

3 Randomisierungs-Tests

3.1 Einführendes Beispiel

- a Hagel-Experiment: („Grossversuch IV“ im Napfgebiet 1978-1983)
 Verringert das „Impfen“ von potenziellen Hagelwolken
 mit Silberiodid die Hagelenergie?
 (Einfache Überlegungen, brauchen nur Kombinatorik und W.)

Zielgrösse: Hagelenergie, gemessen für n Wolken.

Zwei Gruppen: ca. $n/2$ „geimpft“, Rest „Kontrolle“.

$$Y_i : \text{Hagelenergie der Wolke } i$$

$$G_i = \begin{cases} 1 & \text{falls Wolke } i \text{ geimpft,} \\ 0 & \text{sonst.} \end{cases}$$

Hoffnung: Y_i mit $G_i = 1$ fallen tendenziell niedriger aus.

b Beobachtet:

$Y_i = y_i^*$		16672	25	855	0	152	0	46	1219
$G_i = g_i^*$		1	1	0	0	0	1	1	0

g_i^* : Zufallsauswahl der zu impfenden Wolken.

(In Wirklichkeit 216 Wolken; davon wurden 94 geimpft.)

Statistischer Test! H_0 : Keine Wirkung.

(\longrightarrow Widerspruchsbeweis!)

Ungeparter Zwei-Stichproben-Problem. \longrightarrow t-Test ?

Keine Annahmen über die Verteilung der Y_i !!

3.2 Statistische Überlegung

a **Nullhypothese** = Wahrscheinlichkeitsmodell.

Üblich: Verteilung für Y_i ; $G_i = g_i^*$ fest vorgegeben.

Randomisierungstests: G_i zufällig; $Y_i = y_i^*$ als fest betrachtet
(Analyse „bedingt auf die y_i^* “.)

Falls das Impfen keinen Einfluss auf die Hagelenergie hat,
würden wir die genau gleichen Werte y_i^* erhalten,
wenn die Wolken entspr. $\underline{g}^{(1)} = [0, 1, 0, 0, 1, 1, 0, 1]$
oder entsprechend irgendeiner anderen Auswahl
geimpft worden wären.

Zufallsauswahl:

Jede Auswahl von $n/2 = 4$ Elementen aus $n = 8$ hat gleiche Wahrscheinlichkeit

$$p = \binom{8}{4}^{-1} = \frac{1}{70}$$

Damit ist die **Nullhypothese** festgelegt.

- b **Teststatistik:** Soll extreme Werte annehmen, wenn Alternative gilt.

Alternative: y_i^* mit $g_i^* = 1$ sind tendenziell kleiner.

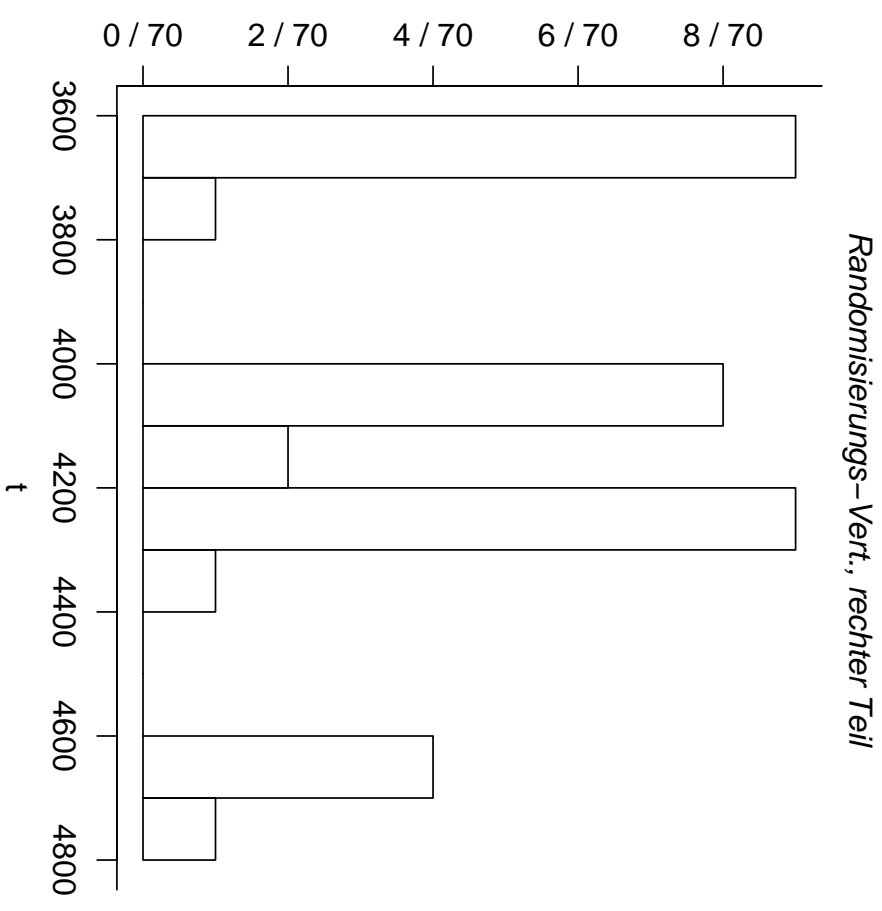
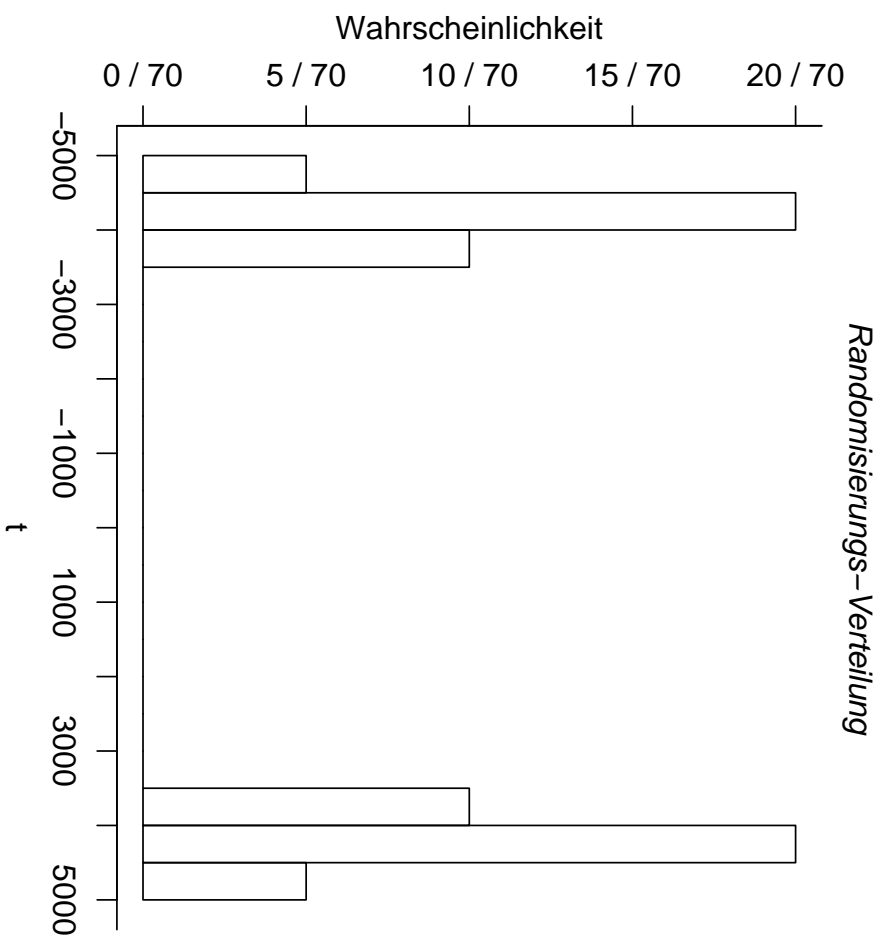
$$T\langle \underline{g}, \underline{y}^* \rangle = \frac{1}{n/2} \sum_{i:g_i=0} y_i^* - \frac{1}{n/2} \sum_{i:g_i=1} y_i^* = \frac{2}{n} \sum_i y_i^* (1 - 2g_i).$$

- c Wie ist T unter der Nullhypothese verteilt?

y_1^*, \dots, y_n^* gegeben $\longrightarrow \leq \binom{n}{n/2}$ mögliche Werte für T .

$$P\langle T\langle \underline{G}, \underline{y}^* \rangle = t \rangle = \frac{\#\{\underline{g} \mid T\langle \underline{g}, \underline{y}^* \rangle = t\}}{\binom{n}{n/2}}$$

„Randomisierungs-Verteilung“



- d **Verwerfungsbereich**: $\alpha = 5\%$ extremste Werte
(so genau als möglich).

Beispiel: $\{t \mid t \geq 4643.25\}$ (einseitig).

- e Experiment:

$$T(\underline{g}^*, \underline{y}^*) = \frac{1}{4}(855 + 0 + 152 + 1219) \\ - \frac{1}{4}(16672 + 25 + 0 + 46) = -3629.25$$

Effekt in die unerwartete Richtung!

Nullhypothese nicht verworfen; Effekt nicht nachgewiesen.

(Auch nicht in umgekehrter Richtung.)

* Voraussetzung des Tests: **Unabhängigkeit**

→ Randomisierung über 76 „potentielle Hageltage“

Davon 33 als Impftage ausgewählt. Anzahl Impftage zufällig.

→ Analyse bedingt auf Anzahl Hageltage mit Impfung.
Eingeschränkte Randomisierung.

g $\binom{76}{33} = 36 \cdot 10^{20}$ mögliche Auswahlen

→ Simulation der Randomisierungs-Verteilung.

3.3 Tests für das Zwei-Stichproben-Problem

- a Beispiel lässt sich leicht verallgemeinern:
- b Randomisierungstests sind auch dann anwendbar, wenn die Durchführung des Versuchs keinen Randomisierungsschritt enthält.

Voraussetzungen, die dann gelten müssen:

- Die Beobachtungen müssen unter H_0 gleich verteilt und
- unabhängig sein.

Dann stimmt die gewählte Irrtumswahrscheinlichkeit α exakt.

Die Randomisierungstests bilden in diesem Sinne den „Goldstandard“ unter den statistischen Tests.

(* Schwächere Voraussetzung: „Austauschbarkeit“.)

c Wenn **Beobachtungen zufällig**:

Stichprobe $[Y_1, \dots, Y_n]$ \longrightarrow geordnete St. $Y_{[1]}, \dots, Y_{[n]}$
oder empirische Verteilungsfunktion \hat{F}_n (s. Bootstrap)

Vert. der Teststatistik, bedingt auf \hat{F}_n , = Randomisierungs-Vt.

Bedingte W. eines Fehlers erster Art, gegeben \hat{F}_n , = α
— für jede Bedingung \hat{F}_n , und deshalb auch ohne Bedingung.

d **Beliebige Teststatistik.**

Differenz der Mittelwerte unrobust.

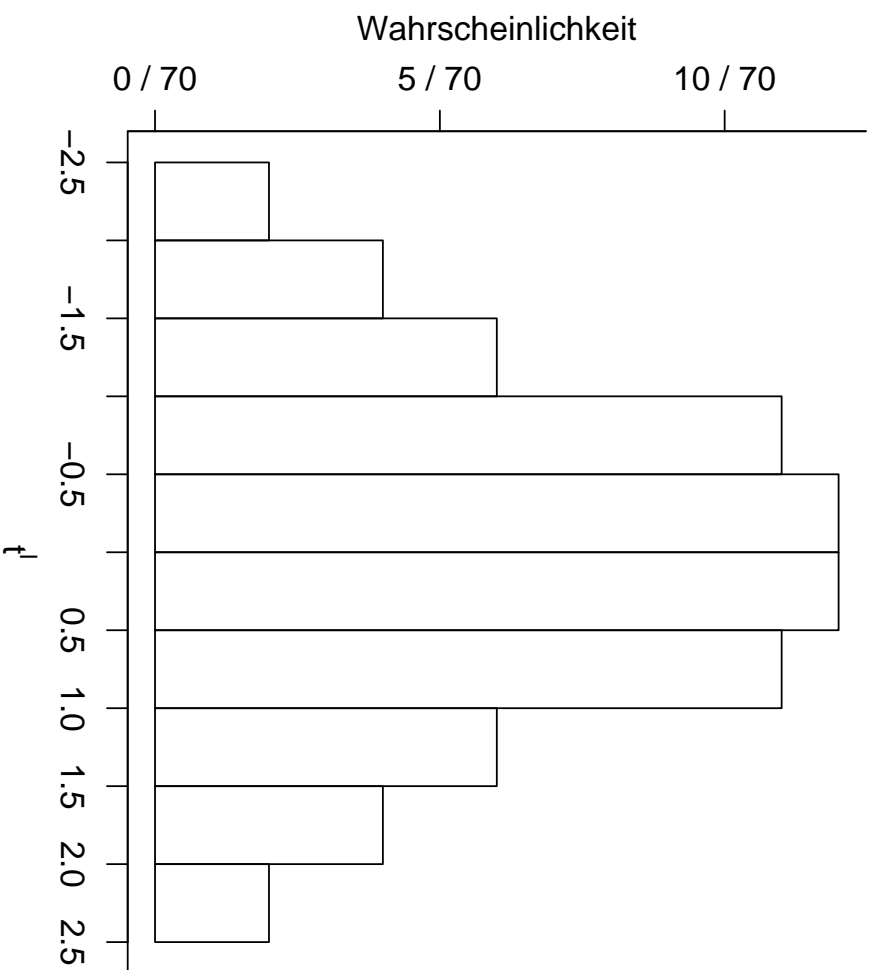
Optimale Teststatistik? → Macht für die Alternative(n) opt.!

Braucht **bestimmte** Verteilung(s-Familie)

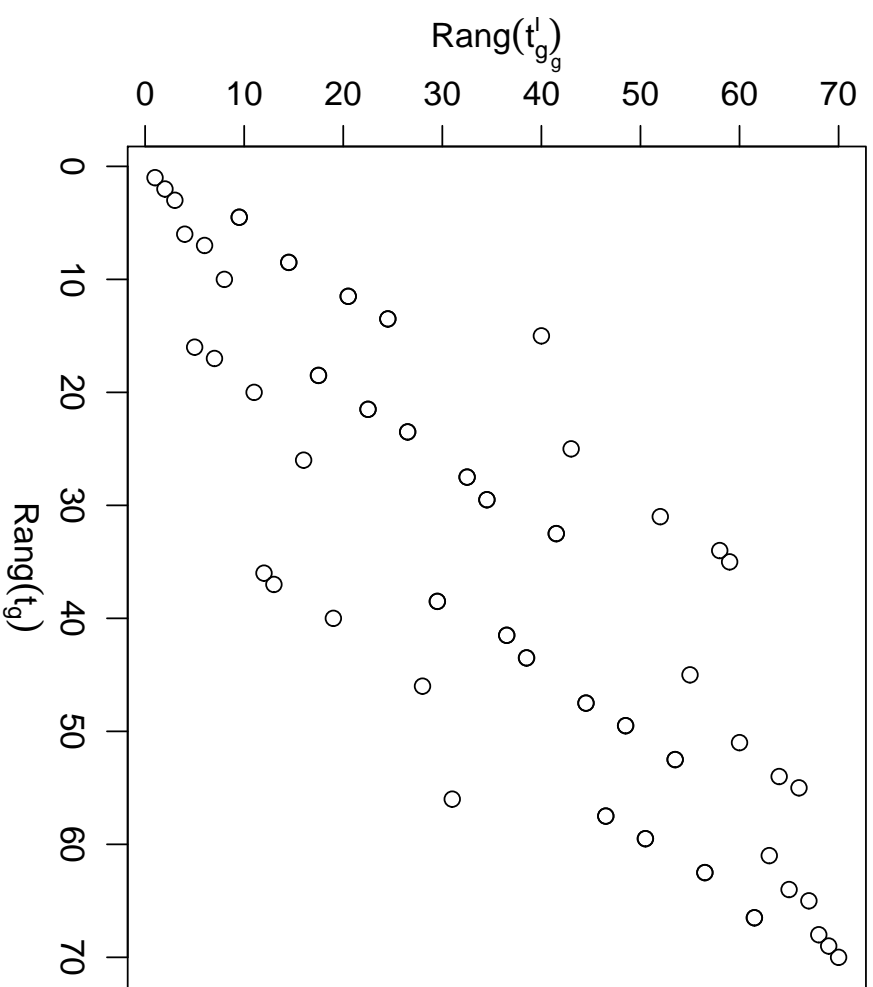
→ optimale Teststatistik (Likelihood-Ratio-Test)

e Beispiel: Logarithmus-Transformation, dann
Mittelwertsdifferenz (robustifiziert).

Rand. Vert. für log. Werte



Vergleich der Test-Statistiken



- f **Robustheit.** Wieso eine robuste Teststatistik verwenden, wenn der Test auch ohne diese „Vorsichtsmassnahme“ die Irrtumswahrscheinlichkeit genau einhält?

- g **Rangsummentest** von Wilcoxon, Mann und Whitney (U-Test),

$$T(\underline{g}, \underline{y}) = \sum_{g_i=1} R_i = \sum_i g_i R_i ,$$

Recht robust \rightarrow **Test der Wahl für das 2-Stichpr.-Problem**
Verteilung der Teststatistik unter H_0 wie gehabt.

- h* **Hagel-Experiment:** Komplizierte Teststatistik, zweidimensional
 \rightarrow zweidim. Verwerfungsbereich.

3.4 Eine Stichprobe oder zwei verbundene

a Beispiel Tranquilizer.

Zielgrösse: „Hamilton depression scale factor IV“.
9 Patienten, vor und nach Anwendung des Tranquilizers.

vorher ($X_i^{(1)}$)	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
nachher ($X_i^{(2)}$)	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
Abnahme ($-Y_i$)	0.952	-0.147	1.022	0.43	0.62	0.59	0.49	-0.08	0.01

b **Verbundene Stichproben.**

Differenzen $Y_i = X_i^{(2)} - X_i^{(1)}$ **symmetrisch um 0** verteilt?

H_0 : Für jedes Y_i ist + und - Vorzeichen gleich wahrscheinlich.

$G_i =$ Vorzeichen, $|Y_i| =$ „ Y_i “ im Zwei-Stichproben-Problem.

Für jede Vorzeichen-Konstellation $\underline{g}^{(\ell)} = [g_1^{(\ell)}, \dots, g_n^{(\ell)}]$

ist Wahrsch. $= 1/2^n$.

c Teststatistik $T(\underline{g}, \underline{z})$ festlegen,

$$g_i = +1 \text{ oder } = -1, z_i > 0.$$

Rand.-Vert. $P(T(\underline{G}, \underline{z}) = t) = \#\{g \mid T(\underline{g}, \underline{z}) = t\} / 2^n$

- $T(\underline{g}, \underline{z}) = (1/n) \sum_i g_i z_i = \text{ave}_i(y_i)$
entspricht dem t-Test für gepaarte Stichproben.
- $T(\underline{g}, \underline{z}) = \#\{i : g_i = 1\}$: Vorzeichentest.
- $T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i$, R_i : Rang von z_i :
Vorzeichen-Rangsummen-Test von Wilcoxon.

e Beispiel:

```
> wilcox.test(d.tranquillizer[,1], d.tranquillizer[,2],  
             paired=TRUE)  
      Wilcoxon signed rank test  
data:  d.tranquillizer[, 1] and d.tranquillizer[, 2]  
V = 40, p-value = 0.03906  
alternative hypothesis: true mu is not equal to 0  
knapp signifikant.
```

Achtung: „Vorher-Nachher-Vergleich“!

Richtig: Vergleich mit Kontrollgruppe oder Crossover-Versuch.

3.5 Schätzungen und Vertrauensintervalle

a **Modell:** Testfrage war: Ist Verteilung symmetrisch um 0?

Allgemeineres Modell: Verteilung symmetrisch um μ

$\Leftrightarrow Y_i - \mu$ symmetrisch um 0.

Test: Teststatistik $T\langle \underline{g}, \underline{y} - \mu \underline{1} \rangle$.

Grosse Werte = Abweichung von $H_0 : \mu$.

b Daraus ergibt sich eine **Schätzung:**

$$\hat{\mu} = \arg \min_{\mu} \langle T\langle \underline{g}, \underline{y} - \mu \underline{1} \rangle \rangle$$

- c Vorzeichen-Rangsummen-Test \longrightarrow Hodges-Lehmann-Schätzer.

Betrachte Walsh averages $(X_h + X_i)/2$.

$$\hat{\mu} = \text{med}_{h \leq i} \langle (X_h + X_i)/2 \rangle .$$

Beispiel Tranquilizier: 45 Walsh-Mittelwerte

-0.1470, -0.1135, -0.0800, -0.0685, -0.0350, 0.0100, ..., 1.022

Median $\hat{\mu} = 0.46$

d* Herleitung: $X_{[k]}$ k -t-kleinsten Wert.

$$X_{[k]} > 0, Z_{hk} = (X_{[h]} + X_{[k]})/2, h < k$$

$$Z_{hk} < 0, \text{ wenn } |X_{[h]}| > |X_{[k]}|.$$

$$\#\{Z_{hk} < 0\} = \#\{h \mid |X_{[h]}| < |X_{[k]}|\} = R_{[k]} - 1$$

$$R_{[k]} = \#\{h \mid Z_{hk} > 0, h \leq k\}.$$

$$X_{[k]} < 0 \implies Z_{hk} < 0, \text{ wenn } h < k.$$

$$T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > 0, h \leq k\}$$

Nullhypothese $\mu = \mu_0$:

$$T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > \mu_0, h \leq k\}$$

Test am wenigsten signifikant, wenn dies = $\frac{n(n+1)}{2}$ ist

$$\longrightarrow \hat{\mu} = \text{median}\langle Z_{hk} \mid h \leq k \rangle.$$

f **Vertrauensintervall** für Vorzeichen-Rangsummen-Test:

Grenzen des Ann.bereichs von T : c und $c' = n(n+1)/2 + 1 - c$
 Vertrauensgrenzen = c -ter und c' -ter Walsh-Mittelwert.

Beispiel Tranquilizer: $c = 6$, $c' = 40$,

Vertrauensintervall $[0.01, 0.786]$.

h Allgemeine Teststatistik $T(\underline{G}, \underline{z}^*; \mu)$: Betrachte

$$Q(\beta) = P(T(\underline{G}, \underline{z}^*; \mu) > T(\underline{g}^*, \underline{z}^*; \mu)) - \beta$$

Schätzung = Lösung von $Q(\beta = 0.5) = 0$.

Vertrauensgrenzen = Nullstellen für

$$Q(\beta = 0.025) = 0 \text{ und } Q(\beta = 0.975) = 0.$$

Lösbar!

3.6 Korrelation und Regression

a **Korrelation und einfache Regression.**

X_i, Y_i (X_i zufällig oder fest)

Nullhypothese: „kein Zusammenhang“

Randomisierung = „Paarung“ = Permutation von \underline{Y} .

Wahrsch. jeder Permutation = $1/n! = 1/(n(n-1)\dots 2 \cdot 1)$.

Teststatistik:

- gewöhnliche Korrelation,
- Rangkorrelation,
- robuste Schätzung des Regressions-Koeffizienten, ...

b **Multiple Regression:**

Permutation von \underline{Y} für Test der Hypothese, dass überhaupt kein Zusammenhang zwischen den Eingangs-Variablen und der Zielgröße besteht.

c **Zeitreihen:** Beobachtungen unabhängig?

Randomisierung: Permutation.

Testgröße: z.B. erste Autokorrelation.

d **Multiple Regression: Einzelner Koeffizient (oder mehrere)**

→ kein strikt richtiges Randomisierungsmodell.

e^* **Permutationen und andere Randomisierungen.**

Regression und Korrelation: Permutationen.

Bei zwei oder mehreren Gruppen: Auswahlen.

Permutationen: viel mehr;

viele führen zur gleichen Gruppenzugehörigkeit

→ gleiche Randomisierungs-Verteilung.

Hagelversuch: Anzahl potentielle Hageltage zufällig,

Anteil geimpfter zufällig.

Randomisierungs-Verteilung: Auswahlen von 33 aus 76 Tagen

→ **bedingter Test**, geg. die Anzahlen Impf- und Kontrolltage.

Merkpunkte

Randomisierungs-Tests

- **Randomisierungstests halten das Niveau exakt ein, ohne Voraussetzungen an die Verteilung. (Unabhängigkeit von Beobachtungen vorausgesetzt.)**
- Die **Teststatistik kann beliebig kompliziert sein.**
Wahl mit (informellen) Überlegungen zur Macht.
Robuste Teststatistik (z.B. aus Rängen) wählen!
- Es können auch Vertrauensintervalle konstruiert werden.