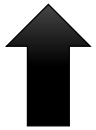
CS/INFO 4154:
Analytics-driven Game Design

Class 31:

A/B Testing Analysis

Mon	Wed	Fri
11/6 A/B Testing Analysis	11/8	11/10
11/13 Newgrounds Release 1	11/15 Newgrounds Release 2	11/17 Newgrounds Release 3
11/20		



Newgrounds Release Report due 10:10am

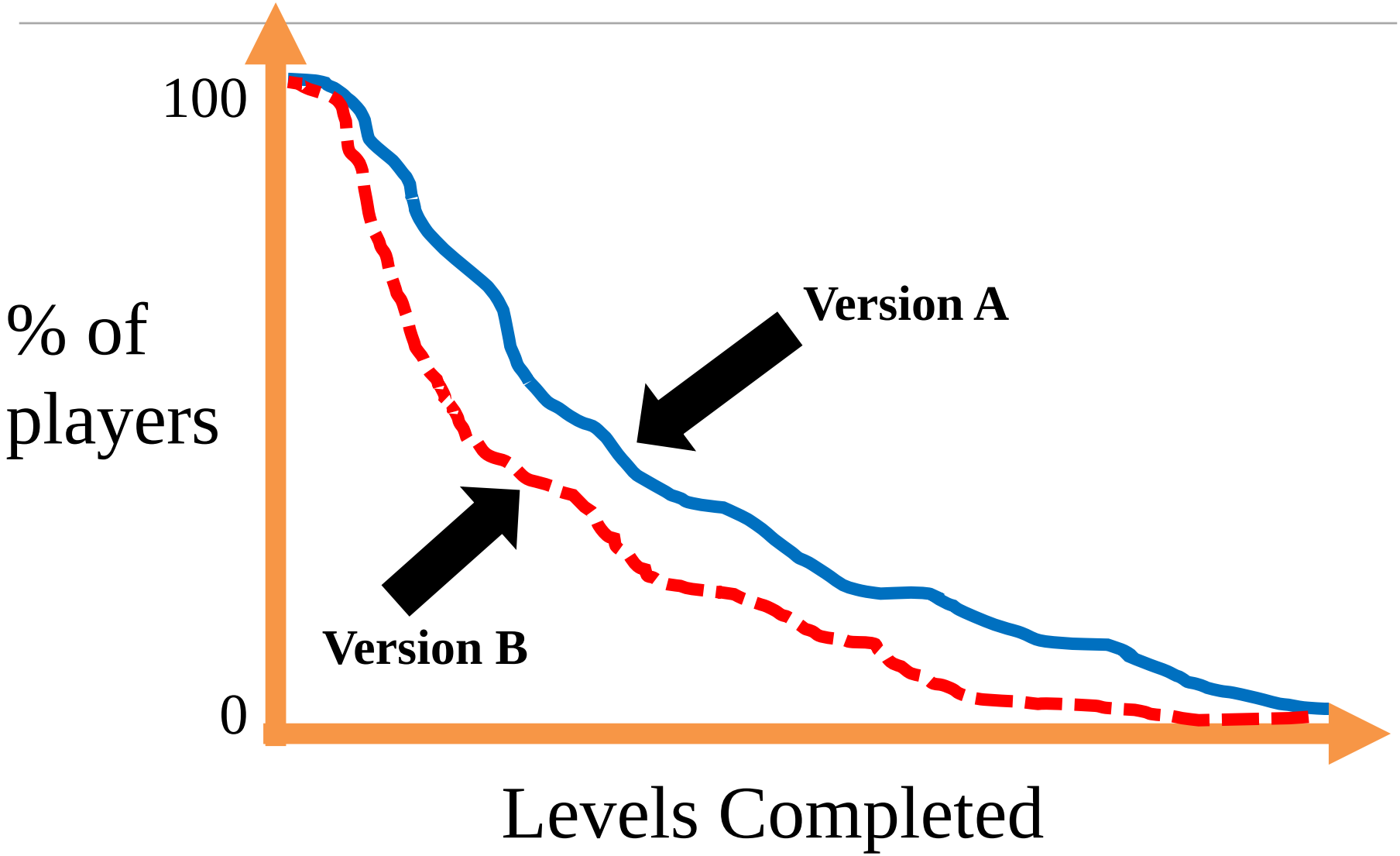
Newgrounds Release 11/13

- New requirements:
  - 12 levels
  - A/B test with 2 conditions and 50/50 player split
  - Privacy policy
- Logistics:
  - Upload to CMS ***before class***
  - Release ***in class***
  - *Strongly suggested*: upload to Newgrounds **test page** before class (but **do not actually release**)

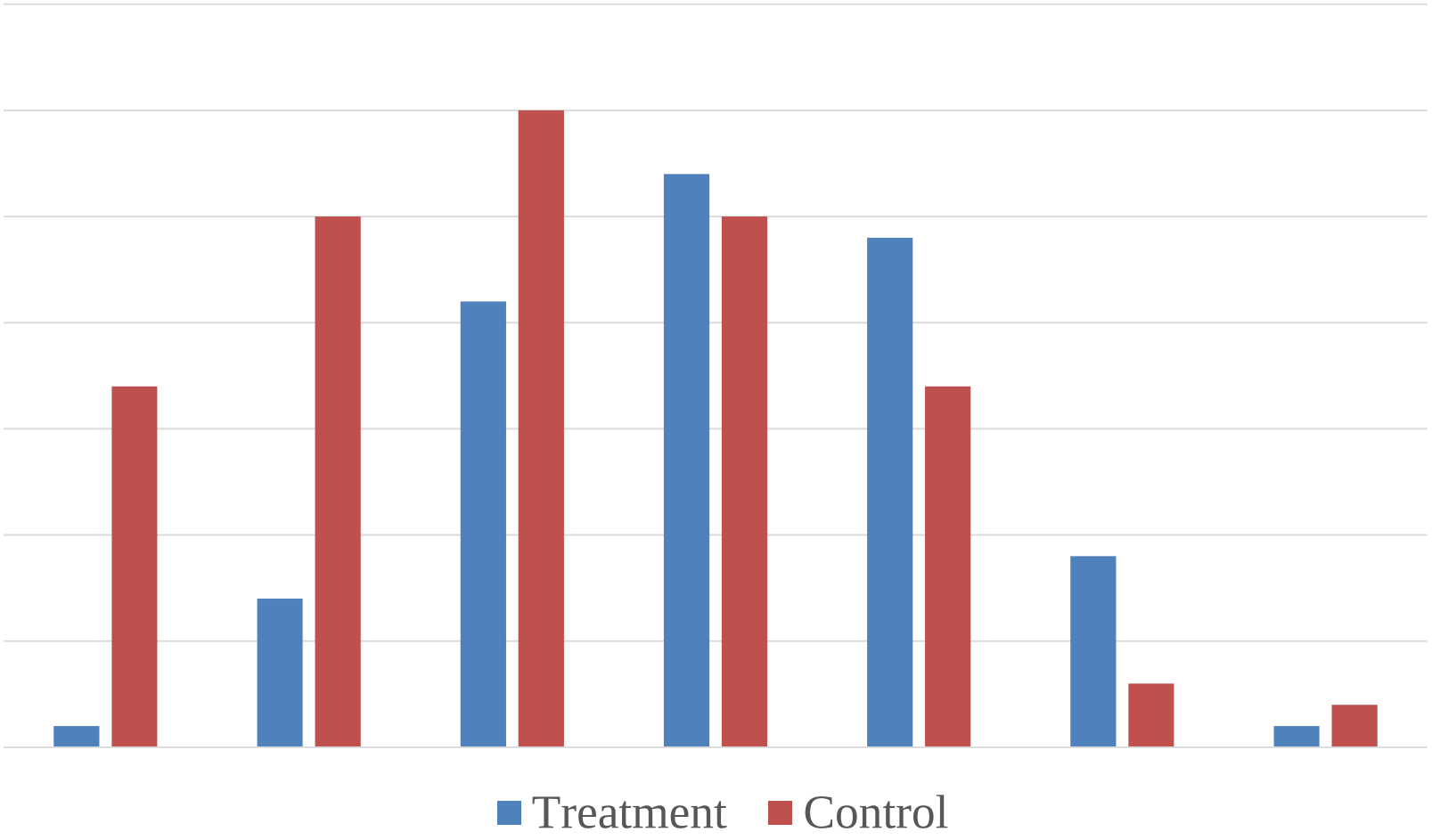
Newgrounds Report 11/20

- What was your A/B test? What did you change? Which condition performed the best? Why?
  - Show and discuss burndown charts for progress (e.g. levels completed) and time played for both the A and B conditions of the Newgrounds Release.
  - Show a heatmap and/or other data that illustrates the impact of the A/B test on player engagement.
  - Test for statistical significance in levels completed using a Wilcoxon-Kruskal-Wallis Two-sample test and report the results.
  - Test for statistical significance in time played using a Wilcoxon-Kruskal-Wallis Two-sample test and report the results.
- Did player engagement improve from Friends to Newgrounds? Why or why not?
  - Show and discuss burndown charts for progress (e.g. levels completed) and time played for the Friends Release and whichever Newgrounds Release condition did the best. (You can show all three if it's difficult to decide.)
  - Show a heatmap and/or other data that illustrates the impact of **the most important** change made from Friends to Newgrounds on player engagement.
- What will you change for Kongregate? **Why does the data support this conclusion?**

Comparing A/B test conditions



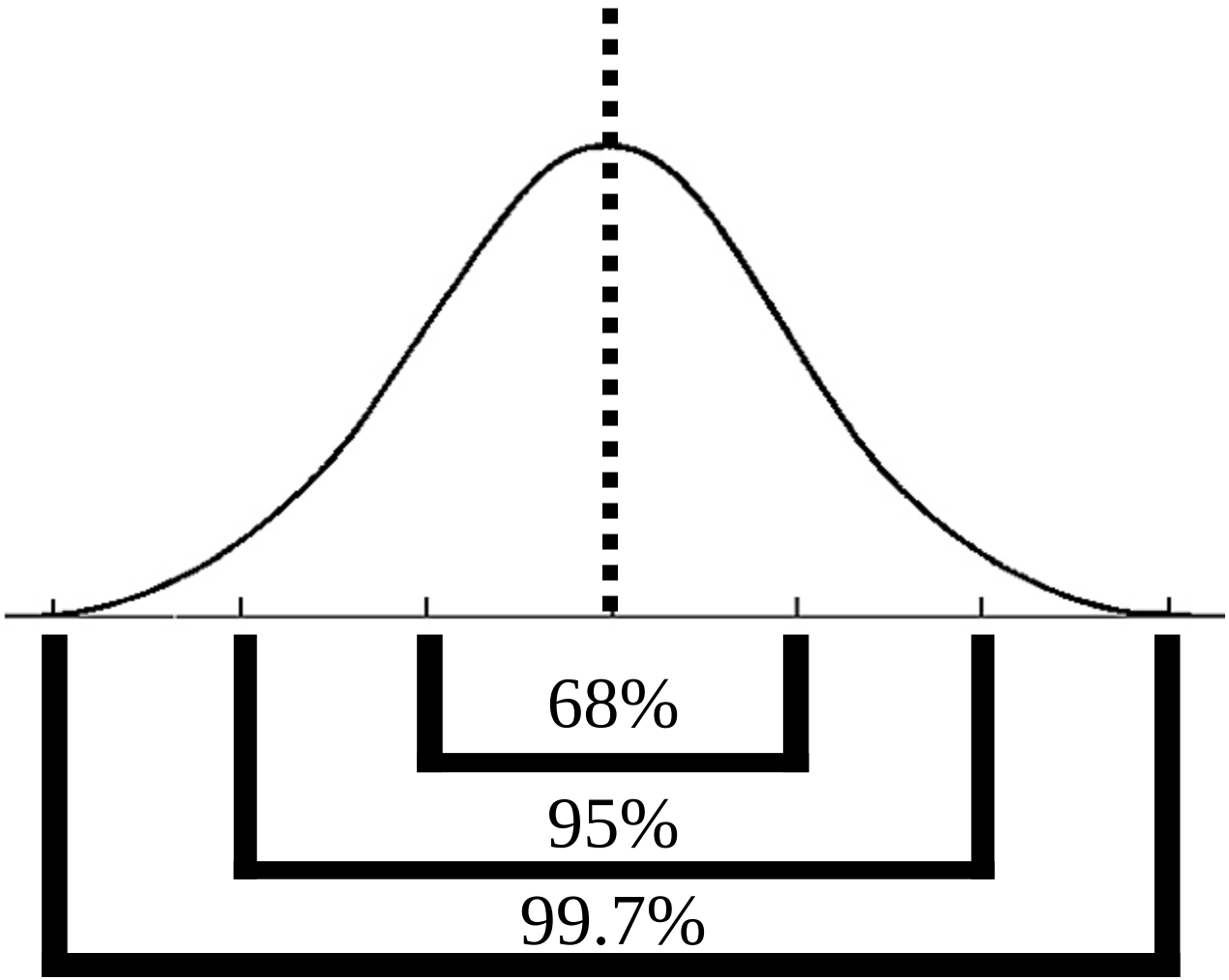
Null hypothesis testing



Typical problem

- Version A (super well tested)
  - Average time played: 120 seconds
  - Standard deviation: 5 seconds
- Version B (new)
  - 100 players
  - Average time played: 135 seconds
- **What is the probability of obtaining a result at least this extreme?**

Bell curve



p-value

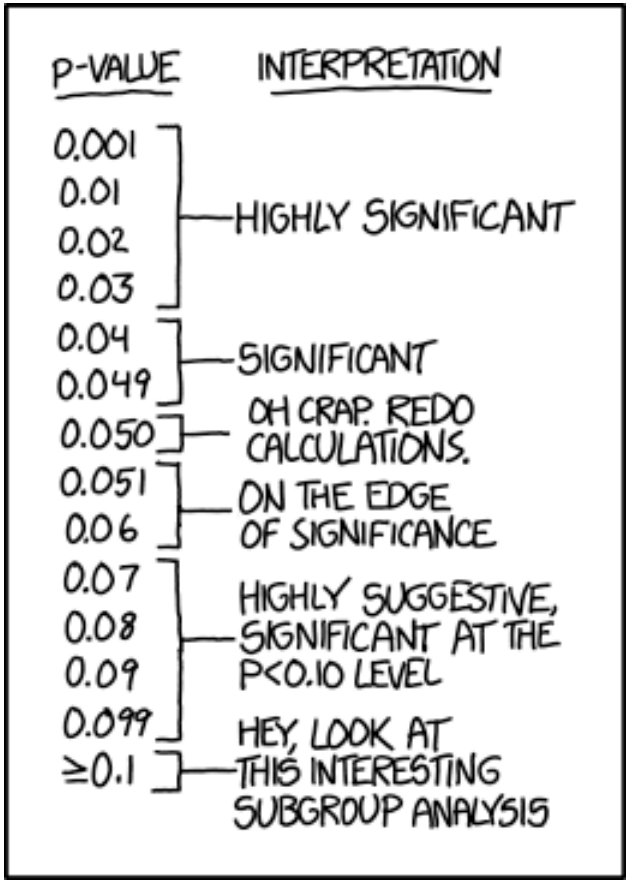
probability the null hypothesis is true

usually we look for: $p < 0.05$

Interpreting null-hypothesis tests

- $p < \alpha$: we *reject the null hypothesis* that the averages are the same.
- $p \geq \alpha$: we *fail to reject* the null hypothesis. Therefore, we conclude **nothing at all**.
- **Common fallacy #1**: $p \geq \alpha$ means no effect
  - (not true)
- **Common fallacy #2**: lower p-value means stronger effect
  - (not true)

Problematic scientific incentives



Source: XKCD

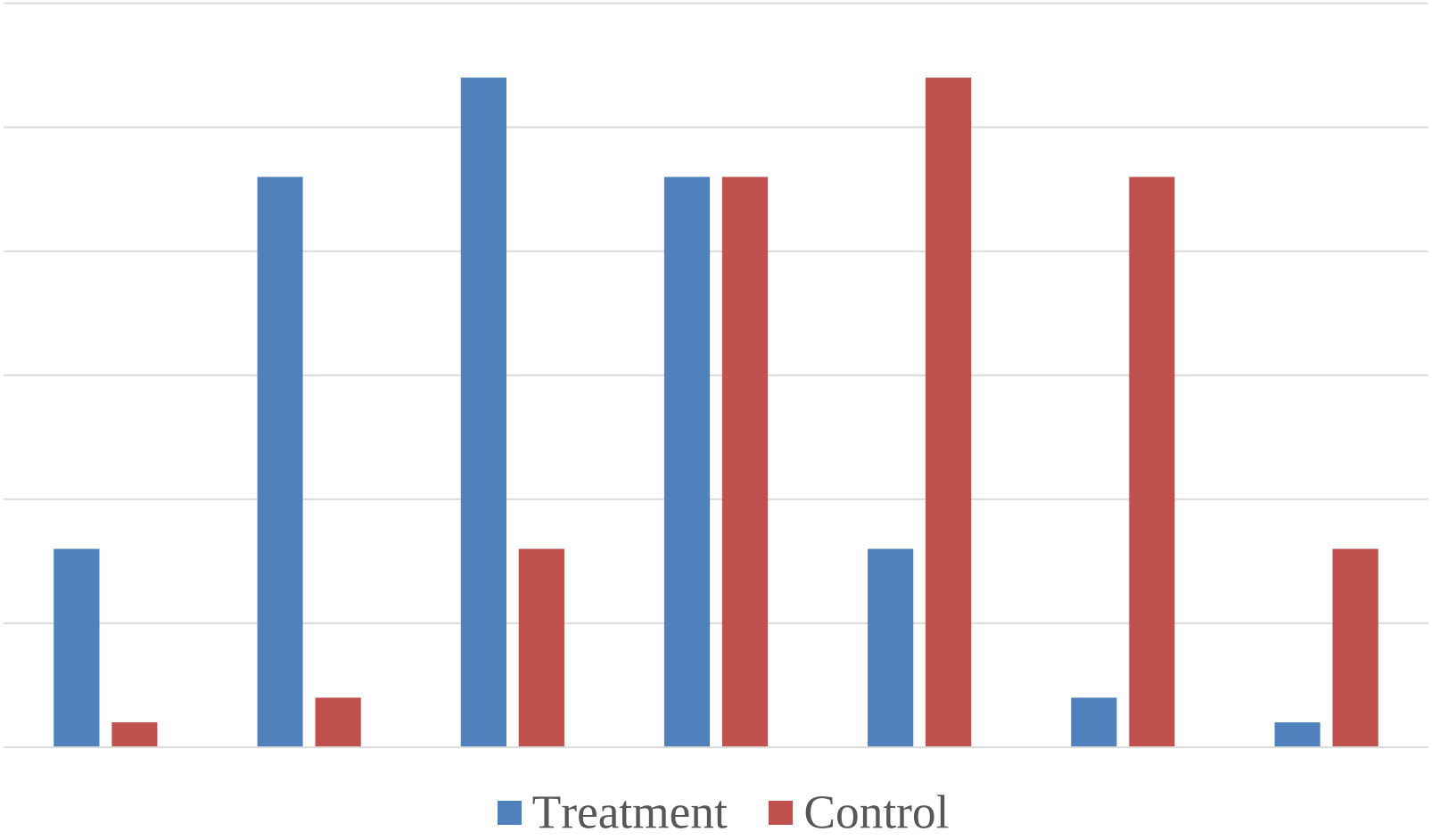
Common tests

- Student’s t-test
- F-test (one-way ANOVA)
- Mann-Whitney U test
  - aka Wilcoxon-Kruskal-Wallis Two-tailed test
- Pearson’s χ^2 test

Student’s t-test

- Compares ***means***
- Assumes normal distribution
- Won’t use in this class, but good to know

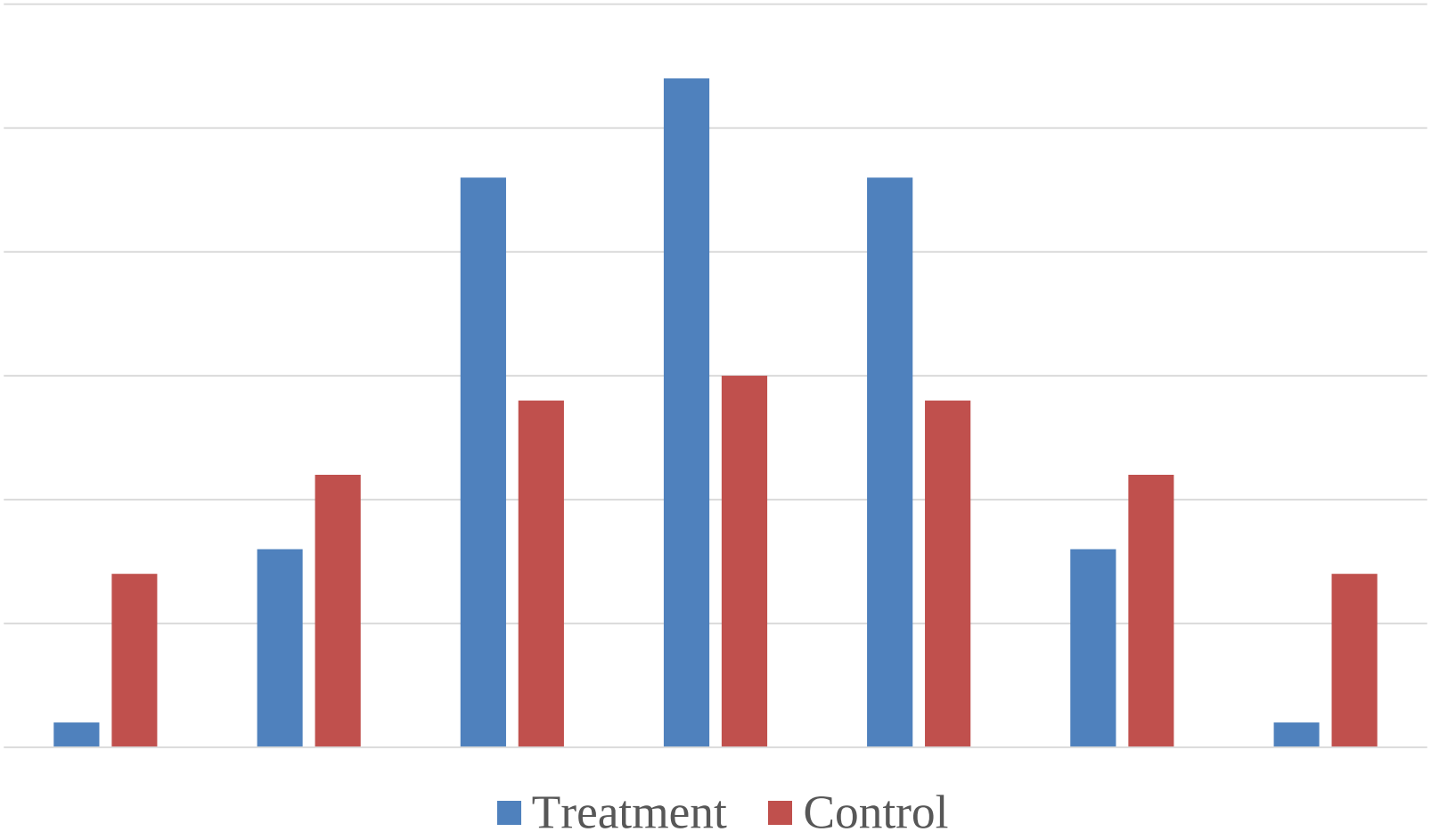
Student’s t-test



F-test

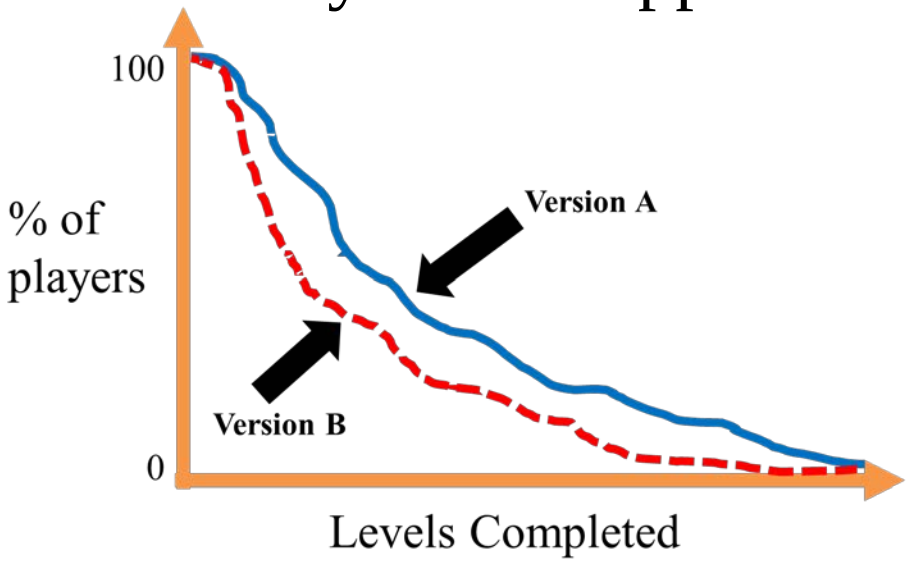
- Compares ***variance***
- Assumes normal distribution
- Won’t use in this class, but good to know

F-test



Up to now

- Everything assumes normal distributions
- This will basically never happen in this class



- So, we will use what are called *nonparametric* statistics

Common tests

- Student’s t-test
- F-test (one-way ANOVA)
- Mann-Whitney U test
  - aka Wilcoxon-Kruskal-Wallis Two-tailed test
- Pearson’s χ^2 test

Common tests

Parametric:

- Student's t-test
- F-test (one-way ANOVA)

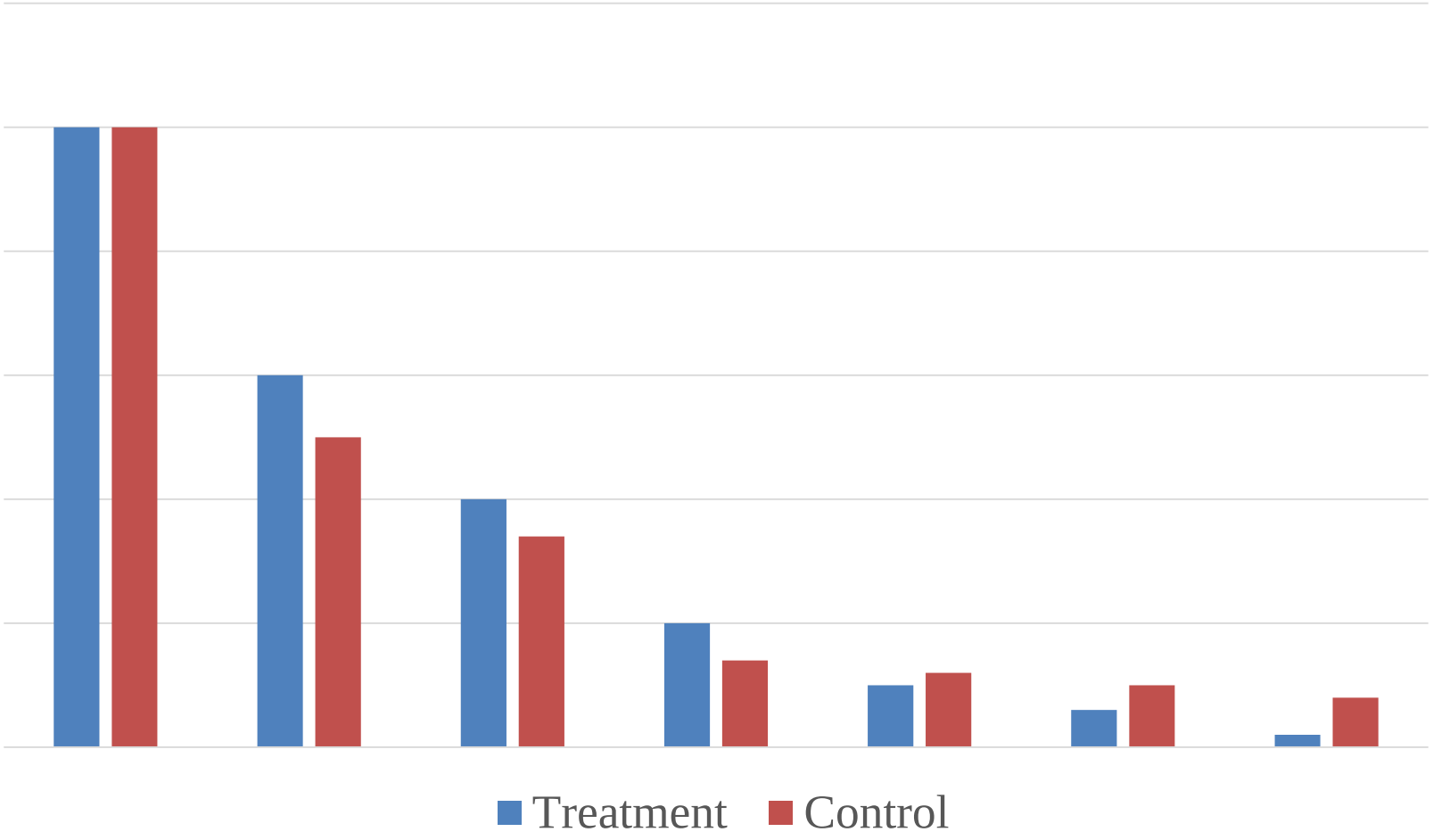
Nonparametric:

- Mann-Whitney U test
  - aka Wilcoxon-Kruskal-Wallis Two-sample test
- Pearson's χ^2 test

Mann-Whitney U-test
(aka Wilcoxon-Kruskal-Wallis two-sample test)

- Compares ***ranks***
- No assumption of normality

Mann-Whitney U-test
(aka Wilcoxon-Kruskal-Wallis two-sample test)



Reporting Conventions

The distribution of the number of distinct days of activity **differed significantly** between the two groups (Mann-Whitney U = 116584.5, Z = 3.41, n_{on} = 516, n_{off} = 515, $p < 0.001$), with the average number of active days for the “badges on” group being 7.01 compared with 6.21 for the “badges off” group.

Denny et al. CHI 2013

Reporting U-test results

- Medians
- Z-statistic
- p-value

Pearson's χ^2-test

- For ***categorical*** data
  - non-continuous distributions
  - true/false
  - baked hamburger vs. fries
- No assumption of normality

Reporting Pearson's χ^2-test results

- Counts of each category (or percentages)
- Pearson's χ^2 statistic
- p-value

The multiple comparisons problem

Source: XKCD

The multiple comparisons problem

- Ethical conventions:
  - Limit number of tests
  - Need to decide on tests ***before*** doing analysis
  - Must have ***rationale*** for running each test
  - Cannot run tests until something is significant
- In order to run unplanned tests, use a ***correction***
  - Bonferroni correction
    - If running m tests, divide significance level α by m
    - For example, to run 10 unplanned tests, look for $p < 0.005$.

Sample file

http://www.cs.cornell.edu/courses/cs4154/2017fa/materials/sample.csv

JMP Demo