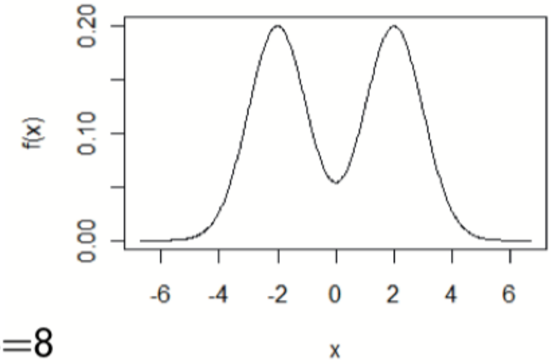
$C_u = 18$ ⟹ Python-Code: st.binom.ppf(0.05, n=50, p=0.5)
↳ 19 ⇒ 18, weil nicht im 5% Bereich liegt)
st.binom.cdf(18, n=50, p=0.5)

- Bsp: $H_0: \mu_0 = 0$
- Beobachte -1.9, 0.2, 2.9, -4.1, 3.9
- Absolutbeträge: 1.9, 0.2, 2.9, 4.1, 3.9

  Der Ränge nach sortiert 2 1 3 5 4
- Ränge der Absolutbeträge: 2,1,3,5,4
- Rangsumme der positiven Gruppe: 1+3+4=8

  Minimale Rangsumme: 0 alle wären Negetaiv

  Maximale Rangsumme: 1+2+3+4+5 = 15 Alle wären Postiv




```
import scipy.stats as st
import numpy as np

x = np.array([-1.9, 0.2, 2.9, -4.1, 3.9])

st.wilcoxon(x, correction=True)

## : UserWarning: Warning: sample size too small for normal approximation.
##   warnings.warn("Warning: sample size too small for normal approximation.")
## WilcoxonResult(statistic=7.0, pvalue=1.0)
```

Übersicht der Tests

Wichtig!

Test	Annahme				n_{min} bei $\alpha = 0.05$	Macht für ein Beispiel (1)
	σ_X bekannt	$X_i \sim N$	Symm. Verteilung	iid		
z	×	×	×	×	1	89%
t		×	×	×	2	79%
Wilcoxon			×	×	6	79%
VZ				×	5	48%

(1): $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $n = 10$; $H_0: \mu = 0$; $H_A: \mu \neq 0$; $\alpha = 0.05$
Macht berechnet für konkrete Alternative: $X_i \sim \mathcal{N}(1, 1)$