Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Wilcoxon-Test für abhängige Stichproben

Wie bei der Betrachtung des t-Tests für abhängige Stichproben wird die Situation auch hier auf den Ein-Stichproben-Fall zurück geführt

Ausgehend von den Paaren (X_i, Y_i), für $i = 1, \ldots, n$ geht man daher folgendermaßen vor

1. Berechne die Differenzen $D_i = X_i - Y_i$, $i = 1, \ldots, n$
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, ...

Auch in diesem Fall bildet die Summe der Ränge $rg|D_i|$ für alle $D_i > 0$ eine Statistik, die sensitiv hinsichtlich der Unterschiede zwischen den gepaarten Werten ist

Wilcoxon-Test für abhängige Stichproben

Annahmen

- $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$ Stichproben mit $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ unabhängig und identisch verteilt mit stetiger und symmetrischer Verteilungsfunktion

Hypothesen

1. $H_0\colon\ \tilde{\mu}_D = 0 \qquad H_1\colon\ \tilde{\mu}_D \neq 0$
2. $H_0\colon\ \tilde{\mu}_D \geq 0 \qquad H_1\colon\ \tilde{\mu}_D < 0$
3. $H_0\colon\ \tilde{\mu}_D \leq 0 \qquad H_1\colon\ \tilde{\mu}_D > 0$

Teststatistik

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

Wilcoxon-Test für abhängige Stichproben

Verteilung der Teststatistik unter $\tilde{\mu}_D = 0$

- 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 $\sigma_w^2 = n(n+1)(2n+1)/24$

Beispiel: Teststärke und Versuchsplanung

Für das obige Beispiel zweier "paralleler" Formen X und Y eines Intelligenztests bestimmen wir den Stichprobenumfang n, um mit ca. 80% Power einen Effekt von $\tilde{\mu}_D = 7$ IQ-Punkten bei $\sigma_x = \sigma_y = 15$, $\rho = 0.9$ und $\alpha = 0.1$ zeigen zu können,
$H_0\colon\ \tilde{\mu}_D = 0$

```
n <- 8                                                  # Stichprobenumfang
s <- 15; r <- 0.9; sxy <- r*s*s

pval <- replicate(2000, {
  x <- MASS::mvrnorm(n,
            mu = c(100 + 7, 100),                       # Erwartungswertvektor
         Sigma = matrix(c(s^2, sxy, sxy, s^2), 2, 2)    # Kovarianzmatrix
  )
  wilcox.test(x[, 1], x[, 2], mu = 0, paired = TRUE)$p.value
})
mean(pval < 0.1)
```

Beispiel

Anschließend bearbeitet eine Stichprobe von insgesamt acht Probanden die beiden "parallelen" Formen X und Y des Intelligenztests und es resultieren die IQ-Werte

x	97	87	116	137	88	95	111	87
y	97	85	114	134	102	103	108	92

Bei der Funktion `wilcox.test()` wird die Option `paired = TRUE` gewählt (wie bei der Funktion `t.test()`)

```
x <- c(97, 87, 116, 137, 88, 95, 111, 87)
y <- c(97, 85, 114, 134, 102, 103, 108, 92)

wilcox.test(x, y, paired = TRUE, correct = FALSE)
wilcox.test(x - y, correct = FALSE)  # aequivalenter 1-SP-Test
```

Beispiel

Als Ausgabe erhält man

```
    Wilcoxon signed rank test

data:  x and y
V = 10, p-value = 0.4974
alternative hypothesis: true location shift is not equal to 0
```

Bei der Durchführung des entsprechenden t-Tests für abhängige Stichproben bei einer zweiseitigen Fragestellung resultierte für die gegebenen Daten ein p-Wert von 0.369

Auch hier wird die Nullhypothese beibehalten, ein Unterschied zwischen den beiden Formen kann mit den Daten nicht nachgewiesen werden

Allerdings hatte die Studie für kleinere Effekte als $\mu_D = 7$ IQ-Punkte auch nicht genügend Power

Randomisierungstests: Überblick

- Randomisierungstests: Grundlagen
- Beispiel: zwei unabhängige Stichproben
- Beispiel: zwei abhängige Stichproben
- Umsetzung in R
- Schlussbemerkungen und Literatur

Beispiel: unabhängige Stichproben

Die Studienanfänger im Fach Psychologie werden zufällig einer von zwei parallel stattfindenden Einführungsveranstaltungen X und Y zugewiesen, in denen jeweils unterschiedliche Lehrmethoden verwendet werden

Nach Abschluss der Veranstaltung absolvieren alle Studierenden denselben Wissenstest

Jeweils drei Studierende aus jeder Lehrveranstaltung erzielten folgende Punktezahlen

Lehrveranstaltung					
X			Y		
A	B	C	D	E	F
16	11	15	9	10	14
$\bar{x} = 14.00$			$\bar{y} = 11.00$		

Grundlagen parametrischer statistischer Tests

In der klassischen Inferenzstatistik beziehen sich die statistischen Hypothesen auf die Verteilung des interessierenden Merkmals in einer Grundgesamtheit (Population) von Untersuchungseinheiten (bzw. auf deren Parameter)

Die Überprüfung statistischer Hypothesen erfolgt über die Ziehung einer (einfachen) Zufallsstichprobe (random sampling)

- Jedes Element der Grundgesamtheit hat dabei die gleiche Wahrscheinlichkeit in die Stichprobe aufgenommen zu werden

Ausgehend von Stichprobenkennwerten sollen dann Inferenzen bezüglich der Populationsparameter gezogen werden

Hierzu wird aus der Verteilung in der Population die Verteilung der Stichprobenkennwerte abgeleitet (sampling distribution)

Parametrischer Test

Die Frage, ob sich die in beiden Veranstaltungen verwendeten Lehrmethoden in ihrer Effektivität unterscheiden, kann mit einem t-Test für unabhängige Stichproben überprüft werden, gegeben dessen Annahmen sind erfüllt

- Für die angegebenen Stichproben errechnet man

$$t = 1.389$$

  bei $df = 6 - 2 = 4$ Freiheitsgraden
- Der unter der Nullhypothese identischer Erwartungswerte über die entsprechende t-Verteilung ermittelte p-Wert beträgt

$$p = 0.237$$

Problematik parametrischer statistischer Tests

Die Annahmen zur Verteilung in der Population, als zentrale Grundlage parametrischer statistischer Tests, sind häufig nicht erfüllt oder unplausibel

- Die abhängige Variable in obigem Beispiel ist als Anzahl eine diskrete Zufallsvariable und kann daher streng genommen nicht als normalverteilt angenommen werden
- Diesem Problem wird für gewöhnlich in zweierlei Weise begegnet
  - Mit Hilfe von Simulationen soll geklärt werden, wie robust die jeweiligen statistischen Tests auf Verletzungen der unterliegenden Verteilungsannahmen reagieren
  - Nonparametrische Tests werden unter Verzicht auf Information (z. B. ausschließliche Betrachtung der Vorzeichen von Mittelwertsdifferenzen) durchgeführt

Problematik parametrischer statistischer Tests

Die Inferenz von Stichprobenkennwerten auf Populationsparameter, beruht entscheidend auf der Ziehung einer Zufallsstichprobe

> *Häufig sind bei sozialwissenschaftlichen Forschungsfragen nicht alle Untersuchungseinheiten, die zu einer Population gehören, bekannt, so dass die Ziehung einer "echten" Zufallsstichprobe unmöglich oder doch zumindest mit einem unzumutbaren Aufwand verbunden ist. Man begnügt sich deshalb gelegentlich mit sog. "anfallenden" oder "ad hoc"-Stichproben (...) in der Hoffnung, auch so zu aussagefähigen Resultaten zu gelangen. Vor dieser Vorgehensweise sei nachdrücklich gewarnt.* *(Bortz, 1999, S. 87)*

Redefinition der entsprechenden Population als Ausweg?

> *...letztlich lässt sich für jede "Stichprobe" eine fiktive Population konstruieren, für die diese "Stichprobe" repräsentativ erscheinen mag. Die Schlüsse, die aus derartigen Untersuchungen gezogen werden, beziehen sich jedoch nicht auf real existierende Populationen und können deshalb wertlos sein.* *(Bortz, 1999, S. 87)*

Randomisierungstests: Grundlagen

Ebenso wie die klassische Inferenzstatistik geht auch die Logik der Randomisierungstests (oder Permutationstests) auf R. A. Fisher zurück

Logik der Randomisierungstests

- Für die statistische Überprüfung des Effekts einer experimentellen Manipulation wird nicht die Inferenz auf eine Grundgesamtheit der Untersuchungseinheiten als grundlegend erachtet
- Für eine Absicherung des Effekts wird die zufällige oder randomisierte Zuordnung (random assignment) der Untersuchungseinheiten zu den experimentellen Bedingungen als entscheidend betrachtet
- Wenn die experimentelle Manipulation keinerlei Effekt hat, dann hängen die beobachteten Unterschiede nur von der verwendeten Zuweisung ab

Randomisierungstests: Grundlagen

Nullhypothese

- Die experimentellen Bedingungen haben keinen differentiellen Effekt

Vorgehensweise

- Für gegebene Daten wird der Wert einer interessierenden Statistik berechnet (z. B. t-Statistik, Mittelwertsdifferenz, p-Wert, ...)
- Unter der Nullhypothese hängt dieser Wert nur von der vorgenommenen Zuweisung der Untersuchungseinheiten zu den experimentellen Bedingungen ab, und bei einer anderen Zuweisung ergibt sich im Allgemeinen ein anderer Wert
- Man betrachtet daher alle möglichen Zuweisungen der Untersuchungseinheiten zu den experimentellen Bedingungen und berechnet hierfür jeweils den Wert der Teststatistik

Randomisierungstests: Grundlagen

Testentscheidung

- Aus der Betrachtung aller möglichen randomisierten Zuordnungen der Untersuchungseinheiten resultiert für gegebene Daten eine Verteilung der Teststatistik
- Mit dieser Verteilung lässt sich die Wahrscheinlichkeit bestimmen, dass unter der Nullhypothese ein Wert auftritt, dessen Absolutbetrag mindestens so groß ist wie der Absolutbetrag des erhaltenen Wertes der Teststatistik (entsprechend für einseitige Fragestellung)
- Zur exakten Berechnung dieser Wahrscheinlichkeit bestimmt man die relative Häufigkeit derjenigen Zuordnungen, bei denen der Absolutbetrag des entsprechenden Werts größer oder gleich dem Absolutbetrag des erhaltenen Wertes ist (entsprechend für einseitige Fragestellung)

Beispiel: unabhängige Stichproben

Effekt der Zuweisung der Versuchspersonen zu den Lehrveranstaltungen

- Daten

X			Y		
A	B	C	D	E	F
16	11	15	9	10	14
$\bar{x} = 14.00$			$\bar{y} = 11.00$		

- Nach Permutation

X			Y		
B	C	D	A	E	F
11	15	9	16	10	14
$\bar{x} = 11.67$			$\bar{y} = 13.33$		

Wieviele mögliche Zuordnungen existieren?

Wir betrachten das Zufallsexperiment der randomisierten Zuordnung der Versuchspersonen zu den Lehrveranstaltungen als Urnenziehung ohne Zurücklegen

- Für die Lehrveranstaltung X werden aus den insgesamt sechs Versuchspersonen genau drei Versuchspersonen ausgewählt
- Der Lehrveranstaltung Y werden dann die verbleibenden Versuchspersonen zugewiesen
- Die Anzahl der möglichen Zuordnungen ergibt sich daher über den Binomialkoeffizient

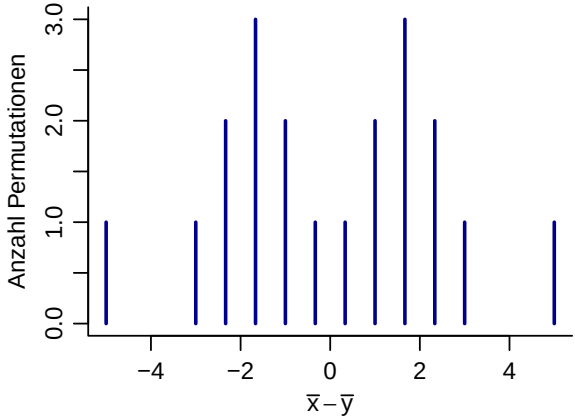
$$\binom{6}{3} = 20$$

Beispiel

Unabhängige Stichproben

Permutation	X	Y	$\bar{x}$	$\bar{y}$	$\bar{x} - \bar{y}$	t	p
1	ABC	DEF	14.00	11.00	**3.00**	**1.389**	**.237**
2	ABD	CEF	12.00	13.00	-1.00	-0.387	.718
3	ABE	CDF	12.33	12.67	-0.33	-0.127	.905
4	ABF	CDE	13.67	11.33	2.33	0.990	.378
5	ACD	BEF	13.33	11.67	1.67	0.668	.541
6	ACE	BDF	13.67	11.33	2.33	0.990	.378
7	ACF	BDE	15.00	10.00	**5.00**	**6.124**	**.004**
8	ADE	BCF	11.67	13.33	-1.67	-0.668	.541
9	ADF	BCE	13.00	12.00	1.00	0.387	.718
10	AEF	BCD	13.33	11.67	1.67	0.668	.541
11	BCD	AEF	11.67	13.33	-1.67	-0.668	.541
12	BCE	ADF	12.00	13.00	-1.00	-0.387	.718
13	BCF	ADE	13.33	11.67	1.67	0.668	.541
14	BDE	ACF	10.00	15.00	**-5.00**	**-6.124**	**.004**
15	BDF	ACE	11.33	13.67	-2.33	-0.990	.378
16	BEF	ACD	11.67	13.33	-1.67	-0.668	.541
17	CDE	ABF	11.33	13.67	-2.33	-0.990	.378
18	CDF	ABE	12.67	12.33	0.33	0.127	.905
19	CEF	ABD	13.00	12.00	1.00	0.387	.718
20	DEF	ABC	11.00	14.00	**-3.00**	**-1.389**	**.237**

Stichprobenverteilung



Beispiel: unabhängige Stichproben

Testentscheidung

- In 4 von insgesamt 20 Fällen ist der Absolutbetrag der t-Statistik größer oder gleich dem Absolutbetrag des erhaltenen Wertes von $t = 1.389$
- Damit erhält man die exakte Überschreitungswahrscheinlichkeit

$$p = 0.2$$

($p = 0.237$ bei t-Test für unabhängige Stichproben mit $df = 4$)

Äquivalenz von Teststatistiken

- Die Ordnung der Permutationen bezüglich der Werte der Mittelwertsdifferenz, der t-Statistik oder des p-Werts ist immer dieselbe
- Die drei Statistiken sind daher unter der Nullhypothese äquivalent

Beispiel: abhängige Stichproben

Wir betrachten das oben dargestellte Beispiel der Werte, die aus der Bearbeitung zweier "paralleler" Formen X und Y eines Intelligenztests durch 8 Probanden resultiert

x	97	87	116	137	88	95	111	87
y	97	85	114	134	102	103	108	92

Geprüft werden soll für $\alpha = 0.10$ wiederum die Hypothese, dass sich die aus beiden Testformen resultierenden IQ-Werte nicht unterscheiden

Unter der Nullhypothese wird angenommen, dass die beiden Testformen für jeden Probanden denselben Wert liefern

- Es ist also unerheblich, ob ein Testwert von der Testform X oder der Testform Y stammt, sodass bei einer Vertauschung der beiden Werte dasselbe Testergebnis resultieren sollte

Vorzeichenwechsel

Eine für jeden Probanden unabhängig durchgeführte Permutation der x- und y-Werte bedeutet für deren Differenzen lediglich einen Vorzeichenwechsel

Ausgehend von den Differenzwerten $d_i = x_i - y_i$ für alle $i = 1, \ldots, 8$

$$0, +2, +2, +3, -14, -8, +3, -5$$

betrachtet man beispielsweise auch die Wertereihen

$$0, -2, -2, +3, +14, -8, +3, -5$$

oder

$$0, -2, +2, +3, -14, +8, -3, +5$$

Anzahl der Permutationen

Da jeder Differenzwert entweder positiv oder negativ sein kann, ergeben sich insgesamt 2^n Wertereihen

- Ist (mindestens) einer der Differenzwerte d_i, $i = 1, \ldots, n$, gleich Null, dann sind die Wertereihen nicht alle voneinander verschieden, was aber die nachfolgenden Berechnungen nicht beeinträchtigt

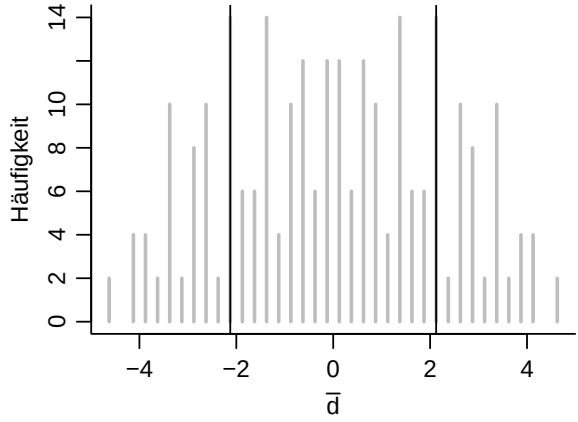
Für jede der 2^n Wertereihen wird die interessierende Statistik (z. B. Summe, Mittelwert, t-Statistik) bestimmt und man betrachtet den Anteil der Wertereihen, für die die Statistik einen mindestens so extremen Wert annimmt, wie in den erhaltenen Daten

Für die gegebenen Daten berechnet man

$$\sum_i d_i = -17, \quad \bar{d} = -2.125, \quad t = -0.959$$

Stichprobenverteilung

Für die Statistik $\bar{d}$ ergeben sich folgende Häufigkeiten



Testentscheidung

Zur Beantwortung der zweiseitigen Fragestellung berechnet man nun den Anteil derjenigen Wertereihen, für die der Absolutbetrag des erhaltenen Mittelwerts mindestens 2.125 beträgt

- Der erhaltene Anteil entspricht dem p-Wert und beträgt im vorliegenden Fall

$$p = 0.453$$

Die Nullhypothese wird also beibehalten, ein Unterschied zwischen den beiden Formen kann mit den Daten nicht nachgewiesen werden

Randomisierungstests: praktische Umsetzung

Die in der Vergangenheit als rechenintensiv qualifizierten Randomisierungstests sind durch die heute zur Verfügung stehenden Rechenkapazitäten im Allgemeinen leicht durchführbar

- Es wurden Algorithmen entwickelt und implementiert, die es gestatten, die Verteilung der betrachteten Teststatistik auch für eine große Zahl von Kombinationen/Permutationen exakt zu berechnen
- Können aus Zeit- oder Kapazitätsgründen nicht sämtliche Kombinationen/Permutationen explizit berechnet werden, so kann man die Verteilung der Teststatistik durch Stichprobenziehung aus der Menge der möglichen Kombinationen/Permutationen approximieren
- Für die Durchführung von Randomisierungstests stehen entsprechende Funktionen z. B. in den R-Paketen `BHH2`, `DAAG` und `coin` zur Verfügung

Randomisierungstests in R

Grundsätzlich kann man derartige Tests mit Hilfe bestehender Funktionen, die sämtliche zu berücksichtigende Kombinationen oder Permutationen berechnen, auch selbst programmieren (zumindest für beschränkte Stichprobenumfänge)

- Alle Kombinationen bzw. Permutationen werden mit den Funktionen `combn()` bzw. `expand.grid()` erzeugt oder mit `gtools::combinations()` bzw. `gtools::permutations()`

Ohne die (aufwendige) exakte Bestimmung aller Permutationen lassen sich gerade bei einfachen Versuchsplänen approximative Tests mit den Funktionen `sample()` und `replicate()` leicht umsetzen

Approximativer Test für unabhängige Stichproben

```
## Wissensscore und Lehrmethode
dat <- read.table(text = "
  score grp
  A 16 X
  B 11 X
  C 15 X
  D  9 Y
  E 10 Y
  F 14 Y
")
p0 <- t.test(score ~ grp, dat)$p.value  # beobachtete Statistik

## Zufaellige Permutation des Scores
pval <- replicate(5000,
                  t.test(sample(score) ~ grp, dat)$p.value
)
mean(pval <= p0)  # p-Wert des Randomisierungstests
[1] 0.194
```

Exakter Test für abhängige Stichproben

```
## Parallele Intelligenztests
x <- c(97, 87, 116, 137, 88, 95, 111, 87)
y <- c(97, 85, 114, 134, 102, 103, 108, 92)

n <- length(x)
d0 <- mean(x - y)  # beobachtete Statistik
d <- d0            # liefert den ersten Fall
```

Exakter Test

```
# Bildung aller Kombinationen moegl. Vorzeichenwechsel
# Mittelwertsdifferenzen in Vektor d speichern

for(k in 1:n) {
  cmb <- combn(n, k)
  for(i in 1:ncol(cmb)) {          # fuer jede Kombination
    sign <- rep(1, n)
    sign[cmb[, i]] <- -1           # Vorzeichen wechseln
    d <- c(d, mean(sign*(x - y)))
  }
}

## Zweiseitige Fragestellung H0: mu_x = mu_y
p <- mean(d <= -abs(d0) | d >= abs(d0))

plot(table(d), xlab = expression(bar(d)), ylab = "Haeufigkeit")
abline(v = c(-d0, d0))
```

Approximativer Test

```
## Entscheidung per Mittelwertdifferenz
d <- replicate(5000, {
  sign <- sample(c(-1, 1), n, replace = TRUE)  # zufaelliger Vorzeichenvektor
  mean(sign*(x - y))
})

## Aequivalente Entscheidung per p-Wert
p0 <- t.test(x - y)$p.value   # beobachtete Statistik
pval <- replicate(5000, {
  sign <- sample(c(-1, 1), n, replace = TRUE)
  t.test(sign * (x - y))$p.value
})
mean(pval <= p0)
```

Schlussbemerkungen

Im Gegensatz zur klassischen Inferenzstatistik ist das zentrale Konzept der Randomisierungstests nicht die Zufallsstichprobe (random sampling), sondern die randomisierte Zuweisung (random assignment) zu einer Bedingung

- Damit wird unmittelbar die (etwa im Kontext von Between-Subjects-Designs verwendete) Randomisierung von Untersuchungseinheiten im zugrunde liegenden Experiment durch den statistischen Test berücksichtigt

Randomisierungstests werden auch als bedingte Tests (conditional tests) bezeichnet, da die Verteilung der Teststatistik von den Daten abhängt und nicht vorab tabelliert werden kann

- Diese Verteilungen können exakt berechnet werden bzw. lassen sich über Stichprobenziehung aus der Menge der Permutationen approximieren
- Darüber hinaus liegen auch theoretische Ergebnisse zur asymptotischen Verteilung derartiger Teststatistiken vor

Schlussbemerkungen

Bereits R. A. Fisher (1935) betrachtete Randomisierungstests als die geeignetsten statistischen Hypothesentests, die jedoch auf Grund der damals zur Verfügung stehenden Rechenkapazitäten nicht praktisch eingesetzt werden konnten

- Zur Approximation der exakten Wahrscheinlichkeiten aus den Randomisierungstests schlug Fisher daher die Durchführung parametrischer statistischer Tests vor

Seit Fisher (1935), Pitman (1937, 1938) und Welch (1938) ist die theoretische Begründung der Randomisierungstests weitgehend vollständig entwickelt

Neuere Entwicklungen (z. B. Strasser & Weber, 1999) verbreitern den theoretischen Rahmen und ermöglichen daher die Anwendung derartiger Tests auf eine Vielzahl von Fragestellungen

Schlussbemerkungen

Die besprochenen nonparametrischen Tests, die sich auf Ranginformationen beziehen, sind eigentlich Randomisierungstests

- Durch Beschränkung auf die Ränge der erhaltenen Stichprobenwerte sind die möglichen Ausgänge des durchgeführten Zufallsexperiments bei festgelegtem Stichprobenumfang vorab bekannt
- Daher lassen sich die darauf bezogenen Verteilungen der Teststatistiken auch tabellieren

Aus heutiger Sicht steht einer Anwendung der Randomisierungstests auf die tatsächlich erhaltenen Stichprobenwerte (ohne Verzicht auf vorliegende Information) nichts mehr im Wege

Literatur

Edgington, E. S. (1995). *Randomization tests* (3rd Ed.). New York: Dekker.

Good, P. I. (2000). *Permutation Tests* (2nd Ed.). New York: Springer.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury.

Lunneborg, C. E. (2000). Random assignment of available cases: Let the inference fit the design.