# Sanity check

Given: each cookie is randomly assigned to the control or experiment group with probability 0.5

Now, total control: 64454, total experiment: 61818

Question: how would you figure out whether this difference is within expectations?

Steps:

1. Compute standard deviation of binomial with probability 0.5 of success

SD=

1. Multiply by z-score to get margin of error

M = SD \* 1.96 = 0.0027

1. Compute confidence interval around 0.5

0.5±0.0027 = (0.4973, 0.5027)

1. Check whether observed fraction is within interval

P\_hat = 64454/64454+61818 = 0.5104 greater than 0.5027

What to do:

1. Talk to engineers (technical problems)
2. Try slicing to see if one particular slice is weird
3. Check age of cookies – does one group have more new cookies

# Analysis with a single metric

Metric: click-through-rate, d\_min = 0.01, alpha = 0.05

Empirical SE: 0.0062 with 5000 pageviews in each group

Control pageviews: 27948 control CTR: 0.1016

Experiment pageviews: 28052 experiment CTR: 0.1132

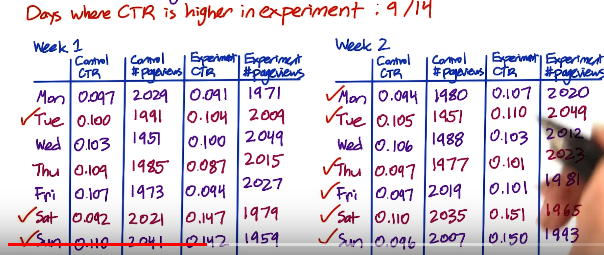
D\_hat = 0.1132 – 0.1016 = 0.0116

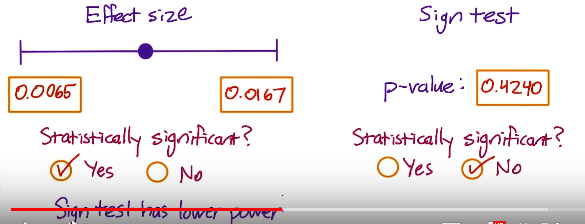
SE = 0.0026

M = 0.0026 \* 1096 = 0.0051

Confidence interval: 0.0065 to 0.0167

Doesn’t include zero -> statistically significant, but include 0.05, so the size of effect might not be big enough





Analysis:

CI for weekdays: -0.0078 to 0.0043

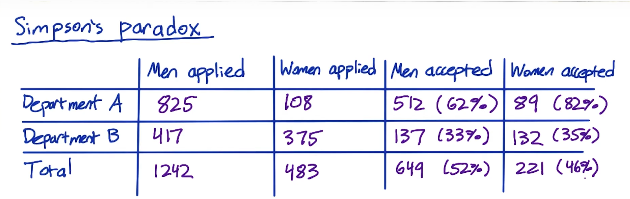
CI for weekends: 0.0361 to 0.0553

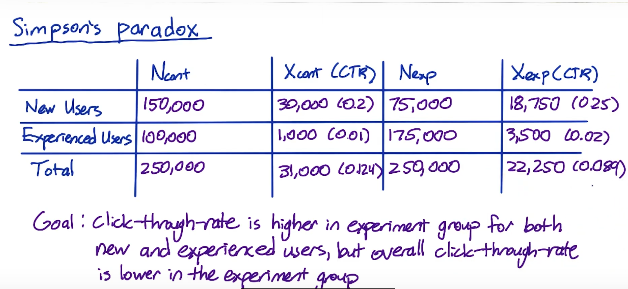
So more significant on the weekends

Recommendation: do not launch (yet).

**Simpson’s paradox**

Difference in behaviour between subgroup





Why are there more pageviews from new users in the control group

1. Something wrong with set-up
2. Change affects new users and experienced users differently

# Tracking multiple metrics

Experiment: prompt students to contact coach more frequently

Metrics:

1. Probability that student signs up for coaching
2. How early students sign up for coaching
3. Average price paid per student

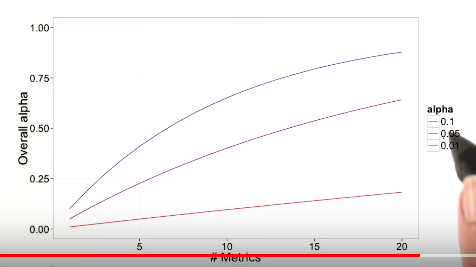
If Audacity tracks all three metrics and does three separate significance tests (alpha=0.05), what is the prob at least one metric will show a significant difference if there is no true difference?

For 3 metrics, whats the chance of at least 1 false positive?

P(FP=0) = 0.95 \* 0.95 \* 0.95 = 0.857 (assuming independence)

P(FP≥1) = 1 – 0.857 = 0.143

Probability of at least one false positive: alpha\_overall = 1 – (1-alpha\_individual)^n



Problem: probability of any false positive increases as you increase number of metrics

Solution: use higher confidence level for each metric

Method 1: assume independence

alpha\_overall = 1 – (1-alpha\_individual)^n

Method 2: Bonferroni correction (simple, no assumption, conservative – guaranteed to give alpha\_overall at least as small as specified)

Alpha\_individual = alpha\_overall / n

e.g. alpha\_overall = 0.05, n = 3, so alpha\_individual = 0.0167

example: update description on course list

3 out of 4 metrics had significant difference at alpha=0.05, but none were significant using Bonferroni correction

Recommendation:

Rigorous answer: use a more sophisticated method

In practice: judgement call, possibly based on business strategy

Different strategies:

1. Control probability that any metric shows a false positive

Alpha\_overall, familywise error rate (FWER)

1. Control false discover rate (FDR)

FDR = E[ #false positives / #rejections ]

Suppose you have 200 metrics, cap FDR at 0.05. this means you are okay with 5 false positives and 95 true positives in every experiment

Statistically and practically significant

Understand why

Factors that cause changes over time

Seasonally: holdback

Novelty effect/ change aversion: cohort analysis

Advertiser budgets