Contrasts and multiple testing

Session 4

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

Outline

Contrasts

Multiple testing

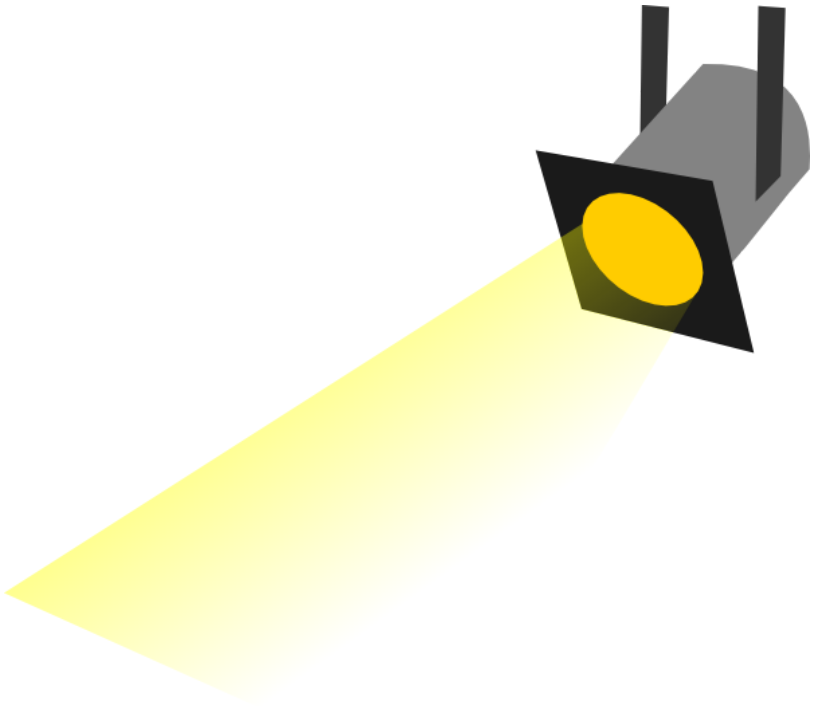
Planned comparisons

- Oftentimes, we are not interested in the global null hypothesis.
- Rather, we formulate planned comparisons *at registration time* for effects of interest

What is the scientific question of interest?

Global null vs contrasts

Global test



Contrasts

Image source: PNGAll.com, CC-BY-NC 4.0

Linear contrasts

With K groups, null hypothesis of the form

$$\mathscr{H}_0: \underbrace{c_1\mu_1 + \cdots + c_K\mu_K}_{\text{weighted sum of subpopulation means}} = a$$

Linear combination of weighted group averages

Examples of linear contrasts

Global mean larger than a?

$$\mathscr{H}_0 : \frac{n_1}{n}\mu_1 + \cdots + \frac{n_K}{n}\mu_K > a$$

Pairwise comparison

$$\mathscr{H}_0 : \mu_i = \mu_j, \quad i \neq j$$

Characterization of linear contrasts

If $c_1 + \cdots + c_K = 0$, the contrast encodes

differences between treatments

rather than information about the overall mean.

- Weights $c_1, \ldots, c_K$ are specified by the **user**.
- Mean response in each experimental group is estimated as sample average of observations in that group, $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$.
- Assuming equal variance, the contrast statistic behaves in large samples like a Student-*t* distribution with $n - K$ degrees of freedom.

Arithmetic example

Setup

group 1 (control)

group 2 (control)

group 3 (praise, reprove, ignore)

Hypothesis of interest

- $\mathscr{H}_{01}$: $\mu_{\text{praise}} = \mu_{\text{reproved}}$ (attention)
- $\mathscr{H}_{02}$: $\frac{1}{2}(\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \mu_{\text{praised}}$ (encouragement)

Contrasts

With placeholders for each group, write $\mathscr{H}_{01} : \mu_{\text{praised}} = \mu_{\text{reproved}}$ as

$$0 \cdot \mu_{\text{control}_1} + 0 \cdot \mu_{\text{control}_2} + 1 \cdot \mu_{\text{praised}} - 1 \cdot \mu_{\text{reproved}} + 0 \cdot \mu_{\text{ignored}}$$

The sum of the coefficients, $(0, 0, 1, -1, 0)$, is zero.

sum-to-zero constraint

Similarly, for $\mathscr{H}_{02} : \frac{1}{2}(\mu_{\text{control}_1} + \mu_{\text{control}_2}) = \mu_{\text{praise}}$

$$\frac{1}{2} \cdot \mu_{\text{control}_1} + \frac{1}{2} \cdot \mu_{\text{control}_2} - 1 \cdot \mu_{\text{praised}} + 0 \cdot \mu_{\text{reproved}} + 0 \cdot \mu_{\text{ignored}}$$

The entries of the contrast vector $(\frac{1}{2}, \frac{1}{2}, -1, 0, 0)$ sum to zero.

Equivalent formulation is obtained by picking $(1, 1, -3, 0, 0)$

Contrasts in **R**

```r
library(emmeans)
linmod <- lm(score ~ group, data = arithmetic)
linmod_emm <- emmeans(linmod, specs = 'group')
contrast_specif <- list(
  controlvspraised = c(0.5, 0.5, -1, 0, 0),
  praisedvsreproved = c(0, 0, 1, -1, 0)
)
contrasts_res <-
  contrast(object = linmod_emm,
                    method = contrast_specif)
# Obtain confidence intervals instead of p-values
confint(contrasts_res)
```

Output

contrast	null.value	estimate	std.error	df	statistic	p.value
control vs praised	0	-8.44	1.40	40	-6.01	<1e-04
praised vs reprove	0	4.00	1.62	40	2.47	0.018

contrast	lower	upper
control vs praised	-11.28	-5.61
praised vs reprove	0.72	7.28

Multiple testing

Post-hoc tests

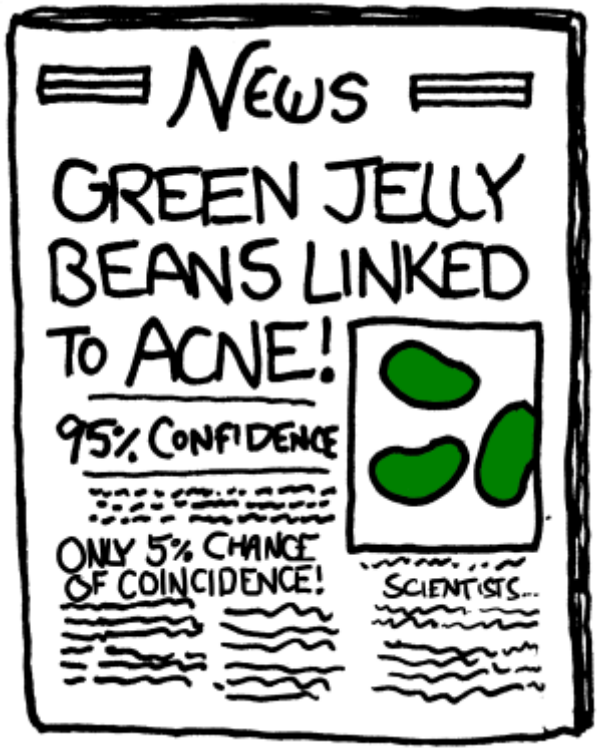
Suppose you decide to look at all pairwise differences

Comparing all pairwise differences: $m = \binom{K}{2}$ tests

- $m = 3$ tests if $K = 3$ groups,
- $m = 10$ tests if $K = 5$ groups,
- $m = 45$ tests if $K = 10$ groups...

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.

There is a catch...

Read the small prints:

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

How many tests?

Dr. Yoav Benjamini looked at the number of 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.

Probability of type I error

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

With $\alpha = 0.05$

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

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

Statistical significance at the 5% level

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.

Family of hypothesis

Consider m tests with the corresponding null hypotheses $\mathscr{H}_{01}, \ldots, \mathscr{H}_{0m}$.

- The family may depend on the context, but including any hypothesis that is 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 indicators

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

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

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

Definition: 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

Sequential Holm-Bonferroni procedure

1. start with the smallest p-value
2. check significance one test at a time
3. stop when the first non-significant p-value is found or no more test.

Conclusion for Holm-Bonferroni

Reject smallest *p*-values until you find one that fails, reject rest

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

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

All *p*-values are lower than their respective cutoff:

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

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

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

Careful: adjust for the real number of comparisons made (often reporter just correct only the 'significant tests', which is wrong).

Controlling the average number of errors

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

We can also look at a more stringent criterion

per-family error rate (PFER)

i.e., the expected number of false positive

Since

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

any procedure that controls the per-family error rate also controls the familywise error rate.

Bonferroni controls both per-family error rate and family-wise error rate.

Confidence intervals for linear contrasts

Given a linear contrast of the form

$$C = c_1\mu_1 + \cdots + c_K\mu_K$$

with $c_1 + \cdots c_K = 0$, we build confidence intervals as usual

$$\widehat{C} \pm \text{critical value} \times \widehat{\text{se}}(\widehat{C})$$

Different methods provide control for FWER by modifying the **critical value**.

All methods valid with equal group variances and independent observations.

FWER control in ANOVA

- **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), Section 4.4 in more details.

Tukey's honest significant difference

Control for all pairwise comparisons

Idea: controlling for the range

$$\max\{\mu_1, \ldots, \mu_K\} - \min\{\mu_1, \ldots \mu_K\}$$

automatically controls FWER for other pairwise differences.

Critical values based on "Studentized range" distribution

Assumptions: equal variance, equal number of observations in each experimental condition.

Scheffé's criterion

Control for all possible linear contrasts

Critical value is $\sqrt{(K-1)F}$, where F is the $(1-\alpha)$ quantile of the $\mathsf{F}(K-1, n-K)$ distribution.

Allows for data snooping (post-hoc hypothesis)

But not powerful...

Adjustment for one-way ANOVA

Take home message:

- same as Wald-based confidence intervals, only with different **critical values**
- larger cutoffs if procedure accounts for more tests

Everything is obtained using software.

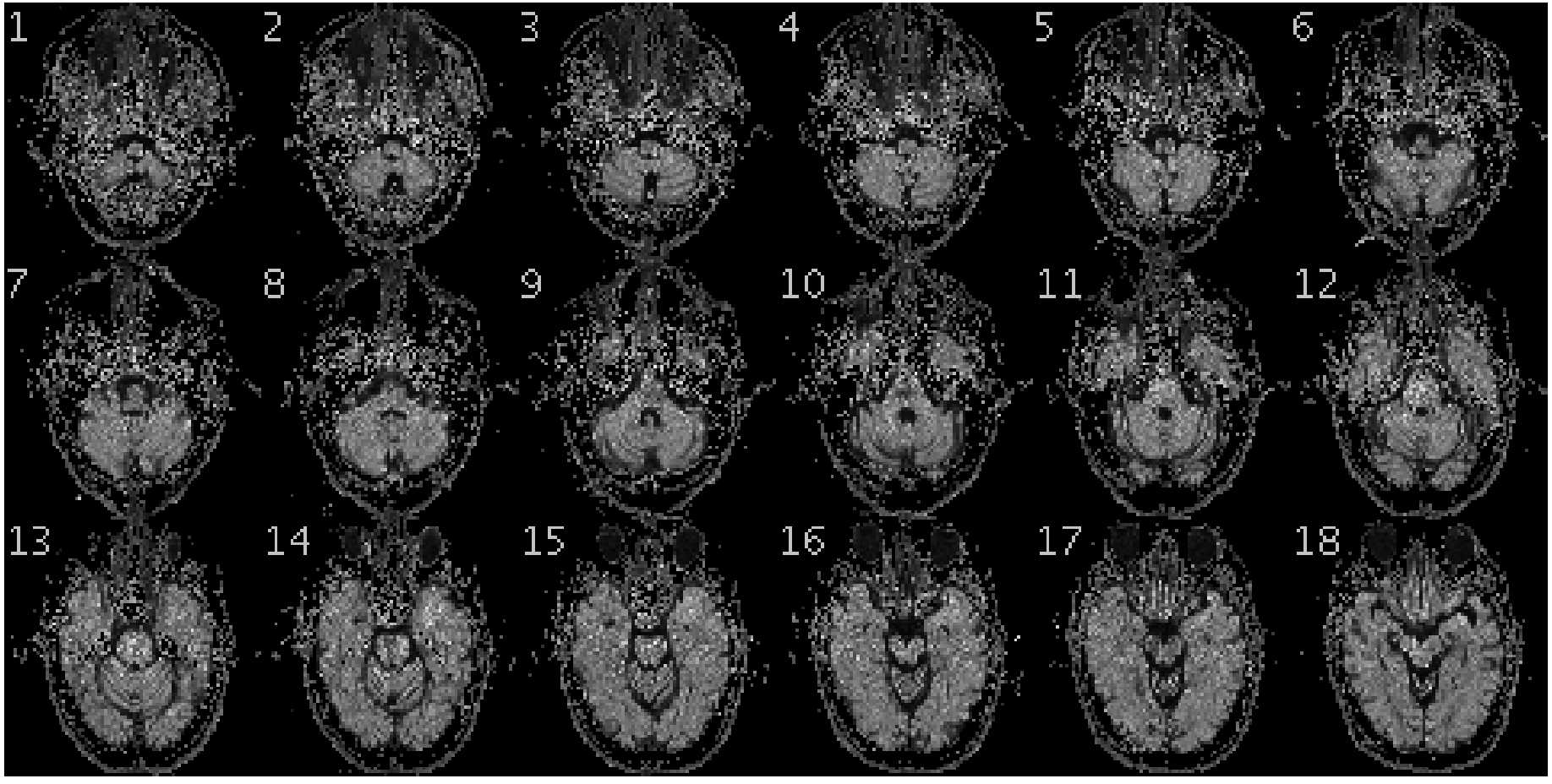
Proceed only if there is a significant difference between groups, i.e., if we reject global null.

Numerical example

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

Sometimes, there are too many tests...



Scaling back expectations...

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

With thousands of tests, this is too stringent a criterion.

The false discovery rate (FDR) provides a guarantee for the proportion **among selected** discoveries (tests for which we reject the null hypothesis).

Why use it? the 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.

False discovery rate

Suppose that m_0 out of m null hypothesis are true

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

$$\text{FDR} = \begin{cases} \frac{V}{R} & R > 0 \text{ (if one or more rejection)}, \\ 0 & R = 0 \text{ (if no rejection)}. \end{cases}$$

Controlling false discovery rate

The Benjamini-Hochberg (1995) procedure for controlling false discovery rate is:

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., $\alpha = 0.05$), set

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

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

Benjamini-Hochberg in a picture

1. Plot *p*-values (*y*-axis) against their rank (*x*-axis)
    - (the smallest *p*-value has rank 1, the largest has rank m).
2. Draw the line $y = \alpha/mx$
    - (zero intercept, slope α/m)
3. Reject all null hypotheses associated to p-values located before the first time a point falls *above* the line.