# One way ANOVA

**Session 3**

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

# Outline

**Recap**

**F tests**

**Parametrizations and interpretation**

**Planned comparisons and *post-hoc* tests**

# Refresher on hypothesis tests

# General recipe of hypothesis testing

**(1) Define variables**

**(2) Write down hypotheses**

**(3) Choose/compute a test statistic**

**(4) Benchmark**

**(5) Compute the *p*-value**

**(6) Conclude (reject/fail to reject)**

**(7) Report findings**

# Level

**Level = probability of judicial error**

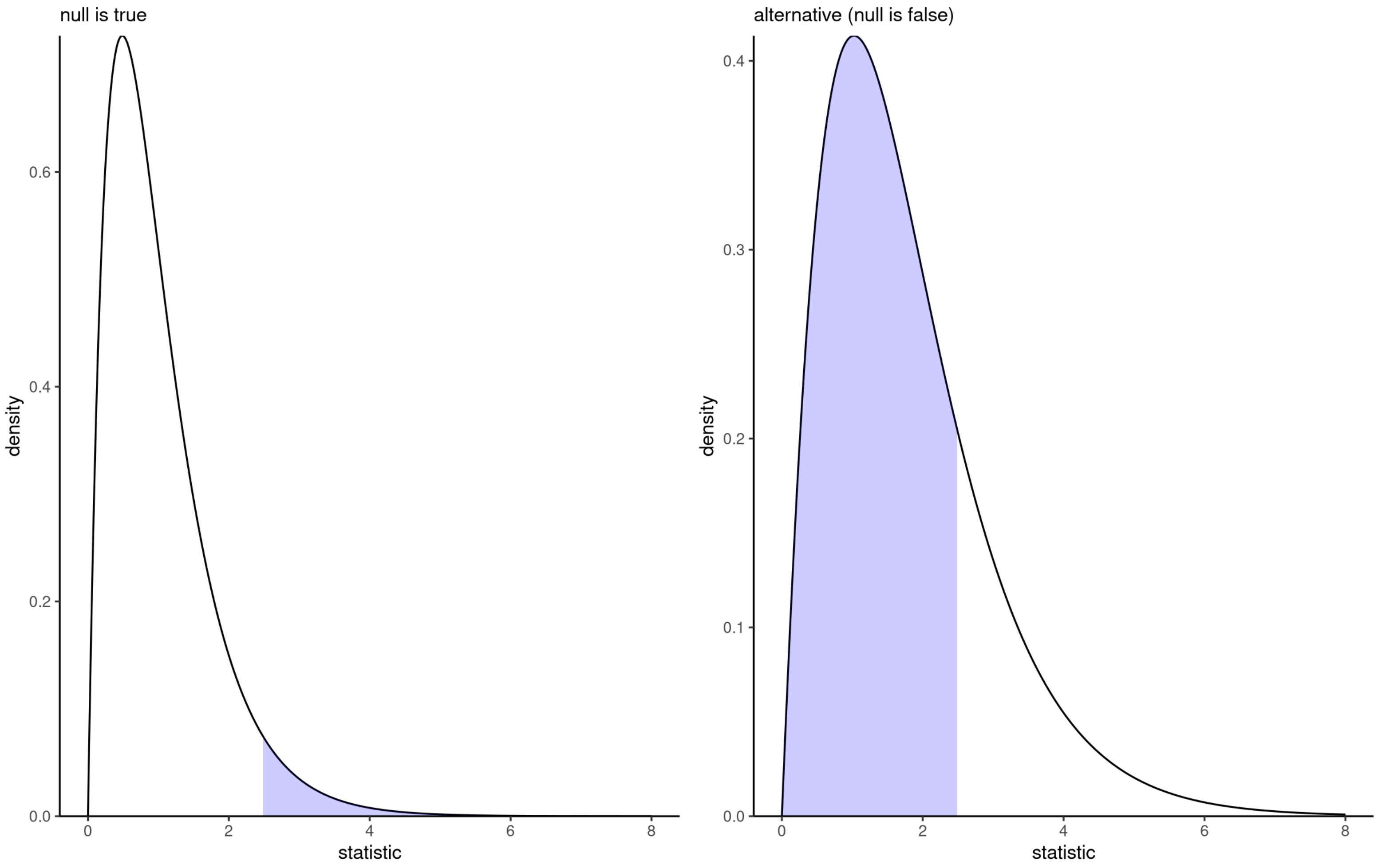
**Analyst fixes level $\alpha$ before the experiment.**

**Choose $\alpha$ as small as possible (don't condemn the innocent!)**

**(typical value is 5%)**

**Reject $\mathscr{H}_0$ if p-value less than $\alpha$**

# Errors



null is true
alternative (null is false)
density
statistic
0.0
0.2
0.4
0.6
0.1
0.3
0
2
4
6
8

# Confidence intervals

Test statistics are standardized,

- Good for comparisons with benchmark
- typically meaningless (standardized = unitless quantities)

Two options for reporting:

- $p$-value: probability of more extreme outcome if no mean difference
- confidence interval: set of all values for which we fail to reject the null hypothesis at level $\alpha$ for the given sample

# Interpretation of confidence intervals

**confidence interval = [lower, upper] units**

Interpretation: under the null, if we repeat the experiments 95%, of intervals will contain the true value (if model correctly calibrated)

**Analogy: coin toss**

# One way analysis of variance

# F-test for one way ANOVA

## Global null hypothesis

No difference between treatments

- $\mathscr{H}_0$ (null): all of the $K$ treatment groups have the same average $\mu$
- $\mathscr{H}_a$ (alternative): at least two treatments have different averages

# $F$-test statistic

## Omnibus test

With $K$ groups and $n$ observations, the statistic is

$$F = \frac{\text{between sum of squares}/(K-1)}{\text{within sum of squares}/(n-K)}$$

The null distribution (benchmark) is $F(K-1, n-K)$.

# Why does it work?

Denote

- $y_{ik}$ is observation $i$ of group $k$
- $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$ the sample average of groups $1, \ldots, K$
- $\widehat{\mu}$ is overall sample mean

**Decomposing variability into bits**

$$\underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu})^2}_{\text{total sum of squares}} = \underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu}_k)^2}_{\text{within sum of squares}} + \underbrace{\sum_k n_i (\widehat{\mu}_k - \widehat{\mu})^2}_{\text{between sum of squares}}.$$

**null model** **alternative model** **added variability**

# Degrees of freedom

The parameters of the null distribution are called **degrees of freedom**

- $K - 1$ is the number of constraints imposed by the null
- $n - K$ is the number of observations minus number of mean parameters estimated under alternative

Idea of *F*-statistic: under the null, both numerator and denominator are variance estimators.

- but the numerator is more variable...
- the $F$ ratio should be approximately one on average

# Parametrizations and interpretation

# Parametrization 1: sample averages

Most natural parametrization, not useful for test

- Sample sizes in each group: $n_1, \ldots, n_K$, are known.
- sample average of each treatment group: $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$.

**$K$ means = $K$ parameters**

Overall mean is

$$n\widehat{\mu} = n_1\widehat{\mu}_1 + \cdots + n_K\widehat{\mu}_K$$

# Parametrization 2: contrasts

In terms of differences, relative to a baseline category $j$

- Intercept = sample mean $\widehat{\mu}_j$
- Coefficient for group $k \neq j$: $\widehat{\mu}_k - \widehat{\mu}_j$
    - difference between averages of group $k$ and baseline

In **R**, the baseline is the smallest alphanumerical value.

```
lm(response ~ group)
```

# Parametrization 3: sum-to-zero

In terms of differences, relative to average of $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$

- Intercept = $(\widehat{\mu}_1 + \cdots + \widehat{\mu}_K)/K$
- Coefficient for group $k$: $\widehat{\mu}_k$ minus intercept

In **R**, the last factor level is dropped by default.

```
lm(response ~ group, contrasts = contr.sum(group))
```

Warning: Intercept $\neq \widehat{\mu}$ unless the sample is balanced

# Comparison for the arithmetic example

| group | mean | contrasts | sum-to-zero |
|---|---|---|---|
| intercept | | 19.66 | 21.00 |
| control 1 | 19.66 | | -1.33 |
| control 2 | 18.33 | -1.33 | -2.66 |
| praised | 27.44 | 7.77 | 6.44 |
| reproved | 23.44 | 3.77 | 2.44 |
| ignored | 16.11 | -3.55 | |

# Planned comparisons and post-hoc tests

# Planned comparisons

Oftentimes, we are not interested in the global null hypothesis.

- Can formulate planned comparisons *at registration time* for effects of interest

**What is the scientific question of interest?**

# Arithmetic example

## Setup

**group 1** (control) | **group 2** (control) | **group 3** (praise, reprove, ignore)

## Hypothesis of interest

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

# Contrasts

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

$$0 \cdot \mu_{\text{control}_1} + 0 \cdot \mu_{\text{control}_2} + 1 \cdot \mu_{\text{praise}} - 1 \cdot \mu_{\text{reprove}} + 0 \cdot \mu_{\text{ignore}}$$

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

**Contrast = sum-to-zero constraint**

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

$$\frac{1}{2} \cdot \mu_{\text{control}_1} + \frac{1}{2} \cdot \mu_{\text{control}_2} - 1 \cdot \mu_{\text{praise}} + 0 \cdot \mu_{\text{reprimand}} + 0 \cdot \mu_{\text{ignore}}$$

The contrast vector $\left(\frac{1}{2}, \frac{1}{2}, -1, 0, 0\right)$ sums to zero.

Equivalent formulation is obtained by picking $(1, 1, -2, 0, 0)$

# Contrasts in 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)
```

# *Post-hoc* tests

Maybe there is some difference between groups?

Unplanned comparisons: go fishing...

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

With $K = 5$ groups, we get 10 pairwise comparisons.

```
emmeans(modlin, pairwise ~ group)
```

If there were no differences between the groups, how many do we expect to find significant by chance with $\alpha = 0.1$?

# Pairwise differences and *t*-tests

## Technical aside

The pairwise differences (*p*-values) and confidence intervals for groups $j$ and $k$ are based on the *t*-statistic:

$$t = \frac{(\widehat{\mu}_j - \widehat{\mu}_k) - 0}{\text{se}(\widehat{\mu}_j - \widehat{\mu}_k)}$$

which has a null $\mathcal{T}(n-k)$ distribution.

The standard error $\text{se}(\widehat{\mu}_j - \widehat{\mu}_k)$ uses the pooled variance estimate, i.e., the within sum of squares divided by $n - K$