Forschungsmethodik: Interventions- und Evaluationsforschung theoretisch begründen

Vorlesung Biostatistik, WS 2026/2027

Dr. Marcus Wagner, IMISE

Vorlesung IV am 03.11.2026

Inhaltsverzeichnis

6. Beziehungen zwischen metrischen Variablen, parametrische Tests

6.8. Wiederholung

6.9. Einfaktorielle ANOVA

6.10. Zur Problematik des multiplen Testens

7. Nichtparametrische Hypothesentests

7.1. Grundlagen

7.2. Stochastic superiority

7.3. Der Mann-Whitney-U-Test

7.4. Wilcoxon-Test für verbundene Stichproben

7.5. Bemerkungen zum Kruskal-Wallis-Test

Wiederholung: Klassischer t-Test

• **t-Test für zwei unabhängige Stichproben $\mathcal{S}'$, $\mathcal{S}''$**

— Voraussetzungen: metrische Variable x in $\mathcal{P}'$, $\mathcal{P}''$ normalverteilt, Standardabweichungen in $\mathcal{P}'$, $\mathcal{P}''$ gleich

— Zutaten: Umfänge N', N'' / Stichproben-Mittelwerte $\widehat{\mu}'$, $\widehat{\mu}''$ / empirische Standardabweichungen $\widehat{\sigma}'$, $\widehat{\sigma}''$

— Prüfgröße: $\widehat{T} = \dfrac{\widehat{\mu}' - \widehat{\mu}''}{\widehat{\sigma}_p \cdot \sqrt{\dfrac{1}{N'} + \dfrac{1}{N''}}}$ *mit* $\widehat{\sigma}_p = \sqrt{\dfrac{(N'-1)\cdot(\widehat{\sigma}')^2 + (N''-1)\cdot(\widehat{\sigma}'')^2}{(N'+N''-2)}}$.

— Verteilung der Prüfgröße: t-Verteilung mit df $= (N' + N'' - 2)$ Freiheitsgraden

— Nullhypothese und Ablehnbereich je nach Fragestellung (zweiseitig, einseitig, ...) und Signifikanzniveau

— Robustheit: mindestens als zweiseitiger Welch-Test mit $N' \approx N'' \geqslant 30$ ziemlich unempfindlich gegen Verletzung der Voraussetzungen

Fitneß bei Herzkrankheit

• **Beispiel 4.1.** *(aus einer abgeschlossenen Interventionsstudie) Der allgemeine Fitneß-zustand herzkranker Personen wird nach der sog. NYHA-Klasse in vier Stufen eingeteilt:*

• *Zitat aus "Leitlinie"* [DGK (2013)]

Klasse I	Keine Einschränkung der körperlichen Aktivität. Normale körperliche Aktivität führt nicht zu Luftnot, Müdigkeit oder Palpitationen.
Klasse II	Leichte Einschränkung der körperlichen Aktivität. Beschwerdefreiheit unter Ruhe-bedingungen; aber bei normaler körperlicher Aktivität kommt es zu Luftnot, Müdigkeit oder Palpitationen.
Klasse III	Deutliche Einschränkung der körperlichen Aktivität. Beschwerdefreiheit unter Ruhebedingungen; aber bereits bei geringer physischer körperlicher Aktivität Auftreten von Luftnot, Müdigkeit oder Palpitationen.
Klasse IV	Unfähigkeit, körperliche Aktivität ohne Beschwerden auszuüben. Symptome unter Ruhebedingungen können vorhanden sein. Jegliche körperliche Aktivität führt zur Zunahme der Beschwerden.

• *Konkreter: 6-Minuten-Gehtest — in 6 Minuten zurückgelegte Wegstrecke in Metern*

• *Daten (Baseline-Untersuchung): 75 von insgesamt 80 Teilnehmern haben den Gehtest absolviert. Verteilung auf die NYHA-Klassen: Klasse I — 14 Personen / Klasse II — 34 Personen / Klasse III — 27 Personen / Klasse IV — niemand.*

• **Frage:** *Gibt es einen signifikanten Unterschied zwischen den mittleren Wegstrecken bei den Teilnehmern mit NYHA-Klassen I, II und III?*

Wie verfährt man bei drei oder mehr Stichproben?

• **Drei unabhängige Stichproben** *(nach dem "Faktor" NYHA-Klasse aufgeteilt)*

— $\mathcal{S}_{(1)}$ *mit* $N_{(1)} = 14$, $\widehat{\mu}_{(1)} = 445.9\,m$, $\widehat{\sigma}_{(1)} = 82.6\,m$ *(NYHA-Klasse I)*
— $\mathcal{S}_{(2)}$ *mit* $N_{(2)} = 34$, $\widehat{\mu}_{(2)} = 437.2\,m$, $\widehat{\sigma}_{(2)} = 62.0\,m$ *(NYHA-Klasse II)*
— $\mathcal{S}_{(3)}$ *mit* $N_{(3)} = 27$, $\widehat{\mu}_{(3)} = 359.9\,m$, $\widehat{\sigma}_{(3)} = 110.9\,m$ *(NYHA-Klasse III)*

• **Warum nicht mehrfach testen?**

— Denkbar wären drei zweiseitige t-Tests, alle zu demselben Signifikanzniveau, etwa $\alpha = 0.05$: $\mathcal{S}_{(1)}$ vs. $\mathcal{S}_{(2)}$ / $\mathcal{S}_{(1)}$ vs. $\mathcal{S}_{(3)}$ / $\mathcal{S}_{(2)}$ vs. $\mathcal{S}_{(3)}$
— Konsequenz: Weil die drei Stichproben unabhängig sind, steigt die Wahrscheinlichkeit für falsch-positives Gesamtergebnis bereits auf 14.3 % (wegen $1 - 0.95^3 = 0.143$)
— Verringert man von vornherein das Signifikanzniveau, so wächst stattdessen die Wahrscheinlichkeit β für einen Fehler II. Art
— Noch ungünstiger für 4 Stichproben (6 Tests), 5 Stichproben (10 Tests), 6 Stichproben (15 Tests), usw. $\Longrightarrow$ anderes Vorgehen nötig!

Einfaktorielle ANOVA: Grundprinzip

• **ANOVA: zweistufiges Verfahren**

Schritt 1: Globaltest: *Sind die Mittelwerte der drei Gruppen alle gleich?*

— *ANOVA ("analysis of variance"): indirekter Vergleich der drei Stichproben-Mittelwerte nach passender Zerlegung der Gesamt-Stichprobenvarianz*

— *Test gegen die Nullhypothese* $\mu_{(1)} = \mu_{(2)} = \mu_{(3)}$

— *zum ursprünglichen Signifikanzniveau* α

Schritt 2: Post-hoc-Tests *zum paarweisen Vergleich der Stichproben: welche Mittelwerte unterscheiden sich tatsächlich, und wie stark?*

— *nur dann, wenn der Globaltest signifikant ausfiel,*

— *hier dürfen jetzt* t*-Tests durchgeführt werden,*

— *aber mit korrigiertem Signifikanzniveau* $\alpha/3$*.*

• **Bezeichnungen:** *Index für die Gruppen:* r *mit* $1 \leqslant r \leqslant R$*, Index für die Teilnehmer in den Gruppen:* n *mit* $1 \leqslant n \leqslant N_{(r)}$*, Gehstrecken:* $Z(n, r)$

Zerlegung der Gesamt-Varianz

• **Gruppen:** $\mathcal{S}_{(1)}$, $\mathcal{S}_{(2)}$, ... , $\mathcal{S}_{(R)}$ *vom Umfang* $N_{(1)}$, $N_{(2)}$, ... , $N_{(R)}$ *mit Mittelwerten* $\widehat{\mu}_{(1)}$, $\widehat{\mu}_{(2)}$, ... , $\widehat{\mu}_{(R)}$ *und Varianzen* $\widehat{\sigma}^2_{(1)}$, $\widehat{\sigma}^2_{(2)}$, ... , $\widehat{\sigma}^2_{(R)}$

• **Gesamt:** *Alle Gruppen poolen:* $\mathcal{S} = \mathcal{S}_{(1)} \cup ... \cup \mathcal{S}_{(R)}$, *Umfang* $N = N_{(1)} + ... + N_{(R)}$, *Mittelwert* $\widehat{\mu} = \sum_{n,\,r} Z(n,r)/N$ *("grand mean") und Gesamtvarianz* $\widehat{\sigma}^2_{\mathit{Gesamt}}$

Bei der Zerlegung rechnet man bequemerweise mit Quadratsummen:

• **"Sum of Squares" (Gesamt):**

$$SS_{\mathit{Gesamt}} \quad = \sum_n \sum_r \big(Z(n,r) - \widehat{\mu} \big)^2 = (N-1) \cdot \widehat{\sigma}^2_{\mathit{Gesamt}} = SS_{\mathit{ZwSubj}}$$

wird zerlegt in

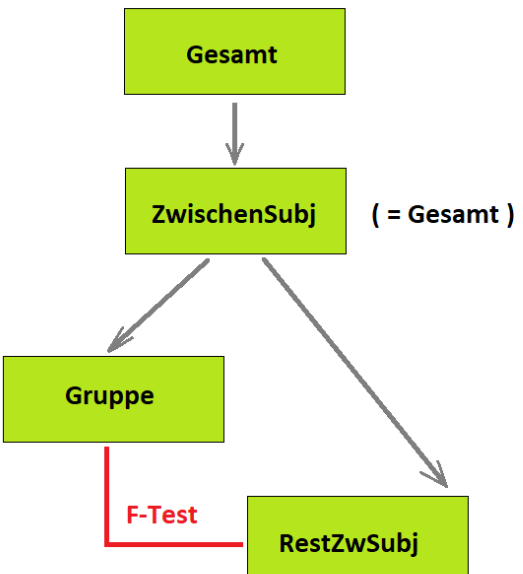
• **Zwischenfaktoranteil:**

$$SS_{\mathit{Gruppe}} \quad = \sum_r N_{(r)} \cdot \big(\widehat{\mu}_{(r)} - \widehat{\mu} \big)^2 = (R-1) \cdot \widehat{\sigma}^2_{\mathit{Gruppe}} \text{ und}$$

• **Innerfaktoranteil:**

$$SS_{\mathit{RestZwSubj}} = SS_{\mathit{Gesamt}} - SS_{\mathit{Gruppe}} = (N-R) \cdot \widehat{\sigma}^2_{\mathit{RestZwSubj}} = \sum_r \big(N_{(r)} - 1 \big) \cdot \widehat{\sigma}^2_{(r)}$$

Zerlegung der Gesamt-Varianz (II)



- **Prüfgröße:** *das Verhältnis*

$$\widehat{T} = \widehat{\sigma}^2_{Gruppe} \,/\, \widehat{\sigma}^2_{RestZwSubj}$$

- **Schlußfolgerung:**

Wenn $\widehat{\mu}_{(1)} = \ldots = \widehat{\mu}_{(R)}$,

dann $\widehat{\mu} = \widehat{\mu}_{(1)} = \ldots = \widehat{\mu}_{(R)}$,

dann SS_{Gruppe}

$$= \sum_r N_{(r)} \cdot \left(\widehat{\mu}_{(r)} - \widehat{\mu}\right)^2$$

$$= (R-1) \cdot \widehat{\sigma}^2_{Gruppe} = 0,$$

also $\widehat{\sigma}^2_{Gruppe} = 0$,

und damit $\widehat{T} = 0$.

- $\widehat{T}$ *ermöglicht den simultanen Vergleich aller Stichproben-Mittelwerte!*

Einfaktorielle ANOVA: Prüfgröße

- **Rechenschema** *zur Bestimmung der Prüfgröße:*

	Zähler	*Nenner*	*Anteil*	*Prüfgröße* $\widehat{T}$
Zwischenfakt.:	$\sum_r N_{(r)} \cdot \widehat{\mu}^2_{(r)} - N \cdot \widehat{\mu}^2$	$(R-1)$	$\widehat{\sigma}^2_{\mathit{Gruppe}}$	$\widehat{\sigma}^2_{\mathit{Gruppe}}/\widehat{\sigma}^2_{\mathit{RestZwSubj}}$
Innerfakt.:	$\sum_r (N_{(r)} - 1) \cdot \widehat{\sigma}^2_{(r)}$	$(N-R)$	$\widehat{\sigma}^2_{\mathit{RestZwSubj}}$	
Kontrolle:	$\sum_n \sum_r Z(n,r)^2 - N \cdot \widehat{\mu}^2$	$(N-1)$		

- **Einsetzen** *der Werte aus Beispiel 4.1.:*

Zwischenfakt.:	111 015.3	2	55 507.6	7.5661
Innerfakt.:	535 551.7	72	7 336.3	
Kontrolle:	646 567.0	74		

Einfaktorielle ANOVA: Globaltest

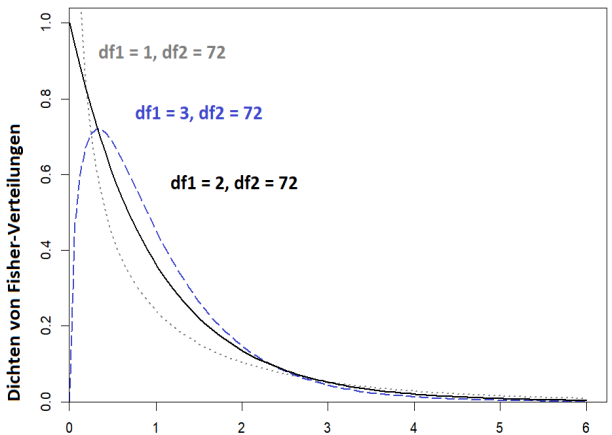
• **Nullhypothese $\mathcal{H}_0$:** *"Für etwa 70-jährige Personen hängt die mittlere Wegstrecke beim 6-Minuten-Gehtest nicht vom allgemeinen Fitneßzustand gemäß NYHA-Klassen I, II oder III ab, also* $\mu_{(1)} = \mu_{(2)} = \mu_{(3)}$*."*

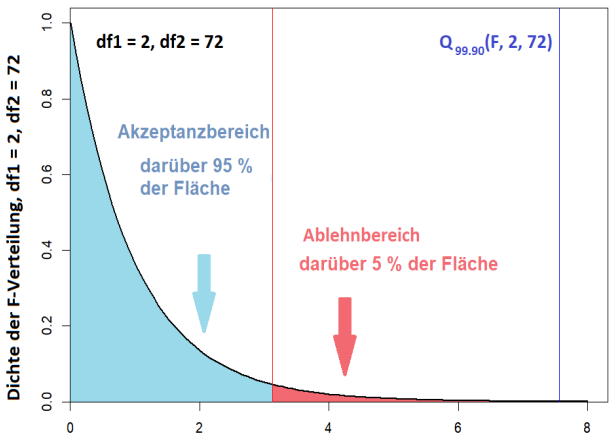
• **Alternativhypothese $\mathcal{H}_A$:** *"Für etwa 70-jährige Personen hängt die mittlere Wegstrecke beim 6-Minuten-Gehtest vom allgemeinen Fitneßzustand gemäß NYHA-Klassen I, II oder III ab, es gibt also mindestens zwei Gruppenmittel mit* $\mu_{(r_1)} \neq \mu_{(r_2)}$*."*

• *Signifikanzniveau festlegen:* $\alpha = 0.05$

• **Lehrsatz 4.2.:** *(Verteilung der Prüfgröße beim Globaltest) Eine metrische Variable* x *in der Gesamtpopulation* $\mathcal{P}$ *sei normalverteilt mit Mittelwert* μ *und Standardabweichung* σ*. Weiter sei* $R > 2$*. Ziehen wir immer wieder Sets von* R *paarweise unabhängigen Stichproben* $\mathcal{S}_{(1)}$*,* $\mathcal{S}_{(2)}$*, ... ,* $\mathcal{S}_{(R)}$ *mit stets denselben Umfängen* $N_{(1)}$*,* $N_{(2)}$*, ... ,* $N_{(R)}$*, so unterliegt die Prüfgröße* $\widehat{T}$ *einer Fisher-Verteilung mit* $df_1 = (R-1)$ *und* $df_2 = (N-R)$ *Freiheitsgraden. Darin ist* $N = N_{(1)} + ... + N_{(R)}$*.*

Die Fisher-Verteilung





• *Die Fisher-Verteilung hängt von zwei Parametern ("Freiheitsgraden") df_1 und df_2 ab. Der erste bestimmt die Grundgestalt der Kurve, der zweite die Anstiege.*

• *Zu $\widehat{T} = 7.5661$ gehört $Q_{99.90}(F, 2, 72)$ sowie $p = 0.0010 < \alpha$. Der Globaltest fällt signifikant aus; also ist $\mathcal{H}_0$ abzulehnen und $\mathcal{H}_A$ anzunehmen, und wir dürfen die Post-hoc-Tests durchführen.*

Drei Post-hoc-Tests mit Bonferroni-Korrektur

• **Lehrsatz 4.3.:** *(Bonferroni-Korrektur beim multiplen Testen) Es werden* $B \geqslant 2$ *unabhängige Hypothesentests durchgeführt. Benutzt man in allen Tests das korrigierte Signifikanzniveau* α/B*, so beträgt die Wahrscheinlichkeit, daß mindestens einer der Tests falsch-positiv ausfällt, höchstens* α*.*

• **Anwendung:** *ursprüngliches Signifikanzniveau* $\alpha = 0.05$*, also Signifikanzniveau mit Bonferroni-Korrektur* $B = 3$ *für die Post-hoc-Tests:* $\alpha' = \alpha/B = 0.0167$*.*

— $\mathcal{S}_{(1)}$ *vs.* $\mathcal{S}_{(2)}$*, zweiseitiger* t*-Test mit df=* 46 *und* $Q_{99.17}(t, 46) = 2.49$*:* $\widehat{T} = 0.3996$*, also ist der Unterschied* $\widehat{\mu}_{(1)} - \widehat{\mu}_{(2)} \neq 0$ *nicht signifikant (*$p = 0.6913$*).*
— $\mathcal{S}_{(1)}$ *vs.* $\mathcal{S}_{(3)}$*, zweiseitiger* t*-Test mit df=* 39 *und* $Q_{99.17}(t, 39) = 2.51$*:* $\widehat{T} = 2.5527$*, also ist der Unterschied* $\widehat{\mu}_{(1)} - \widehat{\mu}_{(3)} \neq 0$ *signifikant (*$p = 0.0147$*).*
— $\mathcal{S}_{(2)}$ *vs.* $\mathcal{S}_{(3)}$*, zweiseitiger* t*-Test mit df=* 59 *und* $Q_{99.17}(t, 59) = 2.47$*:* $\widehat{T} = 3.4482$*, also ist der Unterschied* $\widehat{\mu}_{(2)} - \widehat{\mu}_{(3)} \neq 0$ *signifikant (*$p = 0.0010$*).*

• **Antwort:** *Die Gruppendifferenzen* $\Delta_{\text{I-III}} = 86.0\,\text{m}$ *(95-%-K.I.* $[\,17.8\,,\,154.2\,]$*) und* $\Delta_{\text{II-III}} = 77.3\,\text{m}$ *(95-%-K.I.* $[\,32.4\,,\,122.2\,]$*) zwischen den mittleren Wegstrecken sind zu* $\alpha = 0.05$ *signifikant.*

Einfaktorielle ANOVA: Anmerkungen

• **Robuste Variante:**

— Analog zum t-Test gibt es bei der einfaktoriellen ANOVA eine Variante nach Welch.

— Sind die Umfänge der Stichproben annähernd gleich, dann ist die ANOVA nach Welch robust gegenüber inhomogenen Standardabweichungen bzw. Varianzen.

• **Unterschiedlich große Stichproben:**

— Unterscheiden sich die Gruppengrößen stark, so wird die Gesamt-Stichprobenvarianz von den großen Gruppen dominiert.

— Wenn die Gruppengrößen und empirischen Standardabweichungen stark variieren, ist bei den Post-hoc-Tests auch der Mann-Whitney-U-Test (siehe unten) wenig zu empfehlen. Am besten sollten dann t-Tests nach Welch durchgeführt werden.

• **Bonferroni-Korrektur:**

— Diese Korrektur ist konservativ, d.h. sie verschenkt möglicherweise Power.

Hundertjährige Wetterextreme

• **Beispiel 4.4.:** *Wie groß ist die Wahrscheinlichkeit, daß wir 2026 mindestens ein hundertjähriges Wetterereignis beobachten?*



Wetterextreme und Bonferroni-Korrektur

• *Wahrscheinlichkeit, daß ein einzelnes Ereignis, das im Mittel einmal in 100 Jahren vorkommt, im laufenden Kalenderjahr auftritt:* $\alpha = 0.01$

• *Mindestanzahl möglicher hundertjähriger Wetterextreme:* $R \geqslant 220 \cdot 20 = 4400$

Ereignisse	Kalender-tage (52)	Monate (12)	Jahres-zeiten (4)	Zeilen-summe
extreme Hitze	x	x	x	68
extreme Kälte	x	x	x	68
extreme Trockenheit		x	x	16
extremer Niederschlag	x	x	x	68
Spaltensumme	156	48	16	220

Ort oder Region	Regionen in Deutschland (5)
Großstadt	x
Kleinstadt oder Dorf	x
höchster Punkt	x
tiefster Punkt	x
Summe	20

• *Wahrscheinlichkeit, daß 2026 mindestens eines der genannten hundertjährigen Ereignisse auftritt:* $p = 1 - 0.99^{4400} = 0.9999$, *also 99.99 %.*

• **Was ist wirklich "extrem"?** *Analog zur Bonferroni-Korrektur für das Signifikanzniveau* α *beim multiplen Testen müßte die Wahrscheinlichkeit für Rekordereignisse zu* $p = \alpha/B$ *mit einem passenden* B*-Wert korrigiert werden, damit der Begriff "Extrem" noch sinnvoll bleibt.*

Zwei Grundtypen von Tests

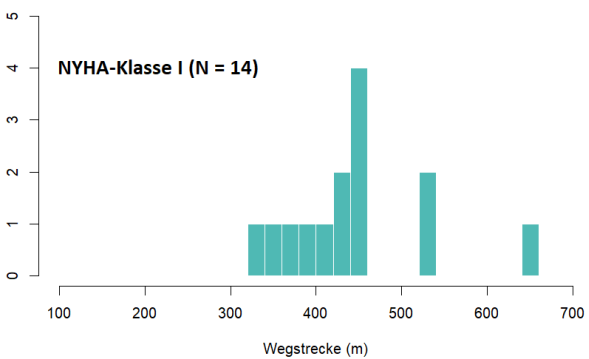
- **Parametrische vs. nichtparametrische Tests**

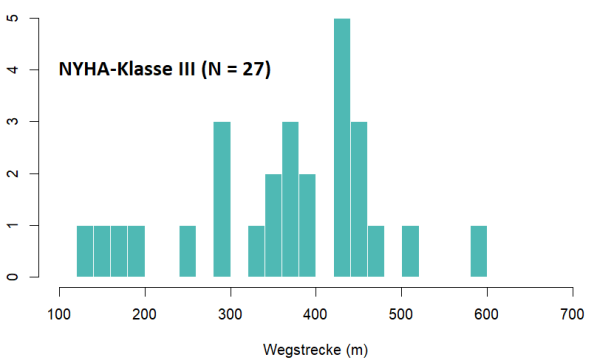
	parametrischer Test	*nichtparametrischer Test*
Teststatistik T	*aus den Werten errechnet*	*aus den Rangfolgen ermittelt*
Verteilungen von x *in* $\mathcal{P}'$, $\mathcal{P}''$	*beide annähernd normalverteilt*	*beliebig, aber beide vom gleichen Typ*
Lagemaße	*Stichproben-Mittelwerte* $\widehat{\mu}'$, $\widehat{\mu}''$ *schätzen* μ', μ''	*Ränge in den Stichproben schätzen u.a.* $m' - m''$
Streuungsmaße	$\widehat{\sigma}'$, $\widehat{\sigma}''$ *schätzen* σ', σ''	*werden nicht geschätzt, aber indirekt beim Test berücksichtigt*
Robustheit	*empfindlich gegen Ausreißer*	*robust*

Nichtparametrisches Testen: Beispiel

• **Voraussetzungen dafür:** *Kategoriale, ordinale Daten oder metrische Daten, jedoch kleine Stichproben und / oder Verteilungen in $\mathcal{P}'$, $\mathcal{P}''$, die stark von der Normalverteilung abweichen oder ganz unbekannt sind.*

• **Beispiel 4.5.:** *Wir untersuchen nochmals die Verteilungen der Wegstrecken beim 6-Minuten-Gehtest (Baseline) bei den Stichproben $\mathcal{S}_{(1)}$ und $\mathcal{S}_{(3)}$ (Probanden mit NYHA-Klassen I und III). Sind die Verteilungen annähernd identisch, oder unterscheiden sie sich signifikant?*





Stochastic superiority

• **Definition 4.6.:** *(Stochastic superiority) Gegeben sind zwei Gesamtpopulationen $\mathcal{P}'$, $\mathcal{P}''$, und darin eine (kategorial-ordinale oder) metrische Variable x. Die Verteilungen von x in $\mathcal{P}'$ und $\mathcal{P}''$ sind beliebig und eventuell völlig verschieden. Zieht man zufällig zwei Werte x', x'' aus $\mathcal{P}'$, $\mathcal{P}''$, so heißt die Wahrscheinlichkeit*

$$A = \mathit{Prob}\{\, x' < x'' \,\} + \frac{1}{2} \cdot \mathit{Prob}\{\, x' = x'' \,\}$$

dafür, daß der Wert x' aus $\mathcal{P}'$ der kleinere ist, "(probability of) stochastic superiority". Kommen die Werte x', x'' aus Stichproben $\mathcal{S}'$, $\mathcal{S}''$, so wird die Größe mit $\widehat{A}$ bezeichnet.

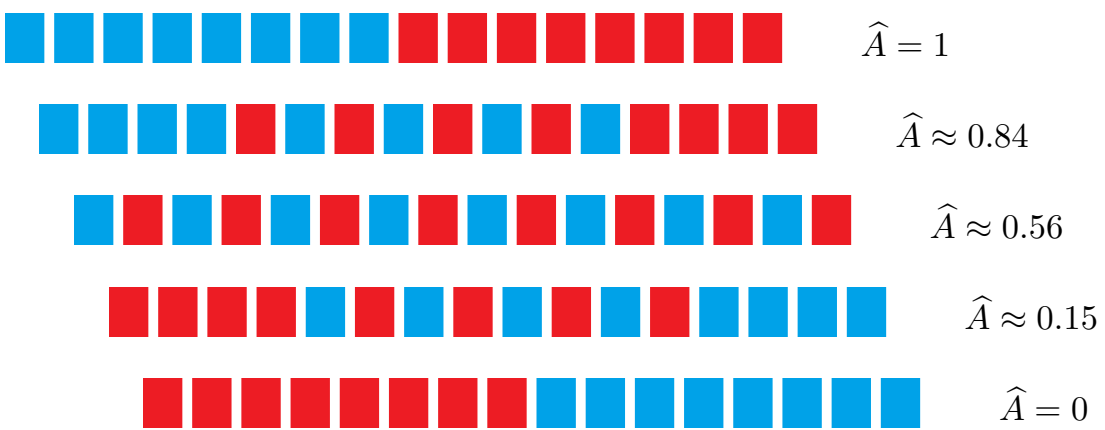
• **Erste Beobachtung:** *Wenn die metrische Variable x in $\mathcal{P}'$ und $\mathcal{P}''$ auf exakt dieselbe Weise verteilt ist, so gilt $A = 0.5$. Denn dann haben beide Verteilungen dieselbe Dichtefunktion $f'(x) \equiv f''(x)$, und wir erhalten $\mathit{Prob}\{\, x' = x'' \,\} = 0$ sowie*

$$\begin{aligned} \mathit{Prob}\{\, x' < x'' \,\} &= \int_{-\infty}^{+\infty} \mathit{Prob}\{\, x' < c \,\} \cdot \mathit{Prob}\{\, x'' > c \,\}\, dc \\ &= \int_{-\infty}^{+\infty} \mathit{Prob}\{\, x' > c \,\} \cdot \mathit{Prob}\{\, x'' < c \,\}\, dc = \mathit{Prob}\{\, x' > x'' \,\}\,. \end{aligned}$$

Stochastic superiority: Ränge in Stichproben

• *Leider darf man diese Schlußfolgerung nicht umkehren: Aus* $A = 0.5$ *folgt die Gleichheit der Dichtefunktionen* $f'(x)$, $f''(x)$ *nur unter zusätzlichen technischen Voraussetzungen.*

• **Zweite Beobachtung:** *Für Werte* x', x'' *aus zwei Stichproben* $\mathcal{S}'$, $\mathcal{S}''$ *kann* $\widehat{A}$ *berechnet werden, indem man beide Stichproben poolt und die Verteilung der Ränge untersucht.*



Zweiseitiger Mann-Whitney-U-Test: Hypothesen

• **Situation I:** *Die Verteilungen von x in $\mathcal{P}'$ und $\mathcal{P}''$ sind beliebig und eventuell völlig verschieden.*

— Zweiseitige Basis-Nullhypothese $\mathcal{H}_0$: "Zieht man zufällig zwei Werte x', x'' aus $\mathcal{P}'$, $\mathcal{P}''$, so treten die Fälle $x' < x''$ und $x' > x''$ mit gleicher Wahrscheinlichkeit auf."

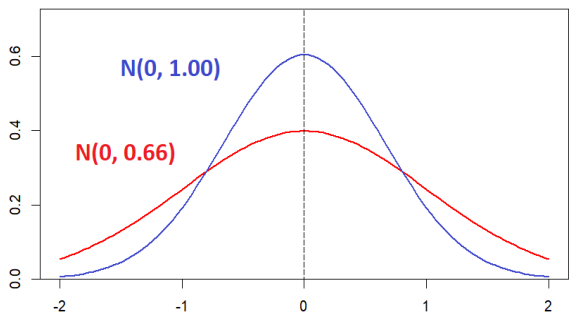
— Zweiseitige Basis-Alternativhypothese $\mathcal{H}_A$: "Zieht man zufällig zwei Werte x', x'' aus $\mathcal{P}'$, $\mathcal{P}''$, so treten die Fälle $x' < x''$ und $x' > x''$ mit verschiedenen Wahrscheinlichkeiten auf."

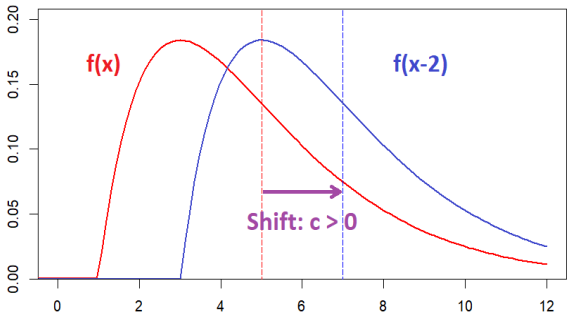
• **Situation II:** *Die Verteilungen von x in $\mathcal{P}'$, $\mathcal{P}''$ haben stetige Dichten $f'(x)$ und $f''(x) = f'(x - c_{\mathsf{Shift}})$ derselben Gestalt, die höchstens gegeneinander verschoben sind.*

— Zweiseitige Shift-Nullhypothese $\mathcal{H}_0$: "Der Shift zwischen den Verteilungsdichten f', f'' von x in $\mathcal{P}'$, $\mathcal{P}''$ beträgt Null, also $c_{\mathsf{Shift}} = 0$ und $f' = f''$."

— Zweiseitige Shift-Alternativhypothese $\mathcal{H}_A$: Der Shift zwischen den Verteilungsdichten f', f'' von x in $\mathcal{P}'$, $\mathcal{P}''$ ist von Null verschieden, also $c_{\mathsf{Shift}} \neq 0$ und $f' \neq f''$."

Was bedeuten die Hypothesen?





• **Basis-Hypothesen:** *Die Testentscheidung hat nichts mit Medianen oder Mittelwerten, sondern unmittelbar mit der Gestalt der Verteilungen zu tun. Unter $\mathcal{H}_0$ haben beide Verteilungen sehr ähnliche Gestalt, unter $\mathcal{H}_A$ sind sie deutlich verschieden.*

• **Shift-Hypothesen:** *Direkter Lagevergleich für Mediane oder Mittelwerte möglich!*
— Die Mediane m', m'' von x in $\mathcal{P}'$, $\mathcal{P}''$ müssen unter $\mathcal{H}_0$ übereinstimmen und unter $\mathcal{H}_A$ verschieden sein.
— Auch die Mittelwerte μ', μ'' von x in $\mathcal{P}'$, $\mathcal{P}''$ müssen unter $\mathcal{H}_0$ übereinstimmen und unter $\mathcal{H}_A$ verschieden sein.

Nichtparametrisches Testen: Ränge bilden

- *Daten zu Beispiel 4.5.* $\longrightarrow$ *Daten poolen und sortieren* $\longrightarrow$ *Ränge bilden* $\longrightarrow$

Strecken	Strecken
450	437
404	376
455	341
649	590
532	372
457	361
345	290
445	425
383	453
334	385
380	255
439	290
440	131
530	342
172	450
199	423
152	422
430	330
462	384
514	443
288	

Strecken sortiert	Strecken sortiert
131	422
152	423
172	425
199	430
255	437
288	439
290	440
290	443
330	445
334	450
341	450
342	453
345	455
361	457
372	462
376	514
380	530
383	532
384	590
385	649
404	

Strecken sortiert	Rang
131	1
152	2
172	3
199	4
255	5
288	6
290	7
290	8
330	9
334	10
341	11
342	12
345	13
361	14
372	15
376	16
380	17
383	18
384	19
385	20
404	21

Strecken sortiert	Rang
422	22
423	23
425	24
430	25
437	26
439	27
440	28
443	29
445	30
450	31
450	32
453	33
455	34
457	35
462	36
514	37
530	38
532	39
590	40
649	41

Nichtparametrisches Testen: Ränge / Rangsummen

- *Ränge für Bindungen korrigieren $\longrightarrow$ Rangsummen bilden*

Strecken sortiert	Rang	Rang mit Bindungen
131	1	1
152	2	2
172	3	3
199	4	4
255	5	5
288	6	6
290	7	7.5
290	8	7.5
330	9	9
334	10	10
341	11	11
342	12	12
345	13	13
361	14	14
372	15	15
376	16	16
380	17	17
383	18	18
384	19	19
385	20	20
404	21	21

Strecken sortiert	Rang	Rang mit Bindungen
422	22	22
423	23	23
425	24	24
430	25	25
437	26	26
439	27	27
440	28	28
443	29	29
445	30	30
450	31	31.5
450	32	31.5
453	33	33
455	34	34
457	35	35
462	36	36
514	37	37
530	38	38
532	39	39
590	40	40
649	41	41

- *Beim nichtparametrischen Testen kommt es nicht mehr auf die Werte der Variablen, sondern nur noch auf die (für Bindungen korrigierten) Ränge an!*
- *Stichprobe* $\mathcal{S}_{(1)} = \mathcal{S}'$
  Umfang: $N' = 14$
  Rangsumme: $R' = 382.5$
- *Stichprobe* $\mathcal{S}_{(3)} = \mathcal{S}''$
  Umfang: $N'' = 27$
  Rangsumme: $R'' = 478.5$
- *Kontrolle:* $R' + R''$
  $= 1+2+ \ldots +40+41 = 861$

Zweiseitiger Mann-Whitney-U-Test: Prüfgrößen

Schritt 1: *Bestimme die Ränge, korrigiere die Bindungen und berechne die Rangsummen.*

Schritt 2: *Rechne die Rangsummen in U-Werte um:*

$$U' = \Big(N' \cdot N'' + \frac{N' \cdot (N'+1)}{2} \Big) - R' \text{ und } U'' = \Big(N' \cdot N'' + \frac{N'' \cdot (N''+1)}{2} \Big) - R'' .$$

• *Beachte: Wie R', R'', so bewegen sich auch U', U'' gegenläufig. Unabhängig von den Werten der Rangsummen bleibt $U' + U'' = N' \cdot N''$ konstant.*

Schritt 3: *Falls $N' + N'' > 30$, rechne die U-Werte nochmals in $\widehat{T}$-Werte um:*

$$\widehat{T}' = \frac{U' - N' \cdot N''/2}{\sigma_U} , \; \widehat{T}'' = \frac{U'' - N' \cdot N''/2}{\sigma_U} \text{ mit } \sigma_U = \sqrt{\frac{N' \cdot N'' \cdot (N' + N'' + 1)}{12}} .$$

• *Beachte: Wie U', U'', so bewegen sich auch $\widehat{T}'$, $\widehat{T}''$ gegenläufig. Wegen $U' + U'' = N' \cdot N''$ gilt sogar $\widehat{T}'' = -\widehat{T}'$.*

Zweiseitiger Mann-Whitney-U-Test: Auswertung

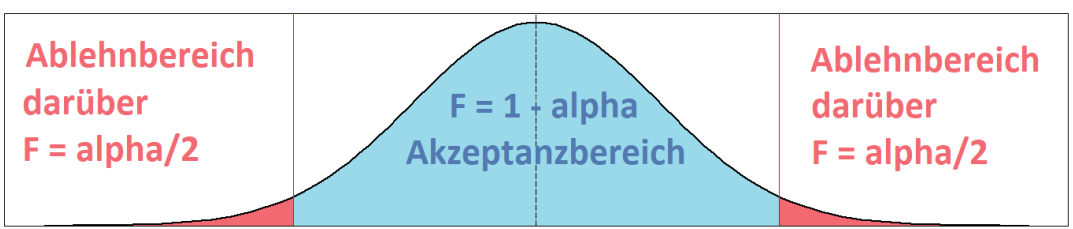
• **Auswertung für $N' + N'' \leqslant 30$ direkt mit U', U''**

$\mathcal{H}_A$ wird angenommen und $\mathcal{H}_0$ abgelehnt, falls entweder $U' \leqslant U_{krit}(N', N'', \alpha)$ oder $U'' \leqslant U_{krit}(N', N'', \alpha)$. Ansonsten wird $\mathcal{H}_0$ beibehalten. Die Werte für $U_{krit}(N', N'', \alpha)$ müssen einer Tafel entnommen werden.

• **Auswertung für $N' + N'' > 30$ mit $\widehat{T}'$, $\widehat{T}''$ nach Lehrsatz 4.7.:**

Sei $N' + N'' > 30$. Wenn die Nullhypothese gilt, so unterliegen beide Prüfgrößen $\widehat{T}'$, $\widehat{T}''$ jeweils in sehr guter Näherung einer Standard-Normalverteilung mit $\mu = 0$ und $\sigma = 1$.

• *Zum zweiseitigen Test gehört ein zweiteiliger, symmetrischer Ablehnbereich, der durch die passenden Quantile der Standard-Normalverteilung begrenzt wird.*



• *$\mathcal{H}_0$ wird abgelehnt, falls sowohl $\widehat{T}'$ als auch $\widehat{T}''$ im zweiteiligen Ablehnbereich liegen. Welche Gestalt der Hypothesen benutzt wird, ist für die Auswertung unerheblich.*

Zweiseitiger Mann-Whitney-U-Test: Durchführung

- *Populationen $\mathcal{P}'$, $\mathcal{P}''$: alle etwa 70-jährigen Personen mit NYHA-Klassen I bzw. III*
- *Situation II, $N' + N'' = 41$, also Auswertung mit $\widehat{T}'$, $\widehat{T}''$*
- *$\mathcal{H}_0$ (zweiseitig, Shift): "Die Verteilungen der Wegstrecken beim 6-Minuten-Gehtest stimmen für $\mathcal{P}'$, $\mathcal{P}''$ mit $c_{\text{Shift}} = 0$ überein, also gilt insbesondere $\mu' = \mu''$."*
- *$\mathcal{H}_A$ (zweiseitig, Shift): "Die Verteilungen der Wegstrecken beim 6-Minuten-Gehtest für $\mathcal{P}'$, $\mathcal{P}''$ sind um $c_{\text{Shift}} \neq 0$ gegeneinander verschoben, also $\mu' \neq \mu''$."*
- *Signifikanzniveau aus Beispiel 4.1. übernehmen: $\alpha = 0.0167$, die zugehörigen Quantile der Standard-Normalverteilung sind $Q_{0.83} = -2.395$ und $Q_{99.17} = 2.395$.*
- *Prüfgrößen berechnen (jeweils beide):*
$U' = 14 \cdot 27 + 14 \cdot 15/2 - 382.5 = 100.5$, $U'' = 14 \cdot 27 + 27 \cdot 28/2 - 478.5 = 277.5$, $\sigma_U = 36.37$, $\widehat{T}' = (100.5 - 189)/36.37 = -2.43$, $\widehat{T}'' = (277.5 - 189)/36.37 = 2.43$, *dazu gehören* $Q_{0.75}$, $Q_{99.25}$ *und* $p = 2 \cdot 0.0075 = 0.0150$.
- *Auswertung / Interpretation: Der Unterschied zwischen $\widehat{\mu}' = \widehat{\mu}_{(1)}$ und $\widehat{\mu}'' = \widehat{\mu}_{(3)}$ ist signifikant zum 1.67-%-Niveau ($p = 0.0150$); $\mathcal{H}_0$ ist abzulehnen und $\mathcal{H}_A$ anzunehmen.*

Zurück zur stochastic superiority

• *Zwischen den Prüfgrößen R' bzw. U' des Mann-Whitney-U-Tests und der stochastic superiority besteht tatsächlich ein unmittelbarer Zusammenhang. Für Werte $x' \in \mathcal{S}'$ und $x'' \in \mathcal{S}''$ kann $\widehat{A}$ wie folgt berechnet werden:*

$$\widehat{A} = \frac{U'}{N' \cdot N''} \,.$$

• **Beispiel / Interpretation:** *In Beispiel 4.5. erhalten wir $U' = 100.5$ und $\widehat{A} = 100.5/(14 \cdot 27) = 0.2659$. Das bedeutet: Die Wahrscheinlichkeit, daß eine zufällig gewählte Wegstrecke aus $\mathcal{S}' = \mathcal{S}_{(1)}$ (Gehtest bei Probanden mit NYHA-Klasse I) kleiner ist als eine zufällig gewählte Wegstrecke aus $\mathcal{S}'' = \mathcal{S}_{(3)}$ (Gehtest bei Probanden mit NYHA-Klasse III), beträgt etwa 26.6 %.*

Ränge bei verbundenen Stichproben

• **Beispiel 4.8.:** *("Vorher-Nachher") Wir untersuchen für die Probanden mit NYHA-Klasse III die Verteilungen der Wegstrecken beim 6-Minuten-Gehtest zu Baseline (Stichprobe $\mathcal{S}'$) und im Follow-up (Stichprobe $\mathcal{S}''$). Sind die Verteilungen annähernd identisch, oder unterscheiden sie sich signifikant?*

• **Grundgedanken:**

— Wir berechnen die Differenzen $d_i = x'_i - x''_i$ der Wertepaare, sortieren sie nach wachsendem Betrag und bilden die Ränge.

— Wertepaare mit Differenz Null werden dabei nicht mit ausgewertet.

— Wenn es bei den Beträgen der Differenzen Bindungen gibt, werden die zugehörigen Ränge auf ihren Mittelwert korrigiert.

— Dann bilden wir die Rangsummen R^- und R^+ für die Differenzen mit negativem und positivem Vorzeichen.

— Mit R^- und R^+ werden Prüfgrößen analog zum Mann-Whitney-U-Test aufgebaut.

Wilcoxon-Rangsummentest: Hypothesen

• **Situation I:** *Die Verteilungen von x in den verbundenen Populationen $\mathcal{P}'$, $\mathcal{P}''$ sind beliebig, jedoch beide annähernd symmetrisch bezüglich ihrer Mediane bzw. Mittelwerte.*

— Zweiseitige Basis-Nullhypothese $\mathcal{H}_0$: "Zieht man zufällig ein verbundenes Wertepaar x', x'' aus $\mathcal{P}'$, $\mathcal{P}''$, so treten die Fälle $x' < x''$ und $x' > x''$ mit gleicher Wahrscheinlichkeit auf."

— Zweiseitige Basis-Alternativhypothese $\mathcal{H}_A$: "Zieht man zufällig ein verbundenes Wertepaar x', x'' aus $\mathcal{P}'$, $\mathcal{P}''$, so treten die Fälle $x' < x''$ und $x' > x''$ mit verschiedenen Wahrscheinlichkeiten auf."

• **Situation II:** *Die Verteilungen von x in $\mathcal{P}'$, $\mathcal{P}''$ besitzen stetige Dichten $f'(x)$ und $f''(x) = f'(x - c_{\mathsf{Shift}})$ gleicher Gestalt, die höchstens gegeneinander verschoben sind. Beide sind symmetrisch bezüglich der Mediane bzw. Mittelwerte.*

— Zweiseitige Shift-Nullhypothese $\mathcal{H}_0$: "Der Shift zwischen den Verteilungsdichten f', f'' von x in $\mathcal{P}'$, $\mathcal{P}''$ beträgt Null, also $c_{\mathsf{Shift}} = 0$ und $f' = f''$."

— Zweiseitige Shift-Alternativhypothese $\mathcal{H}_A$: Der Shift zwischen den Verteilungsdichten f', f'' von x in $\mathcal{P}'$, $\mathcal{P}''$ ist von Null verschieden, also $c_{\mathsf{Shift}} \neq 0$ und $f' \neq f''$."

Wilcoxon-Rangsummentest: Durchführung (I)

Strecke Baseline	Strecke Follow-up	Differenz	Vor-zeichen	Betrag Differenz	Rang	Rang mit Bindungen
437	437	0	entfällt	0	entfällt	entfällt
376	383	7	pos.	7	1	1
341	351	10	pos.	10	2	2.5
443	453	10	pos.	10	3	2.5
199	177	-22	neg.	22	4	4
590	565	-25	neg.	25	5	5.5
330	305	-25	neg.	25	6	5.5
385	412	27	pos.	27	7	7
462	497	35	pos.	35	8	8
372	410	38	pos.	38	9	9
290	339	49	pos.	49	10	10
384	440	56	pos.	56	11	11
425	366	-59	neg.	59	12	12.5
453	394	-59	neg.	59	13	12.5
342	403	61	pos.	61	14	14
430	367	-63	neg.	63	15	15
290	361	71	pos.	71	16	16
450	530	80	pos.	80	17	17
422	506	84	pos.	84	18	18
131	218	87	pos.	87	19	19
288	393	105	pos.	105	20	20.5
361	466	105	pos.	105	21	20.5
172	297	125	pos.	125	22	22
423	553	130	pos.	130	23	23
152	286	134	pos.	134	24	24
514	NA	NA	NA	NA	entfällt	entfällt
255	NA	NA	NA	NA	entfällt	entfällt

- *Es kommt nicht auf die Werte der Differenzen, sondern nur noch auf ihre (für Bindungen korrigierten) Ränge an!*
- *Daten sind bereits nach Betrag der Differenz sortiert*
- *Auswertbare Paare:* $N^* = 24$
- *Rangsummen:*
  $R^- = 55$ *(negative Differenzen),*
  $R^+ = 245$ *(positive Differenzen)*
- *Kontrolle:*
  $$R^- + R^+ = \frac{N^* \cdot (N^* + 1)}{2} = 300$$

Wilcoxon-Rangsummentest: Prüfgröße

• **Voraussetzungen:** *Es wird mindestens $N^* \geqslant 6$ benötigt. Die Voraussetzung über die Symmetrie wird in der Praxis meist ignoriert.*

• **Auswertung für $N^* \leqslant 30$ direkt mit R^-, R^+:**
$\mathcal{H}_A$ wird angenommen und $\mathcal{H}_0$ abgelehnt, falls entweder $R^- \leqslant R_{krit}(N^*, \alpha)$ oder $R^+ \leqslant R_{krit}(N^*, \alpha)$. Anderenfalls wird $\mathcal{H}_0$ beibehalten. Die Werte für $R_{krit}(N^*, \alpha)$ müssen einer Tafel entnommen werden.

• **Auswertung für $N^* > 30$ mit Prüfgrößen $\widehat{T}^-$, $\widehat{T}^+$:**

$$\widehat{T}^- = \frac{R^- - N^* \cdot (N^*+1)/4}{\sigma_R}\,,\ \widehat{T}^+ = \frac{R^+ - N^* \cdot (N^*+1)/4}{\sigma_R} \quad \textit{mit}$$

$$\sigma_R = \sqrt{\frac{N^* \cdot (N^*+1) \cdot (2 \cdot N^*+1)}{24}}\,.$$

• **Lehrsatz 4.9.:** *Sei $N^* > 30$. Dann unterliegen beide Prüfgrößen $\widehat{T}^-$, $\widehat{T}^+$ in sehr guter Näherung einer Standard-Normalverteilung mit $\mu = 0$ und $\sigma = 1$.*

• *Wiederum gilt $\widehat{T}^+ = -\widehat{T}^-$. Welche Gestalt der zweiseitigen Hypothesen $\mathcal{H}_0$ und $\mathcal{H}_A$ benutzt wird, ist für die Auswertung unerheblich.*

Wilcoxon-Rangsummentest: Durchführung (II)

• *Verbundene Populationen $\mathcal{P}'$, $\mathcal{P}''$: alle etwa 70-jährigen Personen mit NYHA-Klasse III vor und nach der 6-wöchigen Studienintervention*

• *Situation II (Symmetrie vernachlässigt),* $N^* = 24 \leqslant 30$*, also direkte Auswertung mit* R^-, R^+

• $\mathcal{H}_0$ *(zweiseitig, Shift): "Die Verteilungen der mittleren Wegstrecken beim 6-Minuten-Gehtest stimmen für* $\mathcal{P}'$, $\mathcal{P}''$ *mit* $c_{\text{Shift}} = 0$ *überein, also gilt* $\mu' = \mu''$*."*

• $\mathcal{H}_A$ *(zweiseitig, Shift): "Die Verteilungen der mittleren Wegstrecken beim 6-Minuten-Gehtest für* $\mathcal{P}'$, $\mathcal{P}''$ *sind um* $c_{\text{Shift}} \neq 0$ *gegeneinander verschoben, also* $\mu' \neq \mu''$*."*

• *Signifikanzniveau festlegen:* $\alpha = 0.05$*, dazu gehört* $R_{\text{krit}}(24, 0.05) = 81$ *als kritischer Wert laut Tabelle.*

• *Prüfgröße und Testentscheidung: Es gilt* $R^- = 55 < 81 = R_{\text{krit}}(24, 0.05)$*, also ist* $\mathcal{H}_0$ *zu verwerfen und* $\mathcal{H}_A$ *anzunehmen: Die Mittelwerte* μ'' *der Wegstrecken nach 6 Wochen Studienintervention und* μ' *der Wegstrecken zu Baseline unterscheiden sich um* $\Delta = 36.5\,\text{m}$ *(95-%-K.I* $[\,12\,,\,70\,]$*). Dieses Ergebnis ist signifikant zum 5-%-Niveau.*

Zum Kruskal-Wallis-Test

• **Voraussetzungen:** *Für eine metrische Variable x liegen mindestens drei Stichproben $\mathcal{S}_{(1)}$, ... , $\mathcal{S}_{(R)}$ mit Umfängen $N_{(1)}$, ... , $N_{(R)}$ aus Gesamtpopulationen $\mathcal{P}_{(1)}$, ... , $\mathcal{P}_{(R)}$ vor. Alle Verteilungsdichten haben dieselbe Gestalt und sind höchstens gegeneinander verschoben: $f_{(1)}(x) = f(x - c_{(1)})$, ... , $f_{(R)}(x) = f(x - c_{(R)})$.*

• **Zweistufiges Verfahren:** *nichtparametrisch-analog zur einfaktoriellen ANOVA*

Schritt 1: Nichtparametrischer Globaltest: *Sind die Shifts aller Verteilungen gleich?*

— Test gegen die Nullhypothese $c_{(1)} = \ldots = c_{(R)}$ (der eigentliche Kruskal-Wallis-Test)
— zum Signifikanzniveau α.

Schritt 2: Nichtparametrische Post-hoc-Tests *zum paarweisen Vergleich der Stichproben: welche Shifts unterscheiden sich tatsächlich, und wie stark?*

— nur dann, wenn der Globaltest signifikant ausfiel,
— und mit korrigiertem Signifikanzniveau α/B, $B = R \cdot (R-1)/2$.

• *Wie bisher: Wir poolen die R Stichproben, bilden in der Gesamtheit die Ränge, korrigieren sie für Bindungen und berechnen für jede Stichprobe die Rangsumme $R_{(r)}$.*

Zum Kruskal-Wallis-Test (II)

• **Prüfgröße im Globaltest:**

$$\widehat{T} = \frac{12}{N \cdot (N+1)} \cdot \Big(\frac{R_{(1)}^2}{N_{(1)}} + \ldots + \frac{R_{(R)}^2}{N_{(R)}} \Big) - 3 \cdot (N+1) \quad \textit{mit} \quad N = N_{(1)} + \ldots + N_{(R)} \,.$$

• **Auswertung für hinreichend große Stichproben nach Lehrsatz 4.10.:** *Falls entweder (i)* $R = 3$ *und* $N_{(r)} \geqslant 9$ *für alle* r *oder (ii)* $R = 4$ *und* $N_{(r)} \geqslant 5$ *für alle* r *oder (iii)* $R = 5$ *und* $N_{(r)} \geqslant 4$ *für alle* r *oder (iv)* $R \geqslant 6$*, so unterliegt die Prüfgröße* $\widehat{T}$ *in sehr guter Näherung einer* χ^2*-Verteilung mit df* $= (R-1)$ *Freiheitsgraden.*

• **Auswertung für zu kleine Stichproben:** *Wenn keine der Voraussetzungen (i)* $-$ *(iv) aus Lehrsatz 4.10. zutrifft, muß der kritische Wert für die Prüfgröße* $\widehat{T}$ *aus einer Tafel entnommen werden.*

• **Post-hoc-Tests:** *Wenn der Globaltest signifikant ausfällt, werden die Stichproben jeweils paarweise mit einem Mann-Whitney-U-Test zum korrigierten Signifikanzniveau* α/B *mit der Bonferroni-Korrektur* $B = R \cdot (R-1)/2$ *untersucht.*