# Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

# Voraussetzungen der parametrischen Tests

Die bisher besprochenen parametrischen Tests setzen voraus, zumindest bei nicht hinreichend großen Stichprobenumfängen, dass die betrachteten Merkmale als normalverteilt angenommen werden können

- Wir werden später Verfahren kennen lernen, die eine Überprüfung dieser Annahme ermöglichen

Bei antizipierten Abweichungen von der Verteilungsannahme (z. B. bei rechtsschiefen Verteilungen, wie sie etwa bei Reaktionszeiten vorkommen) sollten diese Tests nicht verwendet werden

Es kommen verschiedene Alternativen in Frage

- *Nonparametrische Tests* (auch *verteilungsfreie Tests*)
- *Randomisierungstests* (auch *Permutationstests*)
- Bootstrap-Konfidenzintervalle (wurden bereits besprochen)

# Nonparametrische Tests

Begriffsbestimmung

- Das Attribut "nonparametrisch" bezieht sich darauf, dass nicht die Parameter einer Verteilung (wie z. B. $\mu$, $\sigma$) im Vordergrund stehen, sondern allgemeine Charakteristika wie der Median oder Quantile
- Das Attribut "verteilungsfrei" besagt, dass die Verteilung der Teststatistik unter der Nullhypothese $H_0$ nicht von der (stetigen) Verteilungsfunktion des zugrunde liegenden Merkmals abhängt

Häufig verzichten nonparametrische Tests auf die Nutzung von Information, die in den Daten verfügbar ist

Auch nonparametrische Tests machen Voraussetzungen, im Folgenden ist dies vor allem die Annahme, dass das unterliegende Merkmal (und damit die Stichprobenvariablen) eine stetige Verteilungsfunktion besitzt

# Vorzeichentest

Der Vorzeichentest macht Aussagen über den Median $\tilde{\mu}$ der den Stichprobenvariablen $X_1, \ldots, X_n$ unterliegenden stetigen Verteilung

Für einen vorgegebenen hypothetischen Wert $\tilde{\mu}_0$ betrachtet man die Zufallsvariable

$$T = \text{Anzahl der } X_i \text{ mit } X_i < \tilde{\mu}_0$$

Unter Voraussetzung, dass $\tilde{\mu} = \tilde{\mu}_0$ ist, kann man die Bestimmung von $T$ als Folge von Bernoulli-Experimenten mit den Ereignissen $\{X_i < \tilde{\mu}_0\}$ und $\{X_i \geq \tilde{\mu}_0\}$ mit Parameter $p = 0.5$ auffassen

Die Zufallsvariable $T$ ist damit binomialverteilt mit den Parametern $n$ und $p = 0.5$, es gilt also $T \sim B(n, 0.5)$, was sich für die Konstruktion eines statistischen Tests nutzen lässt

# Vorzeichentest

Annahmen

- $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit stetiger Verteilungsfunktion

Hypothesen

1. $H_0\colon\ \tilde{\mu} = \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} \neq \tilde{\mu}_0$
2. $H_0\colon\ \tilde{\mu} \geq \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} < \tilde{\mu}_0$
3. $H_0\colon\ \tilde{\mu} \leq \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} > \tilde{\mu}_0$

Teststatistik

$$T = \text{Anzahl der } X_i \text{ mit } X_i < \tilde{\mu}_0 \text{ für alle } i = 1, \ldots, n$$

# Vorzeichentest

Verteilung der Teststatistik unter $\tilde{\mu} = \tilde{\mu}_0$

- $B(n, 0.5)$, für $n \geq 25$ approximativ $N(n/2, n/4)$

Ablehnungsbereich

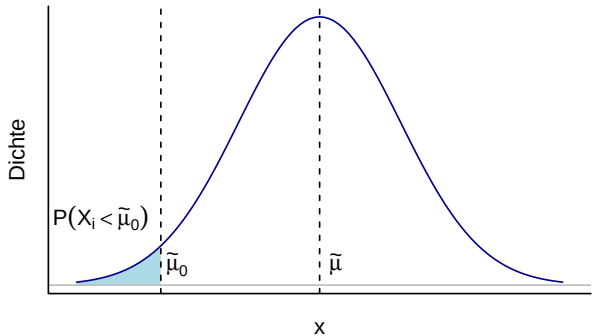
1. $T < b_{\alpha/2}$ oder $n - T < b_{\alpha/2}$
2. $n - T < b_\alpha$ (oder $T > b_{1-\alpha}$)
3. $T < b_\alpha$

mit dem $p$-Quantil $b_p$ der $B(n, 0.5)$-Verteilung, bestimmt durch

$$P(T \leq b_p - 1) < p \leq P(T \leq b_p)$$

# Testentscheidung bei einseitiger Fragestellung

Wahrscheinlichkeit des Ereignisses $X_i < \tilde{\mu}_0$ unter $H_0$: $\tilde{\mu} \geq \tilde{\mu}_0$




Vergleichsweise häufiges Eintreten des Ereignisses $X_i < \tilde{\mu}_0$ (also ein relativ großer Wert von $T$) spricht gegen die Nullhypothese

# Bemerkungen

Gelegentlich treten auch Messwerte auf, die mit dem hypothetischen Median $\tilde{\mu}_0$ identisch sind

Da diese Beobachtungen nichts über die Abweichung von der Nullhypothese aussagen, werden diese Messwerte zumeist ignoriert

- Der Test wird mit den verbleibenden (und von $\tilde{\mu}_0$ verschiedenen) Messwerten durchgeführt

Dadurch reduziert sich der Stichprobenumfang $n$ um die Anzahl der Fälle, in denen der Messwert gleich $\tilde{\mu}_0$ ist

Wegen der vorausgesetzten Stetigkeit der unterliegenden Verteilung sollten derartige Fälle nicht zu häufig auftreten

# Definition einer Funktion `sign.test()`

Es wird eine Funktion definiert, die einen zweiseitigen Vorzeichentest durchführt und den entsprechenden *p*-Wert zurückgibt

```
## Two-sided sign test
sign.test <- function(x, med0) {
  n <- length(x)
  smaller <- sum(x < med0)
  equal <- sum(x == med0)
  count <- min(smaller, n - equal - smaller)
  2 * pbinom(count, n - equal, 0.5)          # return p-value
}
```

# Beispiel

Für $n = 10$ seien folgende Stichprobenwerte gegeben

$$106, 79, 115, 90, 115, 101, 105, 145, 134, 103$$

Geprüft werden soll mit $\alpha = 0.05$ die Hypothese $\tilde{\mu}_0 = 100$

Die Eingabe

```
x <- c(106, 79, 115, 90, 115, 101, 105, 145, 134, 103)
sign.test(x, med0 = 100)
```

liefert als Ausgabe den $p$-Wert 0.109

# Wilcoxon-Vorzeichen-Rangtest

Der Vorzeichentest leitet in sehr einfacher und direkter Weise die Verteilung der Teststatistik ab, was allerdings dadurch erreicht wird, dass nur die Information genutzt wird, ob ein Stichprobenwert größer oder kleiner als der hypothetische Median $\tilde{\mu}_0$ ist

Wird mehr Information aus den Daten genutzt, dann wird die Ableitung der Verteilung der Teststatistik schwieriger, so wie beim Wilcoxon-Vorzeichen-Rangtest

Die Konstruktion der Teststatistik erfolgt in mehreren Schritten

1. Berechne die Differenzen $D_i = X_i - \tilde{\mu}_0$, $i = 1, \ldots, n$
2. Betrachte den Absolutbetrag der Differenzen $|D_i|$, $i = 1, \ldots, n$
3. Ordne den betragsmäßigen Differenzen Ränge zu, indem der kleinste Betrag den Rang 1 erhält, der nächstgrößere Betrag den Rang 2, ...

# Eigenschaften der Teststatistik

Bezeichnet $rg|D_i|$ den Rang von $|D_i|$, dann betrachtet man die Teststatistik

$$W^+ = \sum_{i=1}^{n} rg|D_i| \cdot Z_i \quad \text{mit} \quad Z_i = \begin{cases} 1 & \text{wenn } D_i > 0 \\ 0 & \text{wenn } D_i < 0 \end{cases}$$

Ist der wahre Median tatsächlich $\tilde{\mu}_0$, so sollten bei symmetrischer Verteilung die Summen der Ränge mit $D_i > 0$ bzw. $D_i < 0$ etwa gleich sein

Für den Erwartungswert von $W^+$ gilt

$$E(W^+) = \frac{n(n+1)}{4}$$

also die Hälfte der totalen Rangsumme $\sum_{i=1}^{n} rg|D_i| = n(n+1)/2$

# Wilcoxon-Vorzeichen-Rangtest

Annahmen

- $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit stetiger und symmetrischer Verteilungsfunktion

Hypothesen

1. $H_0\colon\ \tilde{\mu} = \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} \neq \tilde{\mu}_0$
2. $H_0\colon\ \tilde{\mu} \geq \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} < \tilde{\mu}_0$
3. $H_0\colon\ \tilde{\mu} \leq \tilde{\mu}_0 \qquad H_1\colon\ \tilde{\mu} > \tilde{\mu}_0$

Teststatistik

$$W^+ = \sum_{i=1}^{n} rg|D_i| \cdot Z_i \text{ mit } D_i = X_i - \tilde{\mu}_0 \text{ und } Z_i = \begin{cases} 1 & \text{wenn } D_i > 0 \\ 0 & \text{wenn } D_i < 0 \end{cases}$$

# Wilcoxon-Vorzeichen-Rangtest

Verteilung der Teststatistik unter $\tilde{\mu} = \tilde{\mu}_0$

- Verteilung von $W^+$ in einigen Statistikbüchern tabelliert

Ablehnungsbereich

1. $W^+ < w_{\alpha/2}(n)$ oder $W^+ > w_{1-\alpha/2}(n)$
2. $W^+ < w_{\alpha}(n)$
3. $W^+ > w_{1-\alpha}(n)$

mit dem $p$-Quantil $w_p(n)$ der Verteilung von $W^+$ bei einem Stichprobenumfang von $n$

Approximativer Test

- Für $n > 20$ ist $W^+$ approximativ normalverteilt mit $\mu_w = n(n+1)/4$ und $\sigma_w^2 = n(n+1)(2n+1)/24$

# Ties

Bei *Ties* (Bindungen), d. h. $|d_i| = |d_j|$ für zwei oder mehr Differenzen, werden für die identischen Werte Durchschnittsränge vergeben, z. B.

| | | | | | |
|---|---|---|---|---|---|
| $\lvert d_i \rvert$ | 2.5 | 3.8 | 4.1 | 4.1 | 4.1 |
| $rg\lvert d_i \rvert$ | 1 | 2 | 4 | 4 | 4 |

Wenn $d_i = x_i - \tilde{\mu}_0 = 0$ gilt, wird die Beobachtung weggelassen und der für die Verteilung relevante Stichprobenumfang reduziert sich entsprechend um die Anzahl dieser Fälle

# Beispiel

Für eine Messreihe soll mit $\alpha = 0.05$ die Hypothese $\tilde{\mu}_0 = 100$ geprüft werden

Als daten-erzeugendes Modell wird eine gleichmäßige Verteilung angenommen in dem Intervall $[50, 320]$

Zunächst wird der Stichprobenumfang *n* bestimmt, mit dem eine Abweichung von $\tilde{\mu} - \tilde{\mu}_0 = 85$ mit einer Power von ca. 80% gezeigt werden kann

```
n <- 10
pval <- replicate(2000, {
  x <- runif(n, 50, 320)    # gleichmaessige Vtlg., E(X) = 185
  wilcox.test(x, mu = 100)$p.value
})
mean(pval < 0.05)
```

# Beispiel

Anschließend sei folgende Stichprobe vom Umfang $n = 10$ erhoben worden

$$106, 79, 115, 90, 115, 101, 105, 145, 134, 103$$

Geprüft werden soll mit $\alpha = 0.05$ die Hypothese $\tilde{\mu}_0 = 100$

Für die Differenzen und Ränge von deren Beträgen gilt dann

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 106 | 79 | 115 | 90 | 115 | 101 | 105 | 145 | 134 | 103 |
| $d_i$ | 6 | -21 | 15 | -10 | 15 | 1 | 5 | 45 | 34 | 3 |
| $\lvert d_i \rvert$ | 6 | 21 | 15 | 10 | 15 | 1 | 5 | 45 | 34 | 3 |
| $rg\lvert d_i \rvert$ | 4 | 8 | 6.5 | 5 | 6.5 | 1 | 3 | 10 | 9 | 2 |

# Beispiel

Für die Teststatistik erhält man dann

$$\begin{aligned} w^+ &= 4 \cdot 1 + 8 \cdot 0 + 6.5 \cdot 1 + 5 \cdot 0 + 6.5 \cdot 1 \\ &\quad + 1 \cdot 1 + 3 \cdot 1 + 10 \cdot 1 + 9 \cdot 1 + 2 \cdot 1 \\ &= 42 \end{aligned}$$

Als kritische Werte ermittelt man in einer entsprechenden Tabelle der Verteilung die Werte $w_{0.025}(10) = 9$ und $w_{0.975}(10) = 46$

Wegen

$$9 \leq w^+ = 42 \leq 46$$

kann die Nullhypothese $\tilde{\mu}_0 = 100$ nicht abgelehnt werden

Allerdings hatte die Studie für kleinere Effekte als $\tilde{\mu} - \tilde{\mu}_0 = 85$ auch nicht genügend Power

# Funktion `wilcox.test()`

Mit der Eingabe

```
x <- c(106, 79, 115, 90, 115, 101, 105, 145, 134, 103)
wilcox.test(x, mu = 100, correct = FALSE)
```

erhält man die Ausgabe

```
    Wilcoxon signed rank test

data:  x
V = 42, p-value = 0.1392
alternative hypothesis: true location is not equal to 100
```

# Funktion `qsignrank()`

Die Funktionen `dsignrank()`, `psignrank()`, `qsignrank()`, `rsignrank()` geben die Verteilung, die kumulative Verteilungsfunktion, die Quantile, sowie eine Zufallszahlenziehung aus der Verteilung von $W^+$ an

Für das Beispiel bestimmt man die relevanten Quantile durch

```
qsignrank(c(.025, .975), 10)  # 9 46
```

# Wilcoxon-Rangsummentest

Der Wilcoxon-Rangsummentest kann als Analogon des Wilcoxon-Vorzeichen-Rangtests für den Fall des Vergleichs der Werte zweier Stichproben betrachtet werden

Dabei betrachtet man Stichproben aus zwei Verteilungen, die prinzipiell die gleiche Form besitzen, aber möglicherweise um einen bestimmten Betrag gegeneinander verschoben sind

Die Hypothese der Gleichheit der Mediane der beiden Verteilungen impliziert dann die Identität der Verteilungen

Der Grundgedanke des Tests besteht darin, dass unter diesen Voraussetzungen keine der beiden Stichproben im Vergleich zur anderen eine Tendenz zu besonders großen bzw. besonders kleinen Werten aufweist

# Gepoolte Stichprobe

Die Teststatistik wird aus den Rängen der Realisierungen sämtlicher Stichprobenvariablen $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ gebildet, der *gepoolten Stichprobe*

Man erhält hieraus

$$rg(X_1), \ldots, rg(X_n), rg(Y_1), \ldots, rg(Y_m)$$

Die Teststatistik besteht dann aus der Summe derjenigen Ränge, die zur Stichprobe $X_1, \ldots, X_n$ gehören

Auch in diesem Fall werden bei Ties (Bindungen) $x_i = y_j$ Durchschnittsränge gebildet

# Wilcoxon-Rangsummentest

Annahmen

- $X_1, \ldots, X_n$ sowie $Y_1, \ldots, Y_m$ jeweils unabhängig und identisch verteilt mit stetigen Verteilungsfunktionen und $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ unabhängig

Hypothesen

1. $H_0\colon\ \tilde{\mu}_x = \tilde{\mu}_y \qquad H_1\colon\ \tilde{\mu}_x \neq \tilde{\mu}_y$
2. $H_0\colon\ \tilde{\mu}_x \geq \tilde{\mu}_y \qquad H_1\colon\ \tilde{\mu}_x < \tilde{\mu}_y$
3. $H_0\colon\ \tilde{\mu}_x \leq \tilde{\mu}_y \qquad H_1\colon\ \tilde{\mu}_x > \tilde{\mu}_y$

Teststatistik

$$T_W = \sum_{i=1}^{n} rg(X_i)$$

mit den Rängen $rg(X_i)$ aus der gepoolten Stichprobe $X_1, \ldots, X_n, Y_1, \ldots, Y_m$

# Wilcoxon-Rangsummentest

Verteilung der Teststatistik unter $\tilde{\mu}_x = \tilde{\mu}_y$

- Verteilung von $T_W$ in einigen Statistikbüchern tabelliert

Ablehnungsbereich

1. $T_W < w_{\alpha/2}(n, m)$ oder $T_W > w_{1-\alpha/2}(n, m)$
2. $T_W < w_{\alpha}(n, m)$
3. $T_W > w_{1-\alpha}(n, m)$

mit dem $p$-Quantil $w_p(n, m)$ der Verteilung von $T_W$ bei Stichprobenumfängen von $n$ und $m$

Approximativer Test

- Für $n > 25$ oder $m > 25$ ist $T_W$ approximativ normalverteilt mit $\mu_w = n(n + m + 1)/2$ und $\sigma_w^2 = nm(n + m + 1)/12$

# Beispiel

Es seien folgende Stichprobenwerte gegeben

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x_i$ | 96 | 104 | 87 | 101 | 140 |

und

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y_j$ | 118 | 107 | 101 | 103 | 90 |

Für die Ränge von $x_i$ gilt dann

| | $x_3$ | $y_5$ | $x_1$ | $x_4$ | $y_3$ | $y_4$ | $x_2$ | $y_2$ | $y_1$ | $x_5$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 87 | 90 | 96 | 101 | 101 | 103 | 104 | 107 | 118 | 140 |
| $rg(x_i)$ | 1 | | 3 | 4.5 | | | 7 | | | 10 |

# Beispiel

Für die Teststatistik erhält man dann

$$t_w = 1 + 3 + 4.5 + 7 + 10 = 25.5$$

Für die zweiseitige Fragestellung und $\alpha = 0.10$ ermittelt man als kritische Werte in einer entsprechenden Tabelle der Verteilung der Teststatistik $T_W$ die Quantile $w_{0.05}(5,5) = 20$ und $w_{0.95}(5,5) = 35$

Wegen

$$w_{0.05} \leq t_w \leq w_{0.95}$$

wird die Nullhypothese beibehalten

# Funktion `wilcox.test()`

Für den Vergleich zweier unabhängiger Stichproben wird ein Test angeboten, der sich vom dargestellten Wilcoxon-Rangsummentest unterscheidet, aber dazu äquivalent ist

- Dies liegt daran, dass die Literatur nicht ganz einheitlich in der Bezeichnung dieser Verfahren ist

Die Funktion `wilcox.test()` implementiert das auch als Mann-Whitney-Test bekannte Verfahren

Hierauf beziehen sich auch die verfügbaren Funktionen `dwilcox()`, `pwilcox()`, `qwilcox()`, `rwilcox()`

# Mann-Whitney-Test

Die traditionellerweise mit $U$ bezeichnete Teststatistik des Mann-Whitney-Tests erhält man aus der des Wilcoxon-Rangsummentests durch

$$U = T_W - \frac{n\,(n+1)}{2},$$

wobei $n(n+1)/2 = 1 + \ldots + n$ das Minimum der Rangplatzsumme für die erste Stichprobe angibt

Die Zahl $U$ korrespondiert mit der Anzahl aller Paare $(x_i, y_j)$ mit $i = 1, \ldots, n$ und $j = 1, \ldots, m$, für die $x_i \geq y_j$ gilt (Gewichtung mit 0.5 bei Ties)

Für das oben berechnete Beispiel sollte sich also die Teststatistik $U = T_W - 15 = 10.5$ ergeben

# Funktion `wilcox.test()`

Für die betrachteten Daten ergibt sich mit der Eingabe

```
x <- c(96, 104, 87, 101, 140)
y <- c(118, 107, 101, 103, 90)
wilcox.test(x, y, correct = FALSE)
```

als Ergebnis

```
    Wilcoxon rank sum test

data:  x and y
W = 10.5, p-value = 0.6752
alternative hypothesis: true location shift is not equal to 0
```

# Funktion `qwilcox()`

Die Funktionen `dwilcox()`, `pwilcox()`, `qwilcox()`, `rwilcox()` geben die Verteilung, die kumulative Verteilungsfunktion, die Quantile, sowie eine Zufallszahlenziehung aus der Verteilung der Teststatistik des Mann-Whitney-Tests an

Für das Beispiel bestimmt man die relevanten Quantile durch

```
qwilcox(0.05, 5, 5)
qwilcox(0.95, 5, 5)
```

was als Ergebnis die Werte 5 und 20 liefert