Crashkurs

Einführung Biostatistik

Prof. Burkhardt Seifert

Abteilung Biostatistik, ISPM
Universität Zürich

- Deskriptive Statistik
- Wahrscheinlichkeitsrechnung, Versuchsplanung
- **Statistische Inferenz**
  - **Prinzip statistischer Tests**
  - **Konfidenzintervalle**
  - **Stichprobengrösse, Power**
- Korrelation und einfache lineare Regression

Überblick: Statistische Inferenz

1. Prinzip eines statistischen Tests

2. Das 95%-Konfidenzintervall

3. Klassische statistische Tests

4. Power einer Studie und Wahl der Stichprobengrösse

Einführungsbeispiel 1: Stetige Daten

Gewichtsverlust nach einer Diät von 6 Monaten

(Cocco, Pandolfi and Rousson, *Heartdrug*, 2005)

- Rohdaten

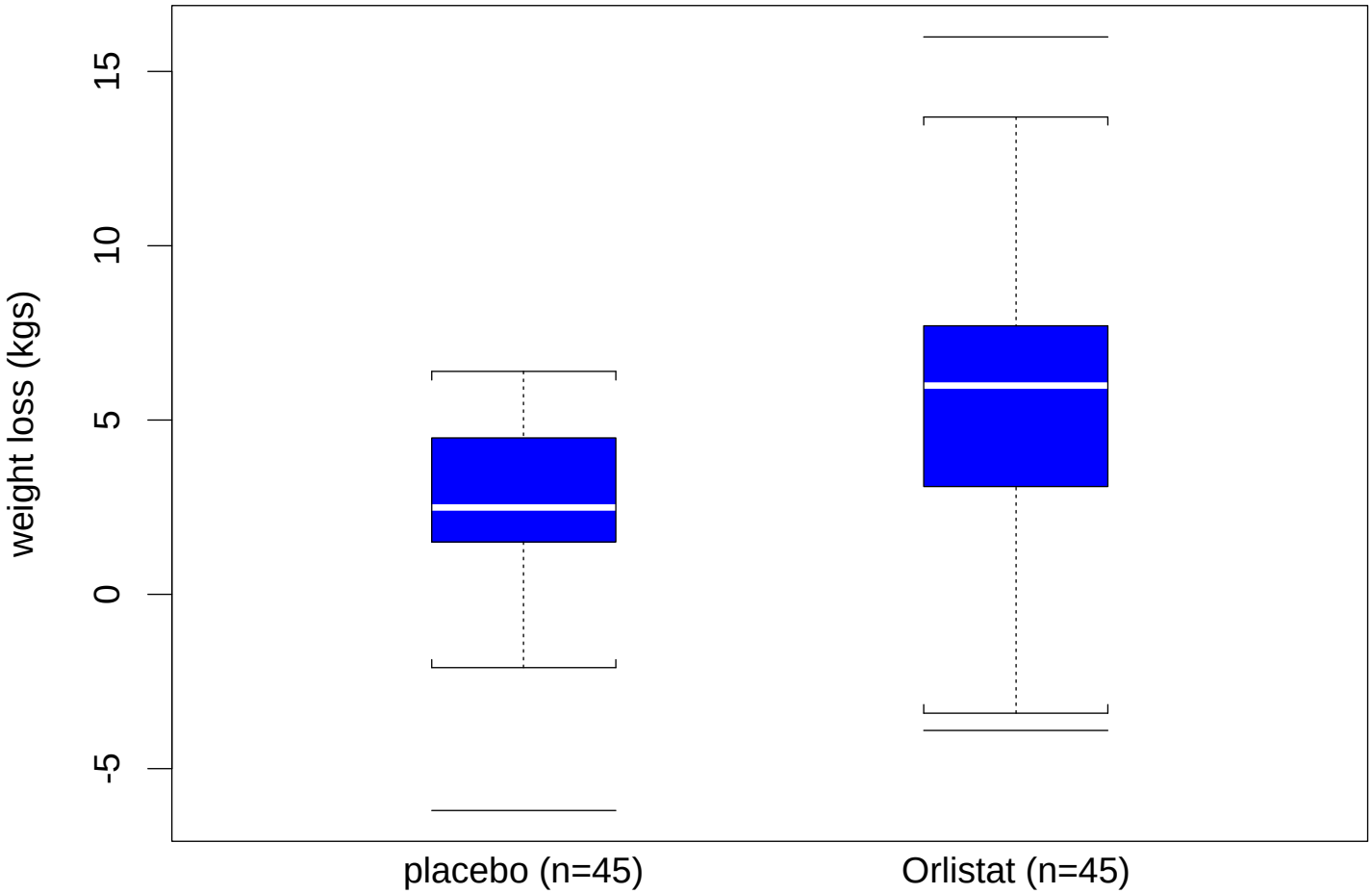
Orilstat: 0.8 3.0 7.4 7.9 8.6 3.1 8.6 10.8 6.0 3.6 6.0 6.1 7.9 6.0 3.0 8.7 6.2 3.2 4.2 7.6 16.0 3.1 6.9 5.6 8.3 7.7 4.4 4.6 7.3 11.6 13.7 7.3 7.8 6.7 -3.9 1.8 3.3 2.3 1.1 5.0 -0.8 7.0 4.3 -2.8 -3.4

Placebo: -0.9 1.5 3.4 5.6 5.2 6.4 2.9 5.6 3.7 1.7 2.2 3.8 5.5 0.7 4.6 1.4 2.0 -0.2 4.9 1.9 5.7 2.0 4.5 3.4 4.5 3.8 2.9 2.5 3.3 1.5 1.5 5.9 4.9 4.5 -2.1 -0.5 2.2 1.6 -0.6 0.0 -1.5 6.3 1.7 -6.2 -1.9

- Zusammenfassung (Behandlungseffekt)

Differenz der Mittelwerte: $5.41 - 2.48 = 2.93$ kg

Boxplots



Einführungsbeispiel 2: Binäre Daten

Lungenkomplikation nach Operation des Magen-Darms
(Petrowsky, Demartines, Rousson and Clavien, *Annals of Surgery*, 2004)

- Rohdaten

Mit Drainage: 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 1 0 1 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1

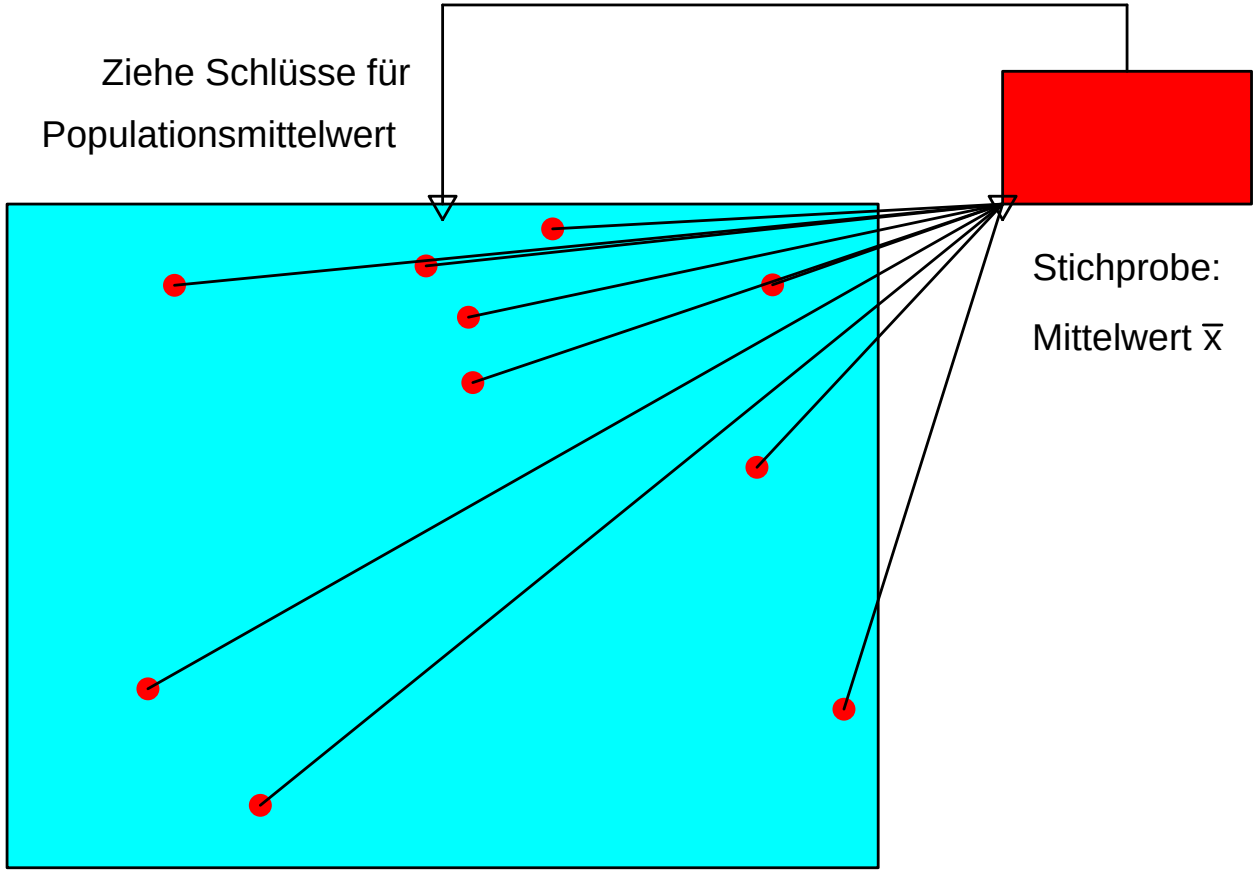
Ohne Drainage: 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 1 0 0 0 0 0 0 0

- Zusammenfassung (Behandlungseffekt)

	Komplikation (1)		Keine Komplikation (0)			
Mit Drainage	26	(16.9%)	128	(83.1%)	154	(100%)
Ohne Drainage	19	(12.7%)	131	(87.3%)	150	(100%)

Differenz der Anteile: $\frac{26}{154} - \frac{19}{150} = 4.2\%$

Grundgesamtheit und Stichprobe



Deskriptive versus schliessende Statistik

- Deskriptive Statistik:

  Daten (Stichprobe) beschreiben

  Kennwerte und Graphiken

  Behandlungseffekt quantifizieren

- Schliessende Statistik:

  Verallgemeinerung von Stichprobe zu Grundgesamtheit

  Tests und Konfidenzintervalle

  Beobachteter Behandlungseffekt als Schätzer des wahren Behandlungseffekts

1. Prinzip eines statistischen Tests

Statistische Frage: Ist der beobachtete Effekt echt oder zufällig?

- Wissenschaftliche Hypothese H_1: Der Effekt ist echt
(es gibt einen Effekt auch in der Grundgesamtheit)

- Nullhypothese H_0: Der Effekt ist zufällig
(es gibt keinen Effekt in der Grundgesamtheit)

Statistischer Test: Um H_1 nachzuweisen, zeigt man, dass H_0 unplausibel ist

p-Wert: Wahrscheinlichkeit, dass der Zufall allein einen so grossen oder grösseren Effekt wie den beobachteten Effekt produzieren kann

Ein kleiner p-Wert widerspricht H_0

Signifikanz versus Nicht-Signifikanz

- $p \leq 0.05$

- Der Effekt sollte echt sein (der Zufall allein kann einen solchen beobachteten Effekt nur selten produzieren)

- Der Effekt ist signifikant

- H_0 wird abgelehnt (Daten nicht kompatibel mit H_0)

- Statistischer Beweis für H_1

- $p > 0.05$

- Der Effekt könnte zufällig sein (man kann es nicht ausschliessen)

- Der Effekt ist nicht signifikant

- H_0 kann nicht abgelehnt werden (Daten kompatibel mit H_0)

- Vorsicht: Kein Beweis für H_0!

Der wahre und der beobachtete Effekt

Der wahre Effekt θ ist der Effekt, der in der Grundgesamtheit beobachtet würde, ist aber unbekannt

Beispiel: $\theta = 3.05$ kg

Der beobachtete Effekt $\widehat{\theta}$ ist ein Schätzer von θ

Beispiel: $\widehat{\theta} = 2.93$ kg

Statistische Gedanken: Wenn wir eine andere Stichprobe (aus derselben Grundgesamtheit) hätten, wäre der Wert von $\widehat{\theta}$ verschieden

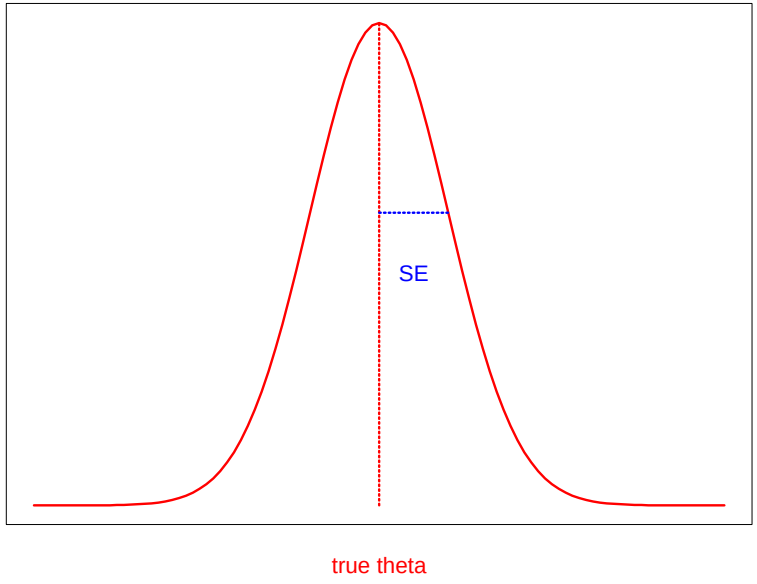
Beispiel: $\widehat{\theta} = 3.51$ kg oder $\widehat{\theta} = 1.57$ kg oder $\widehat{\theta} = -0.30$ kg

$\Rightarrow$ Der Schätzer $\widehat{\theta}$ variiert und hat also eine Verteilung

Die Verteilung eines Schätzers $\widehat{\theta}$

Oft approximativ normal und zentriert um den wahren θ

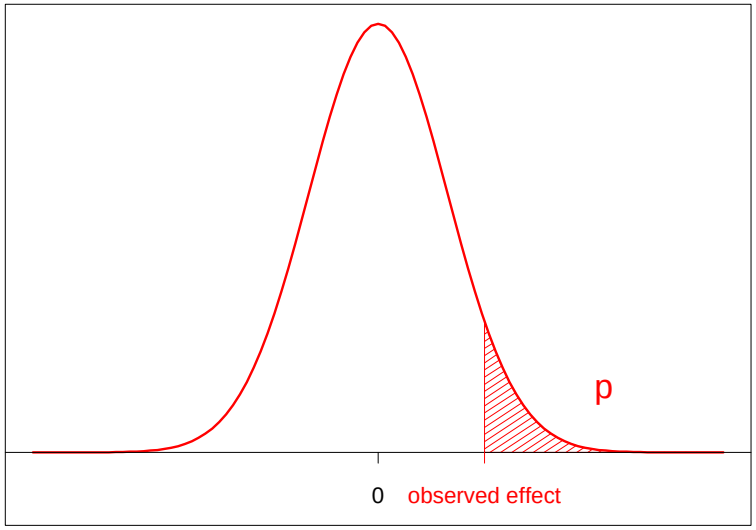
Der Standardfehler $\mathrm{SE}(\widehat{\theta})$ von $\widehat{\theta}$ ist die Standardabweichung dieser Verteilung und wird kleiner wenn n gross ist



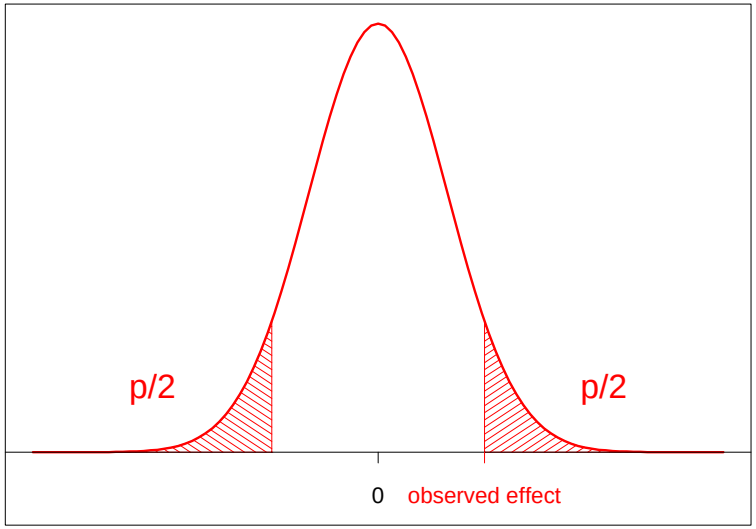
Um einen p-Wert zu berechnen, braucht man die Verteilung von $\widehat{\theta}$ unter H_0 (d.h. wenn $\theta = 0$)

Einseitige und zweiseitige Tests

one-sided test (the exception)



two-sided test (the rule)



Statistische Tests für Einführungsbeispiele

- Gewichtsverlust-Daten

Der beobachtete Effekt (Differenz zwischen zwei Mittelwerten) $\widehat{\theta} =$ 2.93 kg war signifikant mit einem ungepaarten t-Test oder mit einem Mann-Whitney Test ($p < 0.0001$) (die Nullhypothese $\theta = 0$ konnte abgelehnt werden)

- Lungenkomplikations-Daten

Der beobachtete Effekt (Differenz zwischen zwei Anteilen) $\widehat{\theta} = 4.2\%$ war nicht signifikant mit einem Chi-Quadrat Test ($p = 0.30$) oder mit Fishers exaktem Test ($p = 0.34$) (die Nullhypothese $\theta = 0$ konnte nicht abgelehnt werden)

Multiples Testen

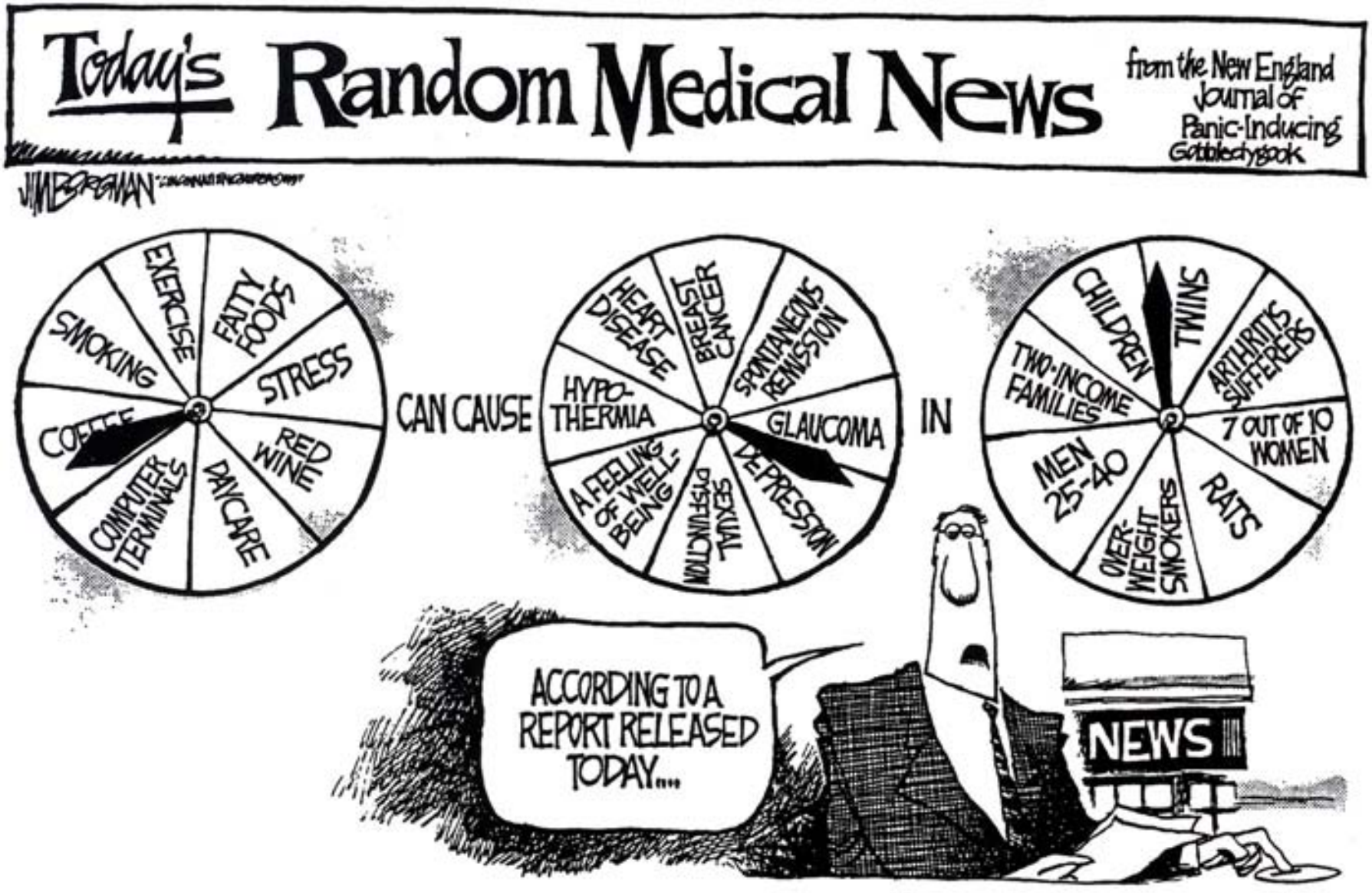
Falls kein Effekt vorliegt:

- 1 Test: Wahrscheinlichkeit einer falschen Signifikanz gleich 5%

- k Tests: Wahrscheinlichkeit mindestens einer falschen Signifikanz gleich $100 \cdot (1 - 0.95^k)\%$ (unter Unabhängigkeit)

Anzahl Tests	falsch sig.	Anzahl Tests	falsch sig.
1	5%	6	26%
2	10%	10	40%
3	14%	20	64%
4	19%	50	92%
5	23%	100	99%

Lösung: Bonferroni-Korrektur ($p \leq 0.05/k$ statt $p \leq 0.05$ als signifikant betrachten)


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
SMOKING
EXERCISE
FATTY FOODS
STRESS
RED WINE
DAYCARE
COMPUTER TERMINALS
COFFEE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
GLAUCOMA
DEPRESSION
SEXUAL DYSFUNCTION
A FEELING OF WELL-BEING
HYPO-THERMIA
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
7 OUT OF 10 WOMEN
RATS
OVER-WEIGHT SMOKERS
MEN 25-40
TWO-INCOME FAMILIES
ACCORDING TO A REPORT RELEASED TODAY...
NEWS

Signifikanz versus Relevanz

- n gross: Der Zufall hat wenig Einfluss

  $\Rightarrow$ Der Effekt kann signifikant sein, auch wenn er tatsächlich klein (nicht relevant) ist

- n klein: Der Zufall kann viel tun

  $\Rightarrow$ Der Effekt kann nicht-signifikant sein, auch wenn er tatsächlich gross (relevant) ist

- $p \leq 0.05$: Man weiss nur, dass der Effekt nicht null ist

  $\Rightarrow$ Information über das Vorzeichen, nicht über die Grösse (die Relevanz) des Effekts

- $p > 0.05$: Man weiss nicht einmal, ob der Effekt positiv oder negativ ist

  $\Rightarrow$ Nicht viel mehr als gar keine Information

2. **Das 95%-Konfidenzintervall** (95% CI)

Intervall, das den wahren aber unbekannten Effekt θ mit
95% Wahrscheinlichkeit enthält

Oft als folgendes berechnet: $\widehat{\theta} \pm 2 \cdot \mathrm{SE}(\widehat{\theta})$

Wird kleiner, wenn n gross wird

Enthält alle "plausiblen Werte" für θ
(alle Werte für θ die kompatibel mit den Daten sind)

Nützlich: $\theta = 0$ ausserhalb des 95% CI $\Leftrightarrow$ Der Effekt ist signifikant

$\Rightarrow$ Informiert über die Signifikanz und die Relevanz des Effekts

Konfidenzintervalle für Einführungsbeispiele

- Gewichtsverlust-Daten

Der beobachtete Effekt ist $\widehat{\theta} = 2.93$ kg und das 95% CI für θ ist [1.50 kg – 4.35 kg]

$\Rightarrow$ Der wahre Effekt ist mindestens 1.50 kg und könnte sogar bis zu 4.35 kg sein

- Lungenkomplikations-Daten

Der beobachtete Effekt ist $\widehat{\theta} = 4.2\%$ und das 95% CI für θ ist [-3.8% – 12.2%]

$\Rightarrow$ Der wahre Effekt könnte dreimal so gross sein wie beobachtet wurde, könnte aber auch negativ sein

Äquivalenz in klinischen Studien

Behandlungsgruppe: Erhält ein neues Medikament

Kontrollgruppe: Erhält ein Placebo oder ein etabliertes Medikament ("aktive Kontrolle")

- Placebo: Man will eine statistische Signifikanz zeigen, um die "Sicherheit" zu haben, dass $\theta > 0$ ist (in der Grundgesamtheit)

- Aktive Kontrolle: Oft ist es genug, eine statistische Äquivalenz zu zeigen

Definition: Zwei Medikamente sind äquivalent wenn θ nah bei null ist, d.h. wenn θ innerhalb eines Äquivalenzbereiches $\mathcal{D}$ liegt

Bemerkung: Die Definition vom $\mathcal{D}$ ist keine statistische Frage!

Beispiel: Man könnte im voraus entscheiden, dass Operation mit und ohne Drainage äquivalent sind wenn $-10\% \leq \theta \leq 10\%$

Statistische Äquivalenz

Definition: Zwei Medikamente sind statistisch äquivalent, wenn man "sicher" ist, dass θ innerhalb $\mathcal{D}$ liegt (in der Grundgesamtheit)

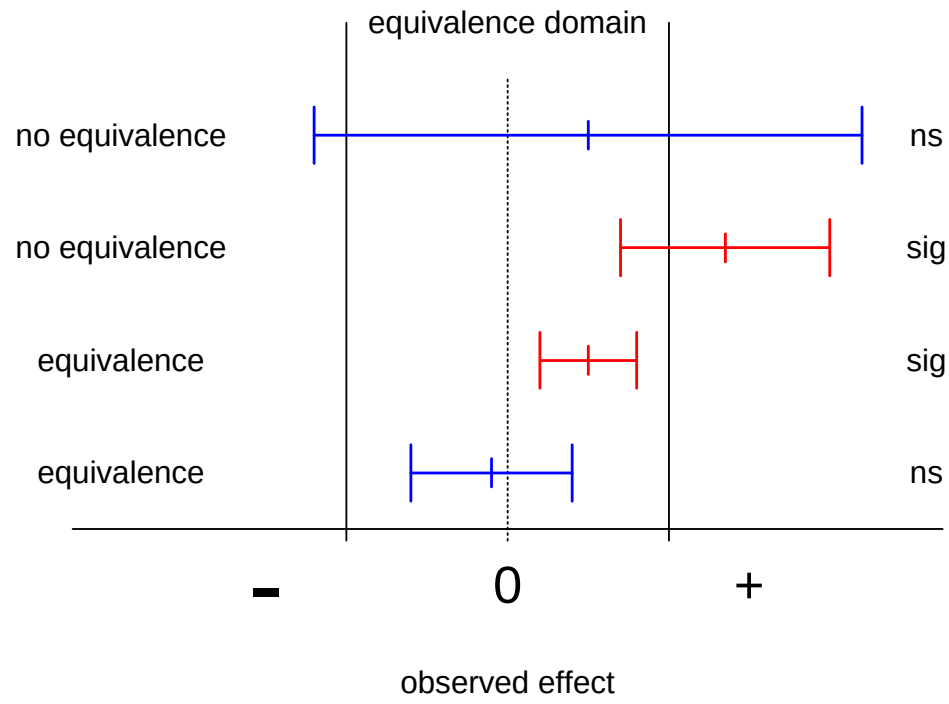
Das ist erreicht, wenn das 95% CI für θ ganz innerhalb $\mathcal{D}$ liegt

Beispiel: Die statistische Äquivalenz zwischen Operation mit und ohne Drainage wurde nicht gezeigt, da [-3.8% – 12.2%] nicht ganz innerhalb [-10% – 10%] liegt

Bemerkung: Für manche Autoren ist es genug ein 90% CI zu berechnen (Schuirmann, *J. of Pharmacokinetics and Biopharmaceutics*, 1987)

Bemerkung: Statistische Äquivalenz ist eine Art H_0 nachzuweisen

Signifikanz versus Äquivalenz



Die Länge des 95% CI ist auch eine Information

95% CI klein: Man kennt den wahren Wert von θ ziemlich genau

95% CI gross: Über θ weiss man nicht viel, aber man weiss, dass man nichts weiss

Konfidenzintervall für einen Mittelwert

Sei μ der Mittelwert in der Grundgesamtheit

Seien die Daten $x_1, x_2, \cdots, x_n$

Sei $\bar{x} = \sum_{i=1}^{n} x_i/n$ der Mittelwert in der Stichprobe (Schätzer von μ)

Standardabweichung (SD): $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Standardfehler von $\bar{x}$: $\mathrm{SE}(\bar{x}) = s_x/\sqrt{n}$

95% CI für μ: $\bar{x} \pm 2 \cdot s_x/\sqrt{n}$

Beispiel von Konfidenzintervall für einen Mittelwert

Normale Kinder laufen im Mittel nach 12 Monaten

Fragestellung: Ist Beginn des Laufens bei herzkranken Kindern verzögert?

Aus Daten berechnet man $\bar{x} = 12.8$ und $s_x = 1.8$

- Falls $n = 10$:

SE$(\bar{x}) = 0.57$ und 95% CI für μ ist $[11.7 - 13.9]$

$\Rightarrow$ Der Beginn des Laufens bei herzkranken Kindern ist nicht signifikant verzögert (Einstichproben-t-Test: 12 im 95% CI)

- Falls $n = 50$:

SE$(\bar{x}) = 0.25$ und 95% CI für μ ist $[12.3 - 13.3]$

$\Rightarrow$ Der Beginn des Laufens bei herzkranken Kindern ist signifikant verzögert (Einstichproben-t-Test: 12 nicht im 95% CI)

Konfidenzintervall für einen Anteil

Sei π der Anteil von "1" in der Grundgesamtheit

Sei p der Anteil von "1" in einer Stichprobe der Grösse n
(Schätzer von π)

Standardfehler von p: $\mathrm{SE}(p) = \sqrt{\frac{p(1-p)}{n}}$

95% CI für π: $p \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$

Beispiel: Zwischen 1950 und 1970 waren n =1'944'700 Geburten in der Schweiz, davon 997'600 Knaben

$\Rightarrow p = 0.513$ (oder 51.3%) und $\mathrm{SE}(p) = 0.0004$

$\Rightarrow$ 95% CI für π ist [51.2% – 51.4%]

$\Rightarrow$ Der Anteil Knaben bei Geburt ist signifikant grösser als 50%

3. Klassische statistische Tests

- Stetige Daten:

	Daten normalverteilt ? ja	nein
1 Stichprobe	Einstichproben-t-Test	Vorzeichentest
2 Stichproben ungepaart	ungepaarter t–Test	Mann-Whitney Test
2 Stichproben gepaart	gepaarter t-Test	Wilcoxon signed rank Test
$>$ 2 Stichproben ungepaart	ANOVA	Kruskal-Wallis Test
$>$ 2 Stichproben gepaart	ANOVA für wiederholte Messungen	Friedman-Test

- Binäre und kategorielle Daten: Chi-Quadrat-Test, Fishers exakter Test, McNemar-Test

Der t-Test

Effekt nicht signifikant wenn 0 innerhalb das 95% CI für θ:

$$\widehat{\theta} - 2 \cdot \mathrm{SE}(\widehat{\theta}) < 0 < \widehat{\theta} + 2 \cdot \mathrm{SE}(\widehat{\theta})$$

oder:

$$-2 \cdot \mathrm{SE}(\widehat{\theta}) < \widehat{\theta} < 2 \cdot \mathrm{SE}(\widehat{\theta})$$

t-Wert: $t = \dfrac{\widehat{\theta}}{\mathrm{SE}(\widehat{\theta})}$

- Effekt signifikant wenn $|t| > 2$

Exakte Tests:

- Kleines n: Effekt signifikant wenn $|t|$ grösser als das 97.5%-Quantil einer t-Verteilung (grösser als 2, bis zu 3 oder 4 für sehr kleines n)

- Grosses n: Effekt signifikant wenn $|t| > 1.96$

t-Wert versus p-Wert

Faustregel:

$$
\begin{array}{lcl}
|t| = 0 & \Leftrightarrow & p = 1/1 \\
|t| = 1 & \Leftrightarrow & p \approx 3/10 \\
|t| = 2 & \Leftrightarrow & p \approx 5/100 \\
|t| = 3 & \Leftrightarrow & p \approx 3/1000 \\
|t| = 4 & \Leftrightarrow & p \approx 1/10000
\end{array}
$$

Exakte Werte: (grosses n)

$$
\begin{array}{rcl}
p = 0.1 & \Leftrightarrow & |t| = 1.64 \\
p = 0.05 & \Leftrightarrow & |t| = 1.96 \\
p = 0.01 & \Leftrightarrow & |t| = 2.58 \\
p = 0.001 & \Leftrightarrow & |t| = 3.29 \\
p = 0.0001 & \Leftrightarrow & |t| = 3.89
\end{array}
$$

Exakte Werte: (kleines n)

t-Werte etwas grösser, berechnenbar mit statistischen Software

Zwei Stichproben: ungepaarte Daten

Daten: $x_1, x_2, \cdots, x_n$ (z.B. Behandlungsgruppe) und $y_1, y_2, \cdots, y_m$ (z.B. Kontrollgruppe)

Mittelwerte: $\bar{x} = \sum_{i=1}^{n} x_i/n$ und $\bar{y} = \sum_{i=1}^{m} y_i/m$

Differenz der Mittelwerte (Behandlungseffekt): $\widehat{\theta} = \bar{x} - \bar{y}$

SD: $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ und $s_y = \sqrt{\frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1}}$

Gemeinsame SD: $s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \approx \sqrt{\frac{s_x^2 + s_y^2}{2}}$ (wenn $n \approx m$)

Standardfehler von $\widehat{\theta}$: $\mathsf{SE}(\widehat{\theta}) = \frac{\sqrt{n+m} \cdot s_p}{\sqrt{nm}} \approx \frac{\sqrt{2} \cdot s_p}{\sqrt{n}}$ (wenn $n \approx m$)

Zwei Stichproben: gepaarte Daten

Daten: $x_1, x_2, \cdots, x_n$ (z.B. vor einer Behandlung) und $y_1, y_2, \cdots, y_n$ (z.B. nach einer Behandlung)

Mittelwerte: $\bar{x} = \sum_{i=1}^{n} x_i/n$ und $\bar{y} = \sum_{i=1}^{n} y_i/n$

Individuelle Differenzen: $d_1 = y_1 - x_1, d_2 = y_2 - x_2, \cdots, d_n = y_n - x_n$

Mittelwert der Differenzen oder Differenz der Mittelwerte (Behandlungseffekt): $\widehat{\theta} = \bar{d} = \sum_{i=1}^{n} d_i/n = \bar{y} - \bar{x}$

SD der Differenzen: $s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$

Standardfehler von $\widehat{\theta}$: $\mathrm{SE}(\widehat{\theta}) = \frac{s_d}{\sqrt{n}}$

Wie für eine Stichprobe (Einstichprobe t-Test)

Beispiel von gepaarten Daten

Fragestellung: Verbessert sich die Herzrate bei Diabetikern bei guter Glukose-Einstellung?

Daten: $n = 8$ Patienten wurden zweimal untersucht
(x: bei schlechter Einstellung und y: bei guter Einstellung)

	x	y	d
	74	66	-8
	72	67	-5
	84	62	-22
	53	47	-6
	75	56	-19
	87	60	-27
	69	63	-6
	71	68	-3
Mittelwert	73.1	61.1	-12.0
SD	10.3	6.9	9.2

Resultate: $\widehat{\theta} = -12.0$, $s_d = 9.2$, $\mathrm{SE}(\widehat{\theta}) = 3.3$, $t = -3.7$ und $p = 0.008$

Beispiel von ungepaarten Daten

Die selben Daten, wenn zwei Gruppen von $n = m = 8$ Patienten (nicht zweimal die selben Patienten) untersucht wurden (Gruppe 1: schlechte Einstellung und Gruppe 2: gute Einstellung)

Gruppe	Herzrate
1	74
1	72
1	84
1	53
1	75
1	87
1	69
1	71
2	66
2	67
2	62
2	47
2	56
2	60
2	63
2	68

Resultate: $\widehat{\theta} = -12.0$, $s_p = 8.8$, $\mathrm{SE}(\widehat{\theta}) = 4.4$, $t = -2.7$ und $p = 0.016$

Rangtests

Idee: Benütze nur Rangordnung der Daten, ähnlich zu Median

Vorteile:

- Ohne Annahme der Normalverteilung
- Unempfindlich gegen Ausreisser und Extremwerte
- Anwendbar für Ordinaldaten
- Gute Power auch für Normalverteilung (für nicht zu kleines n)

Nachteile:

- Keine Signifikanz möglich für sehr kleines n
- Nicht verallgemeinbar für eine multivariate Analyse

Der Mann-Whitney Test (Wilcoxon rank sum Test)

1. Erstelle gemeinsame Rangordnung

2. Berechne getrennt mittlere Ränge der x_i und der y_i

3. Die mittleren Ränge dienen dem Programm dazu, p-Werte zu berechnen

Der Wilcoxon signed ranks Test

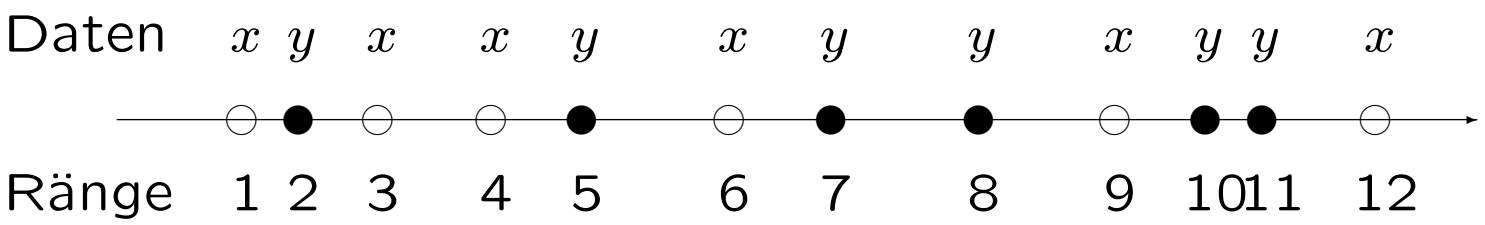
1. Berechne individuelle Differenzen $d_i = y_i - x_i$

2. Berechne Rangsummen der negativen und positiven Differenzen (Null-Differenzen fallen völlig aus der Rechnung heraus)

3. Die Rangsummen dienen dem Programm dazu, p-Werte zu berechnen

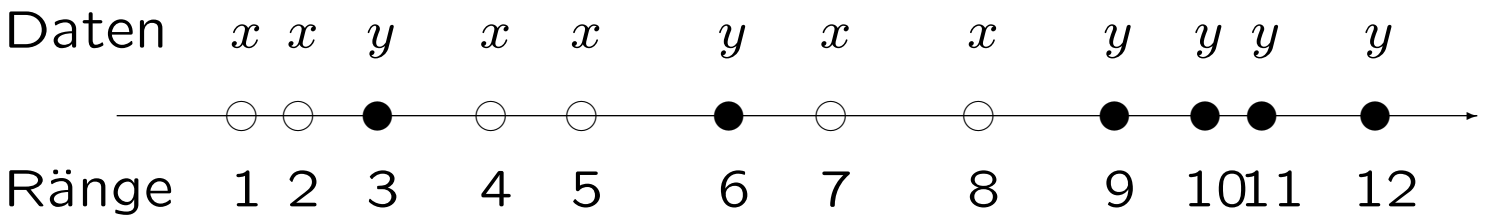
Beispiele für Mann-Whitney Test

- Beispiel 1



Der mittlere Rang der x_i-Werte ist 5.8, derjenige der y_i ist 7.2, so dass sie annähernd gleich sind und $p = 0.59$

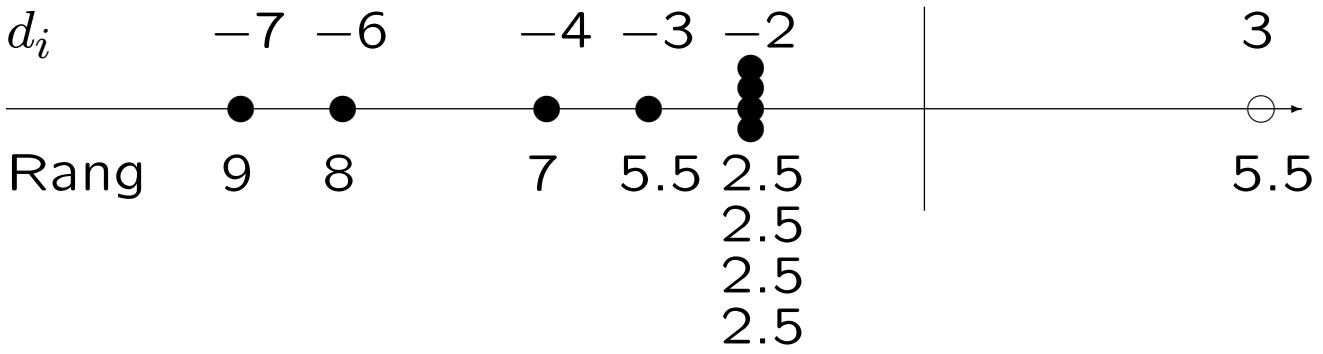
- Beispiel 2



Der mittlere Rang der y_i-Werte ist mit 8.5 deutlich grösser als derjenige der x_i mit 4.5 und $p = 0.065$

Beispiel für Wilcoxon signed ranks Test

x	y	d
19	12	−7
17	15	−2
10	8	−2
15	13	−2
7	10	3
18	12	−6
13	11	−2
12	8	−4
20	17	−3



Die Rangsumme der negativen Differenzen ist 39.5 und diejenige der positiven Differenzen ist nur 5.5

Der Zufall kann solchen Unterschied kaum produzieren ($p = 0.042$)

ANOVA und Kruskal-Wallis Test

Erlaubt mehr als 2 Gruppen zu vergleichen

- $p > 0.05$: Man kann nicht ausschliessen, dass alle Gruppen denselben Mittelwert (dieselbe Verteilung) haben

- $p \leq 0.05$: Man hat statistisch bewiesen, dass nicht alle Gruppen denselben Mittelwert (dieselbe Verteilung) haben

  Problem: Man weiss noch nicht, welche Gruppen sich voneinander signifikant unterscheiden

  Lösung: Post-hoc Tests, wo man jeweils zwei Gruppen vergleicht

  Vorsicht: Man soll das Problem des multiplen Testens berücksichtigen (z.B. Bonferroni Korrektur)

Selbe Problematik für Tests mit mehr als 2 gepaarten Stichproben

Der Chi-Quadrat-Test

Beispiel: In einer randomisierten Studie haben 40 Patienten das Medikament A und 40 Patienten das Medikament B erhalten

	gut		schlecht			
Medikament A	25	(62.5%)	15	(37.5%)	40	(100%)
Medikament B	33	(82.5%)	7	(17.5%)	40	(100%)
	58	(72.5%)	22	(27.5%)	80	(100%)

Falls beide Medikamente genau gleich wären, würden wir erwarten

	gut		schlecht			
Medikament A	29	(72.5%)	11	(27.5%)	40	(100%)
Medikament B	29	(72.5%)	11	(27.5%)	40	(100%)
	58	(72.5%)	22	(27.5%)	80	(100%)

Kann der Zufall die Unterschiede zwischen den zwei Tabellen produzieren? Der Chi-Quadrat-Test liefert $p = 0.045$

Fishers exakter Test ist mehr konservativ und liefert $p = 0.078$

Der McNemar-Test (Vorzeichen Test)

Beispiel: Dieselben Daten, wo dieselben 40 Patienten zweimal untersucht wurden (mit Medikament A und mit Medikament B) und wo die beiden Medikamente für 20 Patienten gut waren

	Med. B gut	Med. B schlecht	
Med. A gut	20	5	25
Med. A schlecht	13	2	15
	33	7	40

Falls beide Medikamente genau gleich wären, würden wir erwarten

	Med. B gut	Med. B schlecht	
Med. A gut	20	9	29
Med. A schlecht	9	2	11
	29	11	40

Kann der Zufall die Unterschiede zwischen den zwei Tabellen produzieren? Der McNemar-Test liefert $p = 0.096$

Chi-Quadrat Test für mehr als 2 Stichproben

Beispiel: An 400 Kindern wird deren Händigkeit geprüft, und ob Vater und Mutter links-oder rechtshändig sind

	Kind rechts		Kind links			
Eltern rechts, rechts	303	(89.1%)	37	(10.9%)	340	(100%)
Eltern rechts, links	29	(76.3%)	9	(23.7%)	38	(100%)
Eltern links, links	16	(72.7%)	6	(27.3%)	22	(100%)
	348	(87.0%)	52	(13.0%)	400	(100%)

Falls die Kinder der drei Gruppen von Eltern genau dieselbe Händigkeit-Verteilung hätten, würden wir erwarten

	Kind rechts		Kind links			
Eltern rechts, rechts	295.8	(87.0%)	44.2	(13.0%)	340	(100%)
Eltern rechts, links	33.1	(87.0%)	4.9	(13.0%)	38	(100%)
Eltern links, links	19.1	(87.0%)	2.9	(13.0%)	22	(100%)
	348	(87.0%)	52	(13.0%)	400	(100%)

Kann der Zufall die Unterschiede zwischen den zwei Tabellen produzieren? Der Chi-Quadrat-Test liefert $p = 0.010$

Fishers exakter Test liefert ebenso $p = 0.010$

Studiendesign und Wahl der Stichprobengrösse

Erster Schritt: Quantifiziere wissenschaftliche Fragestellung!

Primärer Endpunkt: Variable, für die die Stichprobengrösse berechnet wird $\Rightarrow$ "harter" statistischer Schluss nur gültig für primären Endpunkt!

"hart" im Sinn von: halte α ein.

Sekundäre Endpunkte: Weitere interessierende Variablen, werden bei Planung der Stichprobengrösse nicht berücksichtigt.

Wichtige Frage: Welches α wählen wir für die sekundären Endpunkte (Multiples Testen!) $\Rightarrow$ vor Auswertung festlegen!

Bestimmung der Stichprobengrösse

Behandeln wir zuwenige Patienten: unethisch weil

- wissenschaftliche Hypothese kann wahrscheinlich nicht nachgewiesen werden (zuwenig Power),
- behandle Patienten mit möglicherweise unwirksamer Therapie,
- Verschwendung von Ressourcen.

Behandeln wir zuviele Patienten: unethisch weil

- mehr Patienten als nötig erhalten ineffektive Therapie,
- Routineanwendung von effektiver Therapie wird verzögert.

Regulatorische Instanzen (Swissmedic) und seriöse med. Zeitschriften verlangen transparent geplante klinische Studien.

Keine Berechnung der Stichprobengrösse: Eigenschaften des statistischen Tests bzw. des Konfidenzintervalls nicht klar.

Problem: Wie können wir die optimale Stichprobengrösse vor der Studie ermitteln?

Bestimmung der Stichprobengrösse für einen statistischen Test

Vorüberlegungen:

- Primärer Endpunkt?
- Verwendete statistische Methode?
- Klinisch relevanter Effekt?
- Bei stetigem Endpunkt: Variabilität der Daten?
- Bei binärem Endpunkt: Anteil in der Population?

Beispiel mit binärem Endpunkt

Wissenschaftliche Fragestellung: Hat eine neue Behandlung einen Effekt auf das Sterberisiko?

Quantifizierung: Vergleich der Sterberisiken p_1 und p_2 zweier Therapiegruppen.

Klarheit schaffen:

- Wie gross ist der Effekt?
- Studientyp: Prospektive kontrollierte randomisierte Therapiestudie.
- Primärer Endpunkt: Sterblichkeit innerhalb 30 Tagen nach Infarkt.
- Statistische Methode:
  - Statistischer Test: χ^2-Test, Fishers exakter Test oder
  - Konfidenzintervall für Differenz von Sterberisiken.

Festlegen relevanter Grössen

Wissenschaftliche Hypothese H_1: wahre Sterberisiken sind unterschiedlich: $p_1 \neq p_2$.

Nullhypothese H_0: Risiken sind gleich: $p_1 = p_2$.

Signifikanzniveau α: Wahrscheinlichkeit für Signifikanz durch Zufall. üblicherweise $\alpha = 5\%$.

Power (1 - β): Wahrscheinlichkeit, die wissenschaftliche Hypothese zu beweisen, wenn sie wahr ist. Typische Powerwerte 1 - β: 80% oder 90%.

Ziel: Stichprobengrösse so wählen, dass Wahrscheinlichkeit für Fehler 1. und 2. Art kontrolliert wird.

Problem: Power hängt vom Sterberisiko p_1 in der Kontrollgruppe und der Risikodifferenz (der Effektgrösse) RD $= p_1 - p_2$ ab.

Lösung: Wir legen die relevante Effektgrösse, die wir nachweisen möchten, fest.

Wo bekomme ich p_1 und RD her?

RD $= p_1 - p_2$ ist die kleinste klinisch relevante Risikodifferenz

$\Rightarrow$ fachlich diskutieren (nicht statistisch)!

Wir wählen RD $= 5\%$.

p_1: Sterberisiko in der Kontrollgruppe (jene mit Standardtherapie):

- Kontrollgruppe ist sicher schon in der Literatur beschrieben worden (Standardtherapie), sonst Pilotstudie.
- Aus alter Studie wissen wir für Standardtherapie: Anteil der Verstorbenen war $19/151 = 12.6\%$, also setzen wir $p_1 = 12.6\%$.

Beachte: Berechnung der Stichprobengrösse ist nur grobe Abschätzung $\Rightarrow$ die Grössen, die zu ihrer Berechnung benutzt werden, brauchen nicht allzu genau zu sein.

Stichprobengrössen für Beispiel

Notwendige Stichprobengrösse pro Gruppe bei $p_1 = 12.6\%$ für Fishers exakten Test:

p_1	p_2	RD	$\beta = 80\%$	$\beta = 90\%$
0.126	0.116	0.01	16889	22542
0.126	0.076	0.05	601	792
0.126	0.026	0.10	121	151