Notes and Exercises Chapter 11 – Hypothesis Testing
Math 217
Saidi

Chapter 11

You may agree that there is almost always variability in data – one dataset will not be identical to a second dataset even if they are both collected from the same population using the same methods. However, quantifying the variability in the data is neither obvious nor easy to do, i.e., answering the question "how different is one dataset from another?" is not trivial.

Notation

- p denotes a population proportion
- $\hat{p}$ denotes a sample proportion
- μ denotes a population mean
- $\bar{x}$ denotes a sample mean

Three different approaches for quantifying the variability inherent in data

1. Randomization
2. Bootstrapping
3. mathematical models.

Using the methods provided in this chapter, we will be able to draw conclusions beyond the dataset at hand to research questions about larger populations that the samples come from.

Every dataset has some variability in it, so to decide whether the variability in the data is due to:

1. the causal mechanism (the randomized explanatory variable in the experiment) or instead
2. natural variability inherent to the data

Set up a sham randomized experiment as a comparison. That is, assume that each observational unit would have gotten the exact same response value regardless of the treatment level. By reassigning the treatments many times, we can compare the actual experiment to the sham experiment. If the actual experiment has more extreme results than any of the sham experiments, we are led to believe that it is the explanatory variable which is causing the result and not just variability inherent to the data.

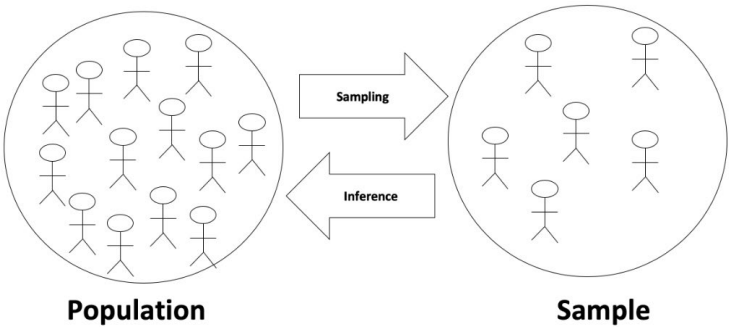
What is statistical inference?

What we are trying to determine is if an effect is statistically significant (signal), rather than an effect due to random chance (noise).

Statistical inference is *the process of making claims about a population based on information from a sample* of data.

Typically, the data represent only a small portion of the larger group which you'd like to summarize. For example, you might be interested in how a drug treats diabetes. Your interest is in how the drug treats all people with diabetes, not just the few dozen people in your study.

At first glance, the logic of statistical inference seems to be backwards, but as you become more familiar with the steps in the process, the logic will make much more sense.



The claim, or **the *null hypothesis***, is denoted **Ho** (pronounced "H-naught") - it is the claim that there is NO DIFFERENCE between two groups.

The claim that corresponds to the research hypothesis, is called ***the alternative hypothesis,*** is denoted by **Ha** (pronounced "H-A") - it is an assertion that there is some difference between the 2 groups.

Almost always, the *goal is to disprove the null hypothesis and claim that the alternative hypothesis is true.*

Example: cheetah speed

Suppose you're conducting research to compare the average running speed of two different subspecies of cheetahs.

The *null hypothesis* is that Asian and African cheetahs run at the same speed, on average.

The *alternative hypothesis* is that African cheetahs are faster than Asian cheetahs, on average.

Randomized distributions

As a way of summarizing each of the null samples, we calculate one statistic from each sample. where each of the sample proportions is denoted "$\hat{\boldsymbol{p}}$**".** The difference in $\hat{p}'s$ changes with each sample.

Understanding the null distribution

We can build a distribution of differences in proportions assuming the null hypothesis, that it is true that there is no link between the proportions in two different groups. That is, the null samples consist of randomly shuffled group variables so that the samples don't have any dependency between the two variables being considered.

Generating a distribution of the statistic from the null population gives information about whether the observed data are inconsistent with the null hypothesis. Always ensure the groups are **independent**.

Example 1: (11.1.1) Consider a study investigating sex discrimination in the 1970s, which is set in the context of personnel decisions within a bank. The research question we hope to answer is, "Are individuals who identify as female discriminated against in promotion decisions made by their managers who identify as male?" (Rosen and Jerdee, 1974) The `sex_discrimination data` can be found in the **openintro** R package.

* This study considered sex roles, and only allowed for options of "male" and "female". We should note that the identities being considered are not gender identities and also that the study allowed only for a binary classification of sex.

decision

sex	not promoted	promoted
female	10	14
male	3	21

	not promoted	promoted
female	0.42	0.58
male	0.12	0.88

We can see that 10/24 (41.7%) women were not promoted and 3/24 (12.5%) men were not promoted. The question is, does this appear to be clear bias, or are the differences in proportions due to random chance?

Figure 11.1: The sex discrimination study can be thought of as 48 red and white cards.

In a randomization distribution, we remove the sex labels and generate a sample of promotion decisions based on the original proportions to explore the difference in decisions when randomized.

We can see that the original (observed) difference in proportions was 41.7% - 12.5% = 29.2%. This observed difference can be thought of as the **point estimate** of the true difference.

The point estimate of the difference in promotion rate is large, but the sample size for the study is small, making it unclear if this observed difference represents discrimination or whether it is simply due to chance when there is no discrimination occurring.

Chance can be thought of as the claim due to natural variability; discrimination can be thought of as the claim the researchers set out to demonstrate. We label these two competing claims, Ho and Ha:

$$H_o: This\ is\ the\ null\ hypothesis - there\ is\ no\ effect$$
$$H_a: This\ is\ the\ alternative\ hypothesis - there\ is\ an\ effect$$

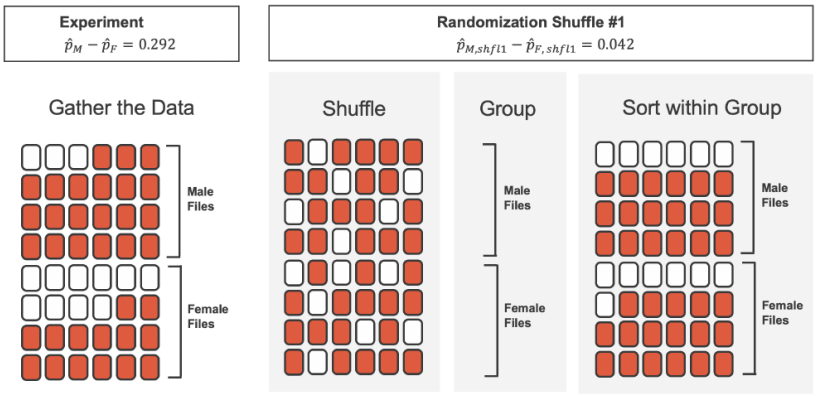
- **Ho:** Null hypothesis. The variables sex and decision are independent. They have no relationship, and the observed difference between the proportion of males and females who were promoted, 29.2%, was due to the natural variability inherent in the population.

- **Ha:** Alternative hypothesis. The variables sex and decision are not independent. The difference in promotion rates of 29.2% was not due to natural variability, and equally qualified female personnel are less likely to be promoted than male personnel.

These hypotheses are part of what is called a **hypothesis test.** A hypothesis test is a statistical technique used to evaluate competing claims using data. Often times, the null hypothesis takes a stance of no difference or no effect. This hypothesis ***assumes that any differences seen are due to the variability inherent in the population and could have occurred by random chance.***

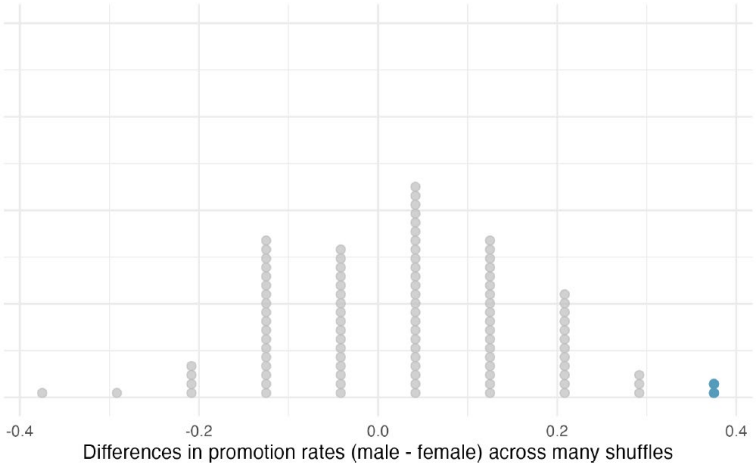
If the null hypothesis and the data notably disagree, then we will reject the null hypothesis in favor of the alternative hypothesis.

11.1.3 Observed statistic vs. null statistics



How often would you observe a difference of at least 29.2% (0.292) according to the figure above? Often, sometimes, rarely, or never?

Randomized data under null model of independence



11.3.2 p-value and statistical significance

• Ho: Sex has no effect on promotion decisions.

• Ha: Female candidates are discriminated against in promotion decisions.

The null hypothesis (Ho) was a perspective of no difference in promotion. The data on sex discrimination provided a point estimate of a 29.2% difference in recommended promotion rates between male and female candidates. We determined that such a difference from chance alone, assuming the null hypothesis was true, would be rare: it would only happen about 2

in 100 times. When results like these are inconsistent with Ho, we reject Ho in favor of Ha. Here, we concluded there was discrimination against female candidates. The 2-in-100 chance is what we call a p-value, which is a probability quantifying the strength of the evidence against the null hypothesis, given the observed data.

Types I and II Errors

1. Type I Error – Occurs if Ho is true and you reject the Ho. *(False positive)*
2. Type II Error – Occurs if Ho is false and you fail to reject the Ho. *(False negative)*

Decision		
Reality about Ho	**Do Not Reject HO**	**Reject Ho**
Ho True	Correct Decision ☺	Type I error
Ho False (Ha True)	Type II error	Correct Decision ☺

Type I error occurs if we reject Ho when it is actually true. (false positive)

Type II error occurs if we do not reject Ho when it is actually false. (false negative)

For example – this may be illustrated in possible errors in a legal trial.

Legal Decision		
Defendant	**Acquit**	**Convict**
Innocent (Ho)	Correct Decision	Type I error
Guilty (Ha)	Type II error	Correct Decision

Load library, set the working directory, and dataset

```
library(tidyverse)
library(knitr)
library(kableExtra)
library(tidymodels)
setwd("C:/Users/rsaidi/Dropbox/Rachel/MontColl/Datasets/Datasets")
sex_disc <- read_csv("sex_discrimination.csv")
set.seed(9753)  # allows to replicate results each time document is rendered
```

This vignette will help you understand HT using tidymodels

https://www.tidymodels.org/learn/statistics/infer/#:~:text=hypothesize()%20allows%20you%20to,to%20form%20the%20null%20distribution.

Two-Way table

```
decision_table  <- table(sex_disc)
kbl(decision_table, caption = "Two-Way Table") |>
  kable_styling() |>
  row_spec(row = 0, color = "dodgerblue")
```

Two-Way Table

	not promoted	promoted
female	10	14
male	3	21

```
kbl(round(proportions(decision_table, "sex"),3)) |> # marginal proportions by sex for pro
motion decision rounded to 3 places
  kable_styling()|>
  row_spec(row = 0, color = "dodgerblue")
```

	not promoted	promoted
female	0.417	0.583
male	0.125	0.875

We can see that 10/24 (41.7%) women were not promoted and 3/24 (12.5%) men were not promoted. The question is, does this appear to be clear bias, or are the differences in proportions due to random chance?

Random permutations

One random permutation

Using the mutate() and sample() functions, the vector of promotion decision is mixed up, or permuted, such that whether someone is male or female can't possibly be causing any difference in proportions. However, due to inherent natural variability, there is also no expectation that the promotion decisions are exactly the same for any sample. We use the sample() function to create the shuffled promotion decisions, and then save that shuffled dataset into a new variable named decision_perm, using the mutate() function.

```
perm1 <- sex_disc |>
  mutate(decision_perm = sample(decision)) |>
  group_by(sex) |>
  summarize(prop_promoted_perm = mean(decision_perm == "not promoted"),
            prop_promoted = mean(decision == "not promoted")) |>
  summarize(diff_perm = diff(prop_promoted_perm),
            diff_orig = diff(prop_promoted))  # not promoted - promoted
kbl(perm1)|>
  kable_styling() |>
  row_spec(row = 0, color = "dodgerblue")
```

diff_perm	diff_orig
0.2083333	-0.2916667

We can see that the original (observed) difference in proportions was 41.7%-12.5%=29.2%

The permutated sample difference in proportions was different. Notice it will change each time you run this chunk due to random chance.

Many random permutations

By repeating the permutation and difference calculations five times, the permuted differences are seen to be sometimes positive, sometimes negative, sometimes close to zero, sometimes far from zero. However, five times isn't quite enough to capture all of the variability in the null differences.

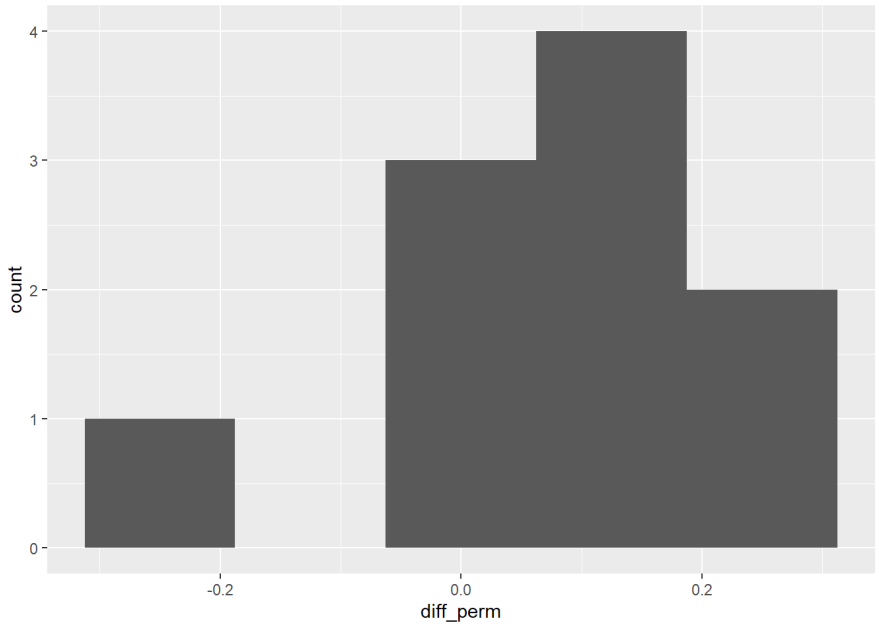
The rep_sample_n() function performs repeated sampling of a dataset, where the size of the sample is specified in the first argument and the number of repetitions is specified in the reps argument. We notice that the sample size is the same size as the sex discrimination dataset, since it equals the number of rows in the original dataset. We are also specifying that the sampling of the datset should be done without replacement (replace = FALSE), since we want each row to only be selected once. You can think of this as creating five copies of the original sex discrimination dataset.

```
perm_reps <- sex_disc |>
  rep_sample_n(size = nrow(sex_disc), reps = 10, replace = FALSE) |>
  mutate(decision_perm = sample(decision)) |>
  group_by(replicate, sex) |>
  summarize(prop_promoted_perm = mean(decision_perm == "not promoted"),
            prop_promoted = mean(decision == "not promoted")) |>
  summarize(diff_perm = diff(prop_promoted_perm),
            diff_orig = diff(prop_promoted))  # not promoted - promoted
`summarise()` has grouped output by 'replicate'. You can override using the
`.groups` argument.
kbl(perm_reps) |>
  kable_styling() |>
    row_spec(row = 0, color = "dodgerblue")
```

replicate	diff_perm	diff_orig
1	-0.0416667	-0.2916667
2	0.1250000	-0.2916667
3	0.2083333	-0.2916667
4	-0.0416667	-0.2916667
5	0.1250000	-0.2916667
6	0.2083333	-0.2916667
7	0.1250000	-0.2916667
8	-0.2916667	-0.2916667
9	0.1250000	-0.2916667
10	0.0416667	-0.2916667

Notice the variation in the permutation differences.

```
ggplot(perm_reps, aes(diff_perm))+
  geom_histogram(bins = 5)
```



Repeat the permutation 1000 times and plot all differences

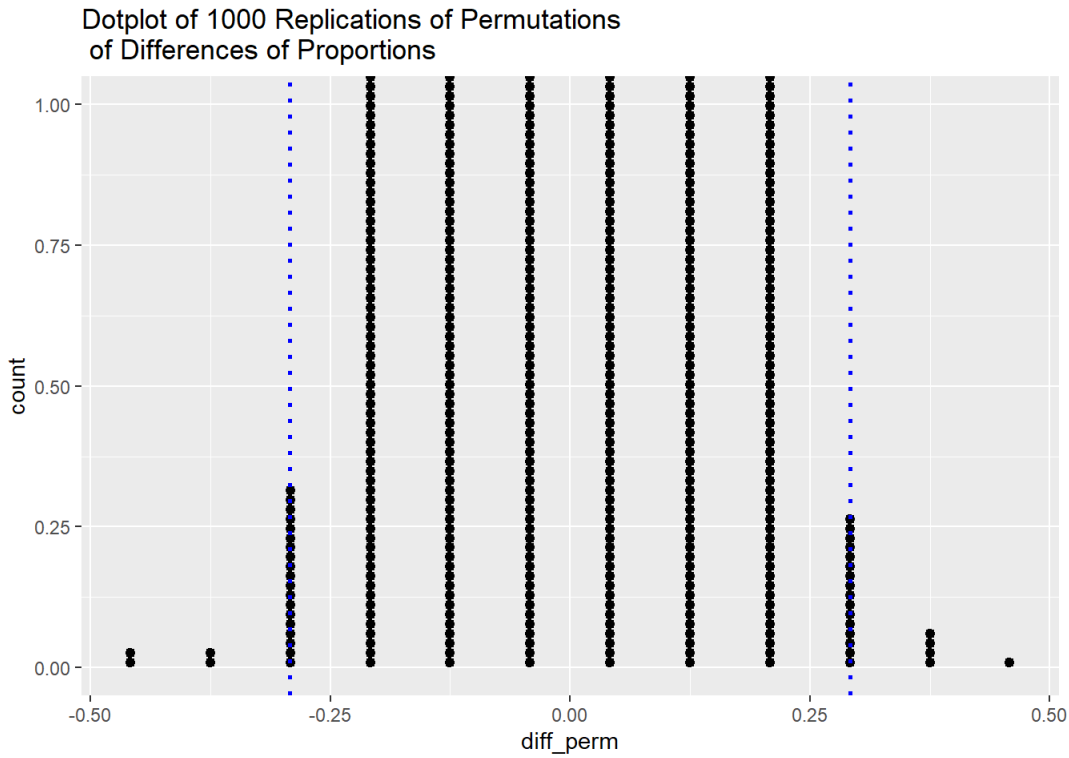
```
perm_1000_reps <- sex_disc |>
  rep_sample_n(size = nrow(sex_disc), reps = 1000, replace = FALSE) |>
  mutate(decision_perm = sample(decision)) |>
  group_by(replicate, sex) |>
  summarize(prop_promoted_perm = mean(decision_perm == "not promoted"),
            prop_promoted = mean(decision == "not promoted")) |>
  summarize(diff_perm = diff(prop_promoted_perm),
            diff_orig = diff(prop_promoted))  # not promoted - promoted
```

```
`summarise()` has grouped output by 'replicate'. You can override using the
`.groups` argument.
```

Create a dotplot of the replications of permutated differences

Show the cut-off for +/- the original difference in proportions of “promoted”

```
plot_perm <- ggplot(perm_1000_reps, aes(diff_perm)) +
  geom_dotplot(binwidth = 0.01) +
  geom_vline(xintercept = c(-0.292, 0.292), color = "blue", linetype = "dotted", linewidth=1) +
  labs(title = "Dotplot of 1000 Replications of Permutations\n of Differences of Proportions")
plot_perm
```



Remember that each dot represents a different permutation of the differences in proportions of promotions between males and females.

Although the plot will appear differently each time due to randomization, notice that very few points are > 0.292 or < -0.292. Visually, this shows that it is less likely that the difference of 29.2% is due to random chance, and more likely that discrimination of promotions was occurring based on sex.

Use the infer framework

Randomized data under null model of independence

- step through specifying the null model and then performing 1000 permutations to evaluate whether decision status differs between the "female" and "male" groups
- specify() that the relationship of interest is decision vs. sex and a success in this context is promotion, set success to "promoted.
- hypothesize() is used with null = "independence" for comparing difference of proportions, null = "point", mu = VALUE for comparing difference of means.

```
# Hypothesize independence (this test is used to compare difference of proportions)
decide_perm <- sex_disc |>
  specify(decision ~ sex, success = "promoted") |>  # syntax is y ~ x
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "diff in props", order = c("female", "male"))
decide_perm
Response: decision (factor)
Explanatory: sex (factor)
Null Hypothesis: independence
# A tibble: 1,000 × 2
   replicate    stat
       <int>   <dbl>
 1         1  0.0417
 2         2  0.0417
```

```
 3         3 -0.125
 4         4  0.125
 5         5  0.292
 6         6 -0.0417
 7         7 -0.125
 8         8  0.0417
 9         9  0.208
10        10 -0.125
# ⓘ 990 more rows
```

Calculate a p-value for the above hypothesis test for two-tail test

```
p_value <- perm_1000_reps |>
  summarize(count = sum(abs(diff_orig) <= abs(diff_perm)),
            proportion = mean(abs(diff_orig) <= abs(diff_perm)))
p_value
# A tibble: 1 × 2
  count proportion
  <int>      <dbl>
1    44      0.044
```

If we wanted only a one-tail test, the resulting p-value is 1/2 the result above

```
p_value <- perm_1000_reps |>
  summarize(count = sum(diff_orig >= diff_perm),
            proportion = mean(diff_orig >= diff_perm))
p_value
# A tibble: 1 × 2
  count proportion
  <int>      <dbl>
1    23      0.023
```

Conclusion based on p-value

This p-value suggests that such a difference from chance alone, assuming the null hypothesis was true, would be rare: it would only happen about 44 in 1000 times. When results like these are inconsistent with Ho, we reject Ho in favor of Ha. Here, we concluded there was discrimination against female candidates. The 44-in-1000 chance is what we call a p-value, which is a probability quantifying the strength of the evidence against the null hypothesis, given the observed data.

Homework Chapter 11

1. Review section 11.5 (the chapter review)
2. **Suggested**: from textbook section 11.6 exercises: 2,3,4,6,7
3. **Suggested**: Unit 4 Tutorial on Foundations of Inference:

    1 - Sampling variability

    2 - Randomization test

3 - Errors in hypothesis testing