Multiple testing

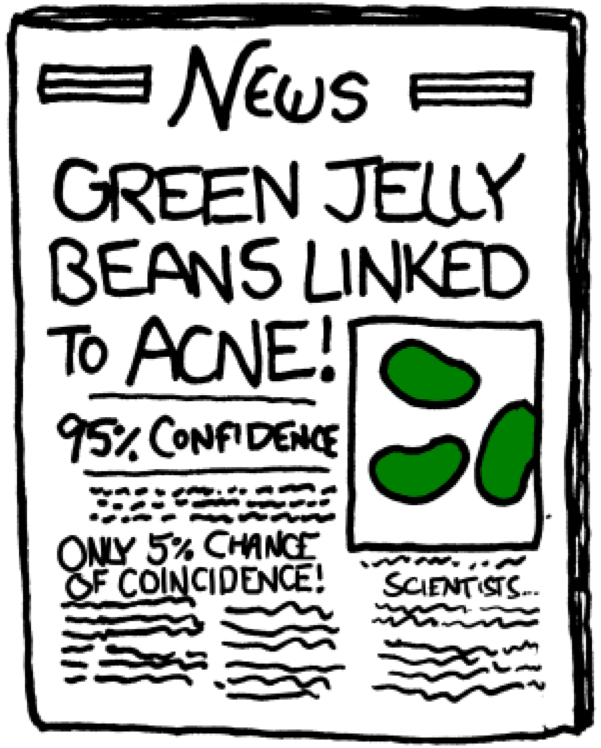
Session 4

MATH 80667A: Experimental Design and Statistical Methods
for Quantitative Research in Management
HEC Montréal

Multiple testing

Scientifist, investigate!

- Consider the Cartoon *Significant* by Randall Munroe (https://xkcd.com/882/)



It highlights two problems: lack of accounting for multiple testing and selective reporting.

How many tests

Consider a one-way ANOVA with K groups.

Having found a significant difference between group means (global null), you proceed to look at all pairwise differences: $\binom{K}{2}$ tests for K groups.

- 3 tests if $K = 3$ groups
- 10 tests if $K = 5$ groups
- 45 tests if $K = 10$

Many tests!

Family-wise error rate

If you do a single hypothesis test and your testing procedure is well calibrated (model assumptions met), there is a probability α of making a type I error if the null is true.

Why $\alpha = 5$%? Essentially **arbitrary**...

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fails to reach this level.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503-513.

How many tests?

Dr. Yoav Benjamini looked at the number of inference / tests performed in the Psychology replication project

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

The number of tests performed ranged from 4 to 700, with an average of 72.

Most studies did not account for selection.

Motivation

- If we do m **independent** comparisons, each one at the level α, the probability of making at least one type I error, say $\alpha^\star$, is

$$\begin{aligned}\alpha^\star &= 1 - \text{probability of making no type I error} \\ &= 1 - (1-\alpha)^m\end{aligned}$$

With $\alpha = 5$%

- $m = 4$ tests, $\alpha^\star \approx 0.185$.
- $m = 72$ tests, $\alpha^\star \approx 0.975$.

Tests need not be independent... but can show $\alpha^\star \leq m\alpha$.

Family of hypothesis

Consider a family of m null hypothesis $\mathscr{H}_{01}, \ldots, \mathscr{H}_{0m}$ tested.

- The family may depend on the context, but all hypothesis that are scientifically relevant and could be reported.

Should be chosen a priori and pre-registered

Keep it small: the number of planned comparisons for a one-way ANOVA should be less than the number of groups K.

Notation

Define

$$
R_i = \begin{cases} 1 & \text{if we reject } \mathscr{H}_{0i} \\ 0 & \text{if we fail to reject } \mathscr{H}_{0i} \end{cases}
$$

$$
V_i = \begin{cases} 1 & \text{type I error for } \mathscr{H}_{0i} \quad (R_i = 1 \text{ and } \mathscr{H}_{0i} \text{ is true}) \\ 0 & \text{otherwise} \end{cases}
$$

with

- $R = R_1 + \cdots + R_m$ the total number of rejections ($0 \leq R \leq m$).
- $V = V_1 + \cdots + V_m$ the number of null hypothesis rejected by mistake.

Decision rule

Classify the decision on the m tests in a table based on whether the null hypothesis is true or false.

We reject the null hypothesis $\mathscr{H}_0$ if the *p*-value is less than the level, $p < \alpha$.

Truth \ Decision	**Reject null hypothesis**	**Fail to reject null**
$\mathscr{H}_0$ is true	$R - V$ correct rejections	$-$
$\mathscr{H}_a$ is true	V type I errors	$-$
Total	R rejections	$m - R$ non-rejections

Familywise error rate

The familywise error rate is the probability of making at least one type I error per family

$$\text{FWER} = \Pr(V \geq 1)$$

If we use a procedure that controls for the family-wise error rate, we talk about simultaneous inference (or simultaneous coverage for confidence intervals).

Bonferroni's procedure

Consider a family of m hypothesis tests and perform each test at level α/m.

- reject ith null $\mathscr{H}_{i0}$ if the associated *p*-value $p_i \leq \alpha/m$.
- build confidence intervals similarly with $1-\alpha/m$ quantiles.

If the (raw) p-values are reported, reject $\mathscr{H}_{0i}$ if $m \times p_i \geq \alpha$ (i.e., multiply reported p-values by m)

Holm's sequential method

Order the p-values of the family of m tests from smallest to largest

$$p_{(1)} \leq \cdots \leq p_{(m)}$$

associated to null hypothesis $\mathscr{H}_{0(1)}, \ldots, \mathscr{H}_{0(m)}$.

Idea use a different level for each test, more stringent for smaller p-values.

Coupling Holm's method with Bonferroni's procedure: compare $p_{(1)}$ to $\alpha_{(1)} = \alpha/m$, $p_{(2)}$ to $\alpha_{(2)} = \alpha/(m-1)$, etc.

Holm-Bonferroni procedure is always more powerful than Bonferroni

Holm-Bonferroni procedure

Sequential testing

- start with the smallest p-value
- check significance one test at a time
- stop when the first nonsignificant p-value is found or no more test in store.

Conclusion

If $p_{(j)} \geq \alpha_{(j)}$ but $p_{(i)} \leq \alpha_{(i)}$ for $i = 1, \ldots, j-1$ (all smaller p-values)

- reject $\mathscr{H}_{0(1)}, \ldots, \mathscr{H}_{0(j-1)}$
- fail to reject $\mathscr{H}_{0(j)}, \ldots, \mathscr{H}_{0(m)}$

If $p_{(i)} \leq \alpha_{(i)}$ for all test $i = 1, \ldots, m$

- reject $\mathscr{H}_{0(1)}, \ldots, \mathscr{H}_{0(m)}$

Numerical example

Consider $m = 3$ tests with raw p-values 0.01, 0.04, 0.02.

i	$p_{(i)}$	Bonferroni	Holm-Bonferroni
1	0.01	$3 \times 0.01 = 0.03$	$3 \times 0.01 = 0.03$
2	0.02	$3 \times 0.02 = 0.06$	$2 \times 0.02 = 0.04$
3	0.04	$3 \times 0.04 = 0.12$	$1 \times 0.04 = 0.04$

Reminder of Holm–Bonferroni: multiply by $(m - i + 1)$ the ith smallest p-value $p_{(i)}$, compare the product to α.

Why choose Bonferroni's procedure?

- simple
- generally applicable (any design)
- but dominated by sequential procedures (Holm-Bonferroni uniformly more powerful)
- low power when the number of test m is large
- m must be prespecified

Alternative measures

The FWER does not make a distinction between one or multiple type I errors.

We can also look at the more stringent criterion **per-family error rate**, $\mathsf{PFER} = \mathsf{E}(V)$, the expected (theoretical average) number of false positive.

One can show that

$$\mathsf{FWER} = \Pr(V \geq 1) \leq \mathsf{E}(V),$$

Any procedure that controls the per-family error rate thus also controls the familywise error rate: Bonferroni does.

Dedicated methods for ANOVA

All methods valid with equal group variances and independent observations.

- **Tukey**'s honestly significant difference (HSD) method: to compare (all) pairwise differences between subgroups, based on the largest possible pairwise mean differences, with extensions for unbalanced samples.
- **Scheffé**'s method: applies to any contrast (properties depends on sample size n and number of groups K, not the number of test). Better than Bonferroni if m is large. Can be used for any design, but not powerful.
- **Dunnett**'s method: only for all pairwise contrasts relative to a specific baseline (control).

Described in Dean, Voss and Draguljić (2017) in more details.

Adjustment for one-way ANOVA

Similar ideas but different **critical coefficients**. All derived using software.

Proceed only if there is a significant difference between groups, i.e. if we reject global null.

With $K=5$ groups and $n=9$ individuals per group (`arithmetic` example), critical value for two-sided test of zero difference with standardized t-test statistic and $\alpha=5$% are

- Scheffé's (all contrasts): 3.229
- Tukey's (all pairwise differences): 2.856
- Dunnett's (difference to baseline): 2.543
- unadjusted Student's t-distribution: 2.021

False discovery rate

Suppose that m_0 out of m hypothesis are true null (so $\mathscr{H}_0$ holds m_0 times).

The **false discovery rate** is the proportion of false discovery among rejected nulls,

$$\mathsf{FDR} = \begin{cases} \frac{V}{R} & R > 0, \\ 0 & R = 0. \end{cases}$$

False discovery rate offers weak-FWER control

> the property is only guaranteed under the scenario where all null hypotheses are true.

False discovery rate vs FWER

A simultaneous procedure that controls family-wise error rate (FWER) ensure any selected test has type I error α.

The false discovery rate (FDR) is less stringent: it's a guarantee for the proportion **among selected** discoveries.

But false discovery rate is scalable:

- 2 type I errors out of 4 tests is unacceptable.
- 2 type I errors out of 100 tests is probably okay.

Controlling false discovery rate

The Benjamini-Hochberg (1995) procedure

1. Order the *p*-values from the m tests from smallest to largest: $p_{(1)} \leq \cdots \leq p_{(m)}$
2. For level α (e.g., $\alpha = 5$%), set

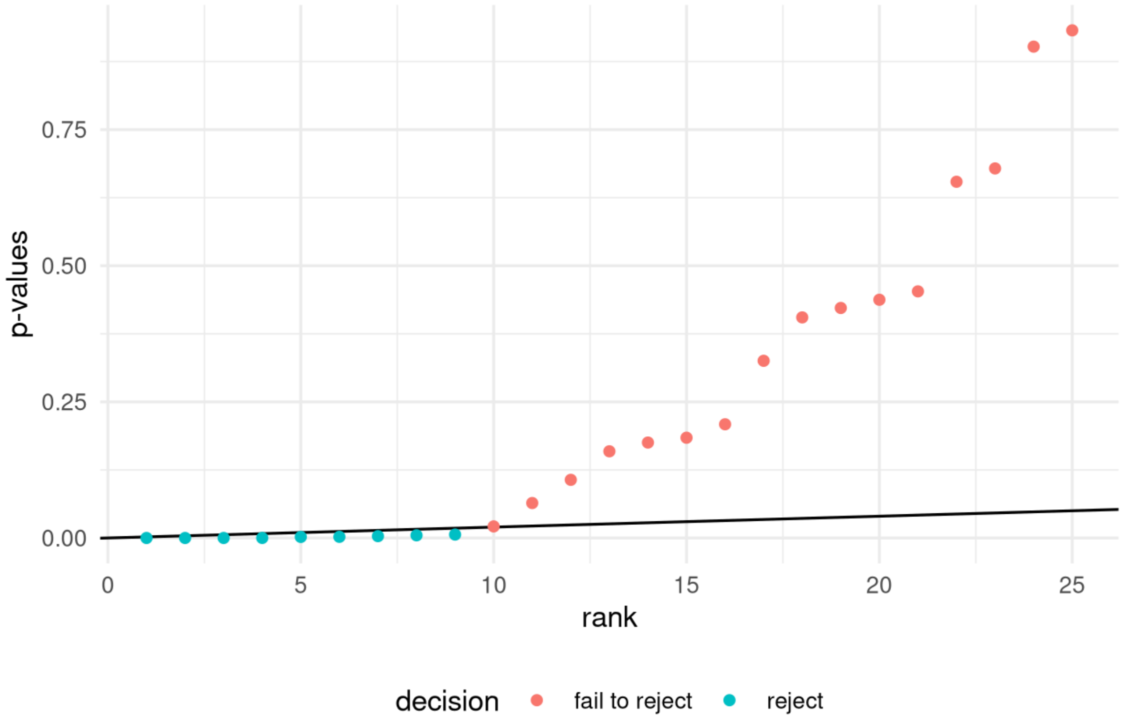
$$k = \max\left\{i : p_{(i)} \leq \frac{i}{m}\alpha\right\}$$

3. Reject $\mathscr{H}_{0(1)}, \ldots, \mathscr{H}_{0(k)}$.

Picture of Benjamini-Hochberg

Plot the m p-values against their rank.

To ensure $\text{FDR} \leq q$, reject null hypotheses corresponding to p-value that fall below the line of slope α/m.



Exercice

Table S3

Planned Comparisons in Study 2

	Other (immersed & distanced) vs. Self-immersed	Self-distanced vs. Self-immersed	Other-distanced vs. Other-immersed	Other (immersed & distanced) vs. Self-distanced
Variables	*t (p-value)*	*t (p-value)*	*t (p-value)*	*t (p-value)*
LIMITS	1.74 (.09)	2.16 (.03)	0.06 (.96)	0.81 (.42)
COMPR	2.02 (.046)	1.95 (.05)	0.05 (.96)	0.31 (.76)
PERSP	4.82 (< .001)	2.83 (.005)	0.74 (.46)	1.28 (.20)
CHANGE	1.80 (.08)	0.06 (.96)	0.15 (.88)	1.63 (.11)

Note. LIMITS - Recognition of limits of knowledge; COMPR - Search for a compromise; PERSP - Consideration of others' perspectives; CHANGE - Recognition of change; Planned comparisons include information from all four cells in the denominator.

> Grossman, I. and E. Kross (2014). Exploring "Solomon's paradox": Self-distancing eliminates the self-other asymmetry in wise reasoning about close relations in younger and older adults, *Psychological Science*, 25(8) 1571-1580

Summary (1/2)

- Researchers often carry lots of hypothesis testing tests
  - the more you look, the more you find!
- Thus want to control probability of making a judicial mistake among all m tests performed
  - (family-wise error rate, FWER)
- Less stringent criterion: control for the **proportion** of condemned (rejections) that were innocent
  - (false discovery rate, FDR)
  - useful if you don't care about making some mistakes, but perform loads of test (potentially millions)

Summary (2/2)

- ANOVA specific solutions: assumes normal data, equal variance, balanced samples...
  - Tukey's HSD (all pairwise differences),
  - Dunnett's method (only differences relative to a reference category)
  - Scheffé's method (all contrasts)
- General methods
  - FWER: Bonferroni (suboptimal), Bonferroni-Holm (more powerful)
  - FDR: Benjamini-Hochberg

Downside of adjustment is loss of power (but more robust findings).

Rant about *p*-values

The American Statistical Association (ASA) published a list of principles guiding (mis)interpretation of *p*-values.

> (2) *P*-values do not measure the probability that the studied hypothesis is true
>
> (3) Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
>
> (4) *P*-values and related analyses should not be reported selectively
>
> (5) *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result