

Von Stichproben, Konfidenzintervallen, statistischen Tests und p -Werten

Kaspar Rufibach

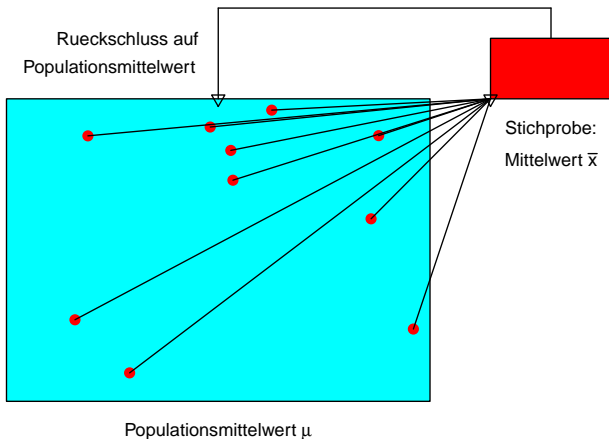
Universität Zürich
Institut für Sozial- und Präventivmedizin
Abteilung Biostatistik

22. März 2010

Inhaltsverzeichnis

- 1 Worum geht es in der Statistik?
- 2 Statistische Unsicherheit
- 3 Testen von wissenschaftlichen Hypothesen
- 4 Der p -Wert

Worum geht es in der Statistik?



⇒ Ziehe Schlüsse über Population aufgrund einer **zufälligen** Stichprobe.

Begriffe

Population, Grundgesamtheit: Interessierende Beobachtungseinheiten (Patienten) für eine gegebene Fragestellung.

Stichprobe: zufällig aus der Population gezogen.

Variable: interessierende Grösse, z.B. Blutdruck, Ansprechen auf Behandlung ja/nein, Zeit von Diagnose bis Tod.

Parameter: Kennzahl (z.B. Mittelwert, Minimum) einer Variablen in der Population.

Schätzer: Kennzahl einer Variablen in der Stichprobe.

Beschreibende vs. schliessende Statistik

Deskriptive Statistik:

- Beschreibe Daten, keine Schlüsse auf Population.
- Kennzahlen, Grafiken.
- (Behandlungs)Effekt quantifizieren.

Inferenz- oder schliessende Statistik:

- Schliesse von Stichprobe auf Grundgesamtheit.
- Tests und Konfidenzintervalle.
- beobachteter (Behandlungs)Effekt als Schätzer des wahren Effekts.

Beispiel für stetige Daten

Population: Alle Personen mit metabolischem Syndrom.

Parameter: Gewichts­differenz Start Diät (=Baseline) bis 6 Monate (=Follow-up). Wir betrachten **Gewichtsverlust**.

Stichprobe: 90 Patienten in Orlistat-Studie (Orlistat = Xenical).

Cocco, Pandolfi und Rousson, (2005)

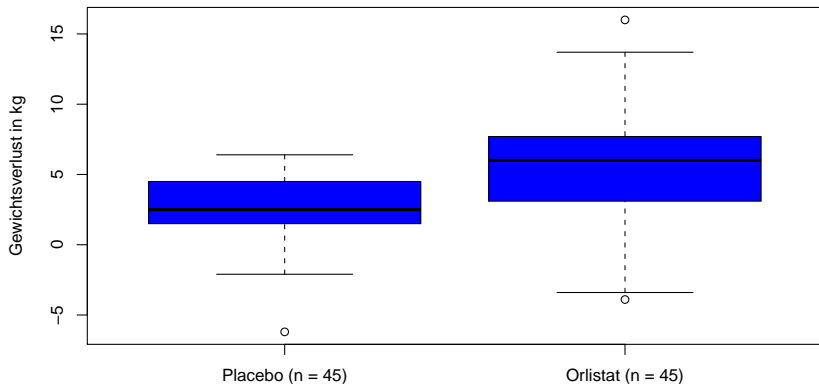
```
> orlistat
```

```
[1] 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  
[17] 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  
[33] 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
```

```
> placebo1
```

```
[1] -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  
[17] 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  
[33] 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
```

Verteilung der Orlistat-Daten



Im Mittel **höherer Gewichtsverlust** mit Orlistat.

Behandlungseffekt = Differenz der Gewichtsverlusts-Mittelwerte.

Schätzung: $\hat{\delta} = 5.41 \text{ kg} - 2.48 \text{ kg} = 2.93 \text{ kg}$.

Inhaltsverzeichnis

- 1 Worum geht es in der Statistik?
- 2 Statistische Unsicherheit**
- 3 Testen von wissenschaftlichen Hypothesen
- 4 Der p -Wert

Statistische Unsicherheit

Stichprobe:

- zufällig gezogen,
- verschiedene Stichproben S_1, S_2, \dots liefern verschiedene Schätzer $\hat{\delta}_1, \hat{\delta}_2, \dots$

Beobachteter Effekt $\hat{\delta} = 2.93$ kg:

- Schätzung des Parameters δ . Quantifizierung der Unsicherheit in dieser Schätzung?
- Tatsächlich vorhanden oder nur durch Zufall verursacht?
- Ist Effekt grösser als bei Gleichheit der Gruppen erwartet würde?

Konfidenzintervall: Enthält plausible Werte für δ .

Statistische Tests: Prüfen wissenschaftlicher Hypothesen.

Konfidenzintervall für Mittelwert

Konfidenzintervall für einen **Mittelwert**:

$$[\bar{x} - f \cdot \text{se}(\bar{x}), \bar{x} + f \cdot \text{se}(\bar{x})]$$

Für $f = 1.96$ überdeckt dieses Intervall den Populationsmittelwert μ mit Wahrscheinlichkeit $\approx 95\%$ (falls n gross genug). Falls n klein: f grösser.

f ist eine Konstante, die von n und der **Vertrauenswahrscheinlichkeit** $1 - \alpha$ (hier 95%) abhängt

95%- t -Konfidenzintervall für Vergleich Gewichtsabnahme Orlistat - Placebo:

$$[(5.4 - 2.5)\text{kg} \pm 1.99 \cdot 0.72\text{kg}] = [1.5\text{kg}, 4.4\text{kg}].$$

Das Intervall [1.5kg, 4.4kg] **überdeckt** die wahre Gewichtsabnahme-Differenz mit Wahrscheinlichkeit 95%.

Beachte Formulierung! Konfidenzintervall ist zufällig, Parameter ist fest.

Es ist **plausibel** zu sagen, dass Orlistat eine **höhere** Gewichtsabnahme zur Folge hat als das Placebo.

Konfidenzintervall

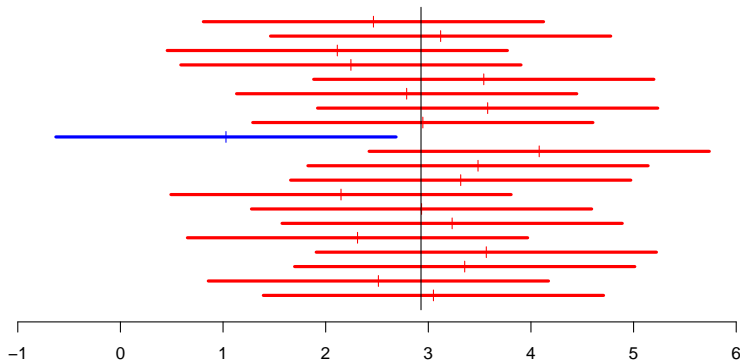
Ein $(1-\alpha)$ -Konfidenzintervall

- ist von der Stichprobe abhängig und deshalb **zufällig**,
- enthält **plausible Werte** für einen Parameter (den wahren, aber unbekanntem Effekt), jene Werte für den Parameter, die „mit den Daten verträglich“ sind.
- **überdeckt** den wahren Parameter mit Wahrscheinlichkeit $1 - \alpha \Rightarrow$ d.h. mit Wahrscheinlichkeit 95% falls $\alpha = 0.05$.
- ist grösser, je grösser die Unsicherheit bezüglich des betrachteten Parameters ist.
- wird kleiner, je grösser die Anzahl Beobachtungen n ist oder
- wird kleiner, je kleiner der Standardfehler ist.
- liefert einen Hinweis zur **Relevanz** eines Effekts.

Vertrauenswahrscheinlichkeit

$(1 - \alpha)$: Gibt an, wie oft der wahre Wert im Mittel überdeckt wird. Keine Garantie, dass KI den wahren Wert enthält!

20 Stichproben aus derselben Verteilung wie Orlistat-Studie. 19 Intervalle überdecken wahren Wert $\mu = 2.93\text{kg}$.



Anwendung von Konfidenzintervallen

Konfidenzintervalle lassen sich für beliebigen interessierenden Parameter berechnen:

- Mittelwert einer stetigen Variablen (Beispiel),
- Differenz der Mittelwerte zweier Gruppen,
- Anteil,
- Differenz der Anteile zweier Gruppen,
- Regressionsparameter,
- etc.

Beachte: „Das“ Konfidenzintervall gibt es nicht. Je nach gewünschten theoretischen Eigenschaften können verschiedene Konfidenzintervalle für denselben Parameter berechnet werden.

Inhaltsverzeichnis

- 1 Worum geht es in der Statistik?
- 2 Statistische Unsicherheit
- 3 Testen von wissenschaftlichen Hypothesen**
- 4 Der p -Wert

Testen von wissenschaftlichen Hypothesen

Ziel: Beurteile **wissenschaftliche Hypothese** anhand der Daten.

Statistischer Test: Wir beurteilen, ob der Unterschied als zufällig betrachtet werden kann oder ob er über das, was man zufällig erwarten würde wenn kein Unterschied bestünde, hinausgeht.

Beispiel: Gewichtsabnahme mit Orlistat verschieden von jener mit Placebo?

Hypothese über Parameter (Populationskennzahl), beurteilt anhand einer Stichprobe \Rightarrow übersetze in **statistische Fragestellung**.

Nullhypothese H_0 : **Gegenteil der wissenschaftlichen Hypothese**, d.h. kein Effekt vorhanden:

$$\delta_{\text{Orlistat}} = \delta_{\text{Placebo}}$$

Alternativhypothese H_1 : Motivation des Experiments, wissenschaftliche Hypothese:

$$\delta_{\text{Orlistat}} \neq \delta_{\text{Placebo}}$$

Statistischer Test

Verwerfe H_0 falls aufgrund der Stichprobe als **unplausibel** beurteilt.

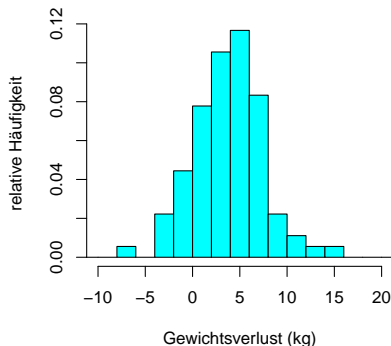
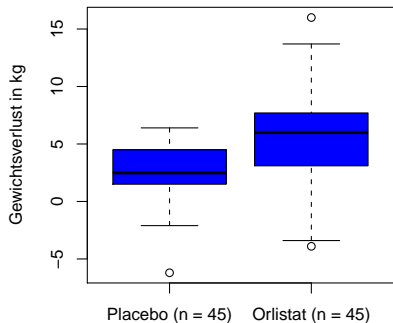
Weder H_0 noch H_1 können „bewiesen“ werden!

Schema eines statistischen Tests:

- Berechne aus Daten **Teststatistik** $t \Rightarrow$ „misst“, wieviele Standardfehler H_0 von H_1 entfernt ist.
- Berechne Verteilung von t unter H_0 .
- Prüfe, ob t mit Nullverteilung verträglich ist.

H_0 ist unplausibel falls Teststatistik „weit weg“ von Nullverteilung.

Beispiel Gewichtsabnahme

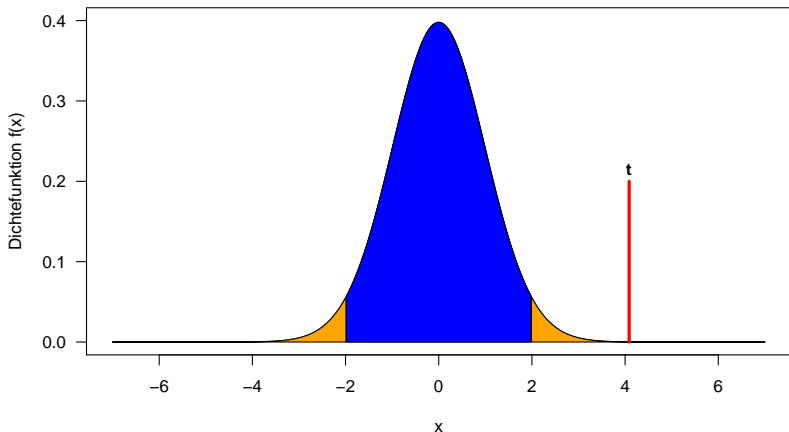


Gewichtsabnahmewerte sind gut normalverteilt \Rightarrow *t*-Test.

Berechne Teststatistik:

$$\begin{aligned} t &= (\bar{\delta}_{\text{Orlistat}} - \bar{\delta}_{\text{Placebo}}) / \text{se}(\bar{\delta}_{\text{Orlistat}} - \bar{\delta}_{\text{Placebo}}) \\ &= \frac{5.41 - 2.48}{0.72} = 4.09. \end{aligned}$$

Unter H_0 : Teststatistik t folgt t -Verteilung mit $n_1 + n_2 - 2 = 88$ Freiheitsgraden:



Wenn kein Effekt vorhanden wäre (d.h. H_0 gälte), würden wir eine Teststatistik t im blauen Bereich erwarten.

t ist aber ausserhalb des blauen Bereichs \Rightarrow lehne H_0 zugunsten von H_1 ab, Test liefert ein **signifikantes** Resultat.

Signifikanzlevel α

Signifikanzlevel: Wahrscheinlichkeit H_0 abzulehnen **obwohl** H_0 gültig ist, d.h. kein Effekt vorhanden ist.

Üblich: $\alpha = 0.05$. Bestimmt den blauen Bereich in der Grafik.

Wenn t im blauen Bereich gelegen hätte \Rightarrow kein Beweis dass H_0 wahr ist! Im Wesentlichen wissen wir **nichts!**

Mögliche Fehler in einem statistischen Test

Fehler erster Art, α -Fehler, („Type I error“):

- Lehne H_0 ab obwohl H_0 stimmt.
- Wahrscheinlichkeit eines Fehlers erster Art: α .
- Diesen Fehler will man **unbedingt** vermeiden! Nutzloses Medikament soll nicht als wirksam deklariert werden.
- Lege α **im voraus** fest, verändere es nicht.
- Hängt nicht von n ab.

Fehler zweiter Art, β -Fehler, („Type II error“):

- Verwerfe H_0 nicht obwohl H_1 wahr ist (beachte Formulierung!).
- Wahrscheinlichkeit eines Fehlers zweiter Art: β .
- Power = $1 - \beta$: „Fähigkeit“ einen vorhandenen Effekt zu entdecken.
- Je grösser α und/oder n , desto kleiner β .
- β ist nur bekannt wenn wahrer Effekt bekannt oder spezifiziert.

Vorsicht bei der Interpretation von Testresultaten

Fehler 1. Art ist nicht **Falsch-Positiv-Rate!**

20 Tests, kein Effekt in allen 20 Fragestellungen \Rightarrow im Mittel ist ein Test **falsch-signifikant**.

Beispiel: 200 klinische Versuche, nur 10% der untersuchten Therapien effektiv.

- Das Signifikanzniveau (Fehler 1. Art) sei $\alpha = 5\%$, Fehler 2. Art sei $1 - \beta = 20\%$.

Ergebnis	Behandlung		Gesamt
	nicht effektiv	effektiv	
Nicht signifikant	171	4	175
Signifikant	9	16	25
Gesamt	180	20	200

- $9/25 = 36\%$ der signifikanten Ergebnisse **falsch positiv!**

$\Rightarrow \alpha, \beta$ geben die Fehlerraten „in the long run“ an.

$\alpha = 0.05$ **bedeutet nicht** dass jedes 20te signifikante Resultat falsch ist!

Inhaltsverzeichnis

- 1 Worum geht es in der Statistik?
- 2 Statistische Unsicherheit
- 3 Testen von wissenschaftlichen Hypothesen
- 4 Der p -Wert**

Der p -Wert

Test: liefert Entscheid auf H_0 oder H_1 . Dazu ist **kein p -Wert nötig!**

Der p -Wert ist

- 1 **NICHT** die Wahrscheinlichkeit der beobachteten Daten unter H_0 ,
- 2 **NICHT** die „beobachtete“ α -Fehlerrate,
- 3 **NICHT** die „false discovery rate“, d.h. die Wahrscheinlichkeit, dass ein signifikantes Resultat „falsch-positiv“ ist,
- 4 **NICHT** die Wahrscheinlichkeit von H_0 .

Ergänzung zu 4: H_0 und H_1 haben je Wahrscheinlichkeit 50%. Der beobachtete p -Wert sei $p = 0.05$ (0.01, 0.001)

⇒ dann ist die posteriori-Wahrscheinlichkeit (unter gew. Annahmen) für H_0 **nicht kleiner** als 32.1% (13.3%, 2.4%)!

Der p -Wert

Aus „Forschung mit Menschen - Ein Leitfaden für die Praxis“ der
Schweizerischen Akademie der medizinischen Wissenschaften (p. 102):

Der p -Wert bezeichnet die Wahrscheinlichkeit (probability) der Nullhypothese, d.h. die Wahrscheinlichkeit, dass das beobachtete Resultat durch Zufall zustande kam. Wird in Dezimalbrüchen angegeben (0.05 entspricht 5%).

Der p -Wert

Der p -Wert ist...die Wahrscheinlichkeit, dass falls H_0 (kein Effekt) gilt, der beobachtete oder ein noch extremerer Effekt beobachtet wird.

Ursprüngliche Idee (Fisher): **Informeller Index** als Mass für Diskrepanz zwischen H_0 und H_1 .

[Goodman, 1999] Fisher suggested that the p -value be used as part of the fluid, non-quantifiable process of drawing conclusions from observations, a process that included combining the p -value in some unspecified way with background information.

Fasse Evidenz gegen H_0 **für das vorliegende Experiment** in einer einzigen Zahl zusammen. Stelle diese Zahl in Zusammenhang!

p -Wert ist Wahrscheinlichkeit unter $H_0 \Rightarrow$ uns interessiert aber eigentlich H_1 !

Der p -Wert

Test und p -Wert sind voneinander unabhängige Konzepte. In med. Literatur als „ein Ganzes“ betrachtet.

Streng genommen: Verwende p -Wert bei stat. Tests **nur**, um zwischen H_0 und H_1 zu entscheiden: $p > 0.05 \Rightarrow H_0$, $p \leq 0.05 \Rightarrow H_1$.

Orlistat: $p < 0.0001 \Rightarrow$ starke Evidenz gegen H_0 .

Vergleiche p -Wert mit α : Wenn $p \leq \alpha \Rightarrow$ Test ist signifikant.

Kein „borderline significant“ oder ähnlicher Quatsch!

[Goodman, 2005] In fact, the p -value is almost nothing sensible you can think of. I tell students to give up trying.

Konfidenzintervall vs. statistischer Test vs. p -Wert

Orlistat-Trial:

- Konfidenzintervall: [1.5kg, 4.4kg].
- Test lehnt H_0 ab.
- p -Wert $< 0.0001 \Rightarrow$ starke Evidenz gegen H_0 .

Konfidenzintervall: liefert Schätzer für Effektgrösse **und Unsicherheit für diese Schätzung**.

Test entscheidet zwischen Hypothesen.

p -Wert: Fasst Evidenz gegen H_0 in einer Zahl zusammen, keine weiteren Informationen.

Empfehlung: Immer **Konfidenzintervall** und **Testresultat** angeben, kann durch p -Wert ergänzt werden.

Statistische Signifikanz vs. klinische Relevanz

Signifikanz und Relevanz sind verschiedene Aspekte!

„signifikant“ bedeutet lange nicht „relevant“!

Orlistat-Trial: [1.5kg, 4.4kg], Test lehnt H_0 ab, $p < 0.0001$.

Resultat ist **statistisch signifikant** und Differenz gross genug um **klinisch relevant** zu sein.

Fähigkeit einen vorhandenen Effekt zu entdecken hängt von n ab \Rightarrow berichte auch nicht-signifikante Resultate, idealerweise mit Konfidenzintervallen!

Vielen Dank für Ihre Aufmerksamkeit.

Abteilung Biostatistik
Institut für Sozial- und Präventivmedizin
Universität Zürich
Hirschengraben 84
8001 Zürich

`kaspar.rufibach@ifspm.uzh.ch`

`http://www.biostat.uzh.ch`

Statistischer Beratungsservice.