Wiederholung: Statistische Tests

Dr. Dirk Hasenclever

University of Leipzig

imise.

Basisverfahren statistischer Inferenz

- **Schätzen von unbekannten Modellparametern**
  - **Punktschätzer**
    - Was ist ein guter Schätzwert?
  - **Konfidenzintervalle** zum Punktschätzer
    - Wie genau wissen wir es eigentlich?
- **Gute biometrische Praxis**:
  - ***Kein Schätzwert ohne Konfidenzintervall!***

- **Beurteilen von statistischen Hypothesen**
  - **Nullhypothese**
    - Behauptung / Theorie über den unbekannten Parameter
  - **Alternativhypothese**
    - Hypothese über Art der Abweichung von der Nullhypothese.
  - **Statistischer Test**
    - Sind die **Daten verträglich mit der Nullhypothese**?
    - *Ergebnis eines statistischen Tests ist ein **Argument**!*

Beispiel für statistische Hypothese

Es sollen zwei Therapien A und B bezüglich ihrer Erfolgswahrscheinlichkeiten p_A und p_B verglichen werden. p_A und p_B sind die "wahren" Parameter der beiden zu Grunde liegenden Bernoulli-Verteilungen.

Parameterraum: $\Omega = [0,1] \; x \; [0,1]$

Nullhypothese: $p_A = p_B$ d.h. formal:

$$H_0 : \quad (p_A, p_B) \in \Omega_0 = \{(p,q) \in \Omega = [0,1] \; x \; [0,1] \quad | \quad p = q\}$$

Wie kann man argumentieren?

- **Der Proponent** macht eine statistische Hypothese über latente Parameter eines identisch wiederholbaren Zufallsexperiments.
- **Der Opponent** spezifiziert eine Zufallsvariable (Stichprobenfunktion), die aus einer Stichprobe des identisch wiederholbaren Zufallsexperiments eine zusammenfassende Teststatistik auszurechen erlaubt.
- **Der Proponent** leitet aus seiner statistischen Hypothese ab, wie die Teststatistik verteilt sein sollte, falls er Recht hat. Diese Nullverteilung repräsentiert die Erwartung des Proponenten über die Teststatistik.
- Gemeinsam generieren **Proponent und Opponent** eine Stichprobe, berechnen die Teststatistik und prüfen, ob die Teststatistik zu der Erwartung des Proponenten (Nullhypothese) passt (-> p Wert).

Statistischer Test als p-Wert Prozedur

Nullhypothese bestimmt statistisches Modell

= ???

Konkretes Experiment

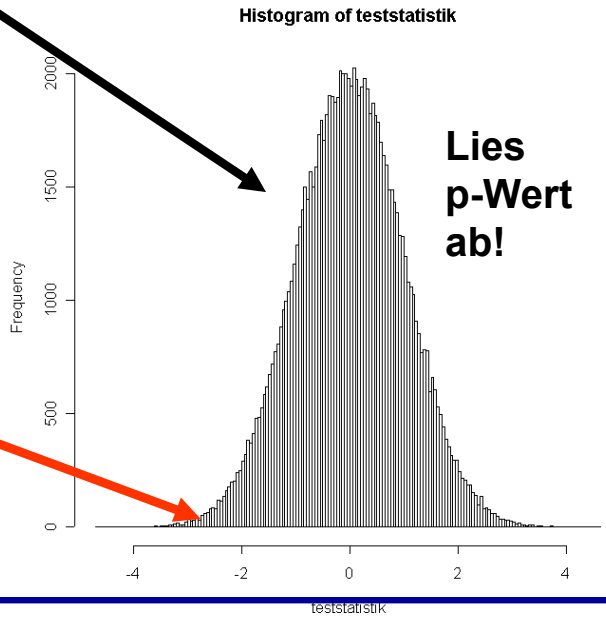
Beobachte einmal

Die eine beobachtete Stichprobe

Simulierte viele Stichproben (oder bestimme analytisch)

Maß für Ab-weichung

Verteilung der Teststatistik



Konkreter Wert der Teststatistik

Lies p-Wert ab!

Liegt der Testwert extremal (=unerwartet)?

Argument gegen die Nullhypothese II
p-Wert

Experiment: Berechne die Realisierung $T = t$ der Teststatistik aus den beobachteten Daten.

Berechne den p-Wert= die Wahrscheinlichkeit unter der Nullverteilung $T = t$ oder einen noch extremeren(=unerwarteteren) Wert der Teststatistik zu beobachten.

Ist der p-Wert klein (z.B. <5% oder 1% oder 0.1%),
so ist der Proponent in Schwierigkeiten:

Entweder es ist etwas Mögliches aber sehr Unwahrscheinliches passiert
ODER es stimmt etwas nicht mit der Nullhypothese.

Verteilung des P Werts unter der Nullhypothese

Die Teststatistik T ist eine Zufallsvariable mit Verteilung $F_T(t)$ und Dichte $f_T(t)\,dt$

Der einseitige p-Wert ist eine Zufallsvariable $P = F_T(T)$, die unter der Nullhypothese gleichverteilt auf [0,1] ist:

Nach Regel für die Dichtefunktion einer transformierten Zufallsvariable gilt:

$$f_P(p)dp = f_T\left(F^{-1}{}_T(p)\right) \cdot \frac{d}{dp}\left(F^{-1}{}_T(p)\right) = 1\ \mathrm{dp}$$

da die Ableitung der Umkehrfunktion
das Inverse der Ableitung ist.

Schätzen Sie p !

Therapie A				Therapie B	
E1=	13			E2=	10
N1=	20			N2=	20
P1=	0.65			P2=	0.5

Erfolgsraten 13/20 versus 10/20

```
 2-sample test for equality of proportions
with continuity correction

data:  c(13, 10) out of c(20, 20)
X-squared = 0.4092, df = 1, p-value = 0.5224
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2028446  0.5028446
sample estimates:
prop 1 prop 2
  0.65   0.50
```

Schätzen Sie p !

Therapie A				Therapie B	
E1=	130			E2=	100
N1=	200			N2=	200
P1=	0,65			P2=	0,5

Erfolgsraten 130/200 versus 100/200

```
 2-sample test for equality of proportions
with continuity correction

data:  c(130, 100) out of c(200, 200)
X-squared = 8.6036, df = 1, p-value = 0.003355
alternative hypothesis: two.sided
95 percent confidence interval:
 0.04923214 0.25076786
sample estimates:
prop 1 prop 2
  0.65   0.50
```

p-Wert Prozedur gegen Neyman-Pearson Entscheidungsprozedur

- **p-Wert Prozedure:**
- **Überprüfung einer wissenschaftlichen Theorie (Nullhypothese)**
- **Theorie mit Daten verträglich?**
- p-Wert = Wahrscheinlichkeit unter H_0 einen so extremen oder extremeren Testwert zu beobachten
- p-Wert als **Argument** gegen die Nullhypothese.

- **Entscheidungsprozedur:**
- Proponent und Opponent vereinbaren hartes **Entscheidungsverfahren** zwischen H_0 und H_A aufgrund noch zu erhebender Daten!
- „**Experimentum crucis**“
  - Z.B. Zulassungsverfahren eines neuen Medikaments
  - Qualitätstest vor Abnahme einer Warenlieferung

Test als Entscheidungsverfahren I

Entscheidungsregel:

Bestimme p-Wert:

Falls $p \leq \alpha$ H_0 wird abgelehnt (Opponent gewinnt)

Falls $p > \alpha$ H_0 wird beibehalten (Proponent gewinnt)

Proponent wählt Signifikanzniveau α. Z.B. $\alpha = 0.05$ oder $\alpha = 0.01$

Damit kontrolliert er vorab seinen

$$\alpha = \textit{Fehler erster Art} = pr(\textit{Test gegen } H_0 \mid H_0 \textit{ richtig})$$

Test als Entscheidungsverfahren II

Opponent wählt eine spezifische Alternativhypothese H_A^*
(eine Punkthypothese in Ω_A, damit er die Verteilung
der Teststatistik unter seiner Erwartung H_A^* berechen kann)

Darauf basierend bestimmt er die nötige Fallzahl.

Damit kontrolliert er vorab (bevor das Experiment durchgeführt wird)
seinen

$$\beta = Fehler\ zweiter\ Art = pr(Test\ behält\ H_0\ bei \mid H_A^*\ richtig)$$

bzw. seine

$$Power = 1 - \beta = pr(Test\ gegen\ H_0 \mid H_A^*\ richtig)$$

Beispiel: Geschmackstest (vgl. Übung) I

Experiment: Proband soll aus N=8 Paaren von Geschmacksproben

einen bestimmten Geschmack herausfinden.

Sei p seine unbekannte Trefferwahrscheinlichkeit

Nullhypothese: H_0 : p=0.5 (Proband rät bloß!)

Spezifische Alternative: H_A : p=0.9 (Proband ist sich ziemlich sicher)

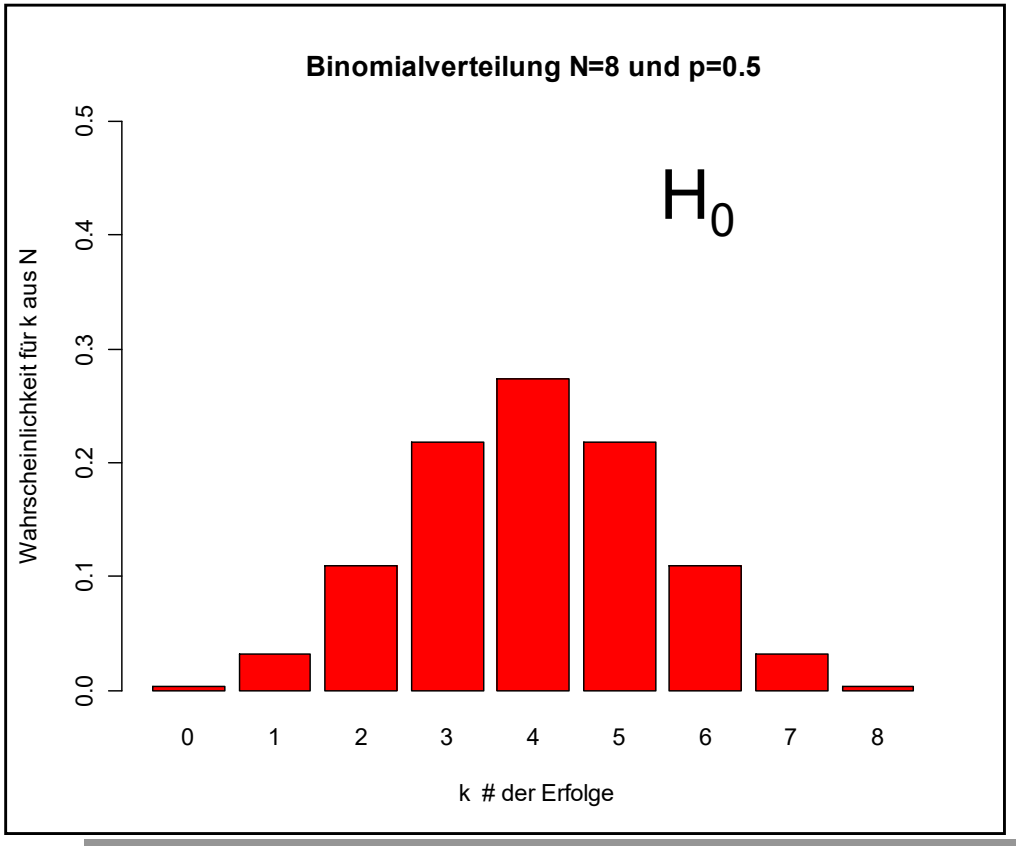
Entscheidungsregel: Sei k die Anzahl der Treffer

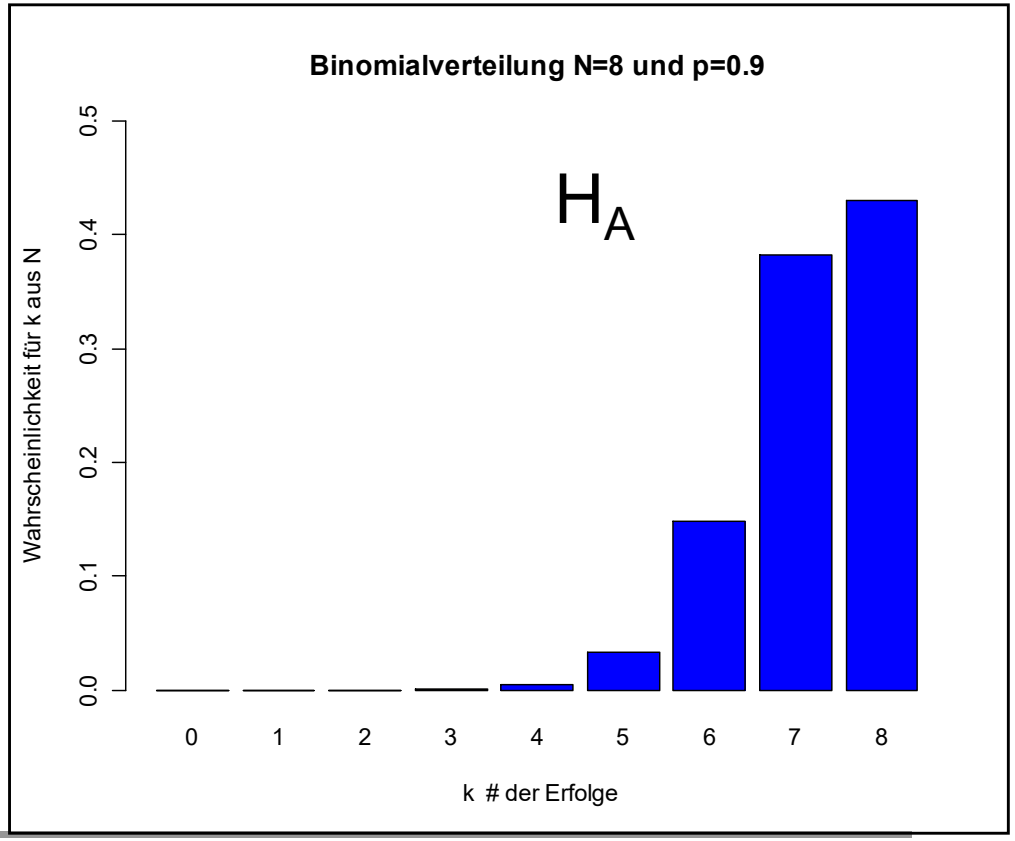
Lehne H_0 ab wenn k= 7 oder 8, sonst behalte H_0 bei.

Binomial-Verteilung

Wahrscheinlichkeit unter N Versuchen k Erfolge zur beobachten, wenn die Erfolgswahrscheinlichkeit p zugrunde liegt.

$$\text{pr}(X=k)=\binom{N}{k} p^k (1-p)^{N-k}$$





Beispiel: Geschmackstest (vgl. Übung) II

Der Versuch ist binomial mit N=8.

Fehler 1.Art:

$pr(k \geq 7 | H_0 : p=0.5) = dbinom(7,8,0.5)+dbinom(8,8,0.5) \approx$

$0.03125 + 0.00390625 \approx 0.03515625 < 0.0352$

Fehler 2.Art:

$pr(k < 7 | H_A^* : p=0.9) = sum(dbinom(0:6,8,0.9)) < 0.187$

Also Power $\approx$ 81.3%

Proponent und Oppenent können sich auf den Versuch einlassen.

Zwei Gruppen t-test

```
> t.test(Wr.Hnd~Sex)

        Welch Two Sample t-test

data:  Wr.Hnd by Sex
t = -10.6187, df = 215.265, p-value < 2.2e-16
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -2.544482 -1.747753
sample estimates:
mean in group Female   mean in group Male
            17.59576             19.74188
```

Zwei-Stichproben (ungepaarter) t-Test I

$$X_i \sim N(\mu_X, \sigma_X^2) \quad i = 1..n$$

$$Y_k \sim N(\mu_Y, \sigma_Y^2) \quad k = 1..m$$

$$\overline{X}_n \sim N(\mu_X, \frac{\sigma_X^2}{n}) \qquad \text{und} \qquad \overline{Y}_m \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$$

$$(\overline{X}_n - \overline{Y}_m) \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(\frac{\mu_X - \mu_Y}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}, 1)$$

Zwei-Stichproben (ungepaarter) t-Test II

Unter $H_0 : \mu_X = \mu_Y$ und falls σ_X und σ_Y bekannt sind, gilt :

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Wenn die Varianzen aus den Daten

geschätzt werden müssen, gilt approximativ :

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}} \sim t_{df=\min(n,m)-1}$$

bzw. eine noch genauere Approximation nach Welch

welche gebrochene Freiheitsgrade (df) liefert.

VORSICHT

- Ein **Test als Entscheidungsprozedur setzt voraus, dass Proponent und Opponent des Versuch gemeinsam geplant haben** (bzw. mit der Kontrolle der Fehler 1. und 2. Art zufrieden sein können).

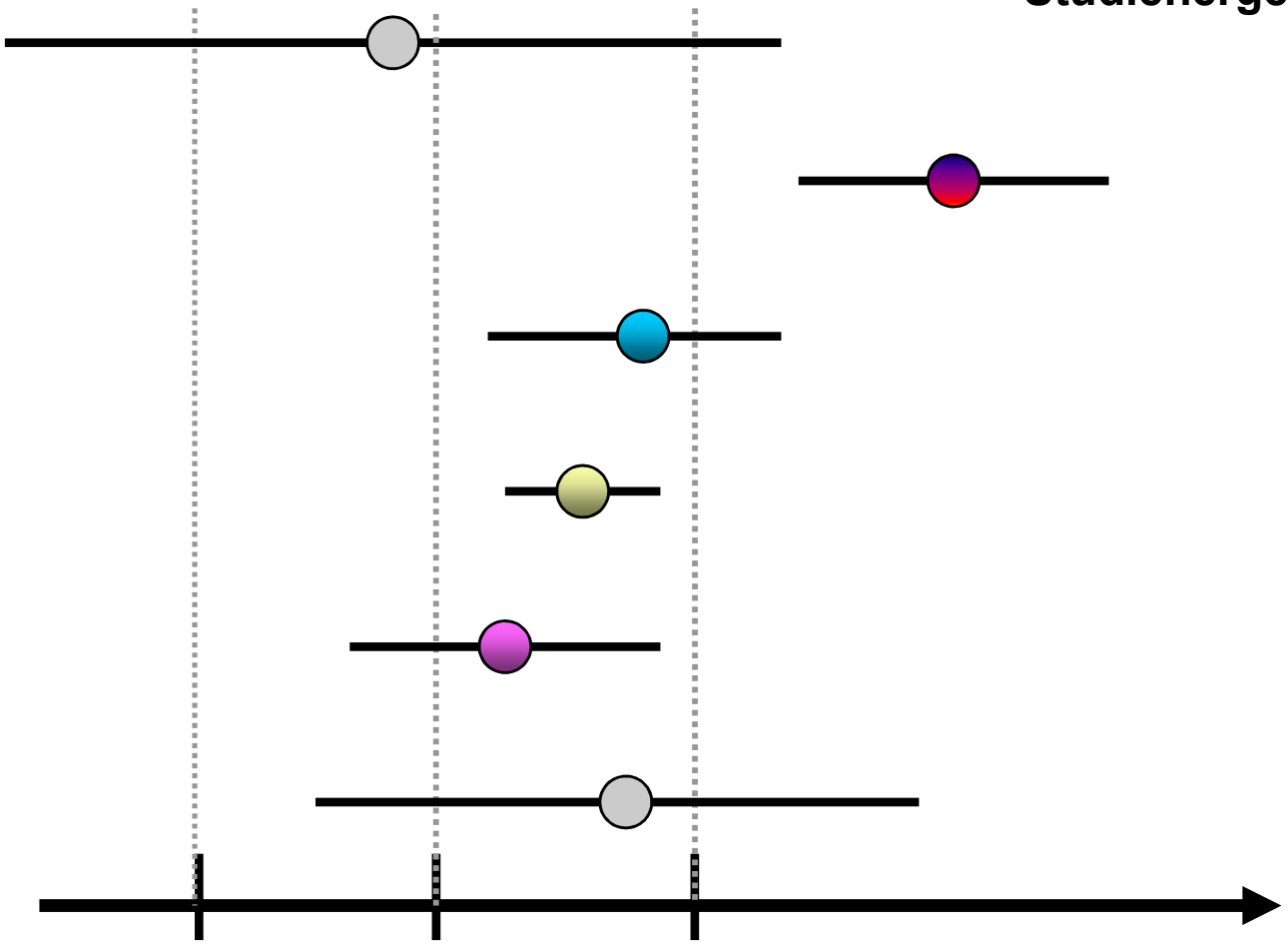
- Ein Test als p-Wert Prozedure beurteilt nur, **ob die Evidenz aus den Daten ausreicht, die Nullhypothese in Frage zu stellen**. Eine Entscheidung fällt nicht automatisch.

Logik des statistischen Tests

„Unterschied nicht signifikant“
heißt nicht:
„Es gibt keinen Unterschied!“

- Wenn Sie einen Zufallsbefund
  nicht ausschließen können,
  d.h. einen Skeptiker (noch) nicht überzeugen können,
  bedeutet dies nicht, dass
  die Nullhypothese wahr ist,
  oder Sie an sie glauben sollten.

Punktschätzer & Konfidenzintervall ≈ statistischer Test + quantitative Information



Studienergebnis – Vergleich zweier Therapien

inkonklusiv

relevant besser

besser
unklar, ob relevant besser

besser,
aber nicht relevant besser

weder relevant besser
noch relevant schlechter

nicht relevant schlechter



Schätzer mit Konfidenzintervall