Week 8

Multiple testing

key ideas

- hypothesis testing / significance analysis is commonly overused
- correcting for multiple testing avoids false positives or discoveries
- two key components:
  - error measure (what you'd like to control)
  - correction (statistical method used to control error measure)

Types of errors

Suppose you're testing a hypothesis that a parameter $\beta = 0$ vs $\beta \neq 0$

possible outcomes

true state	$\beta = 0$	$\beta \neq 0$	hypothesis
claim $\beta = 0$	U	T	$m - R$
claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m

Type I error (false positive) - V

says that $\beta \neq 0$ but it equals to 0

Type II error (false negative) T

says that $\beta = 0$ when $\beta \neq 0$

Error rates

- False positive rate
  the rate at which false results are called significant

$$E\left[\frac{V}{m_0}\right]$$

- Family wise error rate
  the probability of at least one false positive

$$Pr(V \geq 1)$$

- False Discovery Rate
  The rate at which claims of significance are false

$$E\left[\frac{V}{R}\right]$$

Controlling the false positive rate

if P-values are correctly calculated calling all $P < \alpha$ significant will control the false positive rate at level α on average

(all $P < \alpha$ - significant)

but with larger n, the number of false positives is too high!

($10000 \cdot 0.05 = 500$)

⇓

- controlling family-wise error rate (FWER)

The Bonferroni correction — the oldest multiple testing correction

Basic idea

- suppose you do m tests
- you want to control FWER at level α, so $Pr(V \geq 1) < \alpha$
- calculate P-values normally
- $\alpha_{fewer} = \frac{\alpha}{m}$
- call all P-values less than α_{fewer} significant

$\alpha = 0.05$, $\alpha_{fewer} = \frac{0.05}{10000}$ = (no one FP! ← might be too strict (you would want to allow few FP))

- controlling false discovery rate (FDR)

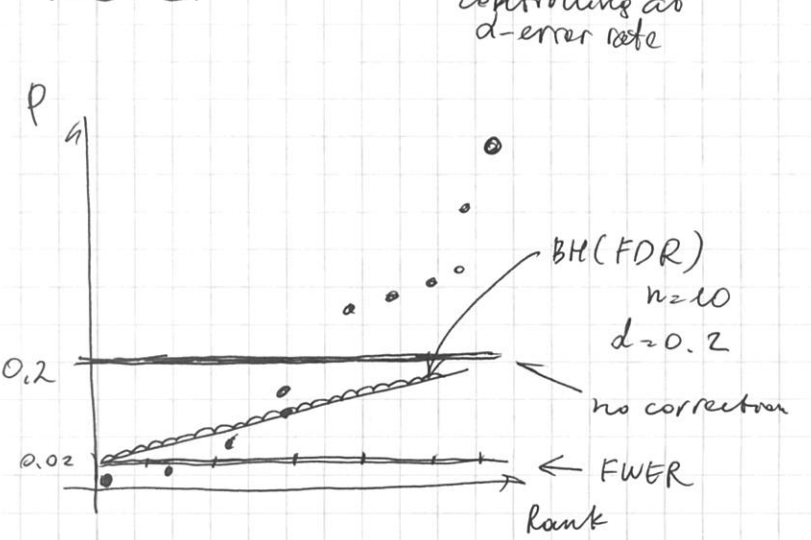
The most popular ~~test when~~ correction when performing lots of tests

Basic idea
- you have m tests
- you want to control FDR at level α, so

$$E\left[\frac{V}{R}\right]$$

- calculate P-values normally
- order the P values from smallest to largest
- call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant

Example



Adjusted P-value

one approach –
– to adjust the threshold α

– another – to calculate "adjusted p-values"

↓

they are not "p-values" anymore!

but they can be used directly, without adjusting α

examples:

- suppose we have $P_1 \dots P_m$
- $P_i^{fwer} = \max(m \times P_i, 1)$ for each P_i
- If $P_i^{fwer} < \alpha$, it's significant

Example:

Suppose we have no relationship at all

```
set.seed(101009)
pValues = rep(NA, 1000)
for (i in 1:1000){
    x = rnorm(20); y = rnorm(20)
    pValues[i] = summary(lm(y~x))$coeff[2,4]
}
sum(pValues < 0.05) == 51!
```

P value ↓

FWER

sum(p.adjust(pValues, method = "bonferroni") < 0.05)

⇓

0

FDR

sum(p.adjust(pValues, method = "BH") < 0.05)

⇓

0

Simulation for model checking

regression model

$$Y_i = b_0 + b_1 X_i + e_i$$

for generating:

x = rnorm(50), y = rnorm(50), b1 = 1, b2 = 2

y = b0 + b1 * x + e

Monte Carlo error

monte carlo variability – ~~when error term is normal~~

when you do more and more simulations, the accuracy increases

1000 → 10000 better

Simulation based on data set

```
library(UsingR)
data(galton)

lm1 = lm(child ~ parent)

— generating new parent heights:

parent0 = rnorm(len, sd = sd(galton$parent),
                mean = mean(galton$parent))

child0 = lm1$coeff[1] + lm1$coeff[2] * parent0 +
  + rnorm(len, sd = summary(lm1)$sigma)
```

measuring error (equal to variability in residuals)

Notes

- simulation can be applied to missing data problems – simulate what missing data might be
- sensitivity analysis – trying different simulations with different assumptions, and seeing how estimates change