Multiple Hypothesentests

Markus Pauly and Thilo Welz

Wintersemester 2021

Regularien

- **Vorlesung:** 1x pro Woche 2h
- **Voraussetzung** Mind. einen Schein aus Wahrscheinlichkeitstheorie oder Entscheidungstheorie
- **Materialien:** Auf Moodle
- **Prüfung:**
  - Die Prüfungsform wird zu Beginn des Semester festgelegt

Literatur

- Blakesley, et al. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology, 23(2), 255.
- Dmitrienko, A. et al. (2010). Multiple testing problems in pharmaceutical statistics. CRC Press.
- Hochberg, Y. and Tamhane, A. C. (1987). Multiple comparison procedures. John Wiley & Sons.
- Lehmann, E. L. and Romano, J. P. (2006). Testing statistical hypotheses. Springer.
- Pigeot, I. (2000). Basic concepts of multiple tests – a survey. Statistical Papers, 41(1), 3-36.
- Shaffer, J. P. (1995). Multiple hypothesis testing. Annual review of psychology, 46(1), 561-584.
- Westfall, P.H. and Young, S.S.(1993). Resampling-based multiple testing: Examples and methods for p-value adjustment (Vol. 279). John Wiley & Sons.

Ziele der Vorlesung

- MOT: Häufig möchte man (z.B. für einen Datensatz) mehrere Fragestellungen gleichzeitig beantworten. Man spricht von sog.
  - **statistischen Mehrentscheidungsverfahren** bzw. von
  - **simultanen Inferenzverfahren**.
- Problem: Multiplizität bzw.
  Addition der Fehlerwahrscheinlichkeiten
- Genauer: Führt man 100 Tests zum Niveau $\alpha = 0.05$ aus, so lehnt man u.U. allein durch Zufall wahre Nullhypothesen ab.
- Die Wahrscheinlichkeit irgendeinen Fehler 1.Art zu machen wird also i.d.R. nicht durch α kontrolliert.
- Beispiel...

Motivation

Beispiel

- Simulation von 100 t-Tests zum Niveau $\alpha = 0.05$:

```
> set.seed(1)
> x<-rep(100)
> for (i in 1:100){
+      x[i]<-t.test(rnorm(100))$p.value
+ }
> sum(x<0.05)
[1] 5
```

- Ergebnis hier: Insgesamt 5 falsche Ablehnungen (wie auch erwartet).

Beispiel 1.1 (Balanced One-Way ANOVA):
Seien $X_{ij} \overset{iid}{\sim} N(\mu_i, \sigma^2)$ mit $1 \leq i \leq k$, $1 \leq j \leq n$, $\mu_i \in \mathbb{R}$, $\sigma^2 > 0$ unbekannt
In diesem Fall überprüft der F-Test der klassischen Varianzanalyse die Nullhypothese

$$H_0 : \{\mu_1 = \mu_2 = \cdots = \mu_k\}$$

Frage: Was macht man, wenn H_0 zum Niveau α (z.N. α) abgelehnt wird? Dies führt auf folgende Problemstellungen:

Problem 1

Paarweise Vergleiche der μ_i's, d.h. teste

$$H_{i\ell} : \{\mu_i = \mu_\ell\} \text{ vs. } K_{i\ell} : \{\mu_i \neq \mu_\ell\} \qquad \text{für alle} \quad 1 \leq i < \ell \leq k$$

Das gibt uns ... zusätzlich zu testende Nullhypothesen.

Naive Idee: Teste jedes $H_{i\ell}$ mittels t-Test z.N. α!
Aber: Wahrscheinlichkeit für irgendeinen Fehler 1.Art ist i.A. $\gg \alpha$
$\Rightarrow$ Notwendigkeit für multiple Testverfahren zum **multiplen Niveau** α.

Forderung: $\mathbb{P}[\ \textit{mind. 1 Fehler 1. Art}\ ] \leq \alpha$

Problem 1

Paarweise Vergleiche der μ_i's, d.h. teste

$$H_{i\ell} : \{\mu_i = \mu_\ell\} \text{ vs. } K_{i\ell} : \{\mu_i \neq \mu_\ell\} \qquad \text{für alle} \quad 1 \leq i < \ell \leq k$$

Das gibt uns $\binom{k}{2}$ zusätzlich zu testende Nullhypothesen.

Naive Idee: Teste jedes H_{jl} mittels t-Test z.N. α!
Aber: Wahrscheinlichkeit für irgendeinen Fehler 1.Art ist i.A. $\gg \alpha$
$\Rightarrow$ Notwendigkeit für multiple Testverfahren (zum **multiplen Niveau** α).

Forderung: $\mathbb{P}[$ *mind. 1 Fehler 1. Art* $] \leq \alpha$

Problem 2

Konfidenzintervall (KI) für alle $\theta_{il} := \mu_i - \mu_l$
Ziel: Finde ein KI $C_{il}(x)$, so dass die Wahrscheinlichkeit für irgendeine falsche "Uberdeckung $\theta_{ij} \notin C_{ij}(x)$ durch ein vorgegebenes, **multiples Konfidenzniveau** α kontrolliert wird, d.h.:
Für alle festen Erwartungswerte $\mu = (\mu_1, \ldots, \mu_k)^T \in \mathbb{R}^k$, $\sigma^2 > 0$ gilt:

$$\mathbb{P}_{\mu,\sigma^2}[C_{il}(X) \ni \theta_{il} \text{ für alle } 1 \leq i < l \leq k] \geq 1 - \alpha,$$

wobei $X = (X_{11}, \ldots, X_{kn})$.
$\Rightarrow$ Das führt auf sog. multiple Konfidenzbereiche.

Problem 3: Hochdimensionale Daten

In der Genom-Forschung hat man bei sog. *Micro-Array-Studien* mehr Hypothesen als Beobachtungen.
Hier wird z.B. gerne pro Gen eine Hypothese aufgestellt.

$\Rightarrow$ *klassische* multiple Fehrlerkontrolle ist i.d.R. zu restriktiv.

$\Rightarrow$ Dies führt auf sog. **FDR-Kontrolle** (FDR = False Discovery Rate)

Zum Abschluss noch etwas Witziges

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2008) **275**, 1661–1668
doi:10.1098/rspb.2008.0105
Published online 22 April 2008

You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans

Fiona Mathews[1,*], Paul J. Johnson[2] and Andrew Neil[3]

[1]*Hatherly Laboratories, School of Biosciences, University of Exeter, Prince of Wales Road, Exeter EX4 4PS, UK*
[2]*Wildlife Conservation Research Unit, Department of Zoology, University of Oxford, Tubney House, Tubney, Oxon OX13 5QL, UK*
[3]*Division of Public Health and Primary Health Care, Institute of Health Sciences, University of Oxford, PO Box 777, Oxford OX3 7LF, UK*

Facultative adjustment of sex ratios by mothers occurs in some animals, and has been linked to resource availability. In mammals, the search for consistent patterns is complicated by variations in mating systems, social hierarchies and litter sizes. Humans have low fecundity, high maternal investment and a potentially high differential between the numbers of offspring produced by sons and daughters: these conditions should favour the evolution of facultative sex ratio variation. Yet little is known of natural mechanisms of sex allocation in humans. Here, using data from 740 British women who were unaware of their foetus's gender, we show that foetal sex is associated with maternal diet at conception. Fifty six per cent of women in the highest third of preconceptional energy intake bore boys, compared with 45% in the lowest third. Intakes during pregnancy were not associated with sex, suggesting that the foetus does not manipulate maternal diet. Our results support hypotheses predicting investment in costly male offspring when resources are plentiful. Dietary changes may therefore explain the falling proportion of male births in industrialized countries. The results are relevant to the current debate about the artificial selection of offspring sex in fertility treatment and commercial 'gender clinics'.

Und die Antwort

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2009) **276**, 1211–1212
doi:10.1098/rspb.2008.1405
Published online 13 January 2009

Comment

Cereal-induced gender selection? Most likely a multiple testing false positive

The recent paper by Mathews *et al.* (2008) with a provocative title 'You are what your mother eats' generated a lot of attention in the press and over 50 000 Google hits putting forth the genetically implausible claim that women who eat breakfast cereal are more likely to have a boy child. Their result is easily explained as chance. We will not go into other methodological issues such as recall bias and measurement errors, difficulty in measuring cumulative exposures in nutritional data, unmeasured confounders, variable categorization, statistical power and study design, as Pocock *et al.* (2004) recently reviewed the sad state of observational studies and Ioannidis (2005) reports that 80 per cent of observational studies fail to replicate or the initial effects are much smaller on retest. An implausible claim should strongly overcome chance as an explanation even to be considered. We focus on chance as the cause of their finding.

It has been long well-known, Cournot (1843), that multiple testing can easily lead to false discoveries when multiple hypothesis testing or comparisons are not adequately taken into account. Cournot commented, 'One could distinguish first of all legitimate births from those occurring out of wedlock, … one can also classify births according to birth order, according to the age, questions at issue.) There was a third time period, but the authors did not present data from this period (table 2). In our first analysis, we computed 264 *t*-tests and plotted the resulting ordered *p*-values versus the integers giving a *p*-value plot, Schweder & Spjøtvoll (1982); figure 1. Some explanation: suppose we statistically test 10 questions where nothing is going on. By chance alone we expect the smallest *p*-value to be rather small. We actually expect the *p*-values to be nicely spread out uniformly over the interval 0–1. Except for sampling variability, we expect that the ordered *p*-values plotted against the integers, 1, 2, …, 10, to line up along a 45-degree line. With this dataset, we have 264 *p*-values and the plot of the ordered *p*-values against the integers, 1, 2, …, 264 is essentially linear. This plot implies that the small observed *p*-values, indeed all of the *p*-values, are simply the result of chance and not due to any effect of the food items.

In our second analysis, we used simulation to compute multiplicity-adjusted *p*-values. Explanation of the computation of adjusted *p*-values: we would wish to know if the smallest observed *p*-value could have arisen by chance. We take the outcome for each mother, 0/1 for girl/boy, and permute the values assigning the gender of the child at random to the mother. We compute *p*-values