# Vorzeichentest
# Wilcoxontest
# Mann-Whitney-Test


Peter Büchel

HSLU I


Stat: Block 08

# Nicht-Normalverteilte Daten: Vorzeichentest

1. *Modell*:

$$X_1, \ldots, X_n \text{ iid}$$

   wobei $X_i$ eine beliebige Verteilung hat

2. *Nullhypothese*:

$$H_0: \ \mu = \mu_0 \quad (\mu \text{ Median})$$

   *Alternative*:

$$H_A: \ \mu \neq \mu_0 \quad (\text{oder einseitige Variante})$$

3. *Teststatistik*:

$$V\!: \ \text{Anzahl } X_i\text{'s mit } (X_i > \mu_0)$$

   *Verteilung der Teststatistik unter $H_0$*:

$$V \sim \text{Bin}(n, \pi_0) \quad \text{mit } \pi_0 = 0.5$$

4. *Signifikanzniveau*:

$$\alpha$$

5. *Verwerfungsbereich für die Teststatistik*:

$$K = [0, c_u] \cup [c_o, n]$$

falls $H_A: \ \mu \neq \mu_0$.

Grenzen $c_u$ und $c_o$ müssen mit Binomialverteilung oder Normalapproximation berechnet werden

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

# Beispiel: Vorzeichentest

- Beobachtet:

$$x_1 = 13, \quad x_2 = 9, \quad x_3 = 17, \quad x_4 = 8, \quad x_5 = 14$$

- Angenommen:

$$H_0: \ \mu = \mu_0 = 10, \qquad H_A: \ \mu \neq 10$$

- Vorzeichen von $x_i - \mu_0$:

$$+, \ -, \ +, \ -, \ +$$

- Binomialtest mit

$$H_0: \ \pi = 0.5, \ H_A: \ \pi \neq 0.5, \ n = 5, \ x = 3 \ \text{ (Anzahl „+")}$$

# Beispiel: Vorzeichentest

- Antwort: Binomialtest mit **Python**

```python
import scipy.stats as st

st.binom_test(x=3, n=5, p=0.5)

## 1.0
```

- Resultat ist $P$-Wert

- Nullhypothese beim Vorzeichentest wird nicht verworfen

- Vorteil vom Vorzeichentest: Keine Annahme an Verteilung

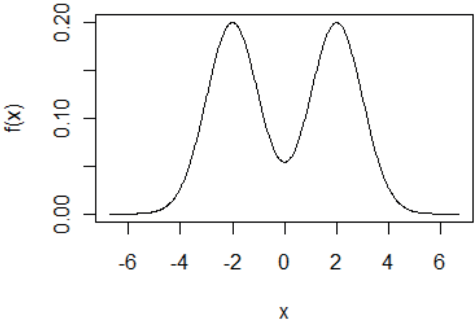
- Nachteil vom Vorzeichentest: Kleinere Macht

# Nicht-normalverteilte Daten: Wilcoxon-Test

- Kompromiss zwischen Vorzeichen- und $t$-Test

- Annahme:

$$X_i \sim F \text{ iid}, \; F \quad \text{ist symmetrisch}$$



- Teste Median $\mu$ : $H_0$ : $\mu = \mu_0$ (einseitig oder zweiseitig)

- Intuition der Teststatistik
    - Rangiere
$$|x_i - \mu_0| \to r_i$$
    - Gib Rängen ursprüngliches Vorzeichen von $(x_i - \mu_0)$ ("‘signed ranks"’)
    - Teststatistik $T$: Summe aller Ränge, bei denen $(x_i - \mu_0)$ positiv ist

- Falls $H_0$ stimmt, sollte Rangsumme nicht zu gross und nicht zu klein sein

# Beispiel: Wilcoxon-Test

- Bsp: $H_0: \mu_0 = 0$
- Beobachte:

$$-1.9, \quad 0.2, \quad 2.9, \quad -4.1, \quad 3.9$$

- Absolutbeträge:

$$1.9, \quad 0.2, \quad 2.9, \quad 4.1, \quad 3.9$$

- Geordnet

$$0.2, \quad 1.9, \quad 2.9, \quad 3.9, \quad 4.1$$

- Ränge der Absolutbeträge:

$$1, \quad 2, \quad 3, \quad 4, \quad 5$$

- Rangsumme der posititven Gruppe: $1+3+4=8$
  - Minimale Rangsumme: 0
  - Maximale Rangsumme: $1+2+3+4+5 = 15$

# Wicoxon-Test mit Python

- Code

```
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)
## /home/bl/.local/lib/python3.6/site-packages/scipy/stats/morestats.py
##   warnings.warn("Warning: sample size too small for normal approxima
## WilcoxonResult(statistic=7.0, pvalue=1.0)
```

# Übersicht der Tests

| Test | Annahme | | | | $n_{\min}$ bei $\alpha = 0.05$ | Macht für ein Beispiel (1) |
|---|---|---|---|---|---|---|
| | $\sigma_X$ bekannt | $X_i \sim N$ | Symm. Verteilung | iid | | |
| $z$ | x | x | x | x | 1 | 89 % |
| $t$ | | x | x | x | 2 | 79 % |
| Wilcoxon | | | x | x | 6 | 79 % |
| VZ | | | | x | 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: $X_i \sim \mathcal{N}(1, 1)$

# Wilcoxon-Test versus t-Test

- *Wilcoxon-Test* meistens *t-Test* oder *Vorzeichen-Test* vorzuziehen
- Er hat in vielen Situationen oftmals wesentlich *grössere Macht*, und selbst in den ungünstigsten Fällen ist er nie viel schlechter
- Wenn man trotzdem den t-Test verwendet, dann sollte man die Daten auch graphisch ansehen, damit wenigstens grobe Abweichungen von der Normalverteilung entdeckt werden
- Insbesondere sollte der *Normal-Plot* angeschaut werden

# Vergleich von zwei Stichproben

- Mögliche Fragestellungen:
  - Vergleich von zwei Messverfahren (Messgerät $A$ vs. Messgerät $B$): Gibt es einen signifikanten Unterschied?
  - Vergleich von zwei Herstellungsverfahren ($A$ vs. $B$): Welches hat die besseren Eigenschaften (z.B. bzgl. einer Festigkeitsgrösse)?
  - Werden männliche Dozenten von weiblichen Studierenden besser als von männlichen Studierenden bewertet?
  - Sammeln jeweils Daten von zwei Gruppen

# Gepaarte Stichproben

- Beispiel Messgeräte: Jeden Prüfkörper mit *beiden* Messgeräten messen

- Pro *Versuchseinheit* (hier: Prüfkörper) zwei Beobachtungen (einmal Gerät $A$ und einmal Gerät $B$)

- Man spricht auch von *gepaarten Stichproben*

- Beide Beobachtungen sind *nicht* unabhängig, da an der *gleichen* Versuchseinheit zwei Mal gemessen wird!

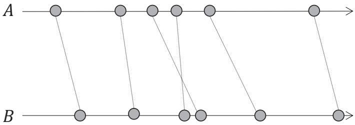
# Ungepaarte (unabhängige) Stichproben

- Beispiel der beiden Herstellungsverfahren: Stichprobe von Verfahren $A$ und eine andere Stichprobe von Verfahren $B$
- Beobachtungen sind hier *unabhängig*; "‘es gibt *nichts*, was sie verbindet"’
- Man spricht von *ungepaarten (oder unabhängigen) Stichproben*

# Unterscheidung gepaart versus ungepaarte Stichproben

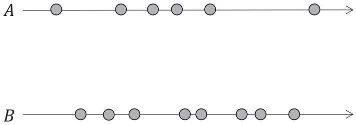
**Gepaarte Stichproben**

- Jede Beobachtung einer Gruppe kann eindeutig einer Beobachtung der anderen Gruppe zugeordnet werden
- Stichprobengrösse ist in beiden Gruppen zwangsläufig gleich




**Ungepaarte Stichproben**

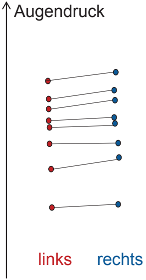
- Keine Zuordnung von Beobachtungen möglich
- Stichprobengrössen können verschieden sein (müssen aber nicht!)
- Man kann die eine Gruppe vergrössern, ohne dass man die andere vergrössert



# Gepaarte versus ungepaarte Stichproben

- Beispiel Augeninnendruck: Ein Auge wird behandelt, das andere nicht (gepaarter Test ist angebracht)
- Gemäss Vorraussetzungen dürfte auch ein ungepaarter Test angewendet werden

$H_0\colon \mu_X = \mu_Y$




Ungepaart:

Intuition Teststatistik: $T = \frac{\overline{X}-\overline{Y}}{\widehat{\sigma_{\overline{X}}}}$

Gepaart:

Differenz $D_i = X_i - Y_i$

Teststatistik $T = \frac{\overline{D}}{\widehat{\sigma_{\overline{D}}}}$

# Statistischer Test für gepaarte Stichproben mit `Python`

- *Gepaarte Stichproben:*

$$X_i \sim \mathcal{N}(\mu_X, \sigma_X^2) \qquad \text{und} \qquad Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

- Betrachten Differenzen $D_i = X_i - Y_i$
- Führen einen t-Test durch mit der Teststatistik $D_i$
- Normalerweise Nullhypothese $\mathrm{E}[D] = \mu_D = 0$, also kein Unterschied
- Falls die Daten nicht normalverteilt, dann kommt ein Wilcoxon- oder Vorzeichentest in Frage.

# Statistischer Test für gepaarte Stichproben mit Python

- Code

```
import scipy.stats as st
from scipy.stats import norm, t, binom
import numpy as np
from pandas import Series

vorher = Series([25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28])

nachher = Series([27, 29, 37, 56, 46, 82, 57, 80, 61, 59, 43])

st.ttest_rel(nachher, vorher)

## Ttest_relResult(statistic=4.271608818429545, pvalue=0.0016328499219
```

- P-Wert ist mit 0.0016 unter Signifikanzniveau und somit wird Nullhypothese verworfen

- Vertrauensintervall:

```
vorher = Series([25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28])
nachher = Series([27, 29, 37, 56, 46, 82, 57,80, 61, 59, 43])
dif = nachher - vorher

t.interval(alpha=.95, df=dif.size-1, loc=dif.mean(),
scale=dif.std()/np.sqrt(dif.size))

## (4.914309935154O7, 15.631144610300478)
```

- Unterschied auf 5 % Signifikanzniveau signifikant, weil P-Wert kleiner 5 %
- 95 %-Vertrauensintervall: Unterschieds in den Gruppenmittelwerten
- Mit 95 % W'keit ist Gruppenmittelwert von `x` um eine Zahl im Bereich

$$[4.91431, 15.63114]$$

grösser als der Gruppenmittelwert von `y`

# Statistischer Test für ungepaarte Stichproben mit `Python`

- *Ungepaarte Stichproben:* Daten $X_i$ und $Y_i$ normalverteilt, aber ungepaart

- Beispiel: Schmelzwärme von Eis: Zwei-Stichproben t-Test für ungepaarte Stichproben mit Nullhypothese $\mu_X = \mu_Y$:

```
x = 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])

y = Series([80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97])

st.ttest_ind(x, y, equal_var=False)

## Ttest_indResult(statistic=2.839932638516127, pvalue=0.018660
```

- P-Wert ist 0.018 unter dem Signifikanzniveau und somit wird die Nullhypothese verworfen.

- Unterschied auf 5 % Signifikanzniveau signifikant, weil P-Wert kleiner als 5 %
- Unterschied in den Gruppenmittelwerten

# Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)

- Falls Daten nicht normalverteilt
- $X_i \sim F,\ i = 1, \ldots, n;\ Y_j \sim G,\ j = 1, \ldots m$
  $H_0:\ F = G$
  $H_A:\ F = G + \delta\ (\delta \neq 0)$ (oder einseitig)
  (d.h., Verteilungen sind verschoben, haben aber gleiche From)
- Code

```
x = 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])

y = Series([80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97])

st.mannwhitneyu(x, y)

## MannwhitneyuResult(statistic=14.5, pvalue=0.007268822472369
```

- Nullhypothese wird verworfen

# Übersicht: Tests für ungepaarte Stichproben

| **Test** | **Annahme** | | | | **$n_{\min}$ falls ($n = m$) bei $\alpha = 0.05$** | **Macht für ein Beispiel (1)** |
|---|---|---|---|---|---|---|
| | $\sigma_X = \sigma_Y$ | $X_i \sim N$ $Y_i \sim N$ | $F, G$ haben gleiche Form | iid pro Gruppe | | |
| $t$ ($\sigma_X = \sigma_Y$) | x | x | x | x | 2 | 57% |
| $t$ ($\sigma_X \neq \sigma_Y$) | | x | | x | 2 | 56% |
| MW U-Test | x | | x | x | 4 | 53% |

(1): $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$, $Y_i \sim \mathcal{N}(\mu_Y, \sigma^2)$ $n = m = 10$; $H_0 : \mu_X = \mu_Y$; $H_A : \mu_X \neq \mu_Y$; $\alpha = 0.05$

Macht berechnet für konkrete Alternative: $X_i \sim \mathcal{N}(0, 1)$, $Y_i \sim \mathcal{N}(1, 1)$

# Übersicht Statistische Tests (stetige Verteilungen)