Analysis of Variance

October 14, 2019

Comparing Many Means

We've spent some time with hypothesis tests for 1 and 2 means... but what if we want to compare 3 or more means?

- We might think to do pairwise comparisons.
- However, eventually we are likely to reject H_0 by chance alone.
- We want a holistic test to examine 3 or more means.

Core Ideas of ANOVA

The holistic test we want is called **ANalysis Of VAriance** or **ANOVA**.

- The ANOVA tests whether means across many groups are equal.
- This relies on a new distribution, F.

ANOVA Hypotheses

The basic hypotheses for the ANOVA are

H_0 : The mean outcome is the same across all groups.

H_A : At least one mean is different.

ANOVA Hypotheses

In statistical notation, we write

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: &\quad \mu_i \neq \mu_j \quad \text{for at least one pair } (i, j) \end{aligned}$$

where k is the number of means being compared and i, j represent the ith and jth groups.

ANOVA Assumptions

There are three core conditions for ANOVA:

1. Observations independent within and between groups.
2. Data within each group are nearly normal.
3. Variability across groups is about equal.

Example

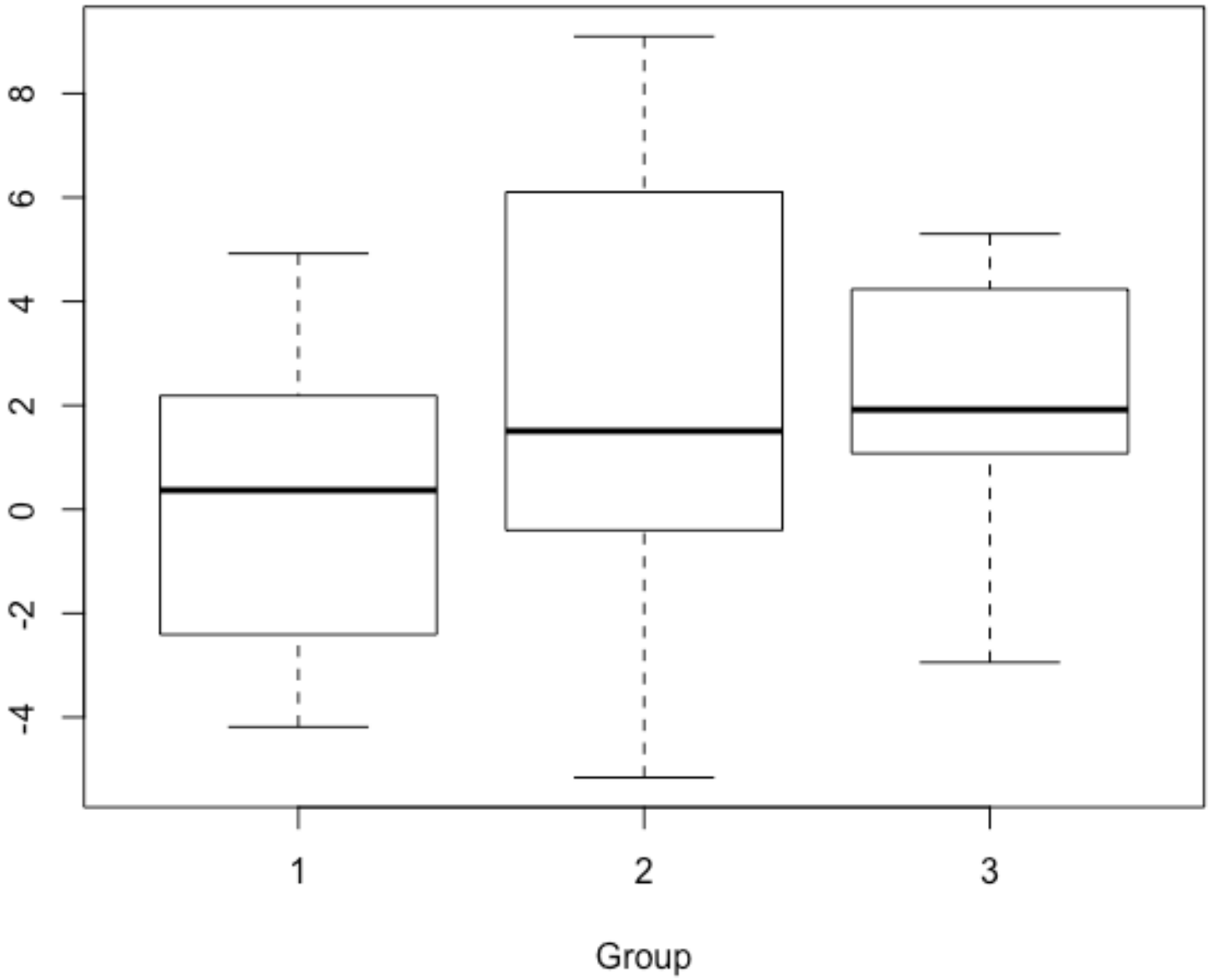
- A university offers 3 lectures for an introductory psychology course.
- A single professor offers 8am, 10am, and 3pm lectures.
- We want to know if the average midterm scores differ between these lectures.

Describe appropriate hypotheses to determine whether there are any differences between the three classes.

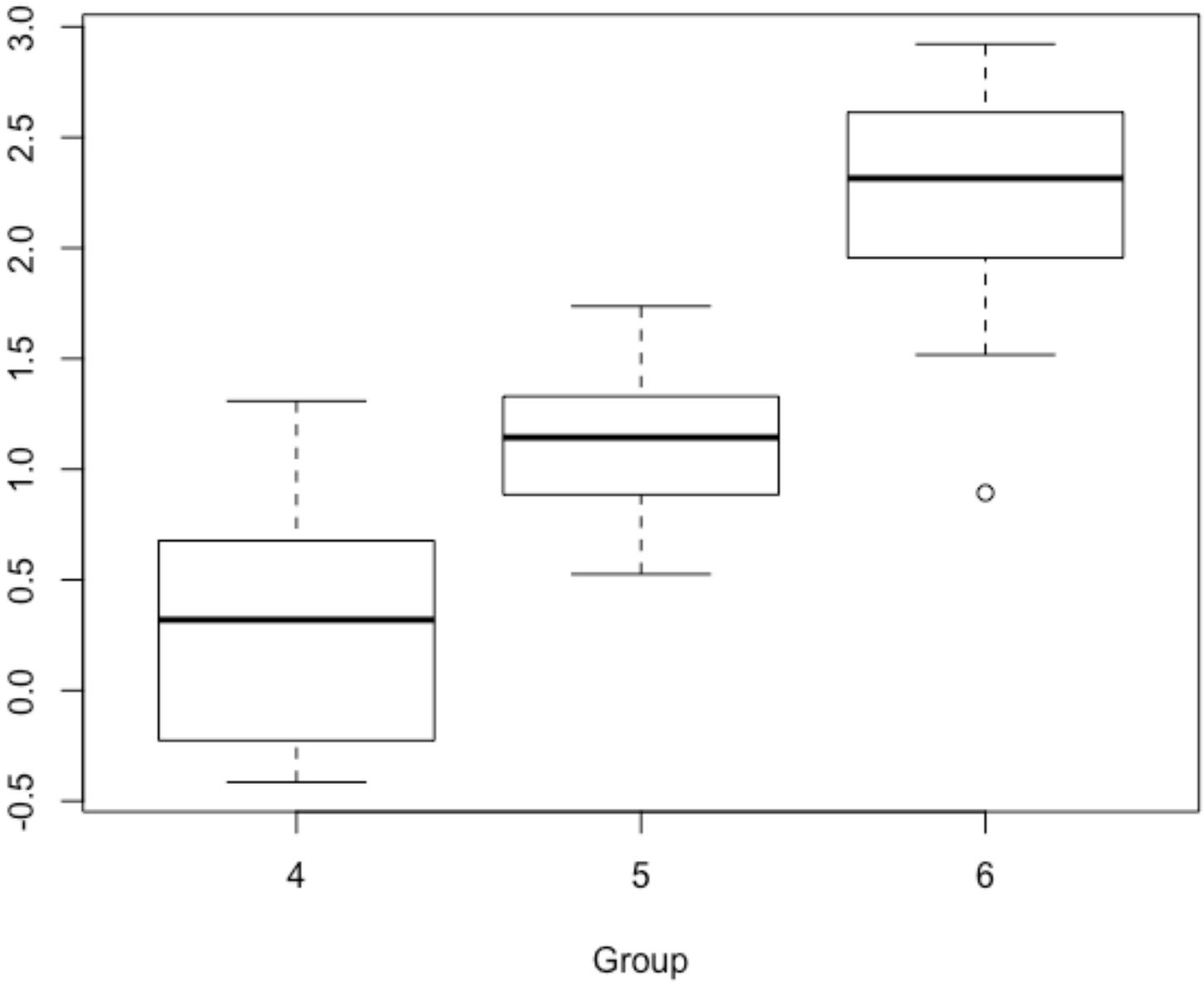
Why "Analysis of Variance"?

- Strong evidence favoring H_A will be unusually large differences between group means.
- It may come as a surprise, but we will quantify this by examining variability.

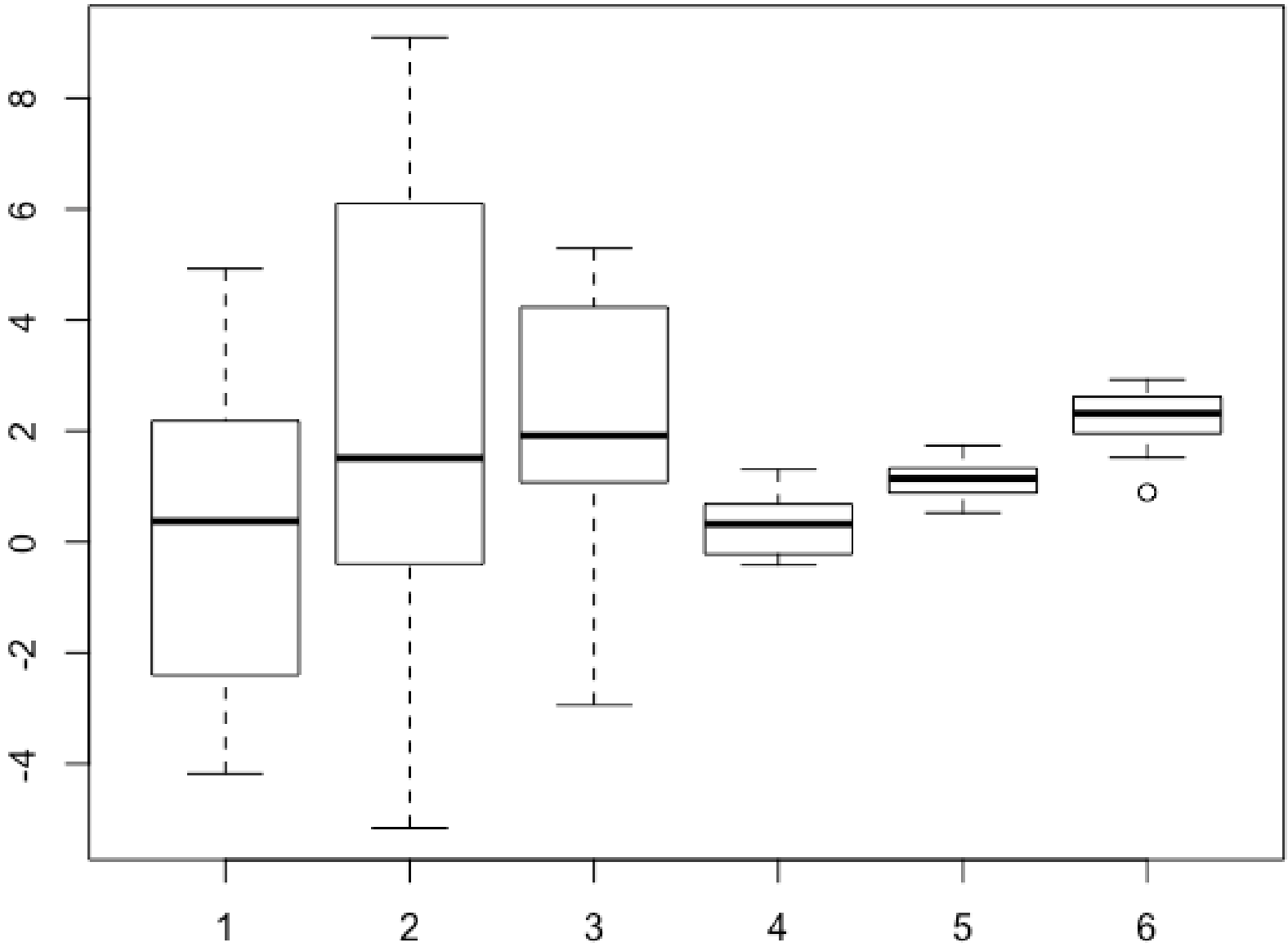
Example: High Within Group Variance



Example: Low Within Group Variance



Example: All 6 Groups



Why "Analysis of Variance"?

- We were able to see the differences in groups 4, 5, and 6 more easily.
- The differences in center are more obvious because the differences are large *relative to the within group variability*.

Example: MLB Batting Performance and Player Position

- We want to determine whether batting performance differs between positions (outfielder, infielder, and catcher).
- We have a dataset from 2018 with batting records of 429 MLB players.
- We will compare on-base percentage, roughly the fraction of times a player gets on base or hits a home run.

Example: MLB

	Name	Team	Position	OBP
1	Abreu, J	CWS	IF	0.325
2	Acuna Jr., R	ATL	OF	0.366
3	Adames, W	TB	IF	0.348
⋮	⋮	⋮	⋮	⋮
427	Zimmerman, R	WSH	IF	0.337
428	Zobrist, B	CHC	IF	0.378
429	Zunino, M	SEA	C	0.259

Example: MLB

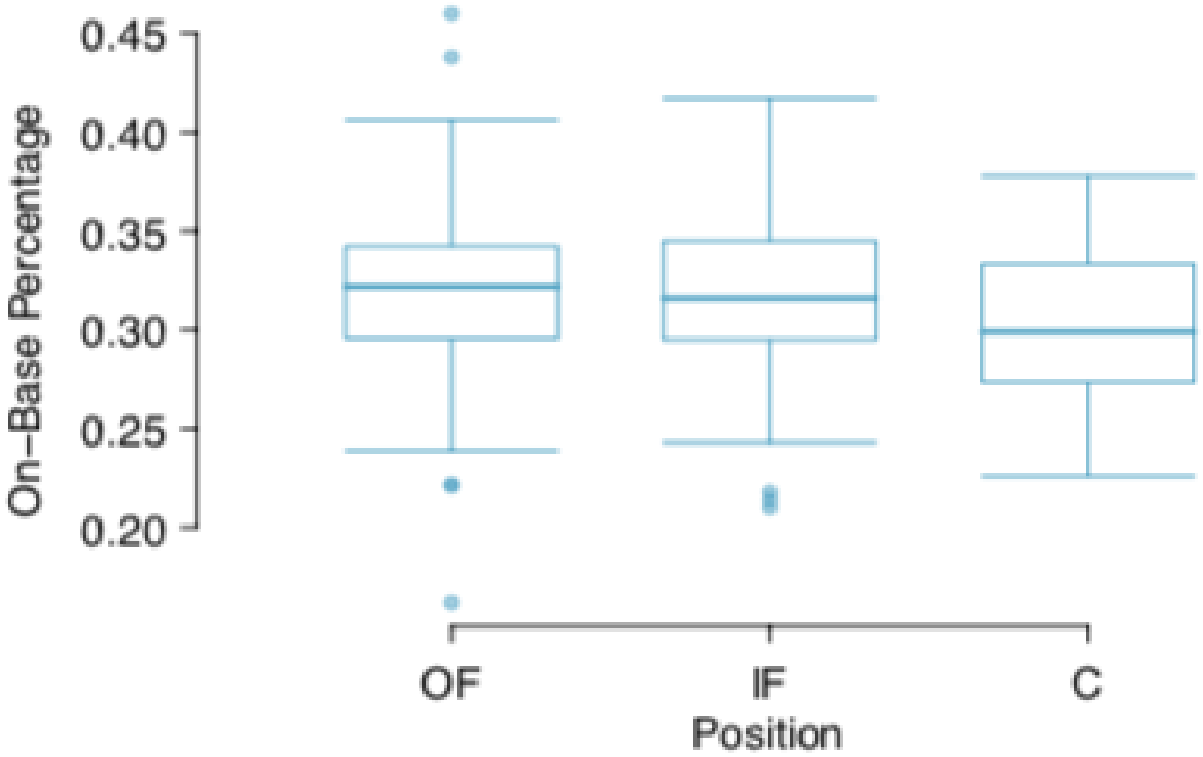
Write the null and alternative hypotheses.

Example: MLB

The by-group summary statistics are

	OF	IF	C
Sample size (n_i)	160	205	64
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302
Sample sd (s_i)	0.043	0.038	0.038

Example: MLB



Example: MLB

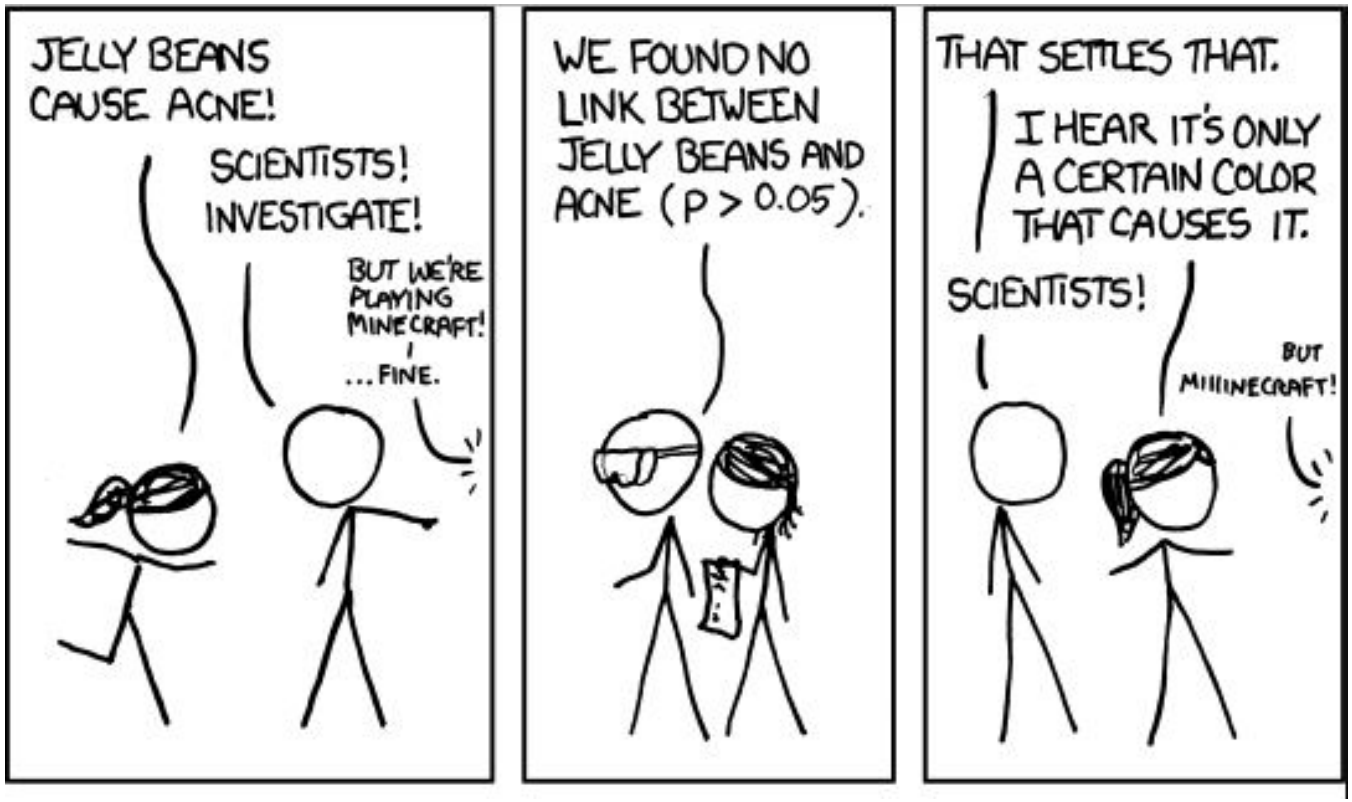
- The largest difference is between the OF and C groups.
- Why not just run a test of $H_0 : \mu_{OF} = \mu_C$?
  - We may miss differences between μ_{OF} and μ_{IF} or μ_C and μ_{IF}.
  - We are inspecting the data before picking comparison groups.

Example

Informal testing (looking at graphs or summary statistics) before choosing tests is called **data snooping**, **data fishing**, or **data hacking**.

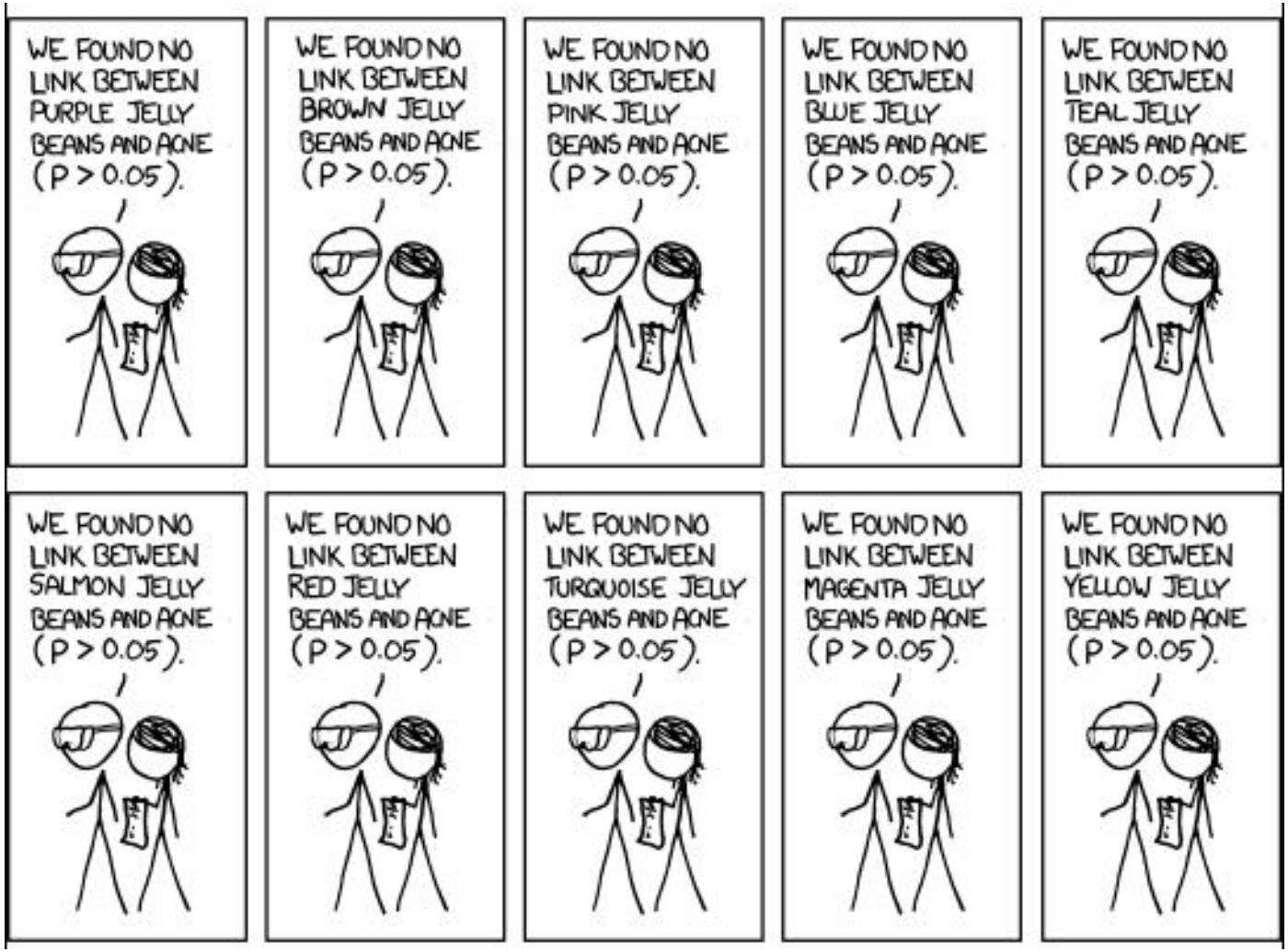
- This will inflate the Type I error rate.

Multiple Comparisons and Type I Error



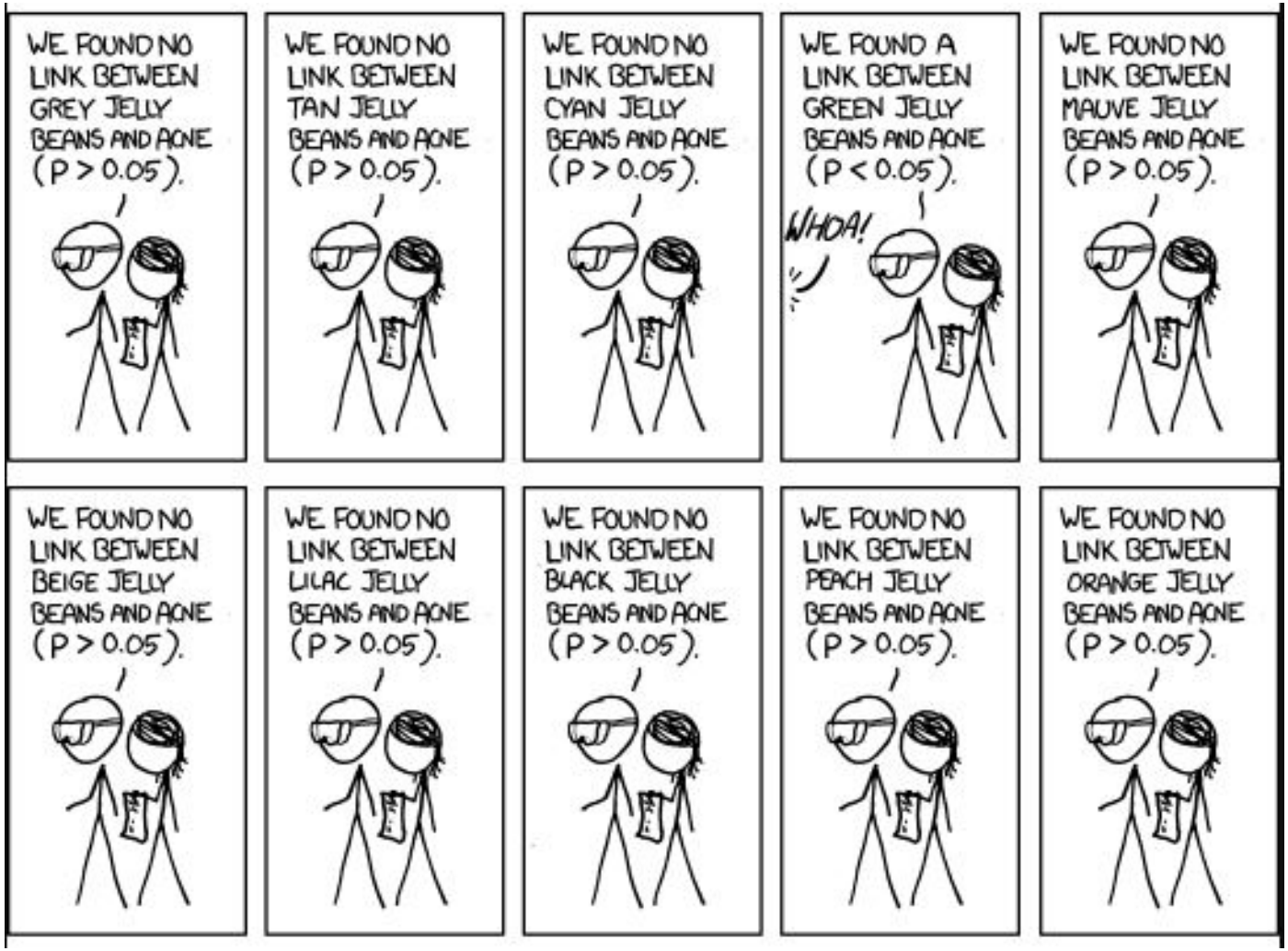
Source: xkcd "Significant" (https://xkcd.com/882/)

Multiple Comparisons and Type I Error



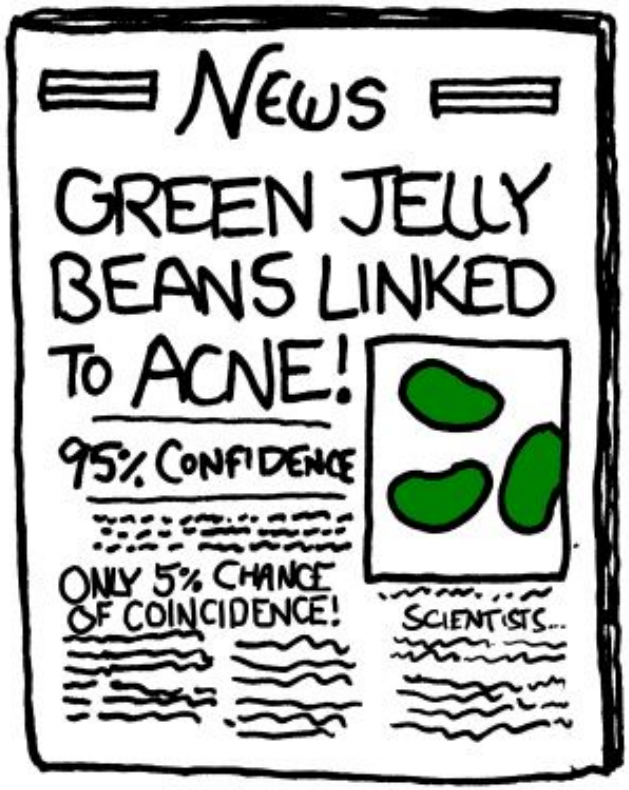
Source: xkcd "Significant" (https://xkcd.com/882/)

Multiple Comparisons and Type I Error



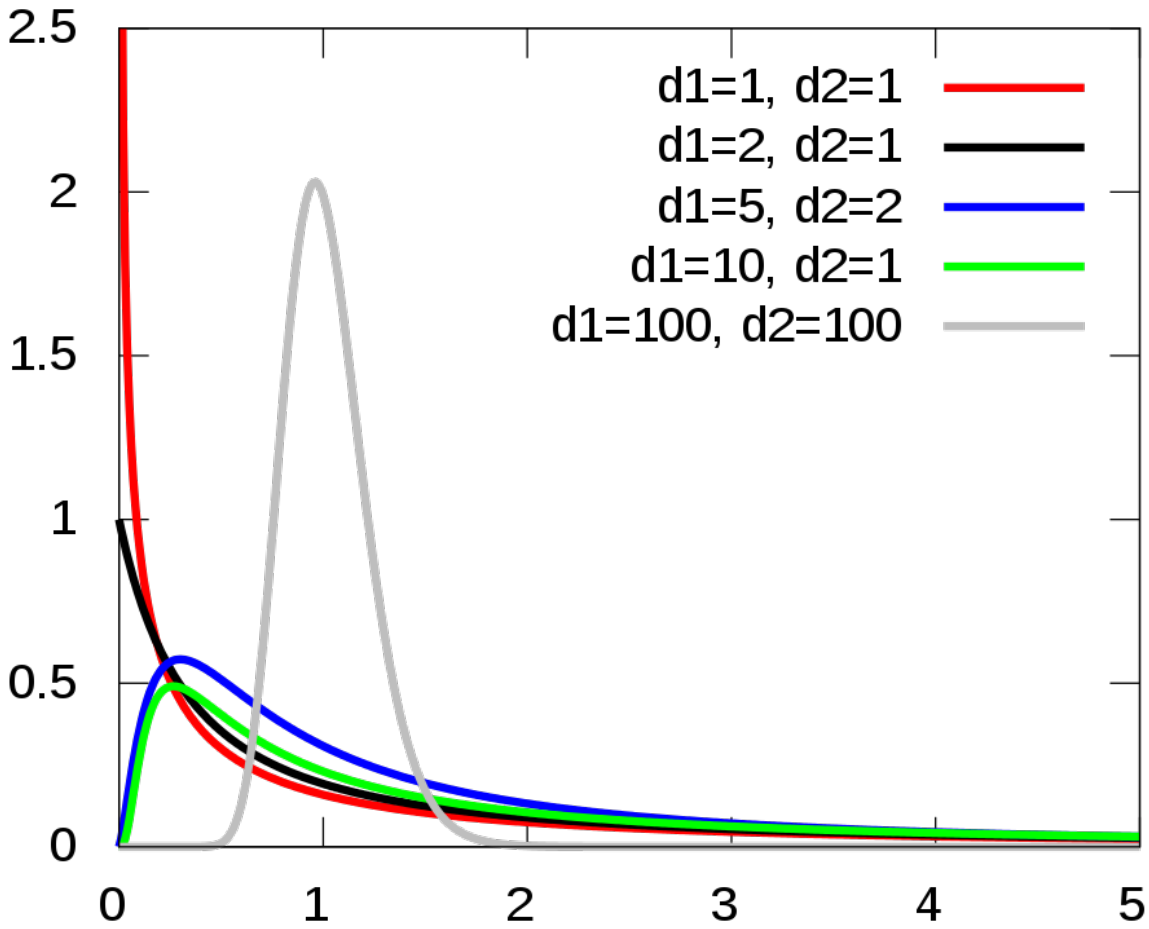
Source: xkcd "Significant" (https://xkcd.com/882/)

Multiple Comparisons and Type I Error



Source: xkcd "Significant" (https://xkcd.com/882/)

The F Distribution



The F Distribution

The F distribution...

- Only takes values ≥ 0.
- Is always right skewed.
- Depends on two sets of degrees of freedom (df_1 and df_2).

We say $X \sim F(df_1, df_2)$.

The F Distribution

The F distribution can be written as the *ratio of two variances*.

$$\frac{s_1/df_1}{s_2/df_2}$$

will have an $F(df_1, df_2)$ distribution.