A/B Testing and Beyond

Designed Experiments for Data Scientists

UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

DATA INSTITUTE
UNIVERSITY OF SAN FRANCISCO

Week 4

Wednesday September 27th, 2017

Outline

- Recap
- Experiments with Two Conditions
    - Evaluating Assumptions
        - Welch's t-test
        - Randomization tests
        - χ^2-tests
    - A discussion of "peeking"

RECAP

Recap

- Experiments with Two Conditions
  - Comparing Means
    - The two-sample *t*-test
    - Power analysis and sample size calculations
  - Comparing Proportions
    - The Z-test for proportions
    - Power analysis and sample size calculations

EXPERIMENTS WITH TWO CONDITIONS

Recall

When comparing means…

- We assume the response variable of interest is measured on a continuous scale

- But this methodology is also commonly applied when the response variable is discrete with a large support set

- We assume that the n_j response measurements in condition $j = 1,2$ follow a normal distribution:

$$Y_{ij} \sim N\left(\mu_j, \sigma^2\right)$$

for $i = 1,2, \ldots, n_j$

Recall

When comparing means…

To formally decide whether $\mu_1 = \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$, we test one or more of the following:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 \leq \mu_2 \text{ vs. } H_A: \mu_1 > \mu_2$$

$$H_0: \mu_1 \geq \mu_2 \text{ vs. } H_A: \mu_1 < \mu_2$$

Evaluating Assumptions

When comparing means…

When testing these hypotheses using the standard *t*-test we saw last week, we make two key assumptions:

- We assume the variances in the two conditions are equal
- We assume that the response measurements follow a normal distribution

When these assumptions are not valid we require alternative approaches

Evaluating Assumptions

When $\sigma_1^2 \neq \sigma_2^2$

In this situation we alter our test statistic, and instead of using

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

we use

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

where $\hat{\sigma}_j^2$ is the sample variance in condition j.

Evaluating Assumptions

When $\sigma_1^2 \neq \sigma_2^2$

However, this version of the test statistic no longer follows a *t*-distribution exactly

It only approximately follows a *t*-distribution with

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{(\hat{\sigma}_1^2/n_1)^2}{n_1 - 1} + \frac{(\hat{\sigma}_2^2/n_2)^2}{n_2 - 1}}$$

degrees of freedom.

Evaluating Assumptions

When $\sigma_1^2 \neq \sigma_2^2$

Thus a conclusion is drawn by comparing the observed value of t with the null distribution: $t_{(\nu)}$ where the degrees of freedom ν are shown on the previous slide.

Note that p-values are calculated as usual.

The *t*-test carried out in this way is called Welch's *t*-test

We can perform this test in `R` by setting `var.equal = FALSE` in the `t.test()` function

Evaluating Assumptions

But how do we know if $\sigma_1^2 \neq \sigma_2^2$ and that Welch's *t*-test is appropriate?

We could formally decide by testing the following hypothesis:

$$H_0\colon \sigma_1^2 = \sigma_2^2 \text{ vs. } H_A\colon \sigma_1^2 \neq \sigma_2^2$$

- If H_0 can be rejected, we should use Welch's *t*-test
- If H_0 cannot be rejected, we should use Student's *t*-test

Evaluating Assumptions

The hypothesis on the previous slide is typically tested using an *F*-test.

This test is so named because the null distribution of the test statistic follows an *F*-distribution.

Let's take a brief detour to familiarize ourselves with this distribution…

Evaluating Assumptions

The *F*-distribution

A $Y \sim F(\nu_1, \nu_2)$ random variable has PDF given by

$$f(y) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} y^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}y\right)^{-\frac{\nu_1+\nu_2}{2}}$$

for $y \geq 0$ and where ν_1 and ν_2 are positive integers.

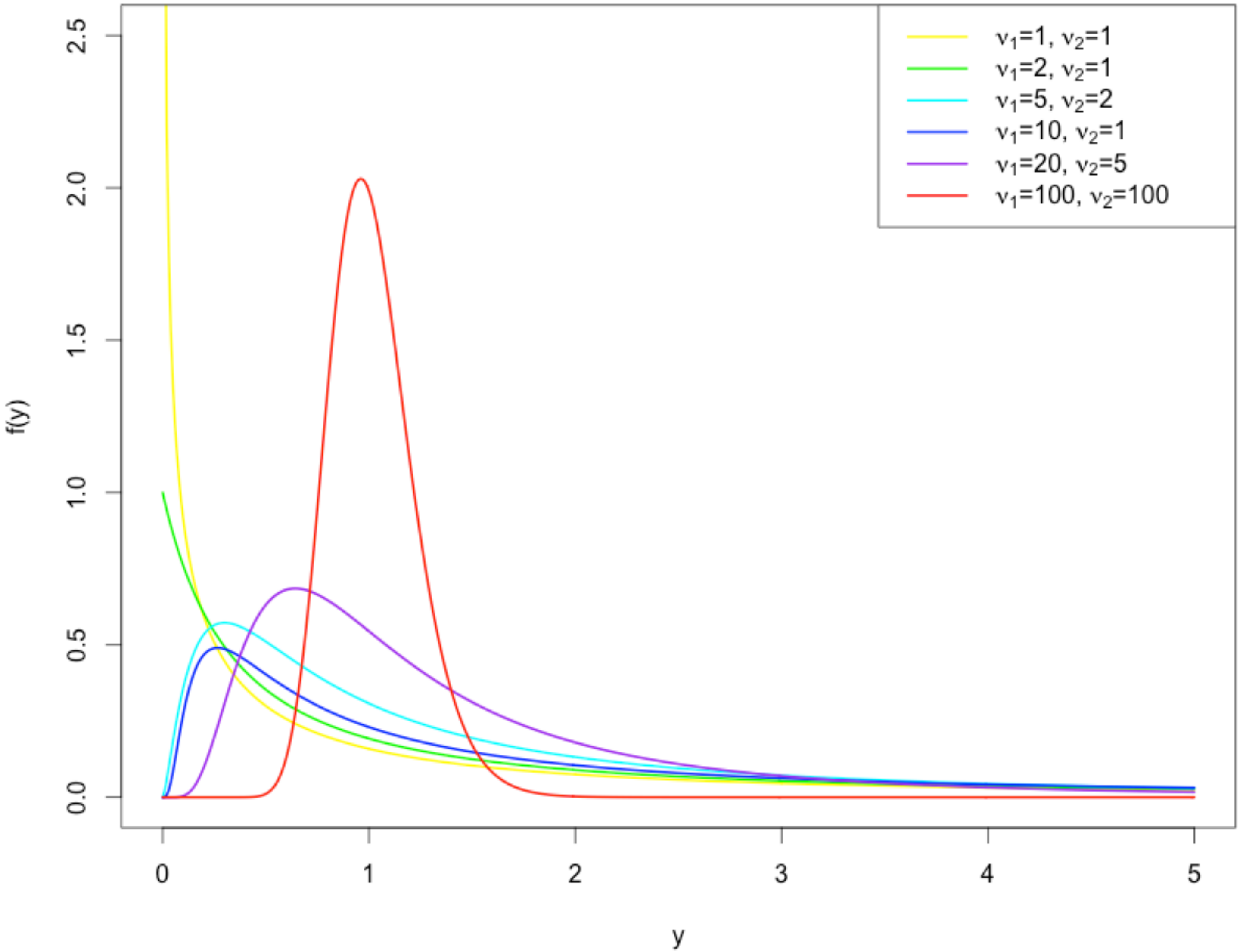
In R:

$P(Y \leq q) =$ `pf(q, df1 = v1, df2 = v2)`

$P(Y \geq q) =$ `1-pf(q, df1 = v1, df2 = v2)`

Evaluating Assumptions

The *F*-distribution



Evaluating Assumptions

The *F*-test for variances

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ vs. } H_A: \sigma_1^2 \neq \sigma_2^2$$

In this situation we assume $Y_{ij} \sim N\left(\mu_j, \sigma_j^2\right)$ for $i = 1,2, \ldots, n_j$, $j = 1,2$ which means that

$$\frac{\left(n_j - 1\right)\hat{\sigma}_j^2}{\sigma_j^2} \sim \chi^2_{(n_j - 1)}$$

and

$$T = \frac{\hat{\sigma}_1^2 / \sigma_1^2}{\hat{\sigma}_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Evaluating Assumptions

The *F*-test for variances

We use T (from the previous slide) as our test statistic whose observed value is

$$t = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

Since, if H_0 is true, $\sigma_1^2/\sigma_2^2 = 1$.

We compare this value to the $F(n_1 - 1, n_2 - 1)$ distribution to determine it extremity

Evaluating Assumptions

The *F*-test for variances

The p-value is calculated to be

$$\text{p} - \text{value} = P(T \geq t) + P(T \leq 1/t)$$

where $T \sim F(n_1 - 1, n_2 - 1)$

Notice that this calculation is slightly different from other two-sided p-values we've calculated

This arises because "at least as extreme" means something different in this setting

Note that this test can be performed in `R` using the `var.test()` function

Evaluating Assumptions

What if Y_{ij} is not normally distributed?

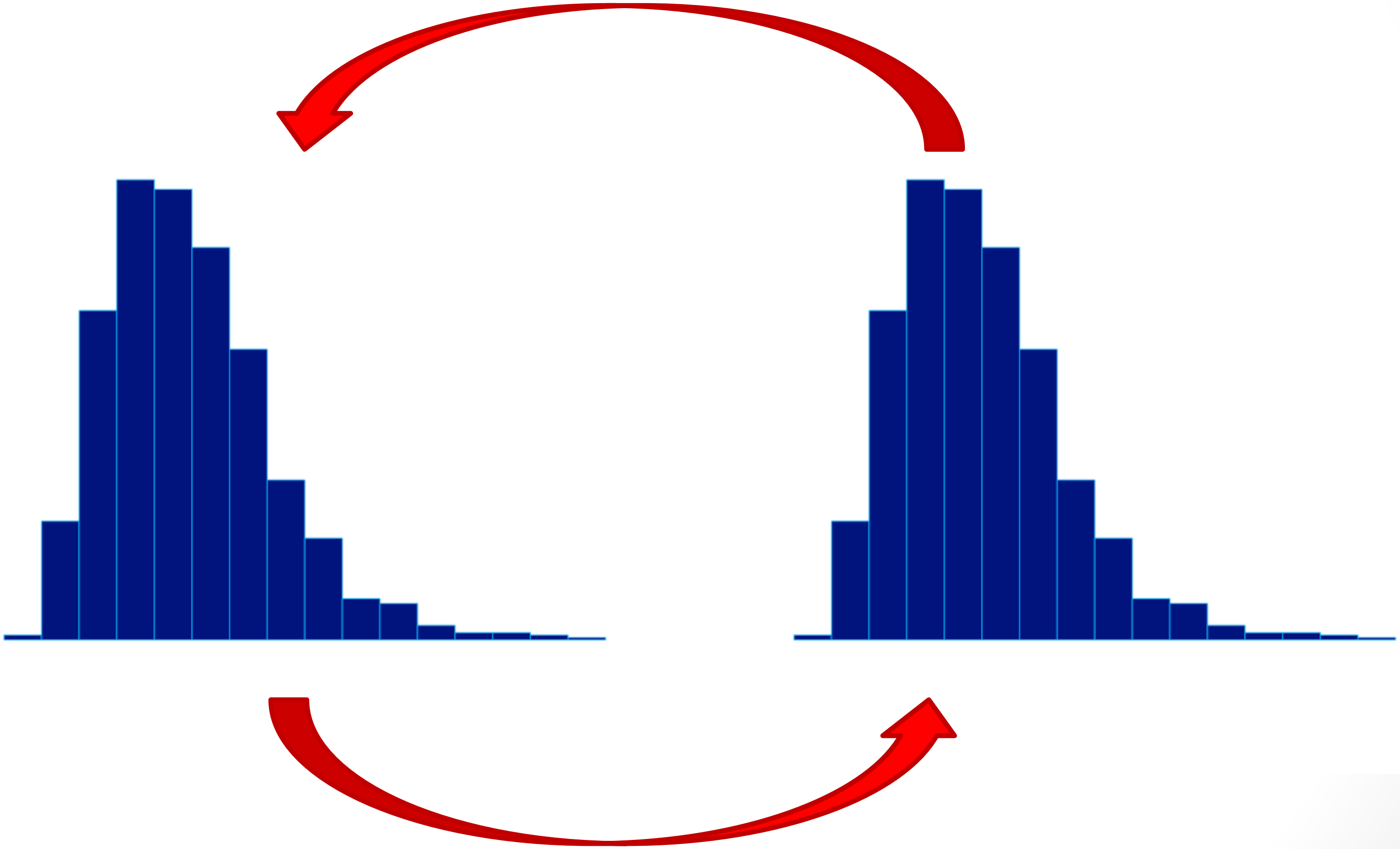
In this case we ran use a permutation or randomization test

Both of these approaches are nonparametric resampling techniques

The motivating idea behind both of these is that, if H_0 is true, any random rearrangement of the data is **equally likely to have been observed**

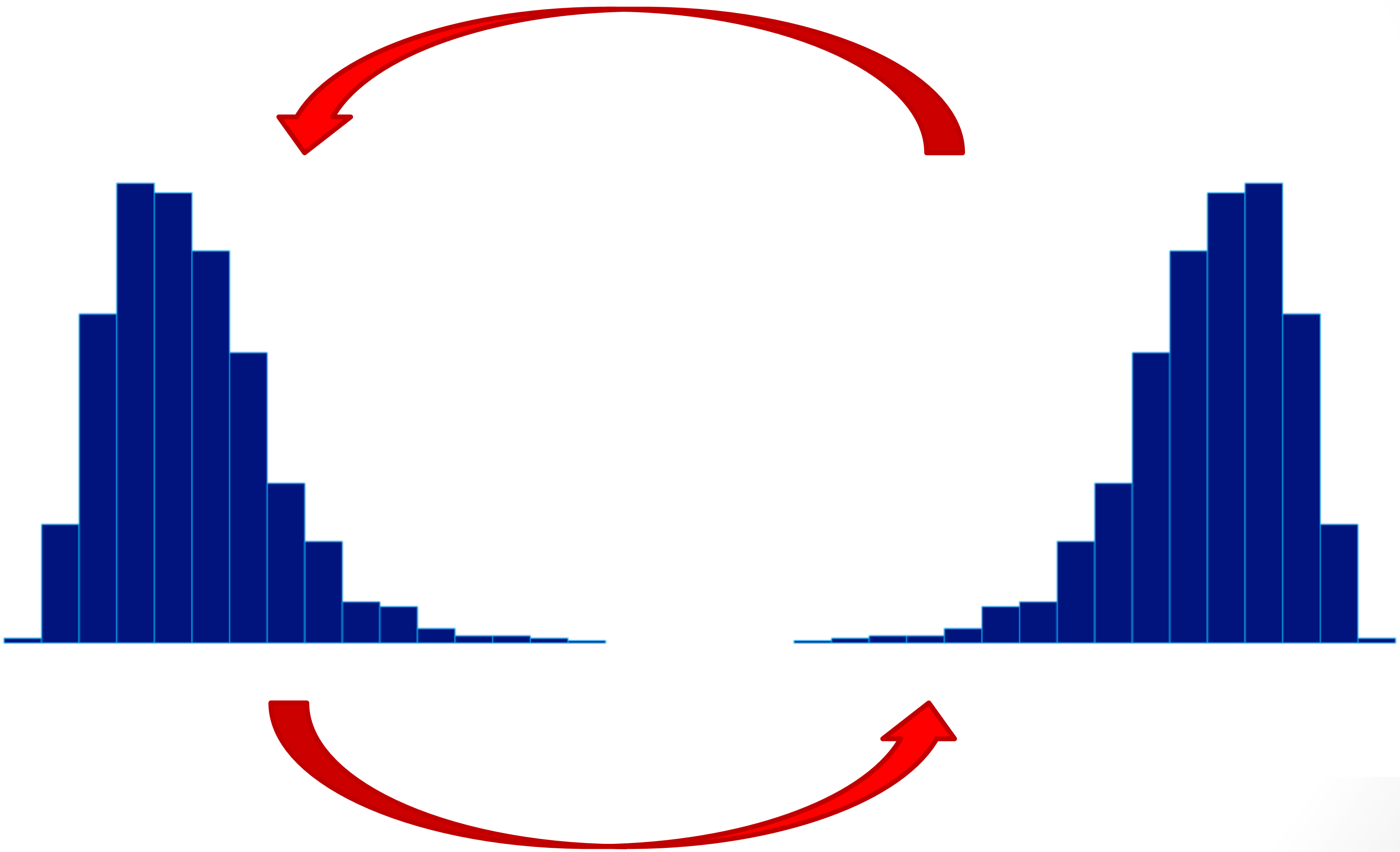
Evaluating Assumptions

Permutation and Randomization Tests



Evaluating Assumptions

Permutation and Randomization Tests



Evaluating Assumptions

Permutation and Randomization Tests

- Note that there are

$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$

  arrangements of the $n_1 + n_2$ observations into two groups (of size n_1 and n_2)

- A **true permutation test** calculates the test statistic on the original sample ($t = \hat{\mu}_1 - \hat{\mu}_2$) and every permuted sample (t^*)

- The distribution of t^* is taken to be the null distribution of the test statistic and the extremity of t is evaluated in this context

Evaluating Assumptions

Permutation and Randomization Tests

- However, even with reasonably small samples,
$$\binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}$$
is a very large number
- If $n_1 = n_2 = 50$, then there are 1.09×10^{29} possible samples to consider
- So the permutation test can be computationally expensive
- As an alternative we use a **randomization test** which investigates a large number of resamples, as opposed to all possible permutations

Evaluating Assumptions

Permutation and Randomization Tests

The algorithm for carrying out a randomization test is as follows:

1. Collect response observations in each condition $\left\{y_{1j}, y_{2j}, \dots, y_{n_j j}\right\}$ for $j = 1,2$
2. Calculate the test statistic $t = \hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2$ on the original sample
3. Resample the data without replacement so that n_1 observations are randomly associated with a resampled 'condition 1' and n_2 observations are randomly associated with a resampled 'condition 2'

Evaluating Assumptions

Permutation and Randomization Tests

4. Calculate the value of the test statistic, labeled t^*, on this resample
5. Repeat steps 3 and 4 a few thousand times
6. Compare t to the distribution which is derived from the resampled values of t^* and calculate the p-value

The p-values of this test are calculated empirically and the calculation depends on whether H_A is one- or two-sided.

Evaluating Assumptions

Permutation and Randomization Tests

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

- p-value = the proportion of resampled test statistics $t^* \geq |t|$ or $t^* \leq -|t|$

$$H_0: \mu_1 \leq \mu_2 \text{ vs. } H_A: \mu_1 > \mu_2$$

- p-value = the proportion of resampled test statistics $t^* \geq t$

$$H_0: \mu_1 \geq \mu_2 \text{ vs. } H_A: \mu_1 < \mu_2$$

- p-value = the proportion of resampled test statistics $t^* \leq t$

Evaluating Assumptions

Permutation and Randomization Tests

Note that we have introduced these tests in the context of comparing means, but they are appropriate for the comparison of any metric that might be compared between two conditions

$$H_0: \theta_1 = \theta_2 \text{ vs. } H_A: \theta_1 \neq \theta_2$$

$$H_0: \theta_1 \leq \theta_2 \text{ vs. } H_A: \theta_1 > \theta_2$$

$$H_0: \theta_1 \geq \theta_2 \text{ vs. } H_A: \theta_1 < \theta_2$$

In this more general case our test statistic is simply calculated to be $t = \hat{\theta}_1 - \hat{\theta}_2$

Evaluating Assumptions

Randomization Test Example

Suppose Niantic is experimenting with two different promotions within Pokémon Go

- The first involves giving users 200 free Pokécoins
- The second involves giving users a 50% discount on in-app shop purchases

What they are interested in is whether, relative to providing no promotion, either of these strategies lead to users spending more of their own money in the shop.

Evaluating Assumptions

Randomization Test Example

To investigate this, a small experiment with $n_1 = n_2 = n_3 = 100$ users is performed in which

- Condition 1 (control): users receive no promotion
- Condition 2: users receive 200 free Pokécoins
- Condition 3: users receive a 50% in the shop

For each user, the amount of real money (in \$) that they spend in the 30 days following the experiment is recorded.

Evaluating Assumptions

Randomization Test Example

$\hat{\mu}_1 = 10.74$

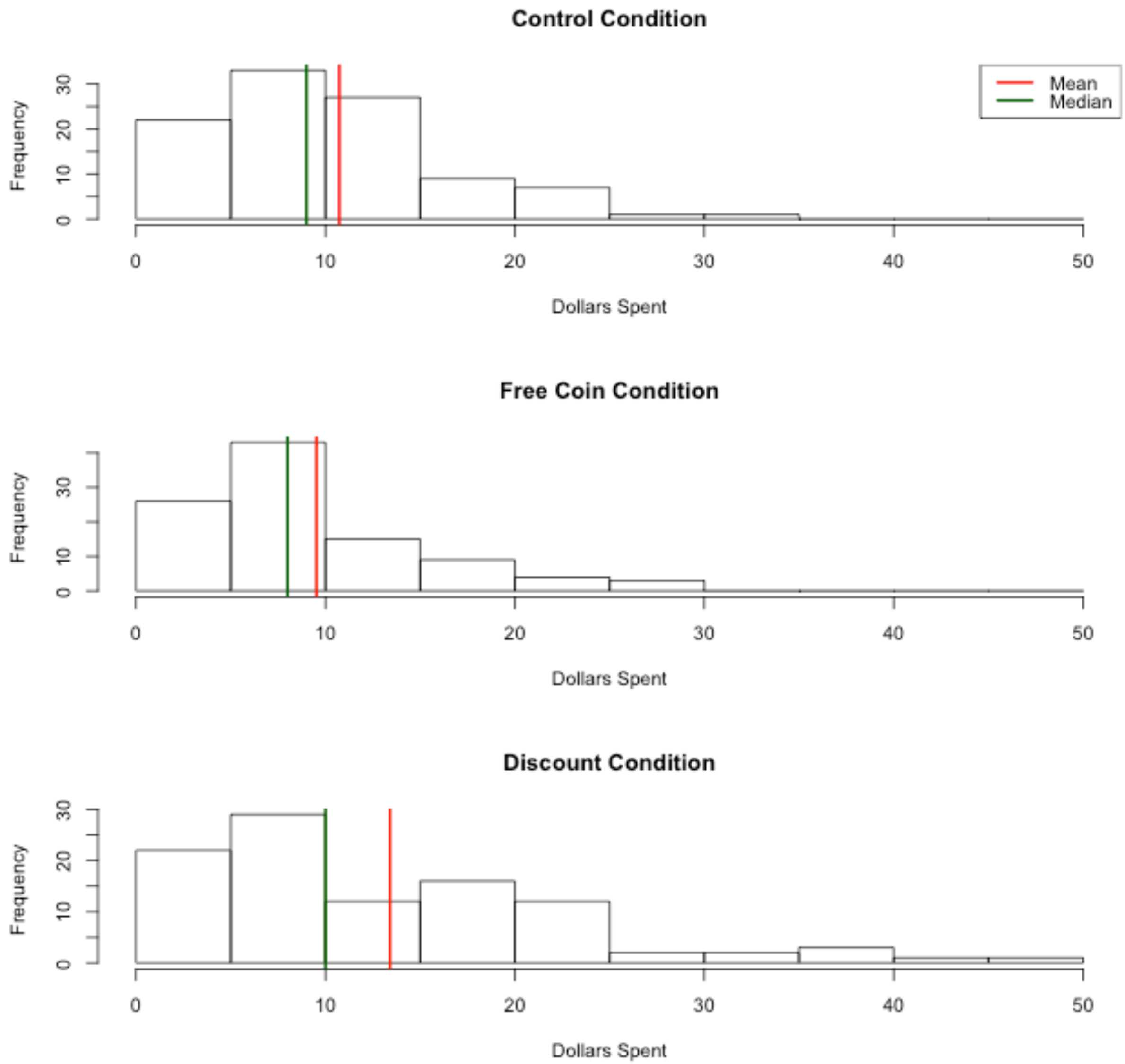
$\check{\mu}_1 = 9$

$\hat{\mu}_2 = 9.53$

$\check{\mu}_2 = 8$

$\hat{\mu}_3 = 13.41$

$\check{\mu}_3 = 10$

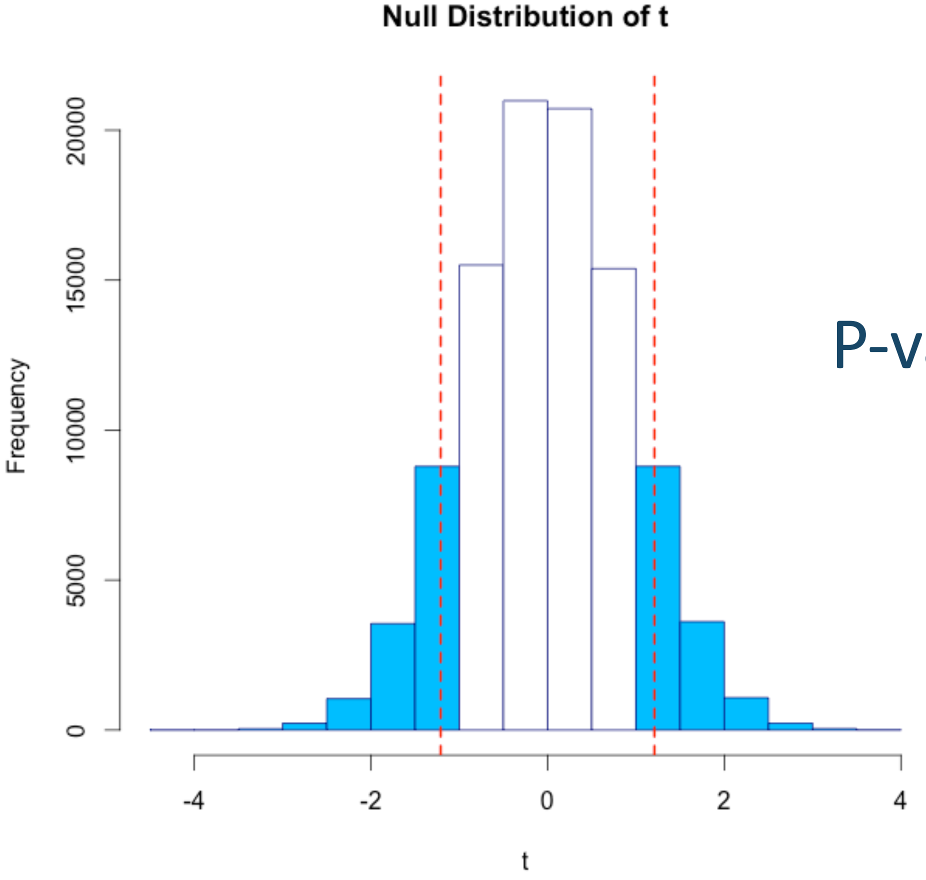


Evaluating Assumptions

Randomization Test Example

Control vs. Free Coins

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$



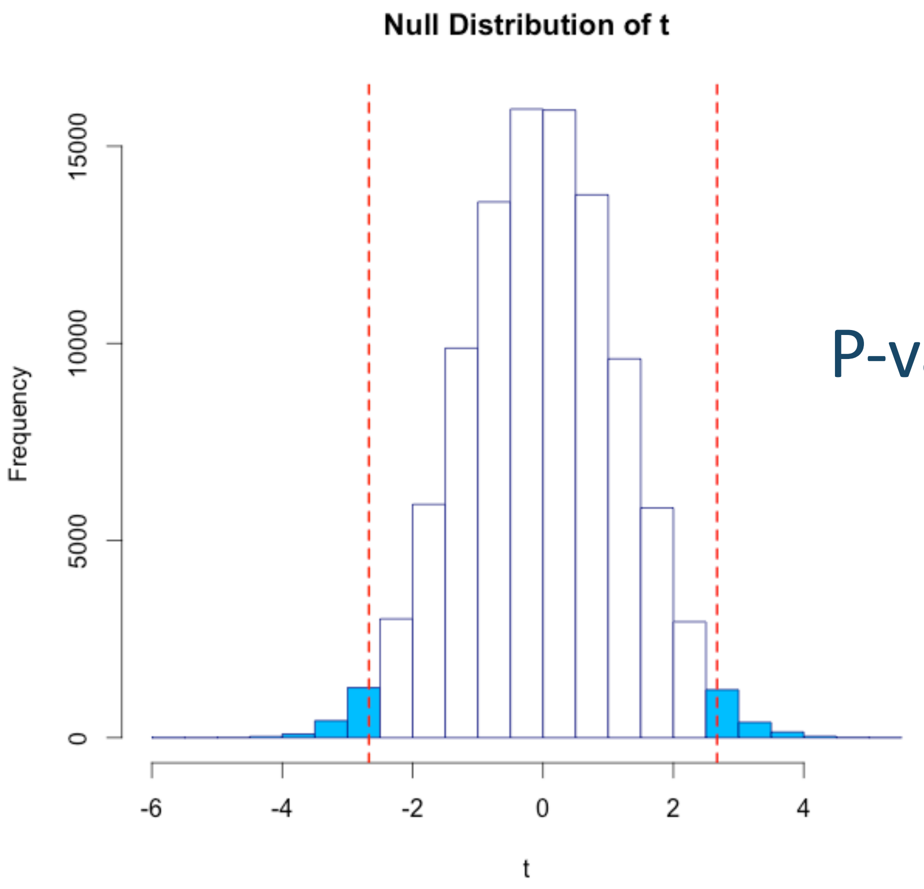
P-value = 0.1882

Evaluating Assumptions

Randomization Test Example

Control vs. Discount

$$H_0: \mu_1 = \mu_3 \text{ vs. } H_A: \mu_1 \neq \mu_3$$



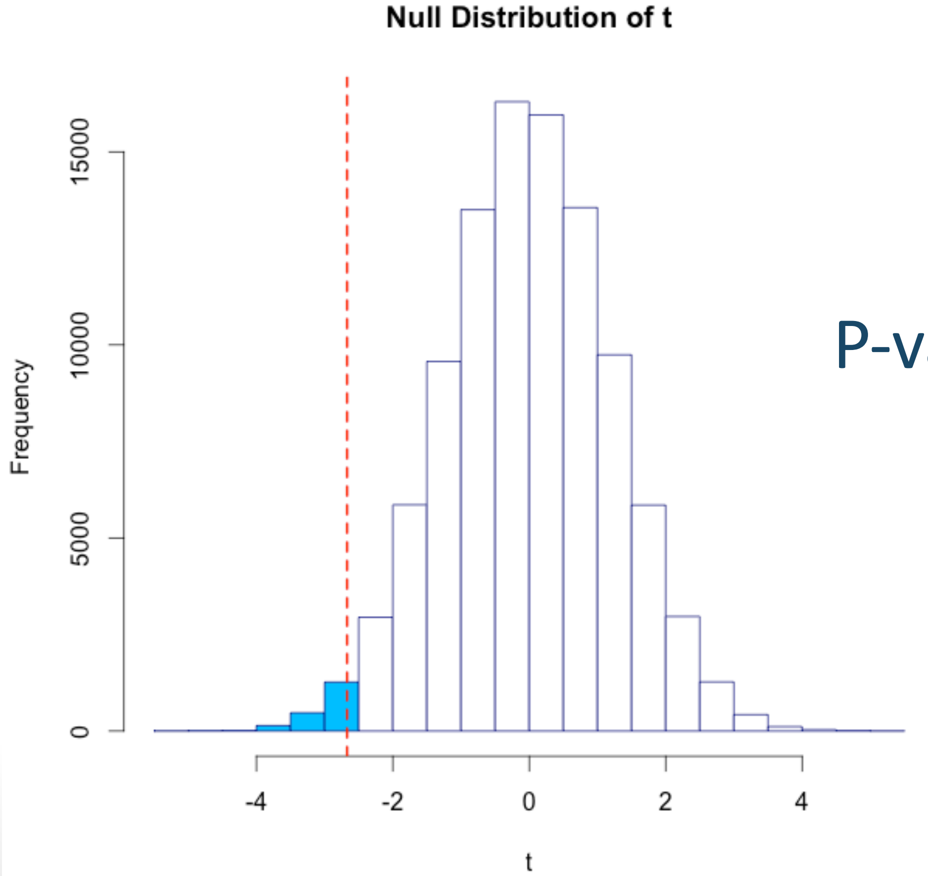
P-value = 0.02515

Evaluating Assumptions

Randomization Test Example

Control vs. Discount

$$H_0: \mu_1 \geq \mu_3 \text{ vs. } H_A: \mu_1 < \mu_3$$



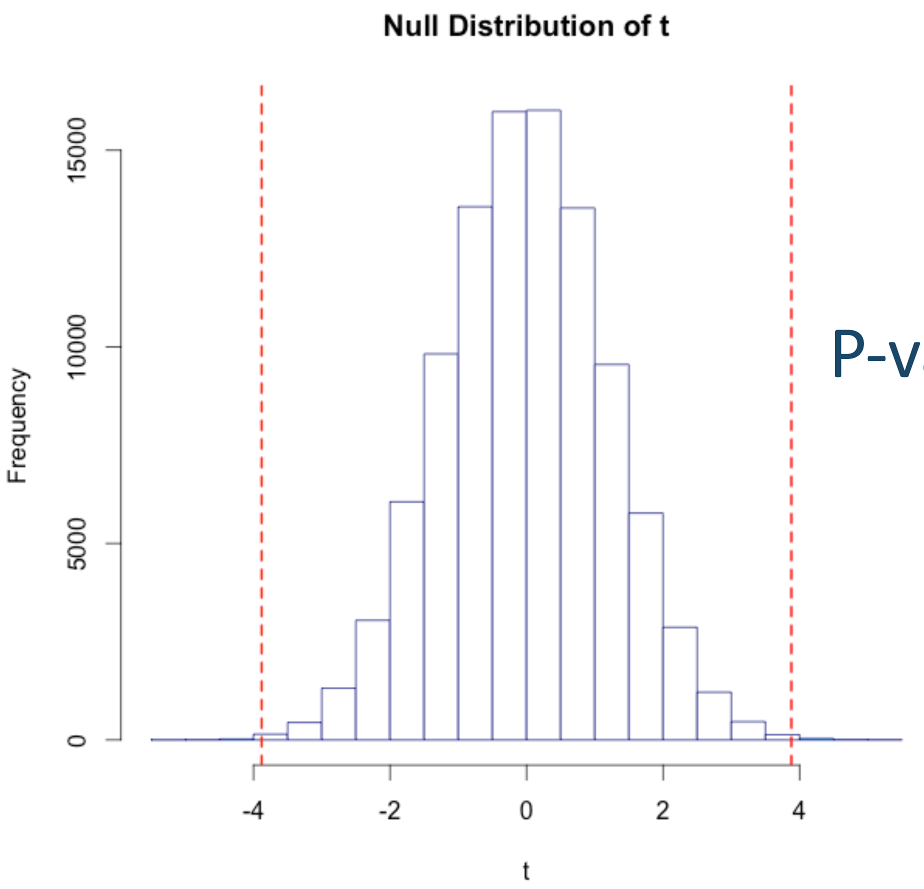
P-value = 0.01349

Evaluating Assumptions

Randomization Test Example

Free Coins vs. Discount

$$H_0: \mu_2 = \mu_3 \text{ vs. } H_A: \mu_2 \neq \mu_3$$



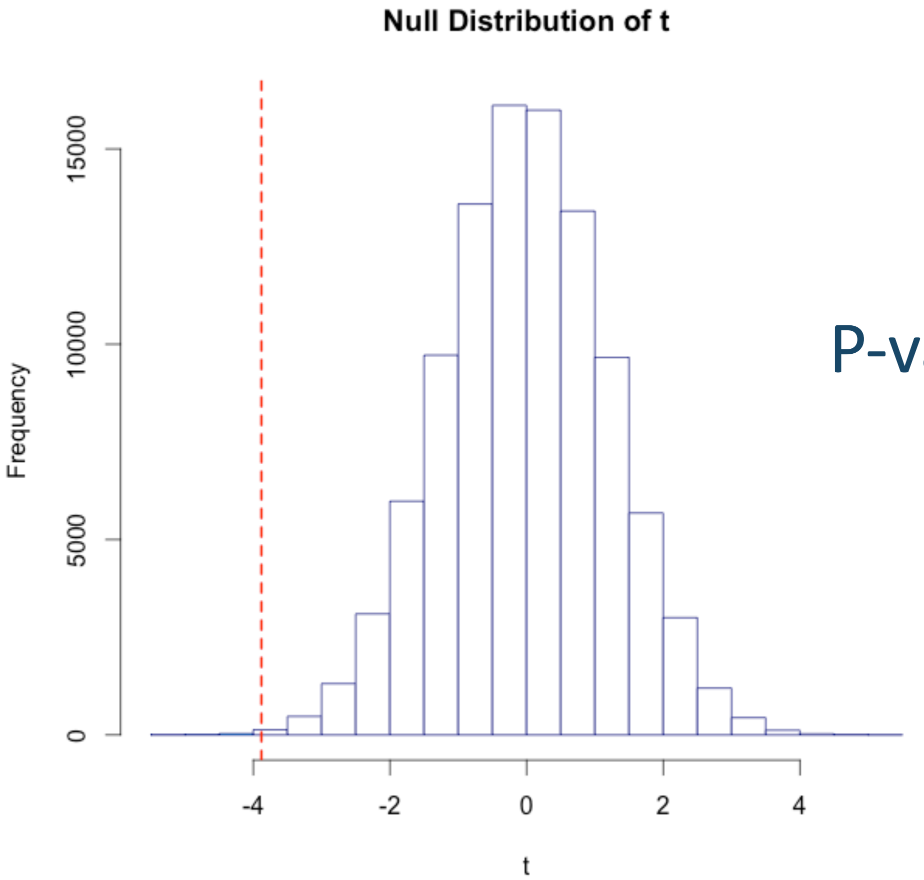
P-value = 0.00112

Evaluating Assumptions

Randomization Test Example

Free Coins vs. Discount

$$H_0\!: \mu_2 \geq \mu_3 \text{ vs. } H_A\!: \mu_2 < \mu_3$$



P-value = 0.00055

Recall

When comparing proportions…

Very often the response variable in an A/B test is binary, indicating whether an experimental unit did, or did not, perform some action of interest

$$Y_{ij} = \begin{cases} 1 \text{ if unit } i \text{ in condition } j \text{ does action} \\ 0 \text{ if unit } i \text{ in condition } j \text{ doesn't do action} \end{cases}$$

for $i = 1,2,\ldots,n_j$, $j = 1,2$

We define $\pi_j = P\left(Y_{ij} = 1\right)$ to be the probability that a unit in condition j performs the action of interest

Recall

When comparing proportions…

The goal of the experiment, then, is to decide whether $\pi_1 = \pi_2$, $\pi_1 > \pi_2$ or $\pi_1 < \pi_2$

We do this formally by testing hypotheses of the form

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_A: \pi_1 \neq \pi_2$$

$$H_0: \pi_1 \leq \pi_2 \text{ vs. } H_A: \pi_1 > \pi_2$$

$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

Evaluating Assumptions

When comparing proportions…

When testing these hypotheses using the *Z*-test we saw last week, we make one key assumption:

- The validity of the method relies on the results of the Central Limit Theorem
- These results, in turn, rely on the assumption that the sample sizes n_1 and n_2 are suitably large
- As a rule of thumb, this method is not valid unless $n_j\pi_j \geq 10$ and $n_j(1-\pi_j) \geq 10$ for $j = 1,2$

In this case we require an alternative approach

Evaluating Assumptions

χ^2-test of Independence

- This test is typically used as a test for 'no association' between two categorical variables
- Here we test the independence of the binary outcome (whether a unit performs the action of interest) and the particular condition they are in
- If the likelihood of performing the action is the same in each condition (i.e., $\pi_1 = \pi_2$) then the response and conditions are not associated
- As such, this test is useful for testing hypotheses regarding $\pi_1 = \pi_2$, $\pi_1 > \pi_2$ or $\pi_1 < \pi_2$

Evaluating Assumptions

χ^2-test of Independence

- The information pertinent to this test can be summarized in a 2x2 contingency table.
- As a concrete example, consider the data from the Optimizely Example last week

		Condition		
		Original	Redesign	
Conversion	Yes	280	399	679
	No	8592	8243	16835
		8872	8642	17514

Evaluating Assumptions

χ^2-test of Independence

We can write this table more generally as

		Condition		
		1	2	
Conversion	Yes	$O_{1,1}$	$O_{1,2}$	O_1
	No	$O_{0,1}$	$O_{0,2}$	O_0
		n_1	n_2	$n_1 + n_2$

where

- $O_{1,j}$ and $O_{0,j}$ respectively represent the observed number of conversions and non-conversions in condition $j = 1,2$, and
- O_1 and O_0 represent the overall number of conversions and non-conversions

Evaluating Assumptions

χ^2-test of Independence

If $\pi_1 = \pi_2 = \pi$ then we would expect the conversion rate in each condition to be the same

Pooled estimates of $\hat{\pi}$ and $1 - \hat{\pi}$ are given by

$$\hat{\pi} = \frac{O_1}{n_1+n_2} \text{ and } 1 - \hat{\pi} = \frac{O_0}{n_1+n_2}$$

With these we can calculate the expected number of observations in each cell of the contingency table:

$$E_{1,j} = n_j\hat{\pi} \text{ and } E_{0,j} = n_j(1 - \hat{\pi})$$

for $j = 1,2$

Evaluating Assumptions

χ^2-test of Independence

The expected frequencies can also be summarized in a contingency table:

		Condition		
		1	2	
Conversion	Yes	$E_{1,1}$	$E_{1,2}$	O_1
	No	$E_{0,1}$	$E_{0,2}$	O_0
		n_1	n_2	$n_1 + n_2$

Note that the margin totals do not change.

The χ^2-test formally compares the what was observed and what is expected under the null hypothesis

Evaluating Assumptions

χ^2-test of Independence

The expected frequencies associated with the Optimizely Example are:

		Condition		
		1	2	
Conversion	Yes	343.96	335.04	679
	No	8524.04	8306.96	16835
		8872	8642	17514

Clearly these don't match what was observed, but we will use the χ^2-test to formally decide whether the observed and expected frequencies are significantly different

Evaluating Assumptions

χ^2-test of Independence

The test statistic that compares the observed count in each cell to the corresponding expected count, and is defined as

$$T = \sum_{l=0}^{1} \sum_{j=1}^{2} \frac{\left(O_{l,j} - E_{l,j}\right)^2}{E_{l,j}}$$

Assuming H_0 is true, T approximately follows a $\chi^2_{(1)}$ distribution

- As a rule of thumb, this approximation may be very poor unless the observed and expected cell frequencies are all greater than 5

Evaluating Assumptions

χ^2-test of Independence

Conclusions about the test are drawn with p-values in according with the following:

$$H_0: \pi_1 = \pi_2 \text{ vs. } H_A: \pi_1 \neq \pi_2$$

- p-value = $P(T \geq t)$

$$H_0: \pi_1 \leq \pi_2 \text{ vs. } H_A: \pi_1 > \pi_2$$

- p-value = $1 - P(T \geq t)/2$

$$H_0: \pi_1 \geq \pi_2 \text{ vs. } H_A: \pi_1 < \pi_2$$

- p-value = $P(T \geq t)/2$

Evaluating Assumptions

χ^2-test of Independence

Returning to the Optimizely Example, the observed test statistic is calculated to be $t = 25.0755$ and so $P(T \geq t) = 5.52\times10^{-7}$

The p-values associated with the three tests are

$$H_0\colon \pi_1 = \pi_2 \text{ vs. } H_A\colon \pi_1 \neq \pi_2$$

- p-value = 5.52×10^{-7}

$$H_0\colon \pi_1 \leq \pi_2 \text{ vs. } H_A\colon \pi_1 > \pi_2$$

- p-value = 0.9999997

$$H_0\colon \pi_1 \geq \pi_2 \text{ vs. } H_A\colon \pi_1 < \pi_2$$

- p-value = 2.76×10^{-7}

The Trouble with Peeking

- The phenomenon whereby you regularly check the results of the experiment before it finishes is known as "peeking"
- This may be tempting, and in some cases impossible to avoid
- Sometimes "peeking" is even a good thing (e.g., to ensure the experiment is not negatively impacting other important metrics)
- The problem, however, arises when, as a result of peeking, you decide to end the experiment early.

The Trouble with Peeking

- Often you might feel pressure to stop the experiment once you see a significant result
- What's the problem? The results tell us that a winner has been found, right?

Wrong

- Well, maybe, but by stopping the experiment early you have not observed enough data to be confident in this conclusion.

The Trouble with Peeking

- Just because the results suggest a winner at one point in time does not mean that the results won't change as more data is collected.

- I might peek at my experiment now and see that condition 1 is significantly out-performing condition 2, but if I peek again in an hour I might find that the opposite is true

- Only until you have observed the pre-specified amount of data should you be sure of your conclusions.

The Trouble with Peeking

- When you stop the experiment you are rejecting the null hypothesis
- Which means you might be making a Type I error
- And by stopping the experiment early the chances you make a Type I error are much higher than the prespecified value of α
- **After all, we did power analyses and sample size calculations for a reason**

The Trouble with Peeking

To illustrate the dire consequences of peeking and ending an experiment early, consider the following simulation.

The set-up:

- $n_1 = n_2 = 1000$ data points are drawn independently from the $N(0,1)$ distribution
- The observations are used to test

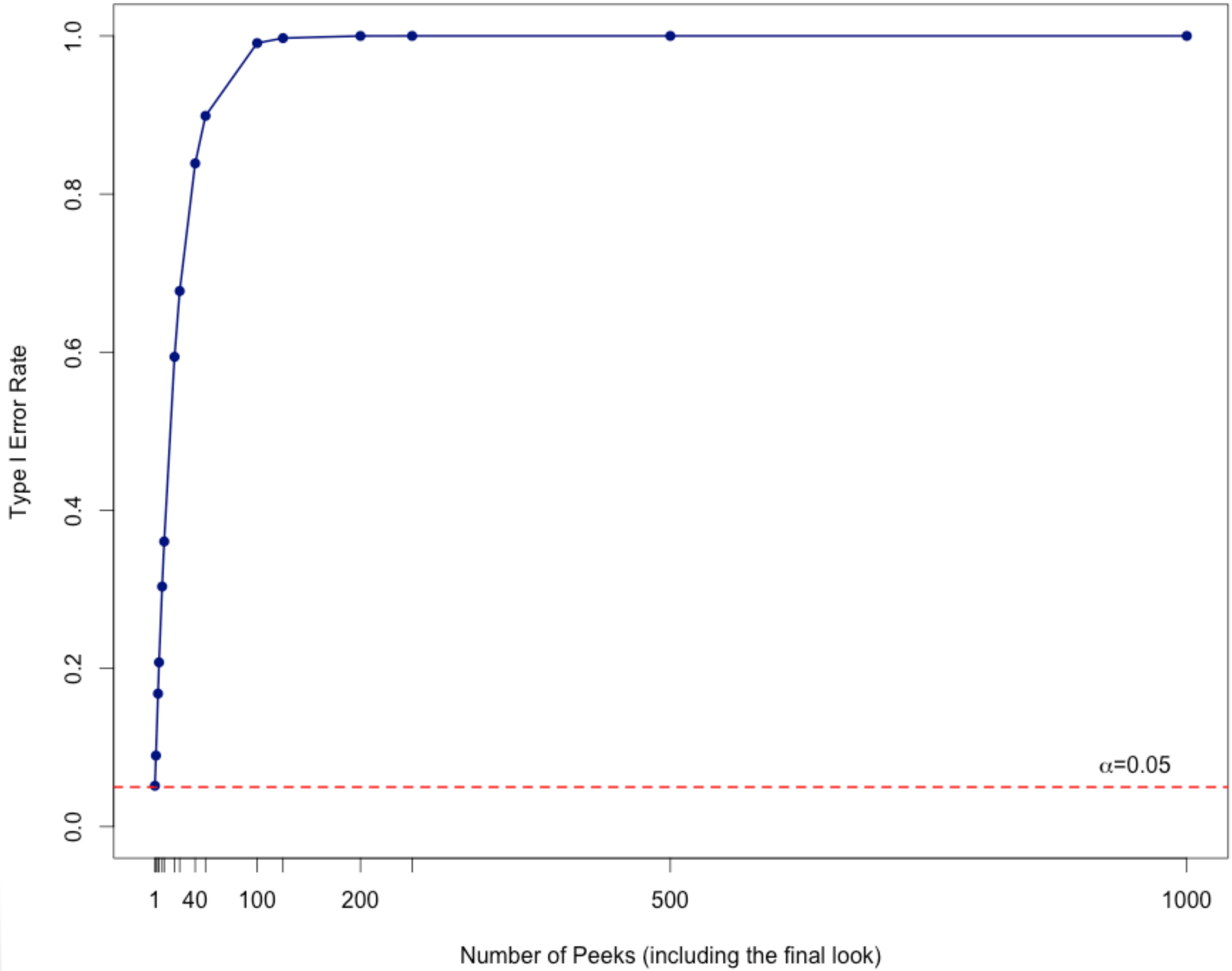
$$H_0: \mu_1 = \mu_2 \text{ vs. } H_A: \mu_1 \neq \mu_2$$

- Because $\mu_1 = \mu_2 = 0$ we should not reject H_0 very often (no more than $\alpha \times 100\%$ of the time)

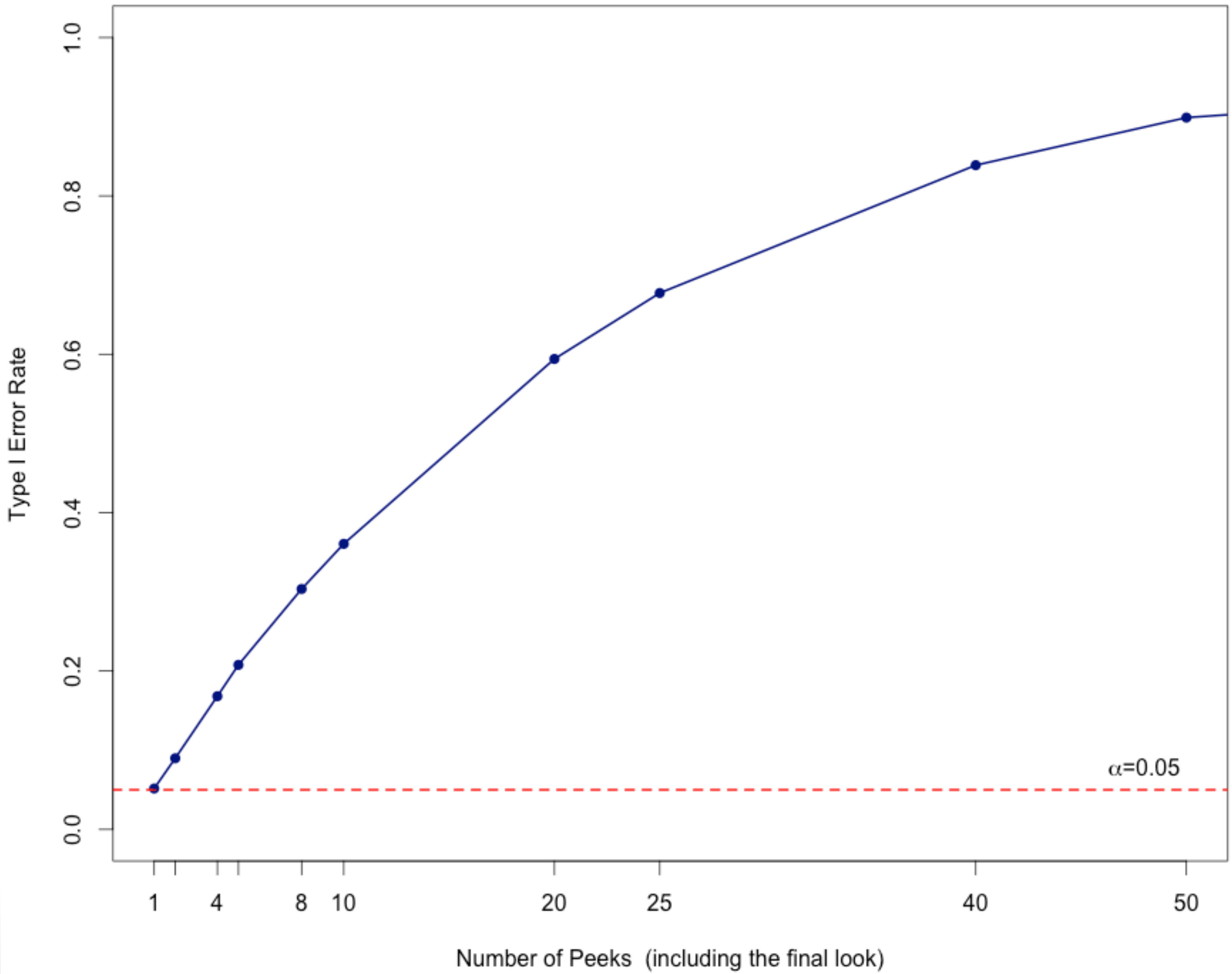
The Trouble with Peeking

- To study the consequences of peeking, we peek – **and end the experiment if a significant result is indicated** – at regular intervals
- We then calculate the Type I Error rate by observing how often an experiment is ended before all 1000 data points are observed in each condition
- We find that by peeking often enough committing a Type I error becomes a certainty

The Trouble with Peeking



The Trouble with Peeking



The Trouble with Peeking

- Note that sequential analysis and sequential testing are statistical disciplines devoted to devising statistically sound methods for performing **repeated significance tests** as more data becomes available.
- Essentially, these techniques that allow you to peek and end an experiment early without increasing Type I error rates.
- However, without adopting one of these techniques, peeking (and ending experiments early) should be avoided at all costs.