Statistics for Data Science
Winter Institute in Data Science

Ryan T. Moore

4 January 2020

Warm Up Quiz

Descriptive Statistics

Descriptive statistics: summarize observed features of data

- *Univariate* statistics: describe single variable
- *Bivariate* statistics describe relationship between two variables
  ("are higher values of X assoc'ed w/ higher values of Y?")
- *Multivariate* statistics summarize several relationships at once
  ("are higher values of X associated with higher values of Y, specifically when $Z = 1$?")

Suppose we measure the number of times each of 12 voters voted in the last 5 presidential elections:

```
times_voted <- c(3, 4, 1, 2, 2, 3, 5, 2, 2, 1, 3, 3)
sort(times_voted)
```

```
##  [1] 1 1 2 2 2 2 3 3 3 3 4 5
```

Summary Statistics with R

```
max(times_voted)
```

```
## [1] 5
```

```
min(times_voted)
```

```
## [1] 1
```

```
range(times_voted)
```

```
## [1] 1 5
```

```
mean(times_voted)
```

```
## [1] 2.583333
```

```r
median(times_voted)
```

```
## [1] 2.5
```

```r
quantile(times_voted, probs = 0.5)
```

```
## 50%
## 2.5
```

```r
median(times_voted)
```

```
## [1] 2.5
```

```r
quantile(times_voted, probs = 0.5)
```

```
## 50%
## 2.5
```

```r
quantile(times_voted, probs = c(1/3, 2/3))
```

```
## 33.33333% 66.66667%
##         2         3
```

```r
quantile(times_voted, probs = c(1/4, 3/4))
```

```
## 25% 75%
##   2   3
```

```r
median(times_voted)
```

```
## [1] 2.5
```

```r
quantile(times_voted, probs = 0.5)
```

```
## 50%
## 2.5
```

```r
quantile(times_voted, probs = c(1/3, 2/3))
```

```
## 33.33333% 66.66667%
##         2         3
```

```r
quantile(times_voted, probs = c(1/4, 3/4))
```

```
## 25% 75%
##   2   3
```

```r
IQR(times_voted)
```

```
## [1] 1
```

Summary Statistics with R

```
summary(times_voted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.500   2.583   3.000   5.000
```

The Root-Mean-Square (RMS)

RMS describes average *magnitude* of variable's values.

The Root-Mean-Square (RMS)

RMS describes average *magnitude* of variable's values.

The RMS takes each value,

1. squares it,
2. takes the mean of these squares, and then
3. takes the square root.

Why take the square, then square root?

The Root-Mean-Square (RMS)

RMS describes average *magnitude* of variable's values.

The RMS takes each value,

1. squares it,
2. takes the mean of these squares, and then
3. takes the square root.

Why take the square, then square root? Why not more intuitive?

Calculate the RMS of `times_voted` "by hand":

```
tv_squared <- times_voted ^ 2
```

```
##  [1]  9 16  1  4  4  9 25  4  4  1  9  9
```

```
mean_tvs <- mean(tv_squared)
```

```
## [1] 7.916667
```

```
root_mean_tvs <- sqrt(mean_tvs)
```

```
## [1] 2.813657
```

Standard Deviation (SD)

SD describes the spread of a variable.

SD: the RMS of the deviations from the average.

Standard Deviation (SD)

SD describes the spread of a variable.

SD: the RMS of the deviations from the average.

Tells us "How far from the average is a typical value of the variable?"

Calculate: take each observation's difference from the average, then take the RMS of those differences.

The $\overline{\text{overline}}$ means "take the mean". For variable x, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

To calculate the SD, for each observation x_i,

1. find $x_i - \bar{x}$,
2. square it $(x_i - \bar{x})^2$
3. take the mean of these squares, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
3. take the square root $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

So,

$$SD(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

To calculate the mean, and the "typical" deviation from the mean:

```
mean(times_voted)
```

```
## [1] 2.583333
```

```
sd(times_voted)
```

```
## [1] 1.1645
```

Variance (SD^2)

Variance: SD squared.

Variance: average of the squared deviations from the mean.

Mathematically easier to work with the variance than the SD, since the variance doesn't have $\sqrt{}$.

Variance (SD^2)

Variance: SD squared.

Variance: average of the squared deviations from the mean.

Mathematically easier to work with the variance than the SD, since the variance doesn't have $\sqrt{\ }$.

$$\begin{aligned} Var(x) &= \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

To calculate it in R,

```
var(times_voted)
```

```
## [1] 1.356061
```

```
(sd(times_voted)) ^ 2
```

```
## [1] 1.356061
```

To calculate it in R,

```
var(times_voted)
```

```
## [1] 1.356061
```

```
(sd(times_voted)) ^ 2
```

```
## [1] 1.356061
```

```
# Not sd(times_voted ^ 2) !
sd(times_voted ^ 2)
```

```
## [1] 6.868351
```

The z-score

For variable X, the z-score of observation x_i tells how far it is from average, in units of the standard deviation.

$$z_i = \frac{x_i - \bar{x}}{SD(x)}$$

The z-score

For variable X, the z-score of observation x_i tells how far it is from average, in units of the standard deviation.

$$z_i = \frac{x_i - \bar{x}}{SD(x)}$$

If y_i is any linear transformation of x_i such that $y_i = ax_i + b$, then the

$$(z\text{-score of } x_i) = (z\text{-score of } y_i)$$

Interpretation: z-score does not depend on units we measure in (as long as linear transformation).

The z-scores for a set of household incomes are the same whether measure in \$, \$1000, CAD, etc.

The z-score can compare variables on different scales, since we divide each variables' values by its own SD. If X is household income and Y is a survey respondent's left-right ideology on a $[0, 10]$ scale, then we might have

The z-score can compare variables on different scales, since we divide each variables' values by its own SD. If X is household income and Y is a survey respondent's left-right ideology on a $[0, 10]$ scale, then we might have

Respondent	Income	Ideology	$z_{i,\text{Inc}}$	$z_{i,\text{Ideol}}$
1	65000	8		
2	20000	3		
Mean (overall)	50000	6		
SD (overall)	15000	2		

The z-score can compare variables on different scales, since we divide each variables' values by its own SD. If X is household income and Y is a survey respondent's left-right ideology on a $[0, 10]$ scale, then we might have

Respondent	Income	Ideology	$z_{i,\text{Inc}}$	$z_{i,\text{Ideol}}$
1	65000	8		
2	20000	3		
Mean (overall)	50000	6		
SD (overall)	15000	2		

The process of calculating the z-scores is called *standardizing* the variable.

Correlation

Are larger values of X associated with larger (or smaller?) values of Y?

This is the question of the *correlation* between X and Y. When X and Y are positively correlated, that means larger values of X are associated with larger values of Y.

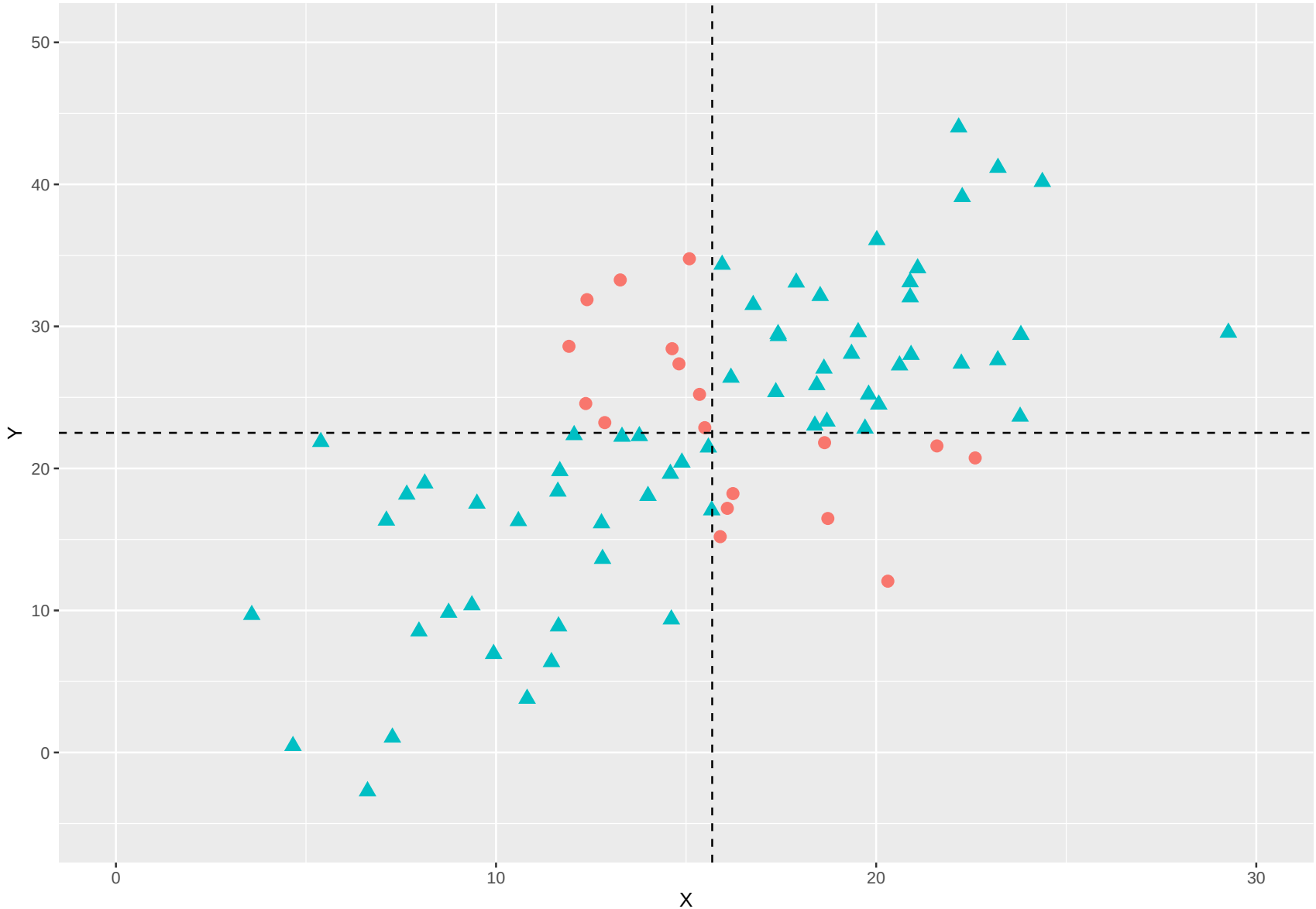


Figure 1: Positive Correlation. Blue triangles outweigh red discs.

On the other hand, when X and Y are negatively correlated, that means larger values of X tend to be associated with *smaller* values of Y.

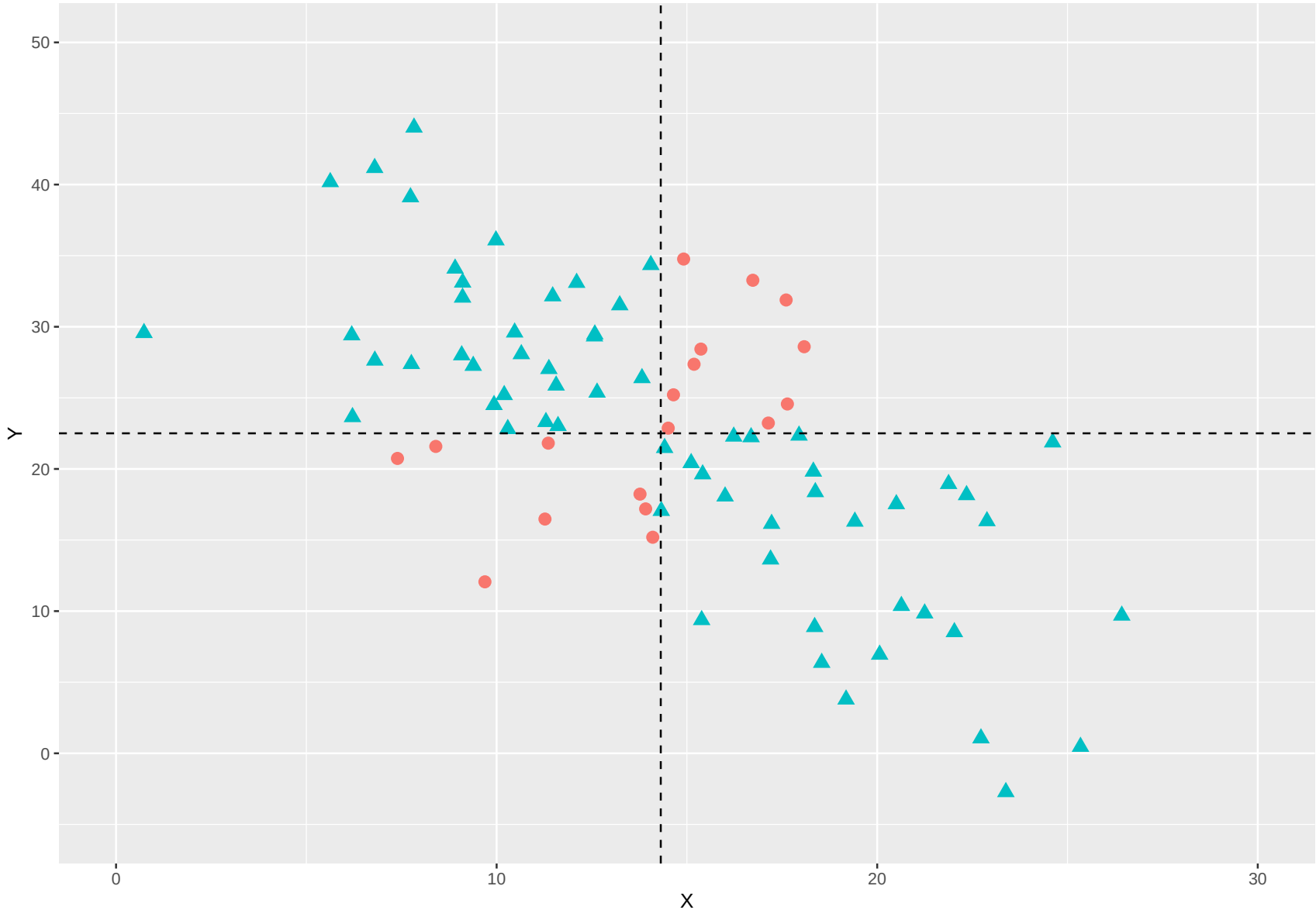


Figure 2: Negative Correlation. Blue triangles outweigh red discs.

Formally, correlation is average of products of z-scores.

Formally, correlation is average of products of z-scores.

That is, a positive correlation is when, on average, $z_i(x_i) \times z_i(y_i) > 0$, which is only true if both scores are positive or both scores are negative. Whether z_i is positive or negative depends on whether unit i is above or below the mean on that variable. Its magnitude is determined by *how far* above or below the average unit i is.

Formally, correlation is average of products of z-scores.

That is, a positive correlation is when, on average, $z_i(x_i) \times z_i(y_i) > 0$, which is only true if both scores are positive or both scores are negative. Whether z_i is positive or negative depends on whether unit i is above or below the mean on that variable. Its magnitude is determined by *how far* above or below the average unit i is.

$$cor(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{SD(x)} \cdot \frac{y_i - \bar{y}}{SD(y)} \right)$$

The correlation always lies in the interval $[-1, 1]$.

```
x <- sample(20)
y <- sample(20)
cor(x, y)
```

```
## [1] -0.2766917
```

QQ Plots

Quantiles describe ranks in a distribution. Percentiles, quartiles, terciles, and the median are all examples of quantiles. Just as the z-score takes a measurement x_i, recenters it, and rescales it, finding a measurement's quantile gives us information about its relative position in the distribution of X – in fact, it tells us x_i's *rank* in the distribution. If a legislator has an ideology score $= 1$, which is the 2^{nd} tercile, we know she ranks above $\frac{2}{3}$ of the legislators in her score.

A QQ plot visually compares the quantiles of two distributions. It gives us information like "Is the median of X greater than or less than the median of Y?" and "Is the first quartile of X greater than or less than the first quartile of Y?" Suppose we have spending in recent House races for some Democratic and Republican candidates, in millions of dollars:

```
rep <- c(0.5, 4, 3, 2.2, 2.2, 2)
dem <- c(1, 2, 2, 1.75, 1.5, 3, 2, 5, 2.1)
```

and we calculate the medians:

```
median(rep)
```

```
## [1] 2.2
```

```
median(dem)
```

```
## [1] 2
```

Republican median a bit higher, representing about 0.2 million dollars more spending.

What happens at other points in the distribution? At the lowest end? At the highest end?

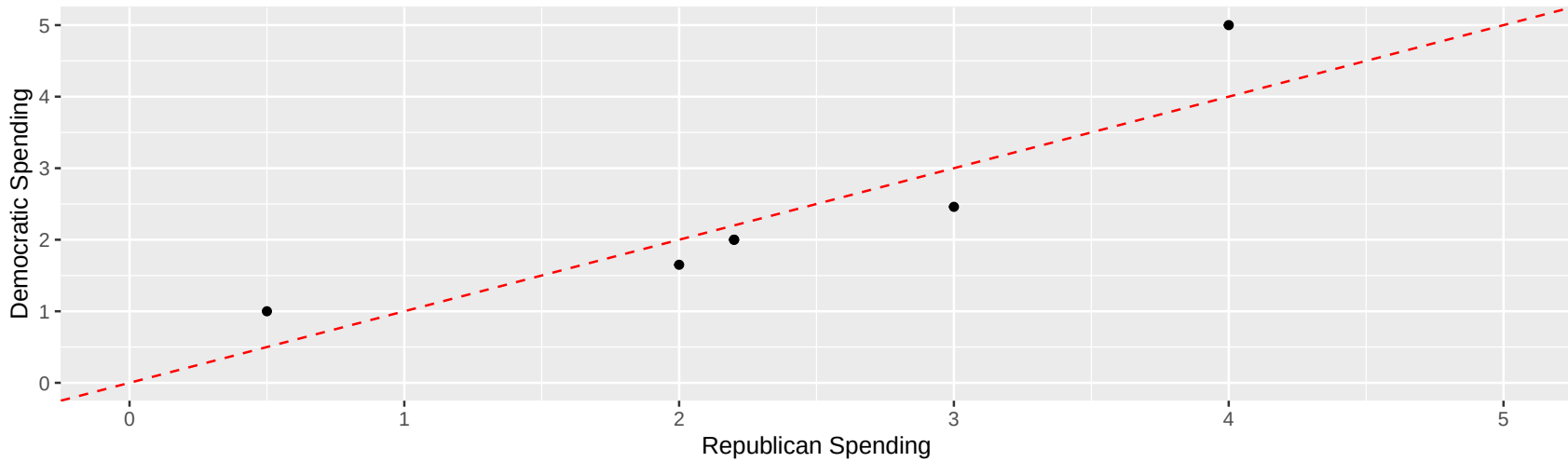


Figure 3: QQ Plot of Hypothetical Republican and Democratic Spending. If points lie on red dotted $y = x$ line, that quantile of the distributions is equal.

Probability

Foundations

1. **Experiment**: a process that yields a probabilistic/stochastic outcome. Not nec. entirely random, but w/ some random component.

Foundations

1. **Experiment**: a process that yields a probabilistic/stochastic outcome. Not nec. entirely random, but w/ some random component.
2. **Outcome space/Sample space**: The set of all possible outcomes of an experiment. Usually Ω = "Omega"

Foundations

1. **Experiment**: a process that yields a probabilistic/stochastic outcome. Not nec. entirely random, but w/ some random component.
2. **Outcome space/Sample space**: The set of all possible outcomes of an experiment. Usually Ω = "Omega"
3. **Event**: a subset of Ω. Usually denoted A, B, etc. The probability of A happening is $P(A)$.

Foundations

1. **Experiment**: a process that yields a probabilistic/stochastic outcome. Not nec. entirely random, but w/ some random component.
2. **Outcome space/Sample space**: The set of all possible outcomes of an experiment. Usually Ω = "Omega"
3. **Event**: a subset of Ω. Usually denoted A, B, etc. The probability of A happening is $P(A)$.
4. **Complement**: the logical negation of an event. For event A, the complement is "not-A happens". Denoted A^C. $P(A^C) = 1 - P(A)$.

Foundations

1. **Experiment**: a process that yields a probabilistic/stochastic outcome. Not nec. entirely random, but w/ some random component.
2. **Outcome space/Sample space**: The set of all possible outcomes of an experiment. Usually Ω = "Omega"
3. **Event**: a subset of Ω. Usually denoted A, B, etc. The probability of A happening is $P(A)$.
4. **Complement**: the logical negation of an event. For event A, the complement is "not-A happens". Denoted A^C. $P(A^C) = 1 - P(A)$.

Assuming all outcomes are equally likely (often false), then

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega}$$

Examples

1. **Experiment**: a voter will vote Dem, vote Rep, vote other, or abstain.
2. **Outcome space/Sample space**: $\Omega = \{\text{Dem}, \text{Rep}, \text{other}, \text{abstain}\}$
3. **Event**:
   - $A =$ abstains. Assuming all equally likely, what is $P(A)$?
   - $B =$ supports a major party candidate. Assuming all equally likely, what is $P(B)$?
4. **Complement**:
   - What is does A^C mean? What is $P(A^C)$?
   - What is does B^C mean? What is $P(B^C)$?

The 3 Axioms

Definitions

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. If events *mutually exclusive* (or, sets *disjoint*), then

$$P(A \text{ or } B) = P(A) + P(B)$$

The 3 Axioms

Examples

1. $P(A)$, the probability of abstaining, cannot be negative.
2. One of $\{\text{Dem}, \text{Rep}, \text{other}, \text{abstain}\}$ must occur.
3. $P(\text{abstains or supports a major candidate}) =$
   $P(\text{abstains}) + P(\text{supports major candidate})$

Probability of Either of 2 Events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Probability of Either of 2 Events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Let $A =$ votes Dem. Then,

$$\begin{aligned} P(\text{votes Dem or major}) &= P(\text{Dem}) + P(\text{major}) - \\ &\quad P(\text{Dem and major}) \\ &= \frac{1}{4} + \frac{1}{2} - \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

Law of Total Probability

We can decompose prob of A into two components: A happening when B also happens, and A happening when "not B" happens:

Law of Total Probability

We can decompose prob of A into two components: A happening when B also happens, and A happening when "not B" happens:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^C)$$

Law of Total Probability

We can decompose prob of A into two components: A happening when B also happens, and A happening when "not B" happens:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^C)$$

Let A = votes for major party candidate. Let B = votes Dem.

Law of Total Probability

We can decompose prob of A into two components: A happening when B also happens, and A happening when "not B" happens:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } B^C)$$

Let $A =$ votes for major party candidate. Let $B =$ votes Dem.

$$\begin{aligned} P(\text{major candidate}) &= P(\text{major and Dem}) \\ &\quad + P(\text{major and not-Dem}) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

Extend this law by splitting A into more components, as long as (like B and B^C) the events form *partition*:

(a) are mutually exclusive, and
(b) cover the entire sample space.

Extend this law by splitting A into more components, as long as (like B and B^C) the events form *partition*:

(a) are mutually exclusive, and
(b) cover the entire sample space.

For such events $B_1, B_2, \ldots, B_N$,

$$P(A) = \sum_{i=1}^{N} P(A \text{ and } B_i)$$

Permutations: Counting orderings

How many ways to **order** k things from a set of n things?

$$ {}_nP_k = \frac{n!}{(n-k)!} $$

Permutations: Counting orderings

How many ways to **order** k things from a set of n things?

$$ {}_nP_k = \frac{n!}{(n-k)!} $$

Suppose you have 5 political news segments to order for the evening broadcast (A, B, C, D, E).

- How many ways to include all 5?
- How many ways to include only 2? (where order matters – AB is a different broadcast than BA)

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

Suppose you select some Senators to be on a committee, from a set of 10 candidates (A, B, ..., J).

- How many ways to include 3 Senators? (where order does **not** matter – ABC is same committee as CBA)
- How many ways to include 9 Senators?
- How many ways to include 2 or fewer Senators?

Conditional Probability

The probability that A will happen, given B has happened:

$$P(A|B)$$

Conditional Probability: Example

In the `FLVoters.csv` data (http://j.mp/2ZOMEeu),
let A = voter is female, B = voter is black.

What is the probability that voter is female, given that we know the voter is black? $P(\text{female}|\text{black})$

Calculation: A perspective on the formula

Suppose we are interested in the **joint** probability that both A and B occur. We can think about this as occurring in two different ways: one has to be assumed, then the other. Either

- A occurs, then B occurs given that we know A already occurred, or
- B occurs, then A occurs given that we know B already occurred.

Calculation: A perspective on the formula

Suppose we are interested in the **joint** probability that both A and B occur. We can think about this as occurring in two different ways: one has to be assumed, then the other. Either

- A occurs, then B occurs given that we know A already occurred, or
- B occurs, then A occurs given that we know B already occurred.

$$\begin{aligned} P(A \text{ and } B) &= P(A)P(B|A) \\ &= P(B)P(A|B) \end{aligned}$$

Simply dividing both sides by the **marginal** probability $P(B)$ yields

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Simply dividing both sides by the **marginal** probability $P(B)$ yields

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

If we want the probability that voter is female, given that we know the voter is black

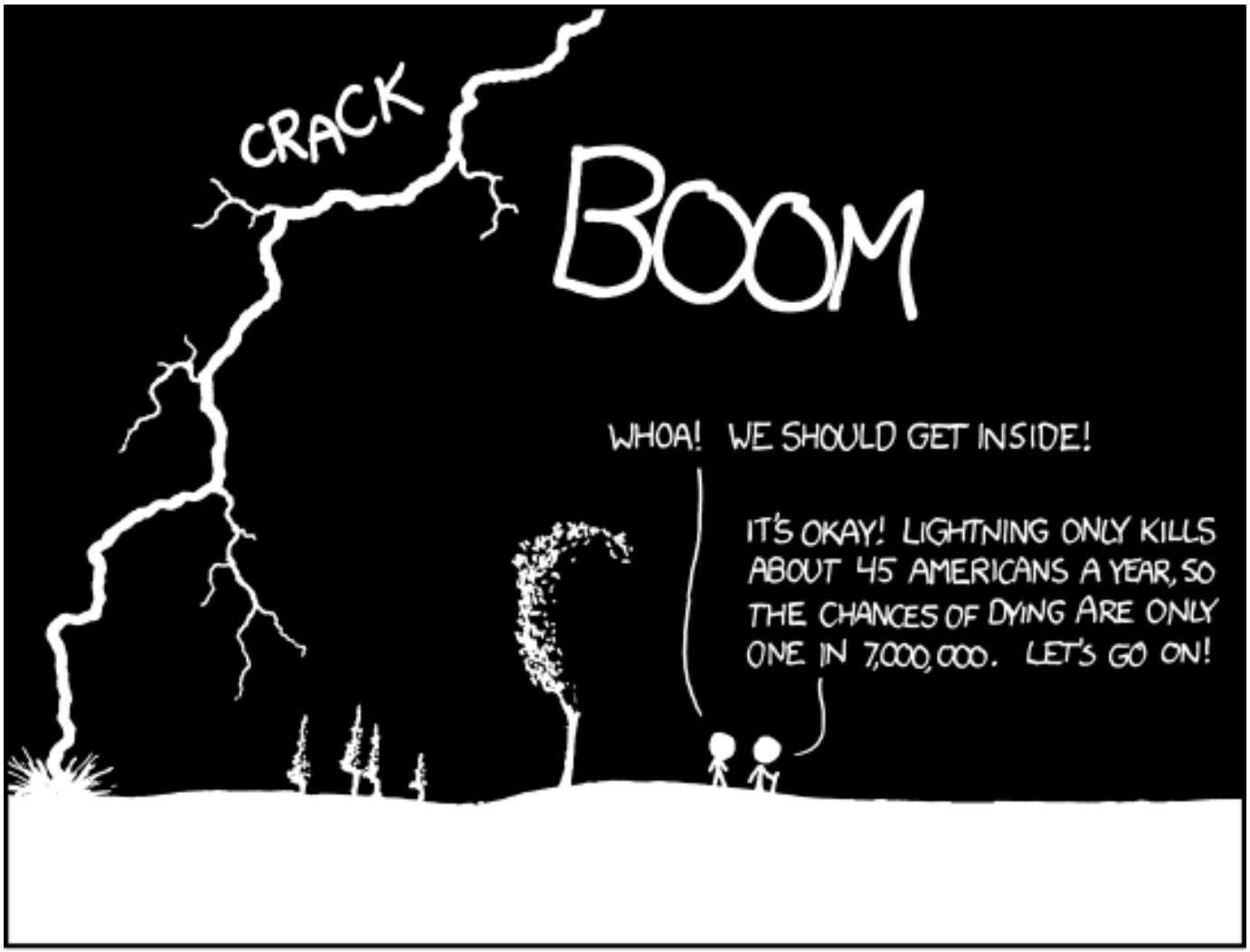
$$P(\text{female}|\text{black}) = \frac{P(\text{female and black})}{P(\text{black})}$$

Conditioning Information can be Subtle

Example 1: Is Lightning Dangerous?

Conditioning Information can be Subtle

Example 1: Is Lightning Dangerous?



Conditioning Information can be Subtle

Example 2: The Older Child Paradox

- Consider picking a two-child family at random, with equiprobable $S = \{FF, MM, FM, MF\}$. Let $B = FF$, so $p(B) = .25$.

Conditioning Information can be Subtle

Example 2: The Older Child Paradox

- Consider picking a two-child family at random, with equiprobable $S = \{FF, MM, FM, MF\}$. Let $B = FF$, so $p(B) = .25$.
- Let A be "at least one girl". Calculate $p(B|A)$.

$$p(B|A) = \frac{p(B \cap A)}{p(A)} = \frac{1/4}{3/4} = \frac{1}{3}$$

Conditioning Information can be Subtle

Example 2: The Older Child Paradox

- Consider picking a two-child family at random, with equiprobable $S = \{FF, MM, FM, MF\}$. Let $B = FF$, so $p(B) = .25$.
- Let A be "at least one girl". Calculate $p(B|A)$.

$$p(B|A) = \frac{p(B \cap A)}{p(A)} = \frac{1/4}{3/4} = \frac{1}{3}$$

- Let C be "older child is girl". Calculate $p(B|C)$.

$$p(B|C) = \frac{p(B \cap C)}{p(C)} = \frac{1/4}{2/4} = \frac{1}{2}$$

Detailed Example

From the Florida voters data,

```
joint_prob <- prop.table(table(race = fl_voters$race,
                               gender = fl_voters$gender))
round(joint_prob, 3)
```

```
##           gender
## race           f     m
##   asian    0.009 0.010
##   black    0.074 0.057
##   hispanic 0.073 0.058
##   native   0.002 0.001
##   other    0.017 0.017
##   white    0.360 0.322
```

- Calculate $P(\texttt{hispanic}|\texttt{female})$
- Calculate $P(\texttt{male}|\texttt{white})$

We can use 3-way tables to answer finer questions:

```
fl_voters$young <- ifelse(fl_voters$age < 30, 1, 0)
joint_3way <- prop.table(table(race = fl_voters$race, gender = f
round(joint_3way, 3)
```

```
## , , young = 0
##
##           gender
## race           f     m
##   asian    0.008 0.009
##   black    0.059 0.043
##   hispanic 0.060 0.046
##   native   0.002 0.001
##   other    0.013 0.012
##   white    0.320 0.287
##
## , , young = 1
##
##           gender
## race           f     m
##   asian    0.001 0.001
##   black    0.015 0.014
```

Independence

Events are **independent** if they don't provide any information about each other. Knowing A happened doesn't change the probability of B happening. Knowing B happened doesn't change the probability of A happening.

Independence

Events are **independent** if they don't provide any information about each other. Knowing A happened doesn't change the probability of B happening. Knowing B happened doesn't change the probability of A happening.

A and B are independent iff both

$$P(A|B) = P(A)$$

and

$$P(B|A) = P(B)$$

Independence

Events are **independent** if they don't provide any information about each other. Knowing A happened doesn't change the probability of B happening. Knowing B happened doesn't change the probability of A happening.

A and B are independent iff both

$$P(A|B) = P(A)$$

and

$$P(B|A) = P(B)$$

Let A = voter is female, B = voter is black. These would be independent if knowing the voter is `female` doesn't change our best estimate of whether the voter is `black`, and knowing the voter is `black` doesn't change our best estimate of whether the voter is `female`.

Independence simplifies the Joint Probability

If A and B are independent, then

$$\begin{aligned} P(A \text{ and } B) &= P(A)P(B|A) \\ &= P(A)P(B) \end{aligned}$$

Independence simplifies the Joint Probability

If A and B are independent, then

$$\begin{aligned} P(A \text{ and } B) &= P(A)P(B|A) \\ &= P(A)P(B) \end{aligned}$$

(Equivalently, $P(A \text{ and } B) = P(B)P(A|B) = P(B)P(A)$.)

Set Independence can be Pairwise, Joint, or Both

Consider two flips of a fair coin. $\Omega = \{HH, HT, TH, TT\}$. Let

- $A_1 = H\square$
- $A_2 = \square H$
- $A_3 =$ exactly one H

These are pairwise independent, but **not** independent as a group. Pick any two, and, for example,

$$\begin{aligned} P(A_1 \text{ and } A_3) &= \frac{1}{4} \\ P(A_1)P(A_3) &= \frac{1}{2} \cdot \frac{1}{2} \end{aligned}$$

However, for the whole set of 3 events,
$P(A_1 \text{ and } A_2 \text{ and } A_3) \neq P(A_1)P(A_2)P(A_3)$:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= 0 \\ P(A_1)P(A_2)P(A_3) &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Independence can be Conditional

It can be that A and B are independent *conditional on* C. In fact, sometimes they are *only* independent given C.

For A and B to be independent conditional on C,

$$P(A \text{ and } B|C) = P(A|C)P(B|C)$$

Independence can be Conditional

It can be that A and B are independent *conditional on* C. In fact, sometimes they are *only* independent given C.

For A and B to be independent conditional on C,

$$P(A \text{ and } B|C) = P(A|C)P(B|C)$$

To test whether `hispanic` status and `female` are independent, conditional on being `young`, we start with

- $P(\texttt{hispanic} \text{ and } \texttt{female}|\texttt{young == 1}) =$
- $P(\texttt{hispanic}|\texttt{young == 1}) \cdot P(\texttt{female}|\texttt{young == 1}) =$

Bayes' Rule

Bayes' Rule: Example

When someone will vote in a local election, a survey correctly predicts that that person will vote in 90% of cases. However, for non-voters, 22% of the time it incorrectly predicts that they will vote. Suppose 30% of all people vote in local elections. What is the probability that a person whom the survey predicts will vote actually does vote?

Bayes' Rule: Example

When someone will vote in a local election, a survey correctly predicts that that person will vote in 90% of cases. However, for non-voters, 22% of the time it incorrectly predicts that they will vote. Suppose 30% of all people vote in local elections. What is the probability that a person whom the survey predicts will vote actually does vote?

Start by assuming 100 people, and classify how many are expected to be in each cell below:

	Survey Pred. "Vote"	Survey Pred. "No Vote"	total
Vote			
No Vote			
total			100

Bayes' Rule: Definition

Recall the joint probability $P(A \text{ and } B)$ is both

$$P(A \text{ and } B) = P(A|B)P(B)$$

and

$$P(A \text{ and } B) = P(B|A)P(A)$$

.

Bayes' Rule: Definition

Recall the joint probability $P(A \text{ and } B)$ is both

$$P(A \text{ and } B) = P(A|B)P(B)$$

and

$$P(A \text{ and } B) = P(B|A)P(A)$$

.

We can set these two equal and write

$$\begin{aligned} P(A|B)P(B) &= P(B|A)P(A) \\ P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \end{aligned}$$

In Bayesian statistics, think of A as the quantity we are trying to make inferences about – the proportion of the population that supports a candidate, the probability of a state failing, the number of voters that will show up on Election Day. B is the data we observe.

$P(A)$ is our *prior* belief about likely values of A.

$P(A|B)$ is our *posterior* estimate – includes both our prior belief $P(A)$, but also updates that belief by how the data look, $P(B|A)$.

posterior $= f($prior, data$)$

Bayes' Rule is an intuitive model for how voters behave, how legislative coalitions strategize, how learning takes place in general. It also yields useful ways to estimate quantities when we have good prior information, when we don't have much data, or when our models are very complicated.

Bayes' Rule is an intuitive model for how voters behave, how legislative coalitions strategize, how learning takes place in general. It also yields useful ways to estimate quantities when we have good prior information, when we don't have much data, or when our models are very complicated.

From the law of total probability, we can write a more detailed, very useful version of Bayes' Rule.

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B \text{ and } A) + P(B \text{ and } A^C)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} \end{aligned}$$

This can be extended to other partitions that split $P(B)$. E.g., suppose A has 3 types – A_1, A_2, and A_3 – not just A and A^C. Then,

This can be extended to other partitions that split $P(B)$. E.g., suppose A has 3 types – A_1, A_2, and A_3 – not just A and A^C. Then,

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \end{aligned}$$

Bayes' Rule: Example

Test correctly detects disease 90% of the time, but incorrectly identifies a person as having it 10% of the time. If 10% of all people have disease at any time, what is prob that person who tests positive actually has disease?

Bayes' Rule: Example

Test correctly detects disease 90% of the time, but incorrectly identifies a person as having it 10% of the time. If 10% of all people have disease at any time, what is prob that person who tests positive actually has disease?

From Bayes,

$$
\begin{aligned}
P(A|B) &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^C)P(B|A^C)} \\
P(Y|+) &= \frac{P(Y)P(+|Y)}{P(Y)P(+|Y) + P(Y^C)P(+|Y^C)} \\
&= \frac{.1 \cdot P(+|Y)}{.1 \cdot P(+|Y) + .9 \cdot P(+|Y^C)} \\
&= \frac{.1 \cdot .9}{.1 \cdot .9 + .9 \cdot .1} \\
&= \frac{.09}{.09 + .09} \\
&= .5
\end{aligned}
$$

We can also solve this with a table. Assume 100 people:

	+ test	− test	total
D	9	1	10
ND	9	81	90
total	18	82	

What proportion of + are real?

For Practice 1

About 5% of the world's nations have nuclear weapons. Suppose that when a country has nuclear weapons, it claims to conduct a successful test 80% of the time. Among countries without nuclear weapons, they claim to conduct a successful nuclear test 7% of the time. Country Q claims to have conducted a successful test. What is the probability Q has nuclear weapons?

For Practice 2

In Boston, 30% of people are conservative, 50% are liberal, and 20% are independent. In the last election, 65% of conservatives, 82% of liberals, and 50% of independents voted. If a person in Boston is selected at random and we learn that she did not vote last election, what is the probability she is a liberal?

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Easy – look at their combined batting averages from 1995-1996:

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Easy – look at their combined batting averages from 1995-1996:

- **Jeter: .310**
- Justice: .270

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Easy – look at their combined batting averages from 1995-1996:

- **Jeter: .310**
- Justice: .270

Case closed. Not so fast . . . if you look at only 1995:

- Jeter: .250
- **Justice: .253**

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Easy – look at their combined batting averages from 1995-1996:

- **Jeter: .310**
- Justice: .270

Case closed. Not so fast . . . if you look at only 1995:

- Jeter: .250
- **Justice: .253**

Well, OK, but look at 1996:

Simpson's Paradox

This phenomenon is also known as the Graduate Admissions Paradox, Batting Average Paradox, Aggregation Bias, etc.

Which 1990's NY Yankee was a better batter, Jeter or Justice?

Easy – look at their combined batting averages from 1995-1996:

- **Jeter: .310**
- Justice: .270

Case closed. Not so fast . . . if you look at only 1995:

- Jeter: .250
- **Justice: .253**

Well, OK, but look at 1996:

- Jeter: .314
- **Justice: .321**

How is that possible??

	1995	1996	Total
Jeter	12/48	183/582	**195/630**
Justice	**104/411**	**45/140**	149/551

(Success correlates with volume.)

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts
- HS grads, no college

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts
- HS grads, no college
- Some college

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts
- HS grads, no college
- Some college
- Bachelor's degrees and higher

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts
- HS grads, no college
- Some college
- Bachelor's degrees and higher

Were 15 years good or bad for American workers?

From 2000 to 2013, US Median wage increased 1%.

But not so fast. Median wage decreased for many groups:

- HS dropouts
- HS grads, no college
- Some college
- Bachelor's degrees and higher

How is that possible??

Distributions, Expectation, Variance, the LLN, and the CLT

Random Variables

A *random variable* X is a function mapping the sample space to the set of real numbers:

$$X : \Omega \to \mathbb{R}$$

Random variables thus (a) summarize the outcome of a probabilistic or *stochastic* trial, and (b) take numerical values. For example, suppose we let $X =$ number of heads in 3 coin flips. This X satisfies both (a) and (b), so X is a random variable.

On the other hand, suppose there will be two coin flips, and Y is one of the outcomes, one of {HH, TH, HT, TT}. Here, Y is not a random variable – it summarizes an experiment, but it isn't a number.

Random variables can be discrete (like X above, which can be 0, 1, 2, or 3) or continuous (like the proportion of Americans who turn out for the 2018 midterm election, which can take any value on [0, 1].)

Data

We often think of data as instantiations of random variables, and we consider the empirical distributions of those random variables. It's important to know your data. In `R`, we've seen ways to describe data like

- `dim()`
- `class()` and `mode()`
- `names()`
- `summary()`
- `table()`
- `str()`
- `head()` and `tail()`

Be sure to know the data for your final project well. Give summaries of the important features of your data. For example, suppose you look at the `social` data:

```r
social <- read.csv("https://raw.githubusercontent.com/kosuk
summary(social[, 1:4])
```

```
##      sex          yearofbirth    primary2004          n
##  female:152702   Min.   :1900   Min.   :0.0000   Civic I
##  male  :153164   1st Qu.:1947   1st Qu.:0.0000   Control
##                  Median :1956   Median :0.0000   Hawthor
##                  Mean   :1956   Mean   :0.4014   Neighbo
##                  3rd Qu.:1965   3rd Qu.:1.0000
##                  Max.   :1986   Max.   :1.0000
```

```r
social$age <- (2006 - social$yearofbirth)
mean(social$age)
```

```
## [1] 49.78558
```

```r
median(social$age)
```

```
## [1] 50
```

```r
sd(social$age)
```

You should summarize the important features only. If my paper were about age and voting, perhaps I would write

"About 40% of those in the experiment voted in the 2004 primary. Registrant ages range between 20 and 106, with a median age of about 50 and a standard deviation of about 14 years. Figure 1 shows the full distribution of registered voters' ages. Figure 2 shows the turnout proportions for each age quintile; higher ages are associated with slightly higher turnout in the 2004 primary."

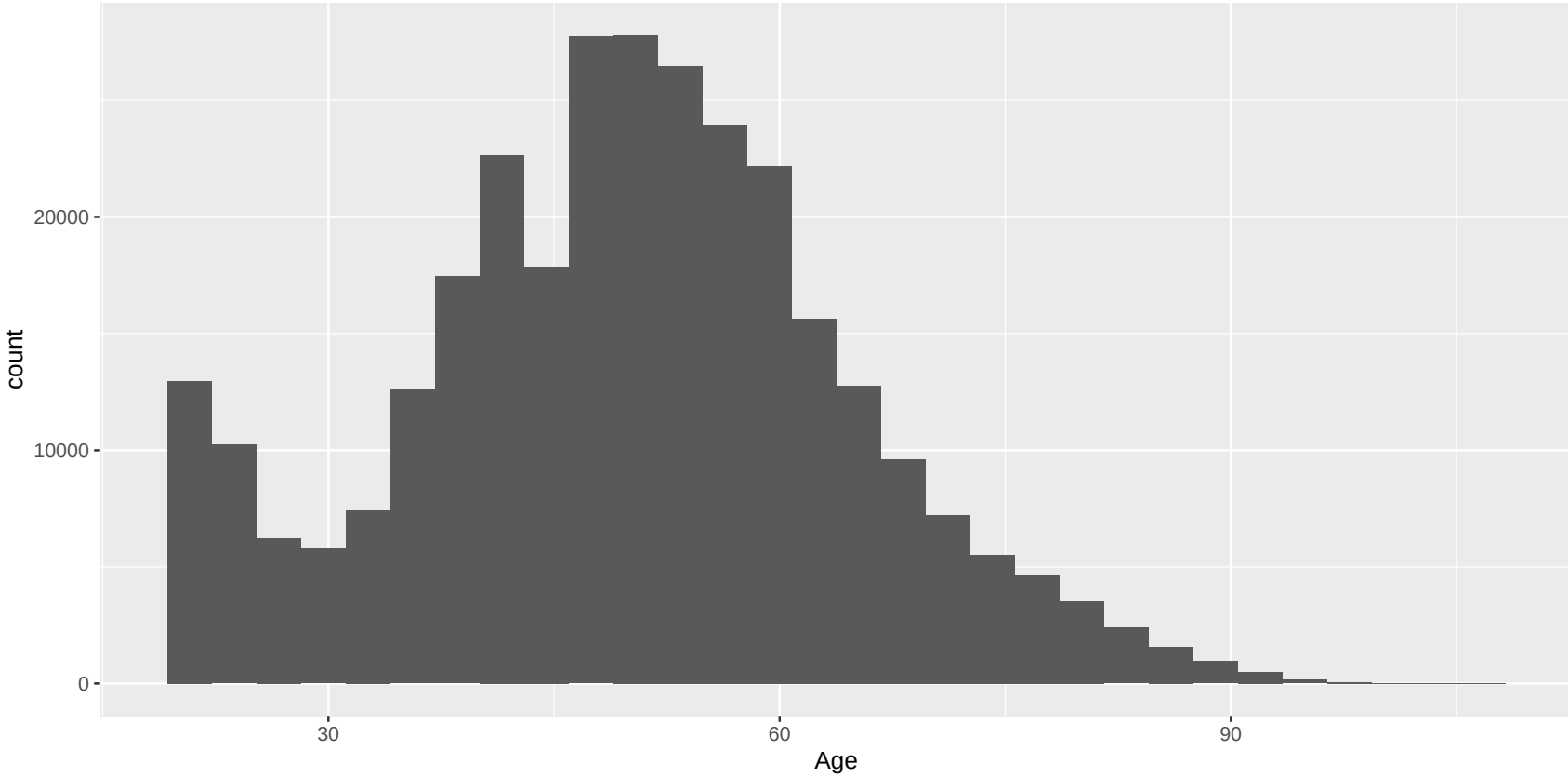


Figure 4: The distribution of voter ages in the social pressure experiment.

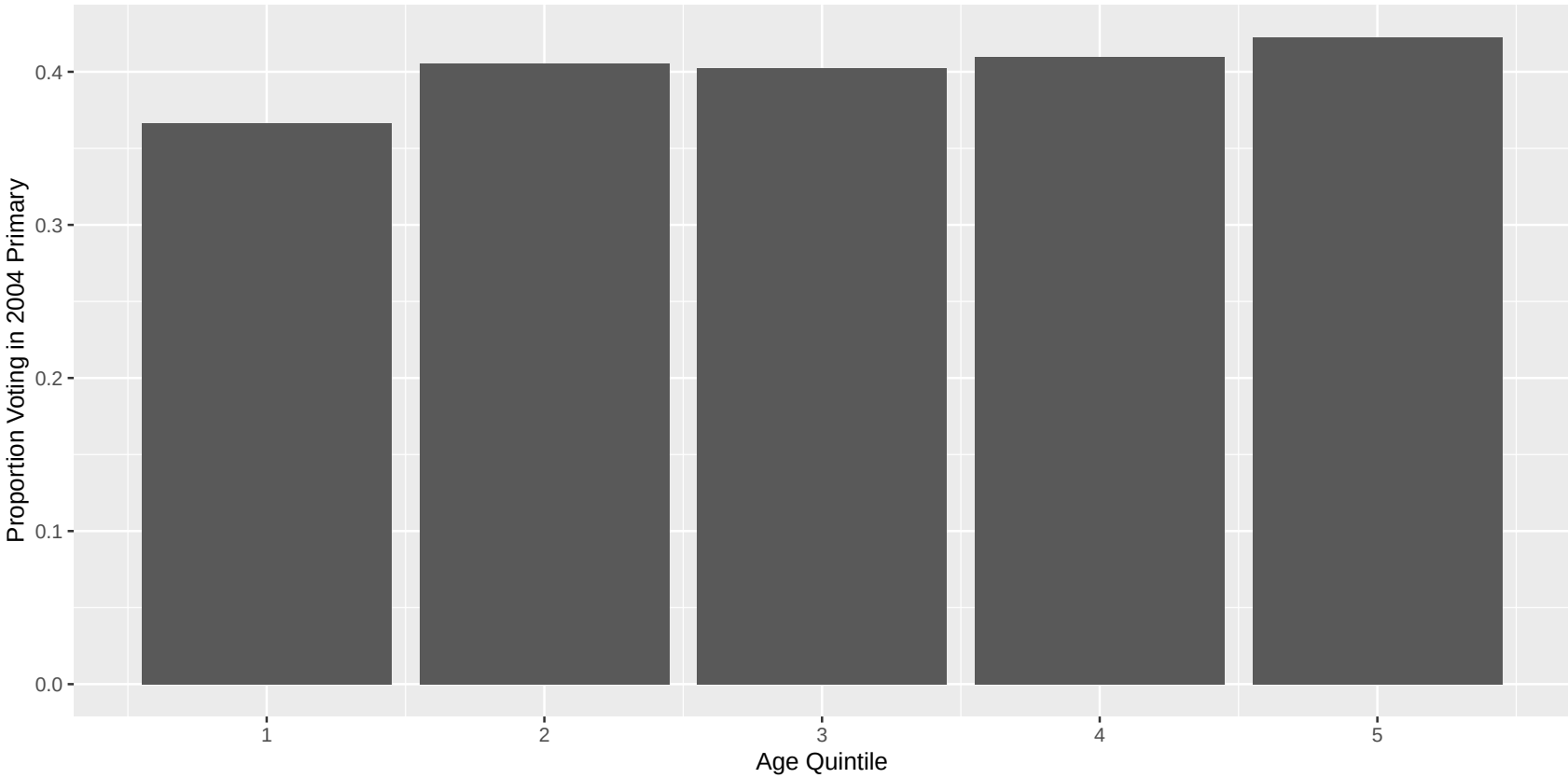


Figure 5: Turnout by age quintile in the social pressure experiment.

Terminology for Statistical Distributions

- *parameter*: unknown numeric value characterizing feature of prob model (Greek)
- *statistic*: quantity calculable from observed data. A function. (Roman)
- *estimator*: statistic used to approximate/guess parameter
- *estimand*: the parameter an estimator attempts to estimate
- *estimate*: application of an estimator function to some obs data

Terminology for Statistical Distributions

- *parameter*: unknown numeric value characterizing feature of prob model (Greek)
- *statistic*: quantity calculable from observed data. A function. (Roman)
- *estimator*: statistic used to approximate/guess parameter
- *estimand*: the parameter an estimator attempts to estimate
- *estimate*: application of an estimator function to some obs data

"The sample statistic $\bar{x}$ is an estimator of true mean parameter μ".

μ is my estimand. 5.1 is my estimate.

PMFs and PDFs

Random variables can be characterized by distributions. The *probability mass function* (PMF) or *probability density function* (PDF) describes the shape of the random variable (in our cases, the data). Discrete distributions have a PMF, but sometimes we just refer to the "PDF" for both discrete and continuous data. We represent them with lower case p.

Since the PDF of a random variable is a *probability* function, it assigns a probability to each value (for a discrete random variable) or each range of values (if X is continuous). The sum of these probabilities must $= 1$. For a general random variable, we write

$$p(X = x|\text{parameters})$$

For example, let X = the number of Gallup survey respondents, out of 3, who disapproved of "the way Donald Trump is handling his job as president" this week. What values can X take?

For example, let $X =$ the number of Gallup survey respondents, out of 3, who disapproved of "the way Donald Trump is handling his job as president" this week. What values can X take?

Suppose we are interested in the probability that 2 respondents disapprove, and we know that the proportion of respondents in the population who disapprove is $\pi = 0.51$ (2019-12-15). Then, we'd write

$$p(X = 2|\pi = 0.51)$$

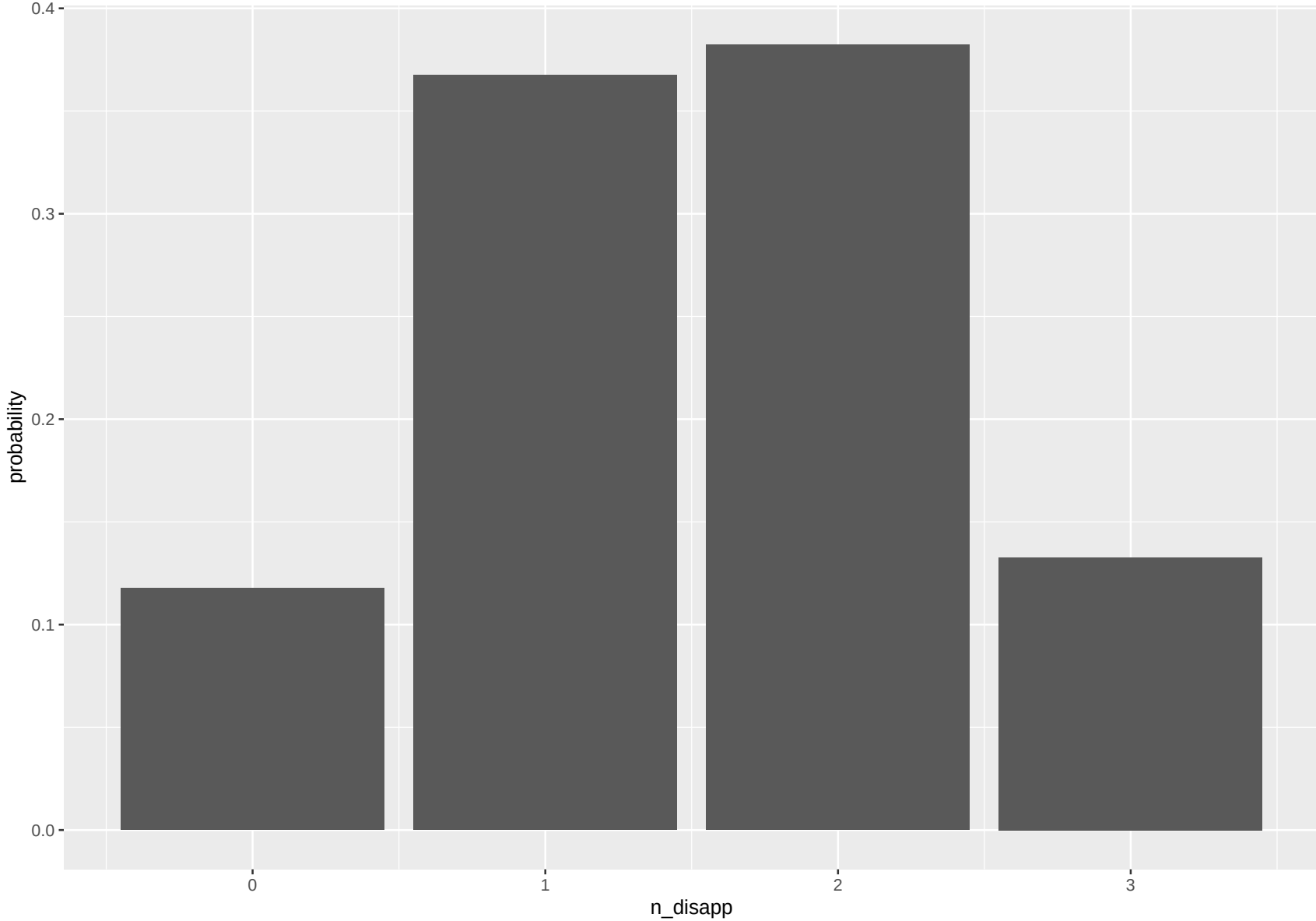


Figure 6: PMF of Number of Respondents Disapproving. Two is most likely, given $\pi = .55$ and 3 resp's.

Since the PDF just shows the distribution of X, it is also called the "marginal distribution of X".

For the 2-dimensional shape that represents a joint distribution of two random variables, we write $p(x, y) = p(X = x, Y = y)$.

CDFs

The *cumulative distribution function* of a random variable summarizes the process of interest, just like the PDF. It uses the same scale on the x-axis. However, the y-axis always covers [0, 1]. For each value of X, the CDF asks "what proportion of the data are below that particular x?"

We use $F(X)$ to represent the CDF. Since the probabilities sum to 1, the value of the CDF at the maximum value of X is always 1.

For a discrete random variable,

$$F(x) = p(X \leq x) = \sum_{x \leq z} p(z) = \sum_{x=0}^{z} p(z)$$

For a continuous random variable,

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} p(z)dz$$

.

To write a joint CDF of two random variables, it would look like $F(x, y) = p(X \leq x, Y \leq y)$.

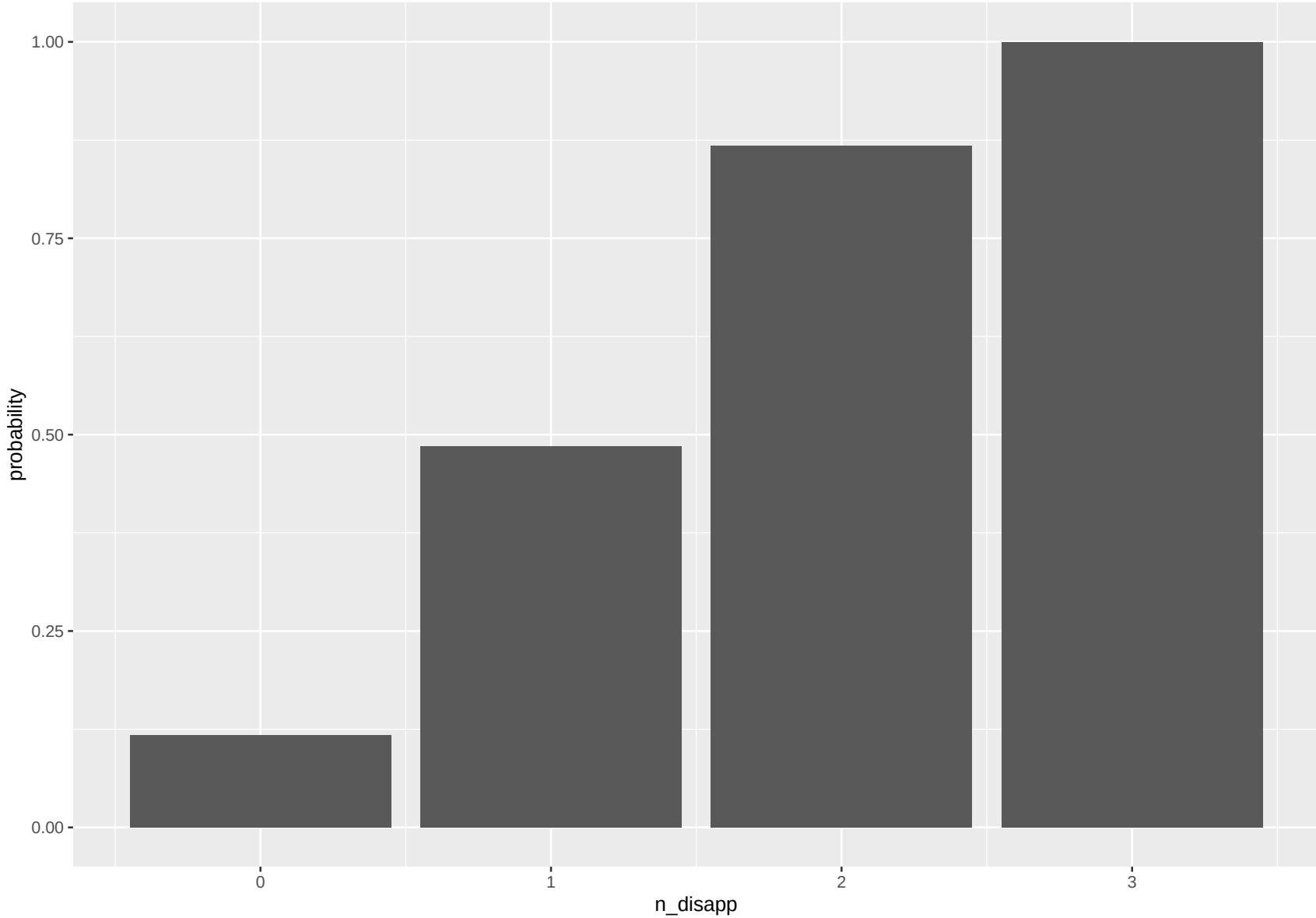


Figure 7: CDF of Number of Respondents Disapproving.

The Quantile Function

Quantiles are just the values of a variable that a certain fraction of the data are at or below. The first quartile is a quantile; it is the value with 25% of the data below it.

The Quantile Function

Quantiles are just the values of a variable that a certain fraction of the data are at or below. The first quartile is a quantile; it is the value with 25% of the data below it.

The quantile function is the inverse of the CDF. For the CDF, you put in a value of X, and you get back the proportion below that value. For the quantile function, you put in a proportion, and you get back the value of X with that proportion below it.

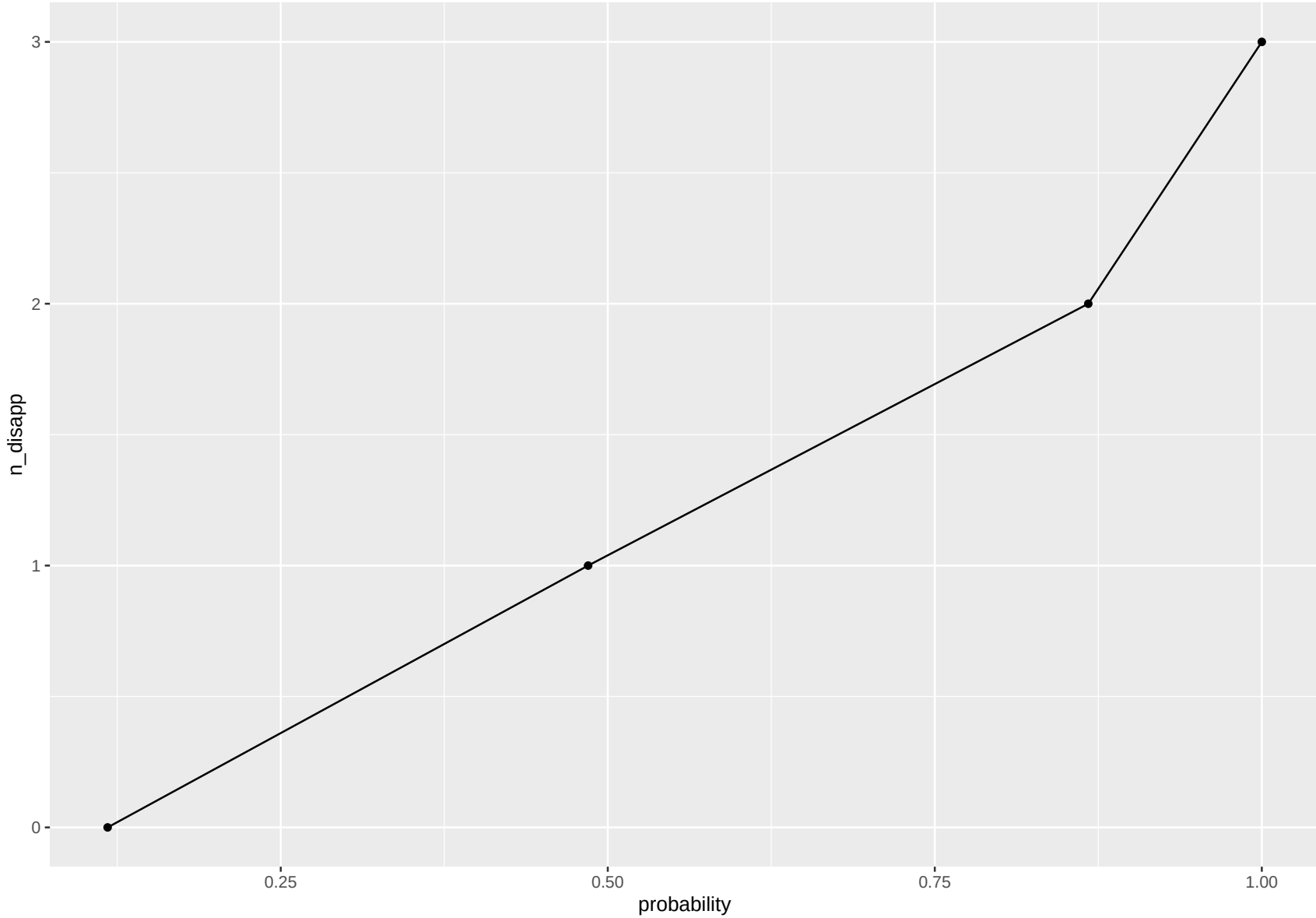


Figure 8: Quantile Function for No. of Resp's Disapproving.

Expectation of a Random Variable

The *expected value* or *expectation* of a random variable is the mean of its outcomes, weighted by their probabilities. For a discrete random variable, we write

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

For a continuous random variable, we write

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx$$

Expectation of a Random Variable

The *expected value* or *expectation* of a random variable is the mean of its outcomes, weighted by their probabilities. For a discrete random variable, we write

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

For a continuous random variable, we write

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx$$

Note especially that the expectation is **not** the sample mean from particular instantiation, a particular data set. We will use the sample mean $\bar{x}$ to *estimate* the expected value.

Variance of a Random Variable

The *variance* of a random variable is the mean of its outcomes' squared deviations from the expectation, weighted by their probabilities:

$$\begin{aligned} V(X) &= E[(X - E(X))^2] \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

Conditional Summaries

- The *conditional expectation* of Y given X is

$$E(Y|X) = \sum_{i=1}^{k} y_i p(y_i|x)$$

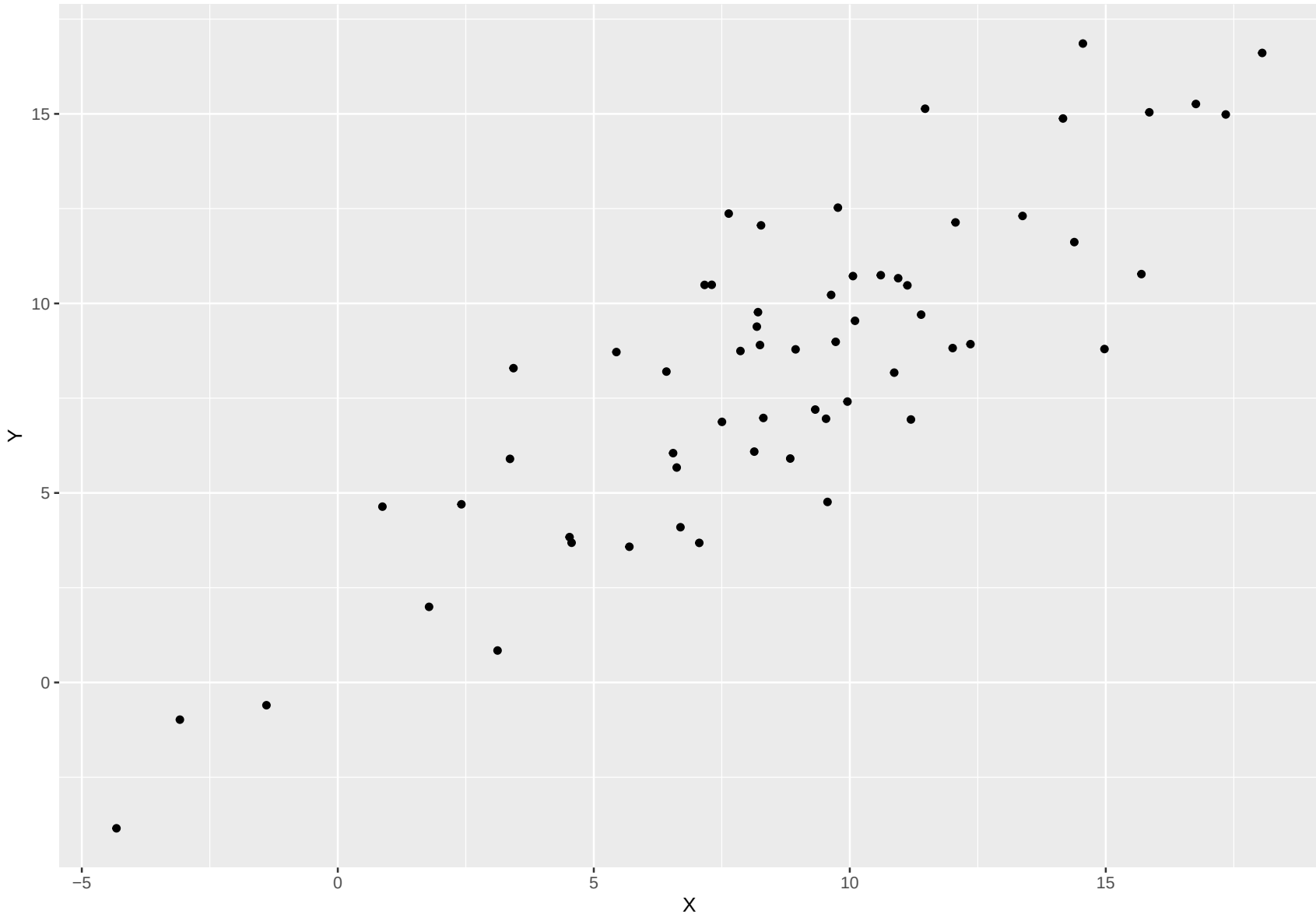
$$E(Y|X) = \int_{-\infty}^{\infty} y_i p(y_i|x)$$

- Regression is estimating a conditional expectation function.
- The *conditional variance* of Y given X is

$$Var(Y|X) = E[(Y - E(Y|X))^2|X]$$

- Regression estimates employ conditional variance assumptions

Estimate average value of Y, conditional on $X \geq 18$, that is, $\bar{Y}|X \geq 18$



Properties of Expectation

- $E(c) = c$
- $E(a + bX) = a + bE(X)$
- $E(X + Y) = EX + EY$
- If X and Y indep., then $E(XY) = E(X)E(Y)$

Properties of Variance

- $Var(X) = E(X - EX)^2$
- $Var(X) = E(X^2) - (EX)^2$
- $Var(c) = 0$
- $Var(Y|X) = E(Y^2|X) - (E(Y|X))^2$
- If X and Y indep., then $Var(X + Y) = Var(X) + Var(Y)$
- If X and Y indep, then $Var(X - Y) = Var(X) + Var(Y)$

Conditional Expectation

Discrete Example

Suppose that X and Y have this joint distribution:

		X			
		-2	**0**	**2**	**3**
Y	**3**	0.27	0.08	0.16	0
	6	0	0.04	0.10	0.35

where the cells give the proportion of data with those values of X and Y. (I.e., 27% of the data have $X = -2$ and $Y = 3$.)

Compute

- $E(Y)$
- $E(Y|X = 2)$
- $Var(Y|X = 2)$ for these data.

The Law of Large Numbers

"As you take larger samples, the sample mean converges to underlying true expected value."

```
set.seed(241)
pop <- 1:20
EX <- mean(pop) ## 10.5
samp_size <- 1:200
samp_means <- NA
for(i in samp_size){
  samp_means[i] <- mean(sample(pop, i, replace = TRUE))
}
head(samp_means) %>% round(2)
```

```
## [1]  7.00 12.00 13.33 13.25 13.00  6.33
```

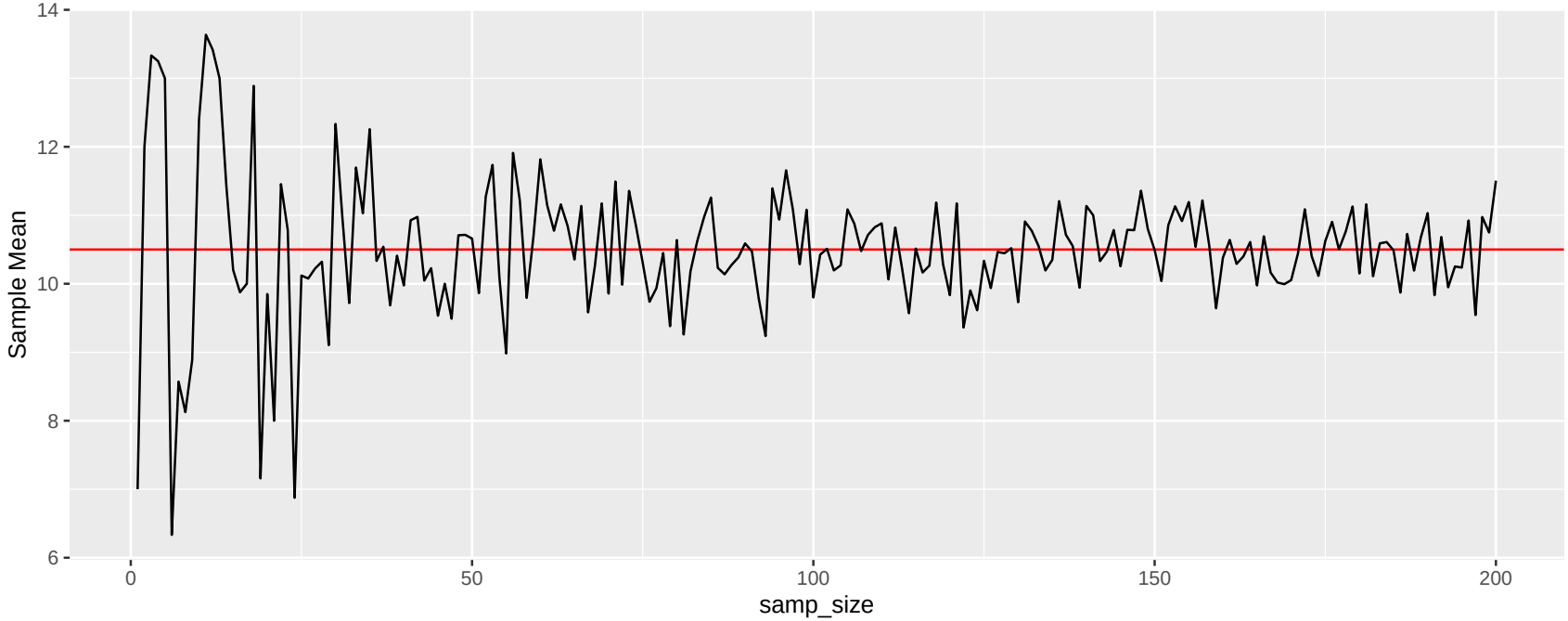


Figure 9: LLN: Bigger sample, closer to true expectation.

Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then, as $n \to \infty$,

$$\bar{X}_n \to E(X)$$

The Central Limit Theorem

"As you take larger samples, the *distribution* of sample means converges to a normal (Gaussian) distribution."

```
samp_size <- 20
how_many_samples <- 40
samp_means <- NA
for(i in 1:how_many_samples){
  samp_means[i] <- mean(sample(pop, samp_size, replace = TR
}
```

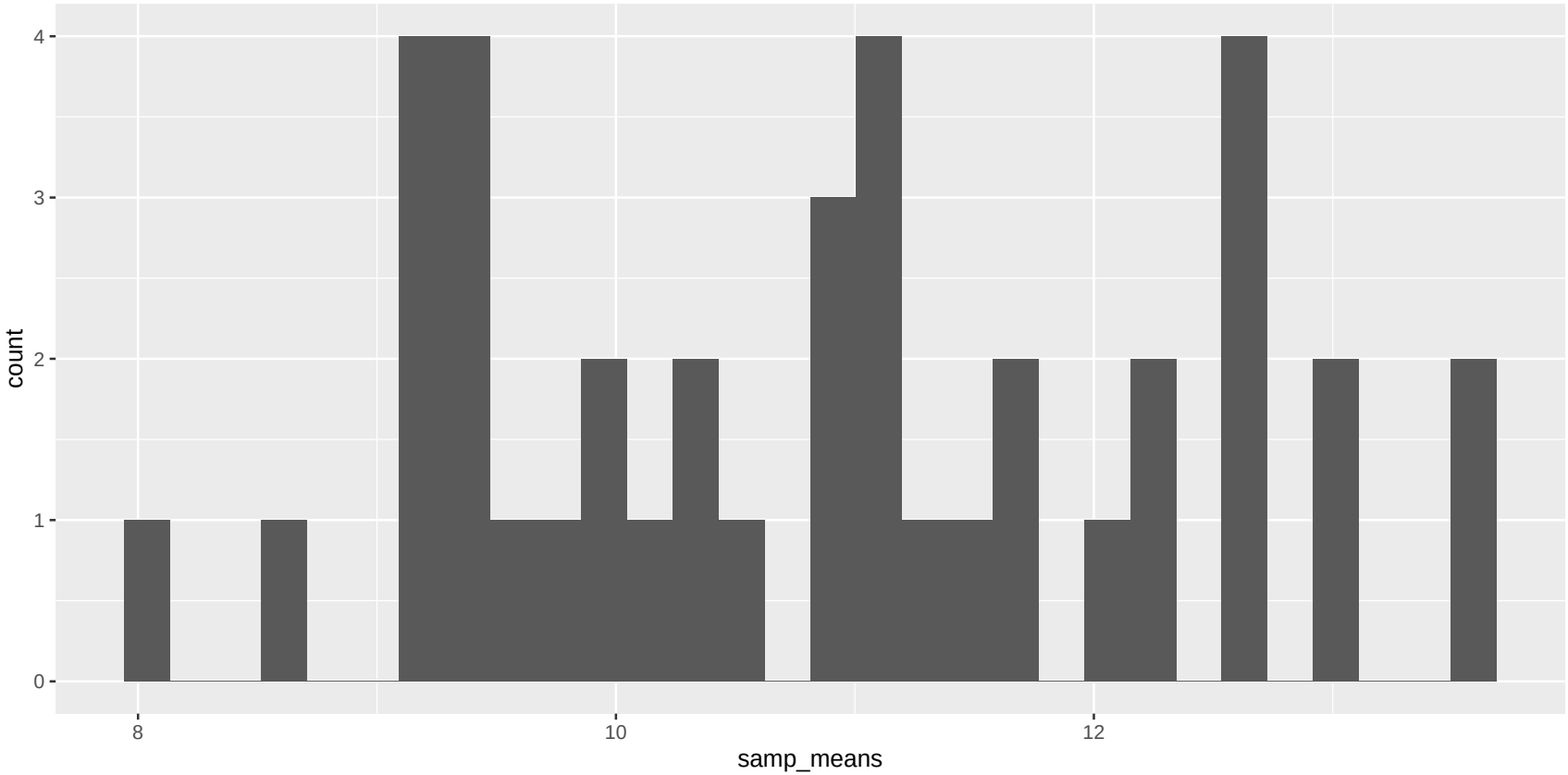


Figure 10: 40 Sample Means

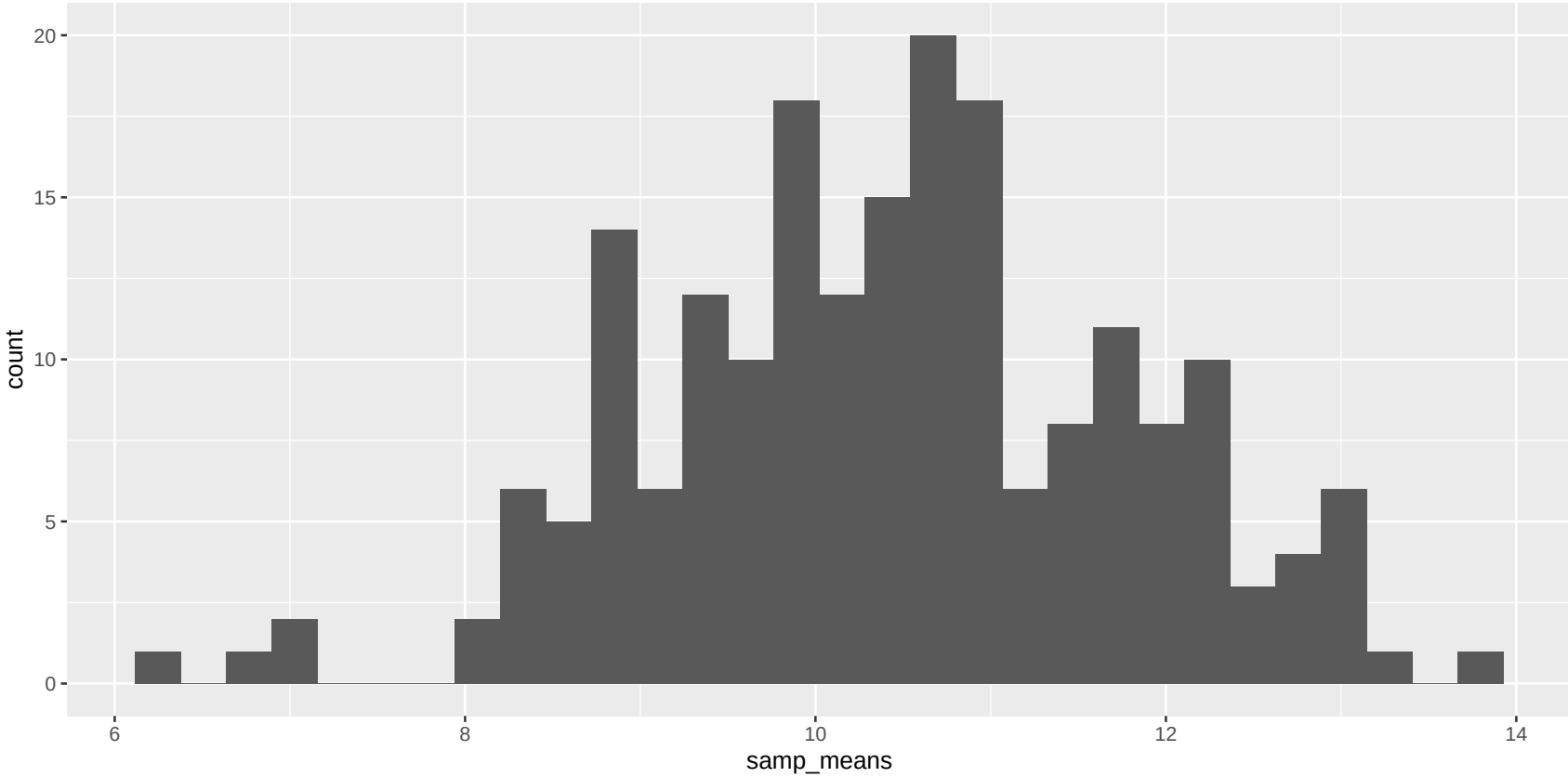


Figure 11: 200 Sample Means

As $n \to \infty$,

$$\frac{\bar{X}_n - E(X)}{\sqrt{\frac{V(X)}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

The Four Key Distributional Methods

For any distribution in R, there are functions

- `rdist`: **r**andom draws from `dist`
- `ddist`: height of **d**ensity of `dist`
- `pdist`: distrib'n function, giving **p**robabilities from `dist`
- `qdist`: **q**uantile function of `dist` (inverse of `pdist`)

Let's look at the standard normal distribution:

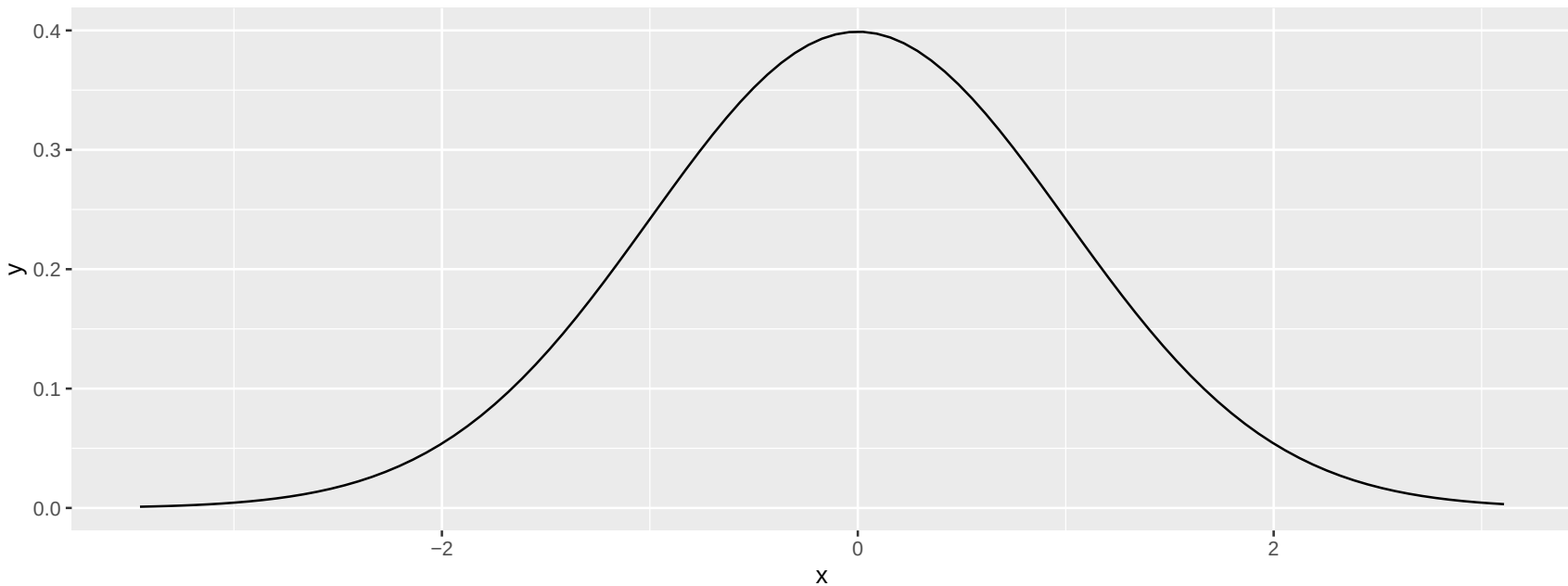


Figure 12: The Standard Normal

Continuous Distributions in R

```
> set.seed(887)
> rnorm(5)
[1]  1.23734933  0.25908638 -0.05131962  1.63780013  0.3399
> rt(5, df = 2)
[1] -0.1525046 -0.6069671 -0.8034640 -4.7216253 -0.7856204
> rchisq(5, df = 2)
[1] 1.191595 2.075456 1.041626 2.044578 1.315084
> runif(5, min = 1, max = 3)
[1] 1.246602 2.301376 2.319596 1.550342 2.694279
```

(But not just `rdist`, also `ddist`, `pdist`, `qdist`)

Discrete Distributions in R

```
> rbinom(5, 2, prob = c(.5, .5))
[1] 0 2 1 1 1
```

Discrete Distributions

Bernoulli Distribution

- Single trial, binary outcome $0, 1$.
- "Prob of success in 1 trial?", $x \in \{0, 1\}$
- $X \sim Bern(p)$
- $p(X = x|p) = p^x(1-p)^{1-x}$
- $F(x) = 1^x(1-p)^{1-x}$

Political examples:

- Let

$$X = \begin{cases} 1 & \text{if you turnout} \\ 0 & \text{if you abstain} \end{cases}$$

Then, $p(X = 1|p = .4) = .4$ prob of you turning out to vote in next election, given underlying true prob $p = .4$;
$p(X = 0|p = .4) = .6$ prob of you abstaining in next election.

- prob of US-NKorea trade agreement during 2018
- "Let X be a Bernoulli-distributed random variable with $p = .4$"
- "Let X be distributed Bernoulli with $\pi = 0.4$"

Binomial Distribution

- n independent, identically distributed (iid) trials, binary outcome $0, 1$.
- "Prob of k successes in n trials?", $k \in \{0, \ldots, n\}$
- $X \sim Bin(n, p)$
- $p(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$
- $X \sim Bin(1, p) \sim p(X = k|1, p) \sim Bern(p)$
- $X_1 \sim Bin(n_1, p)$, $X_2 \sim Bin(n_2, p)$, $X_1 \perp\!\!\!\perp X_2$, then $X_1 + X_2 \sim Bin(n_1 + n_2, p)$
- $F(x) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$
- Sum of n Bernoullis
- Political examples:
  - prob 3 of 6 opposing Senators support an amendment: $p(X = 3|n = 6, p = .3) =$ `dbinom(3, 6, prob = .3)` $\approx .19$
  - prob ≥ 3 of 6 opposing Senators support an amendment: $p(X \geq 3|n = 6, p = .3) =$ `1 - pbinom(2, 6, prob = .3)` = `pbinom(2,6,prob=.3,lower.tail=FALSE)` $\approx .26$

Continuous Distributions

Uniform Distribution

- $X \sim Unif(a, b)$
- $x \in [a, b]$
- $p(x) = \frac{1}{b-a}$
- $F(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad a \leq x < b \\ 1 & \text{for} \quad x \geq b \end{cases}$
- Common: $X \sim Unif(0, 1)$
- Political examples:
  - "Suppose voter's probability of turnout is draw from uniform"

χ^2 Distribution

- $X \sim \chi^2_n$
- $x > 0$
- $n \in \mathbb{Z}_+$
- $p(x) = \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$
- $X_1 \sim \chi^2_{n_1}$, $X_2 \sim \chi^2_{n_2}$, $X_1 \perp\!\!\!\perp X_2$, then $X_1 + X_2 \sim \chi^2_{n_1+n_2}$
- $\chi^2_n \sim Gamma\left(\frac{n}{2}, 2\right)$
- Political examples:
  - Model relationships between table rows/columns

    $X^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$
  - regression statistics

Beta Distribution

- $X \sim Beta(\alpha, \beta)$
- $x \in [0, 1]$
- $\alpha, \beta > 0$, $\alpha - 1$ successes, $\beta - 1$ failures
- $p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
- $Beta(1, 1) \sim Unif(0, 1)$
- Political examples:
  - flexible model for probability
  - conjugate with Binomial

Normal (Gaussian) Distribution

- $X \sim N(\mu, \sigma^2)$
- $x \in \mathbb{R}$
- $\mu \in \mathbb{R}$
- $\sigma > 0$
- $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- Common: $X \sim N(0, 1)$, "standard normal"
- $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- $\Phi(x) =$ standard normal CDF
- Political examples:
  - population quantities, asymptotic/known variance sampling distributions

Student's t Distribution

- $X \sim t_n(\mu, \sigma^2)$
- $x \in \mathbb{R}$
- $n \in \mathbb{Z}_+$
- $p(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sigma\sqrt{n\pi}} \left(1 + \frac{1}{n}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{n+1}{2}}$
- Common: $t_n \sim \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$
- $t_1 \sim Cauchy$ (a very dangerous distribution)
- If $X \sim N(0,1)$, $Y \sim \chi^2_n$, $X \perp\!\!\!\perp Y$, then $\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$
- Political examples:
  - finite sample/unknown variance distributions
  - robust estimation

Solutions to §5.6

$$\begin{aligned} E(Y|X=2) &= \sum_y y \cdot p(Y=y|X=2) \\ &= 3 \cdot p(Y=3|X=2) + 6 \cdot p(Y=6|X=2) \\ &= 3(0.16/0.26) + 6(0.10/0.26) \\ &= 4.15 \end{aligned}$$

Note: $p(Y=3|X=2) = \frac{p(Y=3,X=2)}{p(X=2)} = 0.16/0.26$, and $p(Y=6|X=2) = \frac{p(Y=6,X=2)}{p(X=2)} = 0.10/0.26$. So,

$$\begin{aligned} Var(Y|X=2) &= \sum_y (Y - E(Y|X=2))^2 p(Y|X=2) \\ &= (3-4.15)^2(0.16/0.26) + (6-4.15)^2(0.10/0.26) \\ &= 2.13 \end{aligned}$$

Uncertainty: The Confidence Interval

The Standard Error

The *standard error* (SE) of an estimator is the SD of its sampling distribution. The SE provides a measure of uncertainty around the estimator. Imai (2017) pages 324-5 provides details for sample proportions, means, and differences in means. To begin, quickly calculate the standard errors of the three estimators below. Recall that the `var()` function in R is useful to calculate the variance of a set of numbers.

1. You have a sample of 500 UK survey respondents, 200 of whom supported Brexit. Calculate the sample proportion of supporters, and the SE around the proportion.
2. You have the five Senate committees with 13, 11, 8, 20, and 19 members, respectively. Calculate the sample mean of committee sizes, and the SE around the mean.
3. You have three democratic countries with national election average turnouts of 60, 70, and 80 percent. You also have four autocracies with 50, 80, 90, and 90 percent turnout. Calculate the difference in the sample mean levels of turnout, and the SE around the difference.

The Confidence Interval

A *confidence interval* (CI) characterizes the uncertainty around an estimate that we generate. The CI is a range of values that, if our model is correct, should include the true underlying parameter a specified fraction of the time. For example, if we build many good 90% confidence intervals around a sample proportion, they should include the truth 90% of the time.[1] Because we don't know the true value, we never know whether any *one* confidence interval contains the truth. In other words, we can't tell whether our sample was lucky or unlucky in how well it represents the population; we want to make estimates, but keep in mind that they come with caveats. If $(1-\alpha)\times 100\%$ is our confidence level (like 90% above), then the lower and upper bounds of a CI are

$$[\text{Estimate} - \text{Critical Value} \cdot SE, \quad \text{Estimate} + \text{Critical Value} \cdot SE]$$

[1] Of course, this means that 10% of the time, well-constructed 90% confidence intervals will *not* include the truth.

We know how to get estimates and SE's. (We just did 3 different types.) The only thing left is to get the Critical Value, which quantifies how much uncertainty we want by setting the factor that we multiply the SE by. The Critical Value tells us "how many SE's away from the estimate are we interested in?"

α	Confidence Level	Crit Value $z_{\alpha/2}$	R code
0.01	99%	2.58	`qnorm(0.995)`
0.05	95%	1.96	‘qnorm(0.975)‘
0.1	90%	1.64	‘qnorm(0.95)‘

Table 1: Confidence Levels and (Normal) Critical Values

Sometimes we express our uncertainty with an “alpha level” (α) instead of a confidence level. These are mathematically identical – just different ways to express the same level of uncertainty. $\alpha = 0.05$ corresponds to 95% confidence. In general, the confidence level is

$$\text{Confidence Level} = (1 - \alpha) \cdot 100\%$$

An Example

For a given Congress, suppose that the probability π of bills passing is 0.35. We will calculate an 80% confidence interval for a sample of bills. After you create each object below, look at it and ask if you don't understand it.

1. Take a random sample of 10 bills. `samp <- rbinom(10, 1, .35)`
2. Calculate $\hat{p}$, the prop in your sample that passed. `phat <- mean(samp)`
3. Calculate the SE around the $\hat{p}$ `se <- sqrt(phat * (1 - phat) / 10)`
4. Find the critical value for an 80% interval `critval <- qnorm(.9)`
5. Calculate an 80% confidence interval around p `lower <- phat - critval*se upper <- phat + critval*se`
6. Write a sentence interpreting the interval you calculated.

When you're finished, write your CI on the board.

7. Now, take a sample of 100 bills and calculate an 80% interval. Compare.

Note that the nominal coverage rate (e.g., the "95%" in a 95% confidence interval) is **not** the chance that the true value is in your particular interval.

Sample Means: Inference Using the t Instead of the Normal

Often, we can use the t-distribution to improve upon confidence intervals from the Normal distribution. The t-distribution is somewhat fatter-tailed than the Normal, implying that we expect more variation in the data than the Normal involves. The t is actually a family of distributions that are wider when we have less data, narrower when we have more. When we have n observations, we select the t-distribution with $n - 1$ degrees of freedom.

To calculate an 80% confidence interval using the t-distribution, we get the critical value via `qt(.9, df)`.

Example 1: CI for One Sample Mean

Recall the data from the randomized experiment creating village council seats for women in Indian villages.

```
w_res <- read.csv("https://raw.githubusercontent.com/kosuke
head(w_res, 2)
```

```
##   GP village reserved female irrigation water
## 1  1       2        1      1          0    10
## 2  1       1        1      1          5     0
```

Let's calculate a 95% CI for the number of `water` projects that happen in villages with reservations for women. Our procedure will be the same: get the estimate, the SE, the critical value (from a t this time), and form the lower and upper bounds.

1. Calculate the mean number of water projects in villages with `reserved == 1`.
2. Calculate the SE around the mean.
3. Calculate the critical value with `qt()`
4. Form the interval [Estimate $-$ Critical Value $\cdot$ SE, Estimate $+$ Critical Value $\cdot SE$]

Example 2: CI for a Difference Between Two Sample Means

Calculate the 90% CI for the *difference* between water projects in villages with and without reservations for women. Statistic of interest: difference between average water projects for `reserved == 1` versus `reserved == 0`.

1. Calculate the mean number of water projects for villages without reservations.
2. We will use `t.test()` to calculate the confidence interval directly for us. `water` is the outcome; `reserved` is the treatment. Run the code below

```
t.test(water ~ reserved, data = w_res, conf.level = .9)
```

3. Confirm that your two group means (and thus, their difference) are those reported by `t.test()`.
4. Interpret the interval R provides, noting that `t.test()` always takes "first group − second group", which here

The Margin of Error and Sample Size

In survey sampling, we sometimes refer to the *margin of error* (MoE). This is just a component of the CI calculation:

$$[\text{Estimate} - \underbrace{\text{Critical Value} \cdot SE}_{\text{Margin of Error}}, \quad \text{Estimate} + \underbrace{\text{Critical Value} \cdot SE}_{\text{Margin of Error}}]$$

That is,

$$\text{MoE} = \text{Critical Value} \cdot SE$$

We can use this to find the minimum sample size for a certain level of precision in a survey. Suppose we have a survey asking whether Scottish voters support Brexit, and we want it to be precise to within 0.03 (three percentage points), with 95% confidence. The largest SE for a sample proportion occurs at $\hat{p} = 0.5$. So,

$$
\begin{aligned}
0.03 &= 1.96 \cdot \sqrt{\frac{.5(1-0.5)}{n}} \\
0.03^2 &= 1.96^2 \cdot \frac{.5(1-0.5)}{n} \\
n &= 1.96^2 \cdot \frac{.5(1-0.5)}{0.03^2} \\
n &\approx 3.8416 \cdot 277.8 \\
n &\approx 1067
\end{aligned}
$$

Uncertainty: Null-Hypothesis Significance Testing

Standard errors and confidence intervals quantify the uncertainty around estimates we make. When you report estimates in your final paper, you should provide a measure of the uncertainty around them.

Today, we consider *null hypothesis significance testing* (NHST). Though intimately related to SE's and CI's, NHST differs in two important ways: first, it results in a binary pass/fail of a test; second, the logic of NHST can feel convoluted, so dedicate some time to thinking it through. The idea is *proof by contradiction*: assume a hypothesis is true; then, does the data look too extreme under that assumption? If so, we reject the hypothesis.

NHST asks questions like, "how strange would the data be, if, in fact, there is no relationship between X and Y?" This is a bit odd, since we usually think there might be a relationship – that's what caused us to investigate this political phenomenon in the first place.

Caveat: causal interpretations can be valid when our design is good, but **no test result by itself indicates causality**. Just as with linear regression coefficients and R^2, our p-values below are not inherently imbued with causal meaning. That can only come from good design.

The Logic

The logic of NHST proceeds as follows (see Imai (2017), page 349):

- Assume a null hypothesis value of an underlying parameter (often, but not always, of the form, "no relationship")
- Estimate a test statistic, an estimate of the parameter, from the data
- Find how many SE's the test statistic is from the hypothesized value
- Reject the null hypothesis value if the test statistic is too many SE's away

The Procedure

1. Specify a null hypothesis, H_0.
2. Specify an alternative hypothesis, H_a. Often – but not always – this is the logical negation of H_0. This determines whether the test is one-sided or two-sided. If the alternative includes $\neq$, the test is two-sided; if the alternative includes $>$ or $<$, the test is one-sided.
3. Specify a threshold α. This is how unlikely the data have to be to reject the null hypothesis. It is related to the confidence level – a 95% confidence level is the same as $\alpha = 0.05$.
4. Calculate the test statistic from the data. To test a proportion, $\hat{\pi}$; to test a single mean, $\bar{Y}$; to test the difference between two means, $\bar{Y}_T - \bar{Y}_C$; to test a regression coefficient, $\hat{\beta}$.
5. Calculate the SE for the test statistic.

4. Standardize the test statistic. Subtract off the null hypothesis value and divide by the SE.[2]

5. Calculate the p-value. Using the correct reference distribution (a normal or t from the CLT), calculate the proportion of the probability mass as extreme or more extreme than what you observed.

6. Compare the p-value to α. If $p < \alpha$, the data were unusual if the H_0 were true, so **reject H_0**. If $p > \alpha$, the data were not unusual if the H_0 were true, so **do not reject H_0**.

[2] For our four tests,

$$
\begin{aligned}
z &= \frac{\hat{\pi} - \pi}{SE} \\
t &= \frac{\bar{Y} - \mu}{SE} \\
t &= \frac{(\bar{Y}_T - \bar{Y}_C) - (\mu_T - \mu_C)}{SE} \\
t &= \frac{\hat{\beta} - \beta}{SE}
\end{aligned}
$$

Example: Testing a Sample Proportion

Following the Procedure in §6

1. H_0: the true proportion of Trump approval is $\pi = 0.4$ (40%)
2. H_a: $\pi \neq 0.4$
3. Let $\alpha = 0.05$.
4. On 1 April 2019, fivethirtyeight.com estimates Trump approval to be $\hat{\pi} = 0.421$ (42.1%)
5. Assuming H_0 is true, the standard error is

$$SE(\hat{\pi}) = \sqrt{\frac{.4 \cdot (1 - .4)}{1000 \text{ survey respondents}}} \approx 0.015$$

, or about 1.5 percentage points. So, our estimate is $\frac{42.1-40}{1.5} \approx 1.4$ SE's from the hypothesis.

6. $z = \frac{\hat{\pi} - \pi}{SE} = \frac{0.421 - 0.4}{.0015} \approx 1.4$

7. Find the proportion of the normal distribution to the right of $z \approx 1.4$:

```
z <- (0.421 - 0.4) / sqrt(.4 * .6 / 1000)
pnorm(z, lower.tail = FALSE) # prop to the right of 1.4
```

```
## [1] 0.08762212
```

```
1 - pnorm(z)                 # 1 - prop to the left of 1.4
```

```
## [1] 0.08762212
```

Since H_a is two-sided, double this:

```
pvalue <- 2 * pnorm(z, lower.tail = FALSE)
pvalue
```

```
## [1] 0.1752442
```

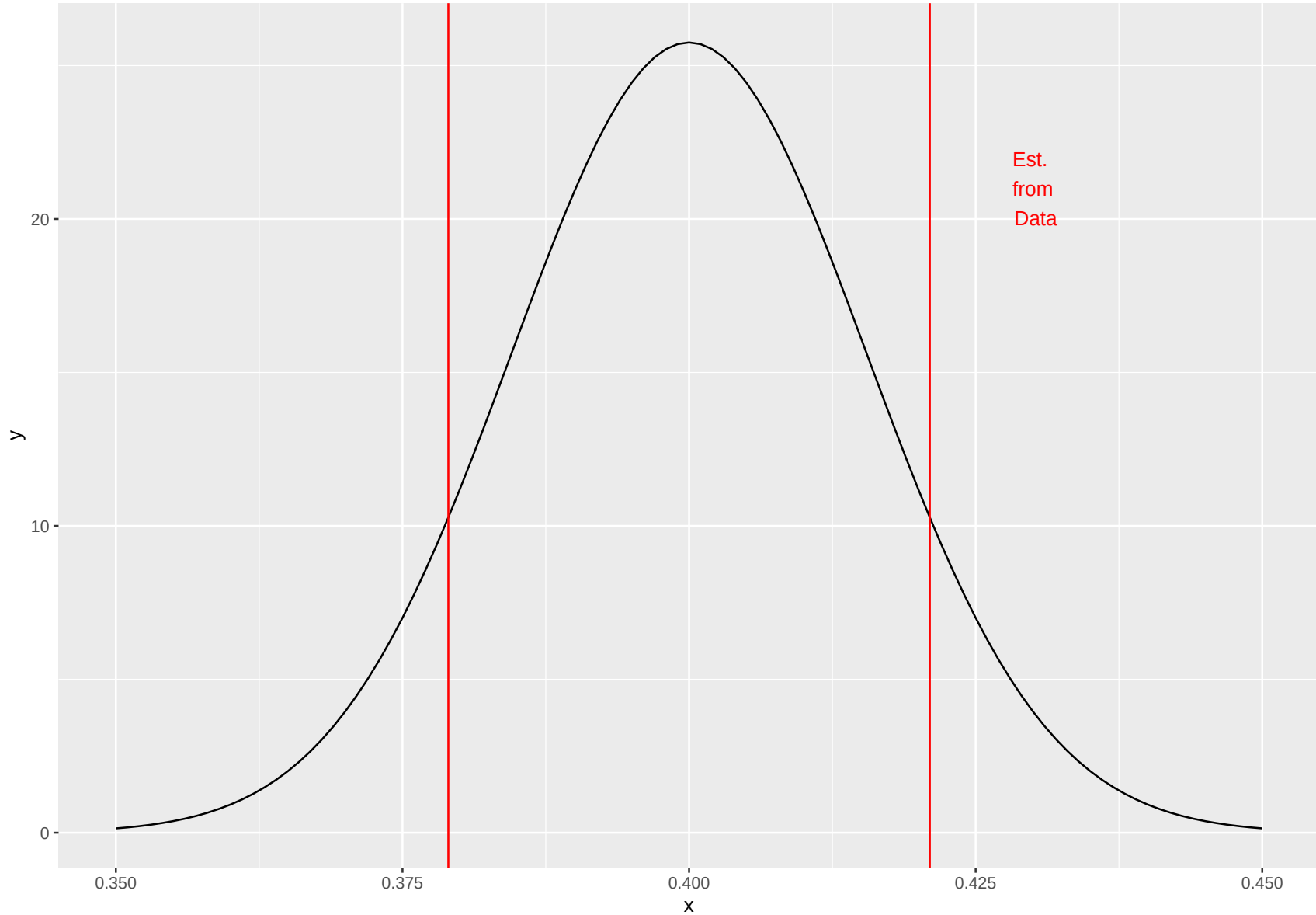
8. Here $0.1752442 > \alpha$, so we **do not reject** H_0. We do **not** have evidence that the true π is different from 0.4.

More Detail

How do we figure out if 1.4 SE's is "too far" away? We calculate the p-value – the proportion of the reference distribution that is at least 1.4 SE's away, and we compare it to a chosen threshold α. If $p < \alpha$, it's too far, and we reject the null hypothesis. If $p > \alpha$, we do not reject it.

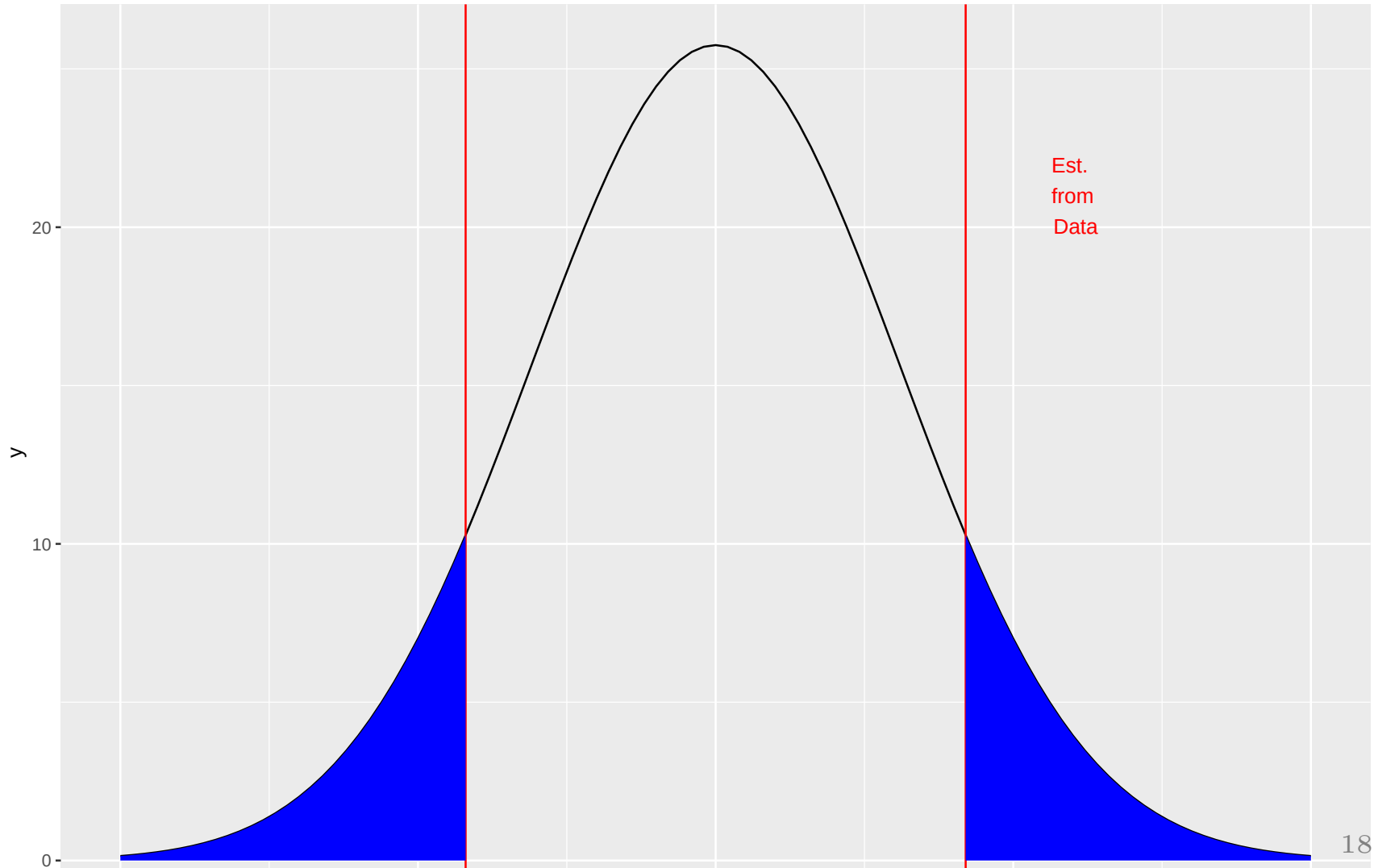
Since we're testing proportions, the reference distribution is the normal distribution.

Here's our normal distribution, marking -1.4 and $+1.4$ SE's from the hypothesis. That is, from $40\% \pm 2.1\%$:



Now, we set a threshold for rejecting H_0. This threshold is α,

Next, we calculate the p-value of our estimate, still assuming H_0 is true. The p-value is the area under our reference distribution that is *as extreme or more extreme* than the data we saw. What fraction of our data is as or more extreme than $40 \pm 2.1\%$? The area that is the p-value looks like this:



To calculate the exact proportion of the probability mass representing how extreme the data appear (i.e., the shaded region), we use `pnorm()`. Our calculation above first standardized, then used the standard normal. Below, we (equivalently) don't standardize, but use more arguments of `pnorm()` to find the same result:

```
se_null <- sqrt(.4 * .6 / 1000)
prop_right_of_data <- pnorm(.421, mean = .4, sd = se_null,
prop_right_of_data
```

```
## [1] 0.08762212
```

So, about 8.8% of the area is less than what we observed. The total area *at least as extreme* includes both sides, though, so we'll double this to get the p-value of 0.1752.

(We can check that the doubling is correct by asking "how much probability is to the left of $.40 - .021$?":)

```
prop_left <- pnorm(.4 - .021, mean = .4, sd = se_null, lowe
prop_left
```

```
## [1] 0.08762212
```

Finally, we compare the p-value to our threshold α. Since this p-value is greater than our threshold, $0.1752 > \alpha$, we **do not reject** the null hypothesis.

Using `prop.test()`

We can get the same result with `prop.test()`:

```
prop.test(421, 1000, p = .4, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  421 out of 1000, null probability 0.4
## X-squared = 1.8375, df = 1, p-value = 0.1752
## alternative hypothesis: true p is not equal to 0.4
## 95 percent confidence interval:
##  0.3907589 0.4518457
## sample estimates:
##     p
## 0.421
```

An Alternative Strategy

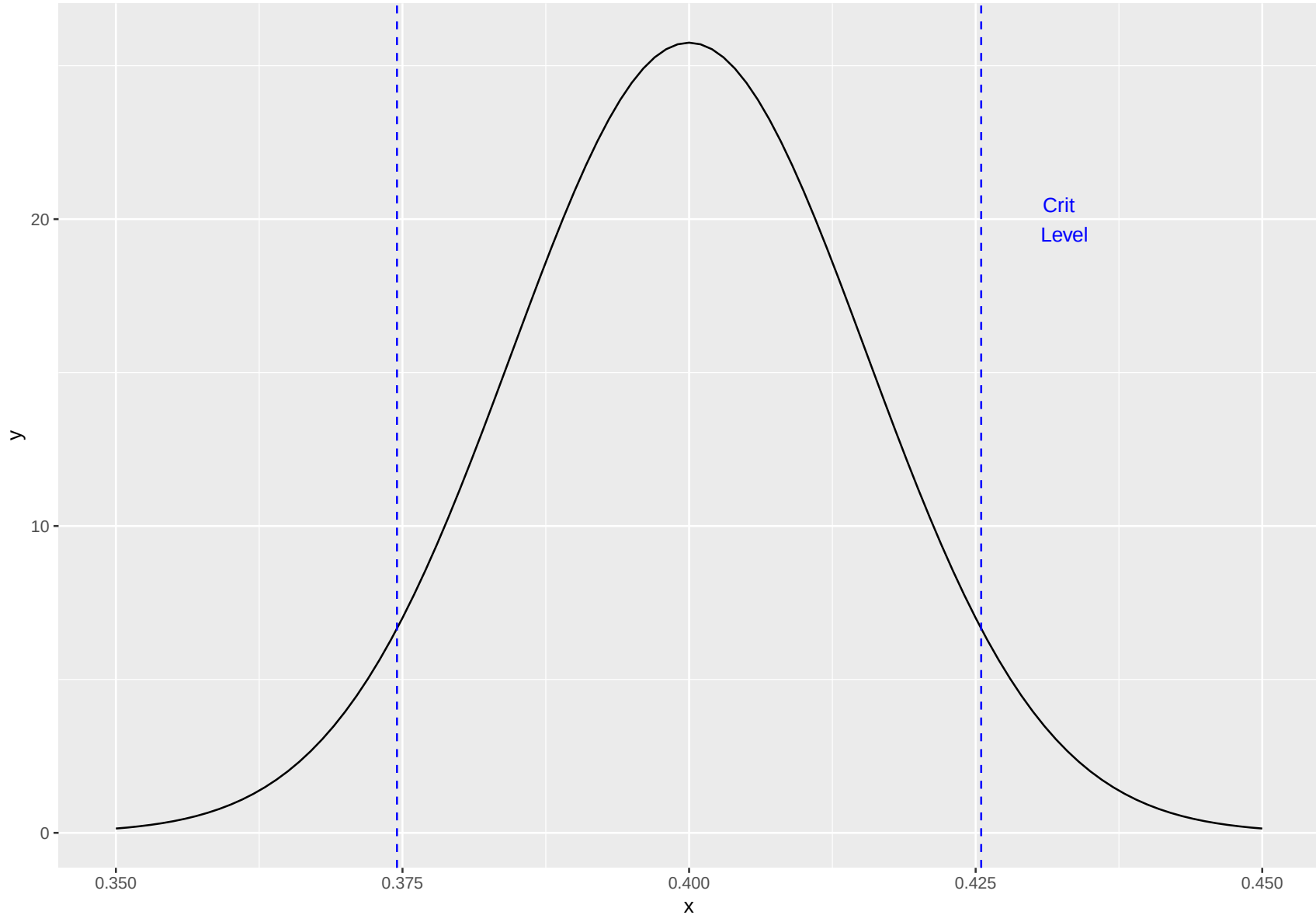
Another way of getting the same result is to ask, "what is the largest number of SE's away from the null we could be to not be considered 'too far'?".

For example, we could say "under the null, for a normal distribution, how many SE's away is associated with $\alpha = 0.1$?" If the data are more extreme than these values, then we **reject** the null. To calculate how many SE's away this is, the *critical value*, we use `qnorm()`. For $\alpha = 0.1$, which is 90% confidence, under the normal, we get a critical value of about 1.64 SE's:

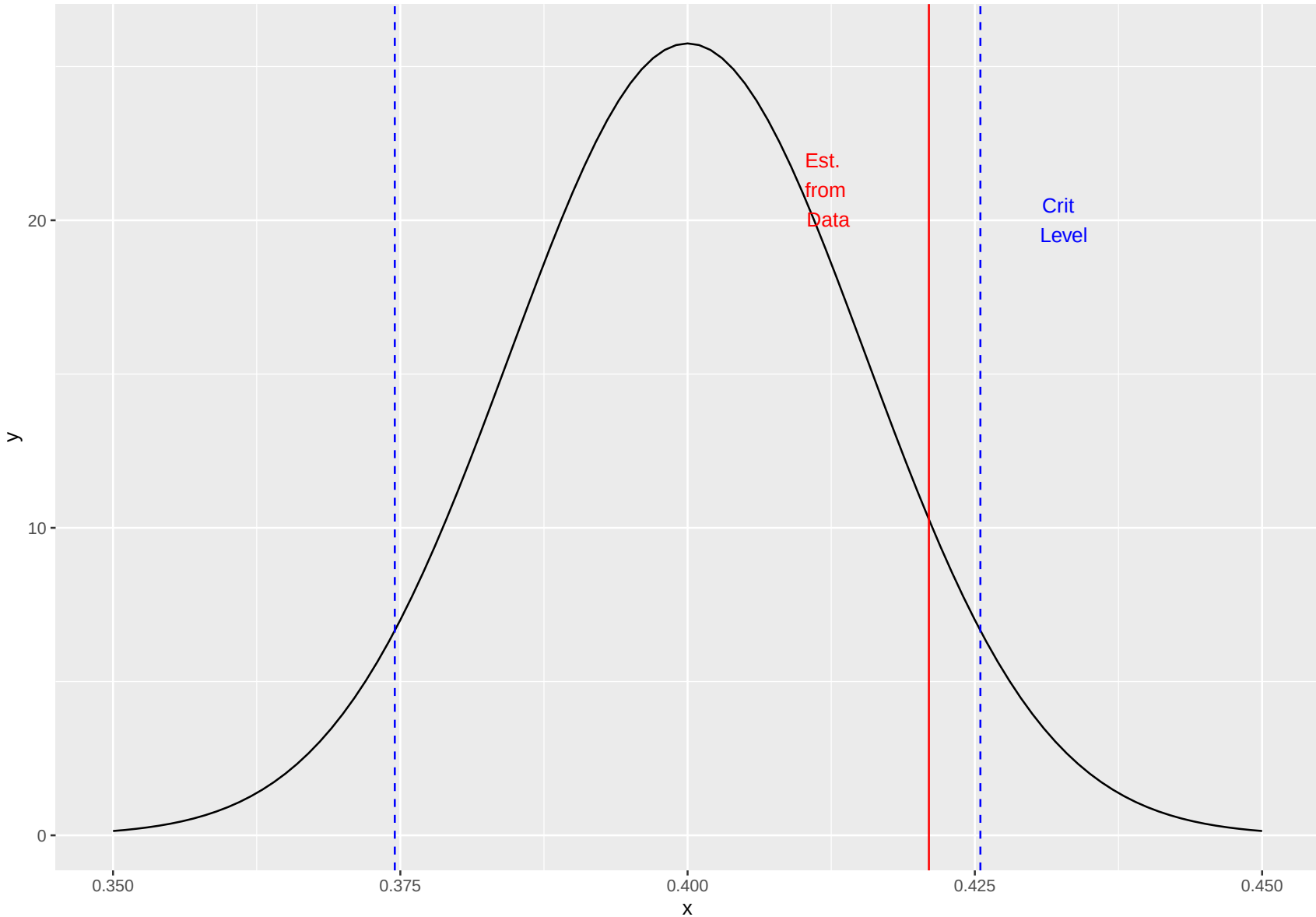
```
alpha <- .1
qnorm(alpha / 2)  ## divide by 2: half prob on each side
```

```
## [1] -1.644854
```

Here are the critical values:



When we add the data, it's clear that the data are slightly less extreme, so we **do not reject** the null hypothesis.



Relationship to the CI

If we formed a 90% confidence interval around our estimate, and the H_0 value is outside our interval, then we **reject** H_0. If the data are inside the CI, we **do not reject** H_0. This is mathematically equivalent to determining whether the p-value is less than $\alpha = 0.1$.

Example: Testing a Single Mean

Recall the data from the randomized experiment creating village council seats for women in Indian villages.

```
w_res <- read.csv("https://raw.githubusercontent.com/kosuke
```

Let's test whether the mean number of water projects for villages without reservations is statistically significantly different than 15.

1. Set the null hypothesis: $H_0 : \mu_{\text{without res}} = 15$
2. Set $H_a : \mu_{\text{without res}} \neq 15$
3. Let $\alpha = 0.1$.
4. Estimate the test statistic

```
water_proj_no_res <- w_res$water[w_res$reserved == 0]

x.bar <- mean(water_proj_no_res)
x.bar
```

```
## [1] 14.73832
```

5. Calculate the SE

Using `t.test()`

We can do this calculation in R with

```
t.test(water_proj_no_res, mu = 15, conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  water_proj_no_res
## t = -0.20156, df = 213, p-value = 0.8405
## alternative hypothesis: true mean is not equal to 15
## 90 percent confidence interval:
##  12.59352 16.88312
## sample estimates:
## mean of x
##  14.73832
```

Example: Testing a Difference in Means

Using `t.test()`

We can use `t.test()` to test whether the mean water projects is different for reserved vs. not reserved councils, as well. The null is "no difference" $H_0 : \mu_{\text{reserved}} - \mu_{\text{not reserved}} = 0$. This is a null hypothesis of "no average treatment effect", and is expressed in the order "*treatment* minus *control*".

```
t_out <- t.test(water ~ reserved, data = w_res, conf.level
t_out
```

```
##
##  Welch Two Sample t-test
##
## data:  water by reserved
## t = -1.8141, df = 122.05, p-value = 0.07212
## alternative hypothesis: true difference in means is not
## 90 percent confidence interval:
##  -17.7058080  -0.7990379
## sample estimates:
## mean in group 0 mean in group 1
```

A Note on Degrees of Freedom

The degrees of freedom for a two-sample test is given by the Welch–Satterthwaite equation (above, `t.test()` returns 122.046). The value depends on the two sample sizes and the two sample variances (just as does the SE). A conservative estimate for the degrees of freedom is given by

$$\text{df} = \min\{n_1 - 1, n_2 - 1\}$$

, the smaller of the two group sizes, minus one. (This is "conservative", in that you assume you have less than half of the data points you actually have, roughly speaking.)

Randomization (Design-based) Inference

A volunteer?

A volunteer?

The task: select the 2 folders with messages

A volunteer?

The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?

A volunteer?

The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?
  - "No x-ray vision. No ESP. Effect of messages on choice = 0."

A volunteer?

The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?
  - "No x-ray vision. No ESP. Effect of messages on choice = 0."
- What is an alternative?

A volunteer?

The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?
  - "No x-ray vision. No ESP. Effect of messages on choice = 0."
- What is an alternative?
  - "Some way to detect messages. Message location → choice."

A volunteer?

The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?
  - "No x-ray vision. No ESP. Effect of messages on choice = 0."
- What is an alternative?
  - "Some way to detect messages. Message location → choice."

A volunteer?

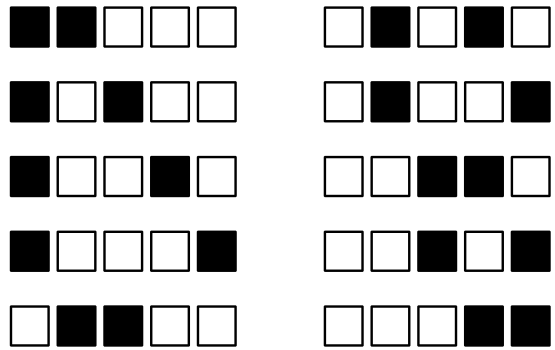
The task: select the 2 folders with messages

- What is our baseline expectation/model for this process?
  - "No x-ray vision. No ESP. Effect of messages on choice = 0."
- What is an alternative?
  - "Some way to detect messages. Message location → choice."

Select!

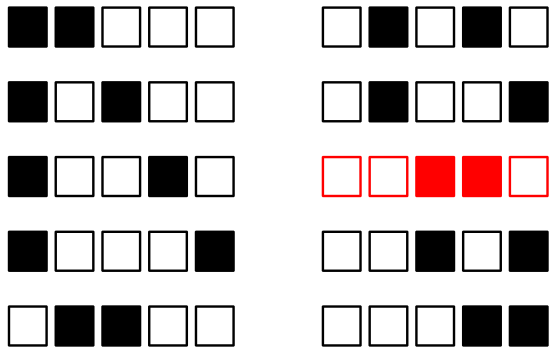
Randomization Inference

The possible choices:



Randomization Inference

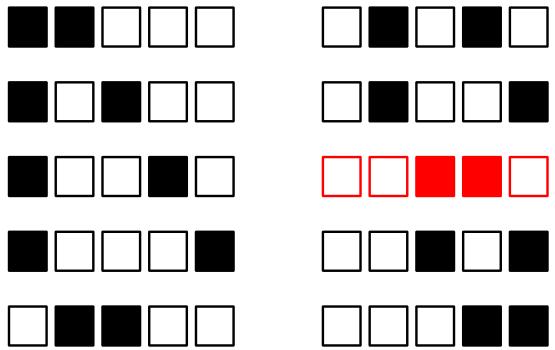
The possible choices:



- You chose _____ and _____. Let X = number found.

Randomization Inference

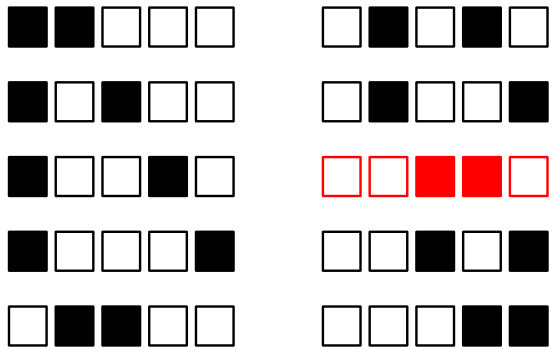
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$?

Randomization Inference

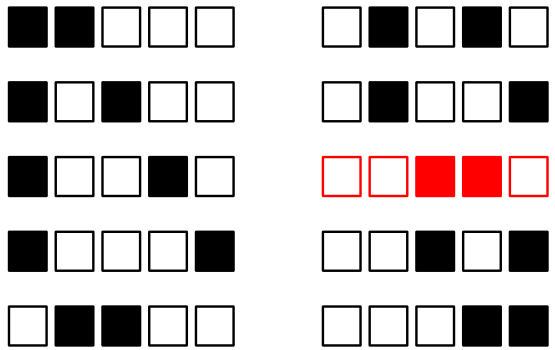
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$

Randomization Inference

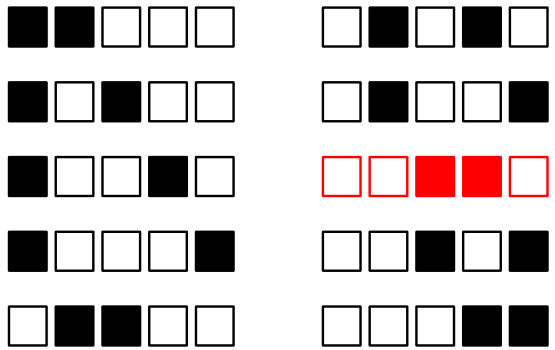
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$?

Randomization Inference

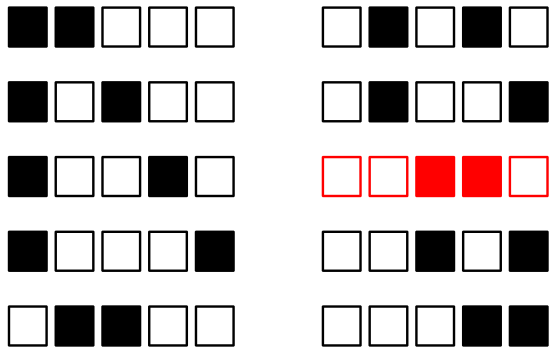
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$? $\frac{7}{10} = 0.7$

Randomization Inference

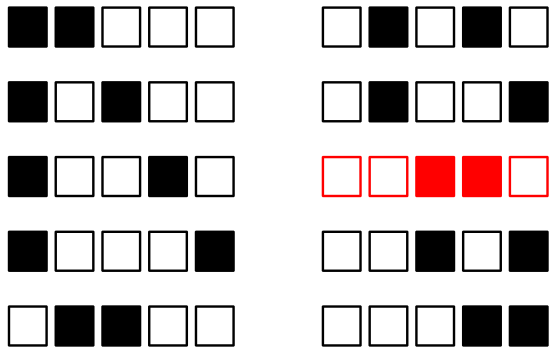
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$? $\frac{7}{10} = 0.7$
- What is "prob result at least this extreme, given model of no effect"?

Randomization Inference

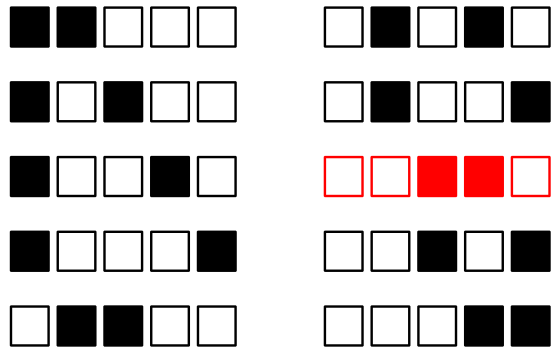
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$? $\frac{7}{10} = 0.7$
- What is "prob result at least this extreme, given model of no effect"?
- Definition of p-value!

Randomization Inference

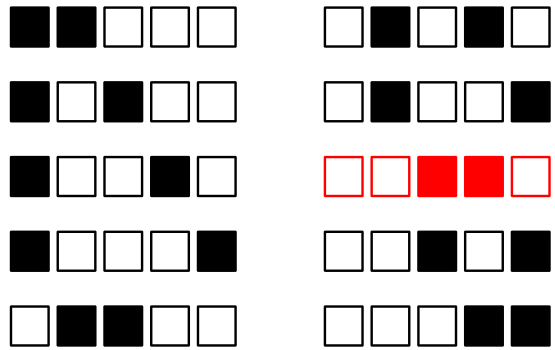
The possible choices:



- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$? $\frac{7}{10} = 0.7$
- What is "prob result at least this extreme, given model of no effect"?
- Definition of p-value!
- Valid, exact, with no distributional assumption, no large n.

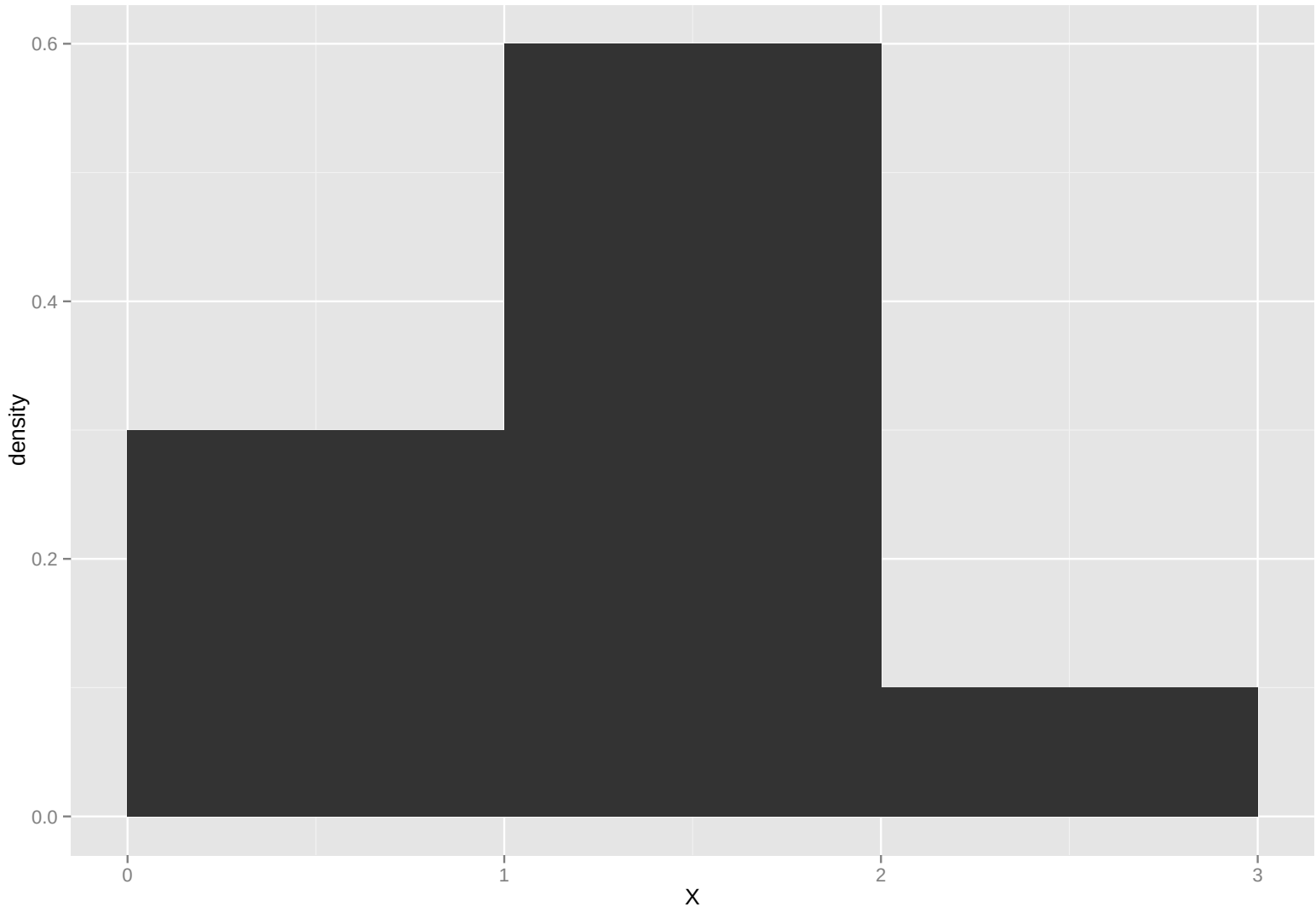
Randomization Inference

The possible choices:

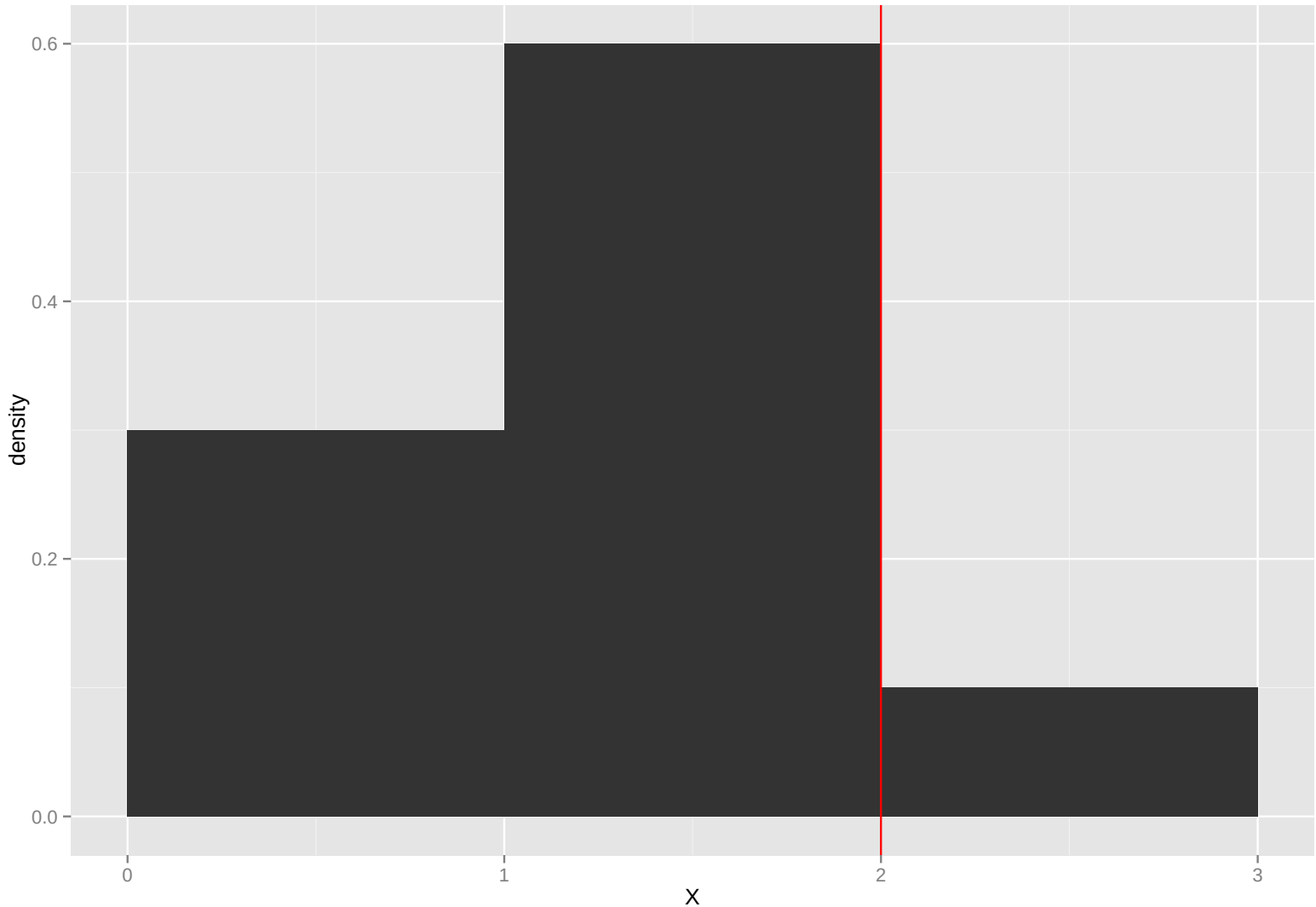


- You chose _____ and _____. Let X = number found.
- What was $P(X \geq 2|\text{no ESP})$? $\frac{1}{10} = 0.1$
- What was $P(X \geq 1|\text{no ESP})$? $\frac{7}{10} = 0.7$
- What is "prob result at least this extreme, given model of no effect"?
- Definition of p-value!
- Valid, exact, with no distributional assumption, no large n.
- *Randomization* creates dist'n of possible numbers correct

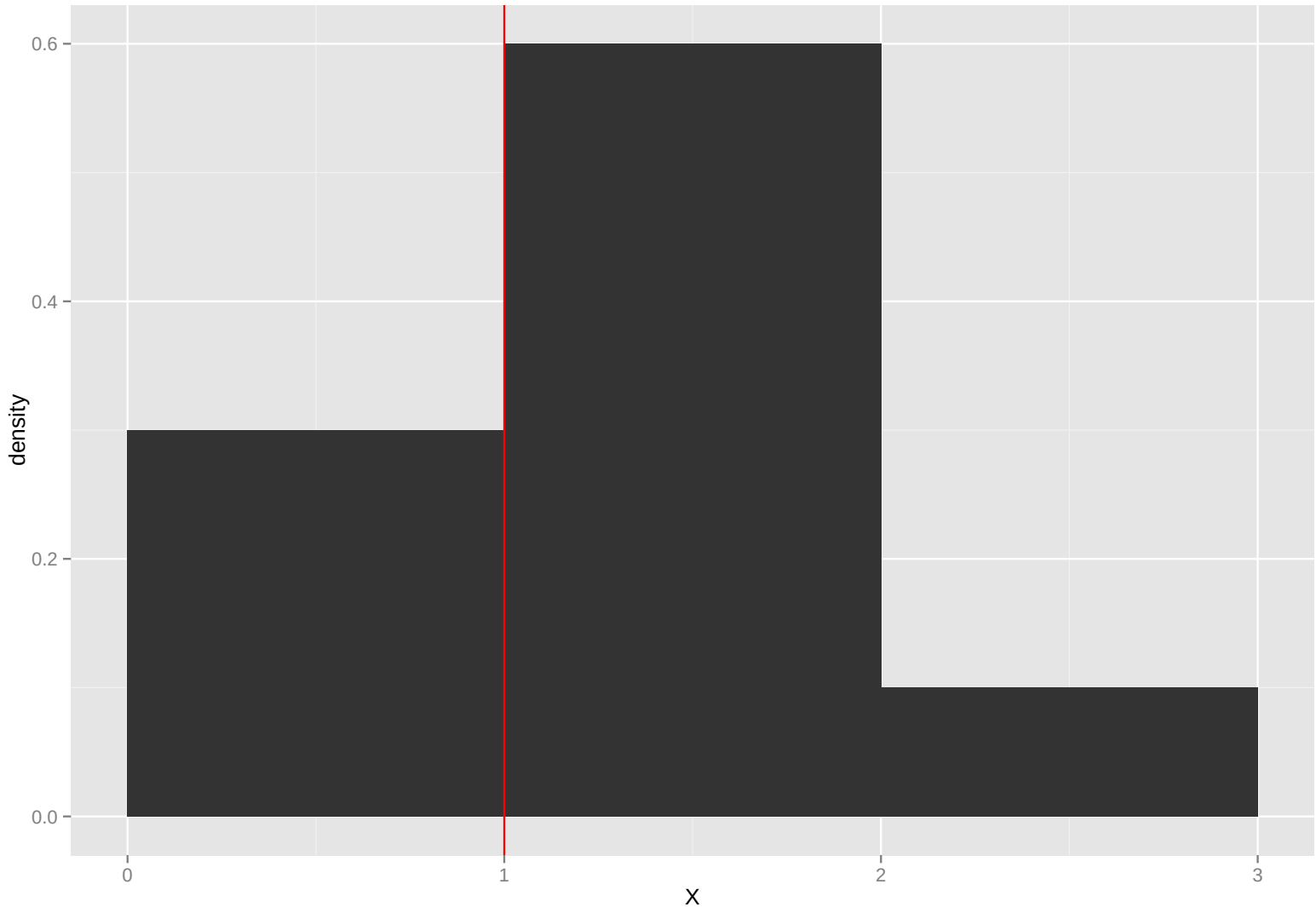
The Randomization Distribution of X



The Randomization Distribution of X



The Randomization Distribution of X



Parametric Null Hypothesis Significance Testing

- Specify and assume H_0
- Define H_A
- Examine reference dist'n (t, χ^2, ...) under H_0
- Calculate p-value
- Compare to some α; reject H_0 if $p < \alpha$

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)
- Define H_A

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)
- Define H_A
- Create reference dist'n from all possible values of X under H_0
  (or at least a big sample of them)

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)
- Define H_A
- Create reference dist'n from all possible values of X under H_0
  (or at least a big sample of them)
- What prop. of possible "at least as extreme as" observed?
  $\rightsquigarrow$ p-value!

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)
- Define H_A
- Create reference dist'n from all possible values of X under H_0
  (or at least a big sample of them)
- What prop. of possible "at least as extreme as" observed?
  $\rightsquigarrow$ p-value!
- Compare to some α; reject H_0 if $p < \alpha$

Randomization Inference

- Specify and assume H_0
  (sharp null of no treatment effect)
- Define H_A
- Create reference dist'n from all possible values of X under H_0
  (or at least a big sample of them)
- What prop. of possible "at least as extreme as" observed?
  $\rightsquigarrow$ p-value!
- Compare to some α; reject H_0 if $p < \alpha$
- CA ballot ordering effects (JASA 2006)

Randomization Inference

The RI p-value is

$$p = \frac{\#\text{ outcomes } \geq \text{as extreme as obs}}{\text{total } \#\text{ outcomes}}$$

Randomization Inference

The RI p-value is

$$p = \frac{\#\text{ outcomes } \geq \text{ as extreme as obs}}{\text{total } \#\text{ outcomes}}$$

or

$$p = \frac{\#\text{ randomizations producing extreme } \widehat{ATE}}{\text{total } \#\text{ randomizations}}$$

Randomization Inference

The RI p-value is

$$p = \frac{\#\text{ outcomes } \geq \text{ as extreme as obs}}{\text{total } \#\text{ outcomes}}$$

or

$$p = \frac{\#\text{ randomizations producing extreme } \widehat{ATE}}{\text{total } \#\text{ randomizations}}$$

How many randomizations are there?

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

How many ways to choose 5 villages of 10 for treatment?

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

How many ways to choose 5 villages of 10 for treatment?

$$ {}_{10}C_5 = \binom{10}{5} = \frac{10!}{5!(10-5)!} $$

Combinations: Counting selected sets

How many ways to **select** k things from a set of n things?

$$ {}_nC_k = \binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} $$

How many ways to choose 5 villages of 10 for treatment?

$$ {}_{10}C_5 = \binom{10}{5} = \frac{10!}{5!(10-5)!} $$

$$ \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252 $$

Common Assumptions, Null Hypotheses

- Constant effect:

$$\tau_i = Y_{i1} - Y_{i0} = \tau \quad \forall i$$

- Null hypothesis of no average effect:

$$ATE = \overline{\tau} = 0$$

- Sharp null hypothesis of no effect:

$$\tau_i = 0$$

An Assignment Mechanism: Perfect Doctor

Calculate RI p-value for Perfect Doctor, under sharp null.

Patient	Y(0)	Y(1)	τ	T
1	(1)	6	(5)	1
2	(3)	12	(9)	1
3	9	(8)	(-1)	0
4	11	(10)	(-1)	0
Mean	10	9	(3)	

An Assignment Mechanism: Perfect Doctor

Calculate RI p-value for Perfect Doctor, under sharp null.

Patient	Y(0)	Y(1)	τ	T
1	(1)	6	(5)	1
2	(3)	12	(9)	1
3	9	(8)	(-1)	0
4	11	(10)	(-1)	0
Mean	10	9	(3)	

(See `03-ri-perfect-dr.R`)

RI versus the t-test

Perfect Doctor:

- RI: $p = 1$
- `t.test()`: $p \approx 0.8$
- "If no tr effect, then this result typical"

RI versus the t-test

Perfect Doctor:

- RI: $p = 1$
- `t.test()`: $p \approx 0.8$
- "If no tr effect, then this result typical"

(Odd logic of NHST: "assume false thing, how strange is data?"")

Randomization Inference

- Resume audit study, Bertrand and Mullainathan (2004)

	0	1
black	2278	157
white	2200	235

Randomization Inference

- Resume audit study, Bertrand and Mullainathan (2004)

	0	1
black	2278	157
white	2200	235

- Only possible values: $\tau_i \in \{-1, 0, 1\}$

Randomization Inference

- Resume audit study, Bertrand and Mullainathan (2004)

	0	1
black	2278	157
white	2200	235

- Only possible values: $\tau_i \in \{-1, 0, 1\}$

```
resume %>% group_by(race) %>% summarise(call_rate = me
```

```
## # A tibble: 2 x 2
##   race  call_rate
##   <fct>     <dbl>
## 1 black    0.0645
## 2 white    0.0965
```

- Assume the sharp null $\tau_i = 0$ for every employer.

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

$$ {}_{4870}C_{2435} = \binom{4870}{2435} = \frac{4870 \cdot 4869 \cdot \ldots \cdot 2436}{2435!} $$

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

$$ {}_{4870}C_{2435} = \binom{4870}{2435} = \frac{4870 \cdot 4869 \cdot \ldots \cdot 2436}{2435!} $$

$$ \approx 1.1 \times 10^{1464} $$

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

$$ {}_{4870}C_{2435} = \binom{4870}{2435} = \frac{4870 \cdot 4869 \cdot \ldots \cdot 2436}{2435!} $$

$\approx 1.1 \times 10^{1464}$

(There are $\approx 10^{86}$ fundamental particles in the universe.)

- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

$$ {}_{4870}C_{2435} = \binom{4870}{2435} = \frac{4870 \cdot 4869 \cdot \ldots \cdot 2436}{2435!} $$

$\approx 1.1 \times 10^{1464}$

(There are $\approx 10^{86}$ fundamental particles in the universe.)

- Let's do 1000, or 100,000 – something reasonable

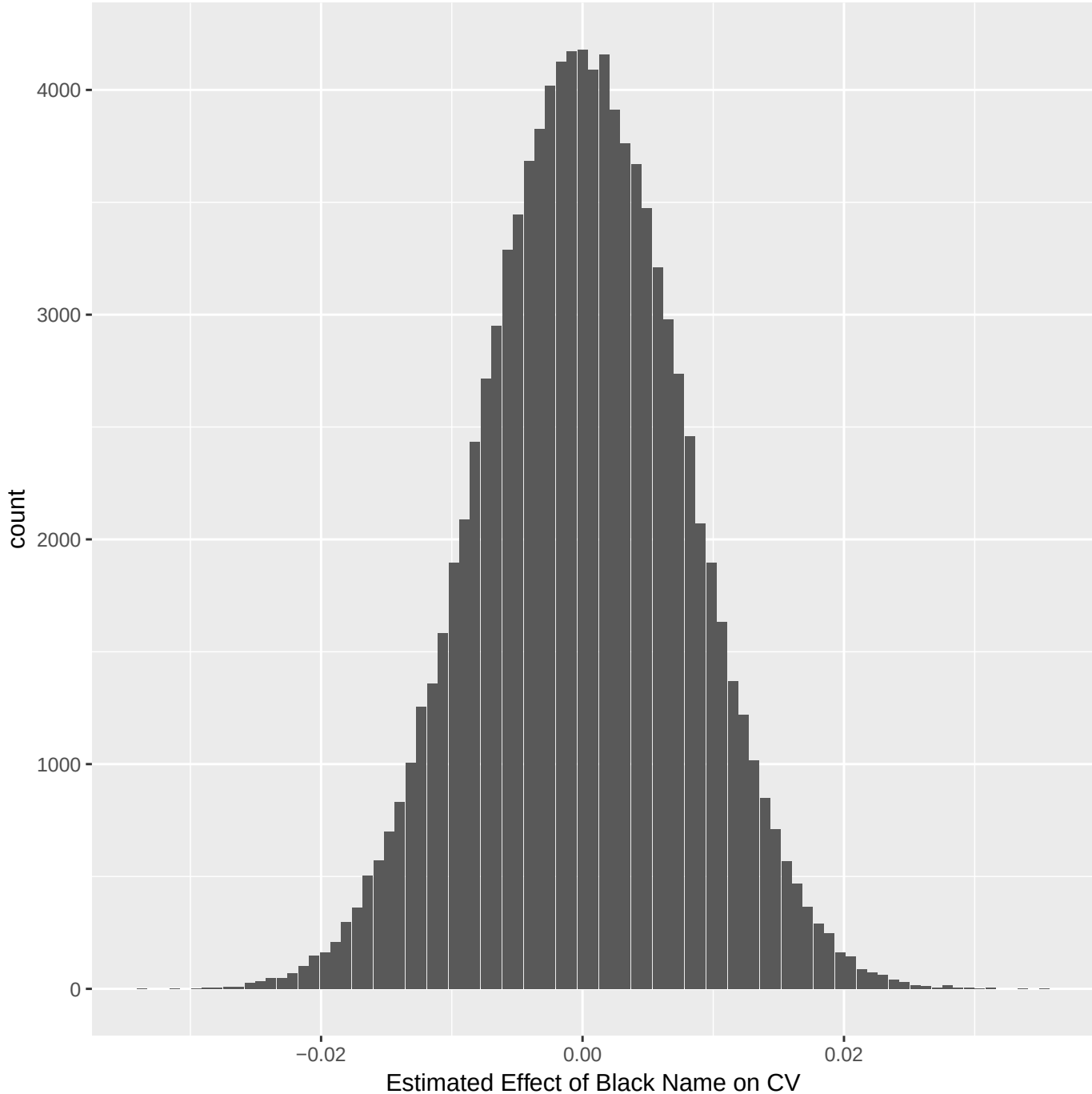
- Assume the sharp null $\tau_i = 0$ for every employer.
- $H_0 : \mu_{\text{black name}} = \mu_{\text{white name}}$
- $H_A : \mu_{\text{black name}} \neq \mu_{\text{white name}}$
- Create reference dist'n of all possible assignments

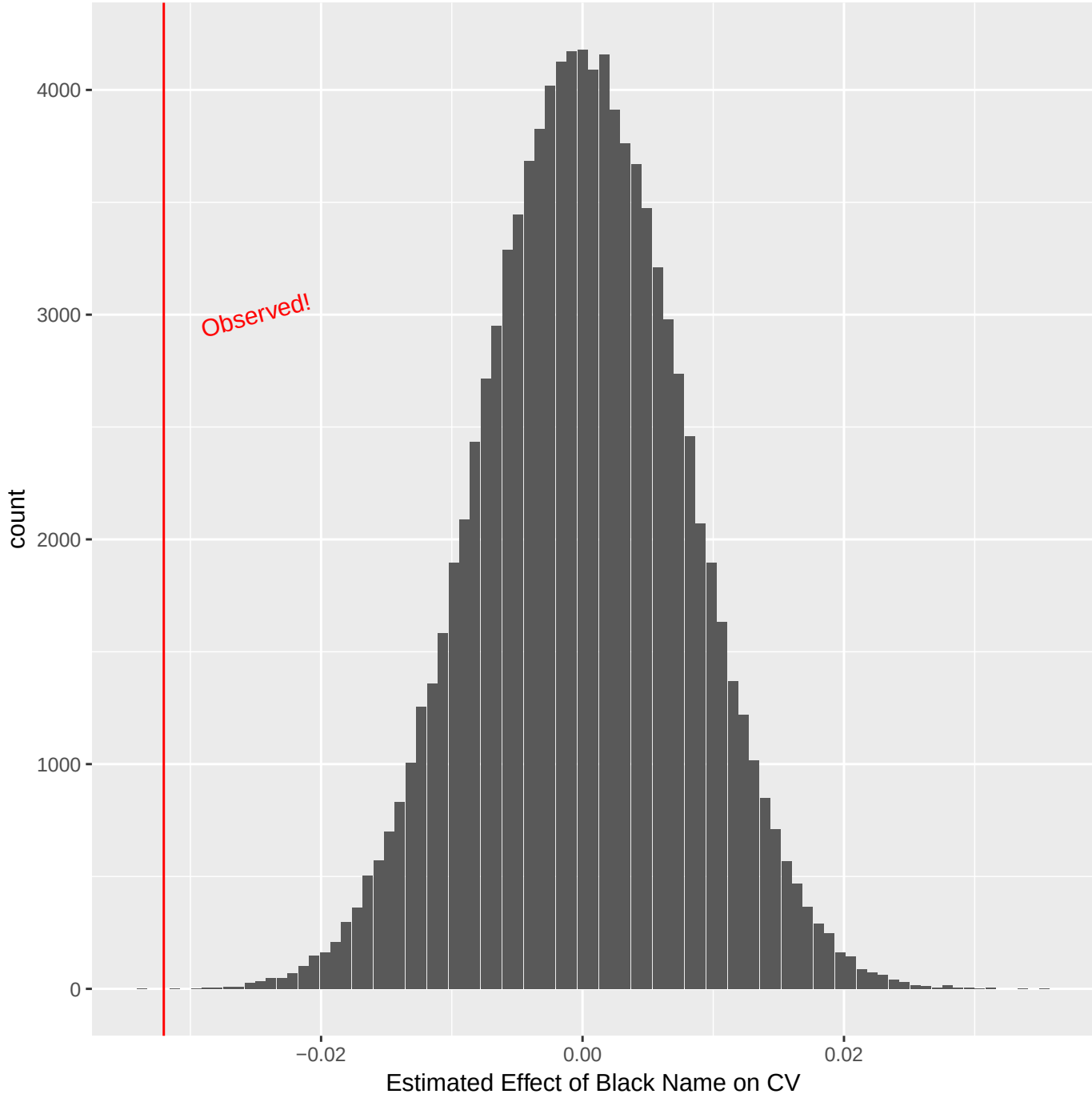
$$_{4870}C_{2435} = \binom{4870}{2435} = \frac{4870 \cdot 4869 \cdot \ldots \cdot 2436}{2435!}$$

$\approx 1.1 \times 10^{1464}$

(There are $\approx 10^{86}$ fundamental particles in the universe.)

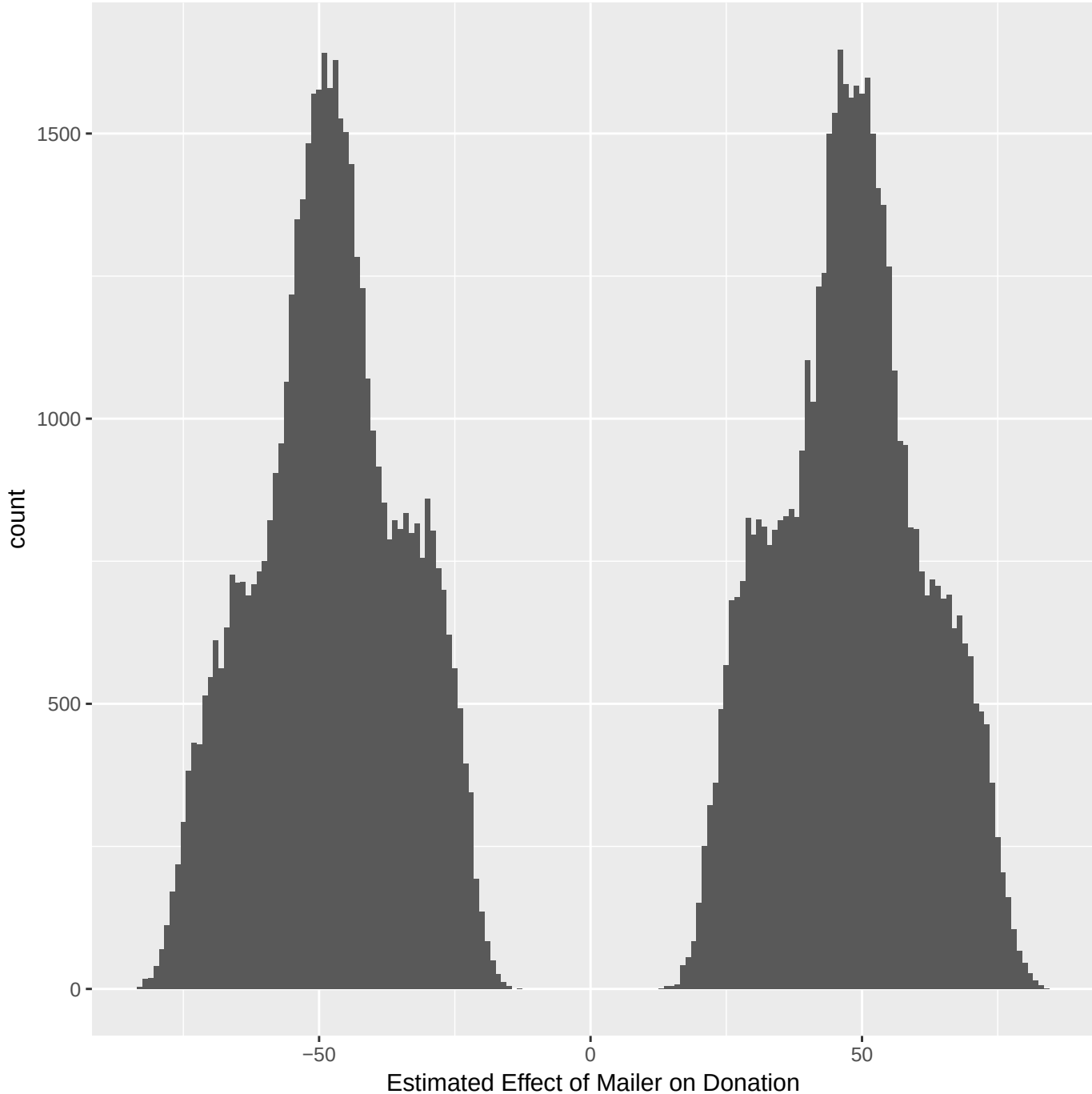
- Let's do 1000, or 100,000 – something reasonable
- See `02-ri-resume-donate.R`

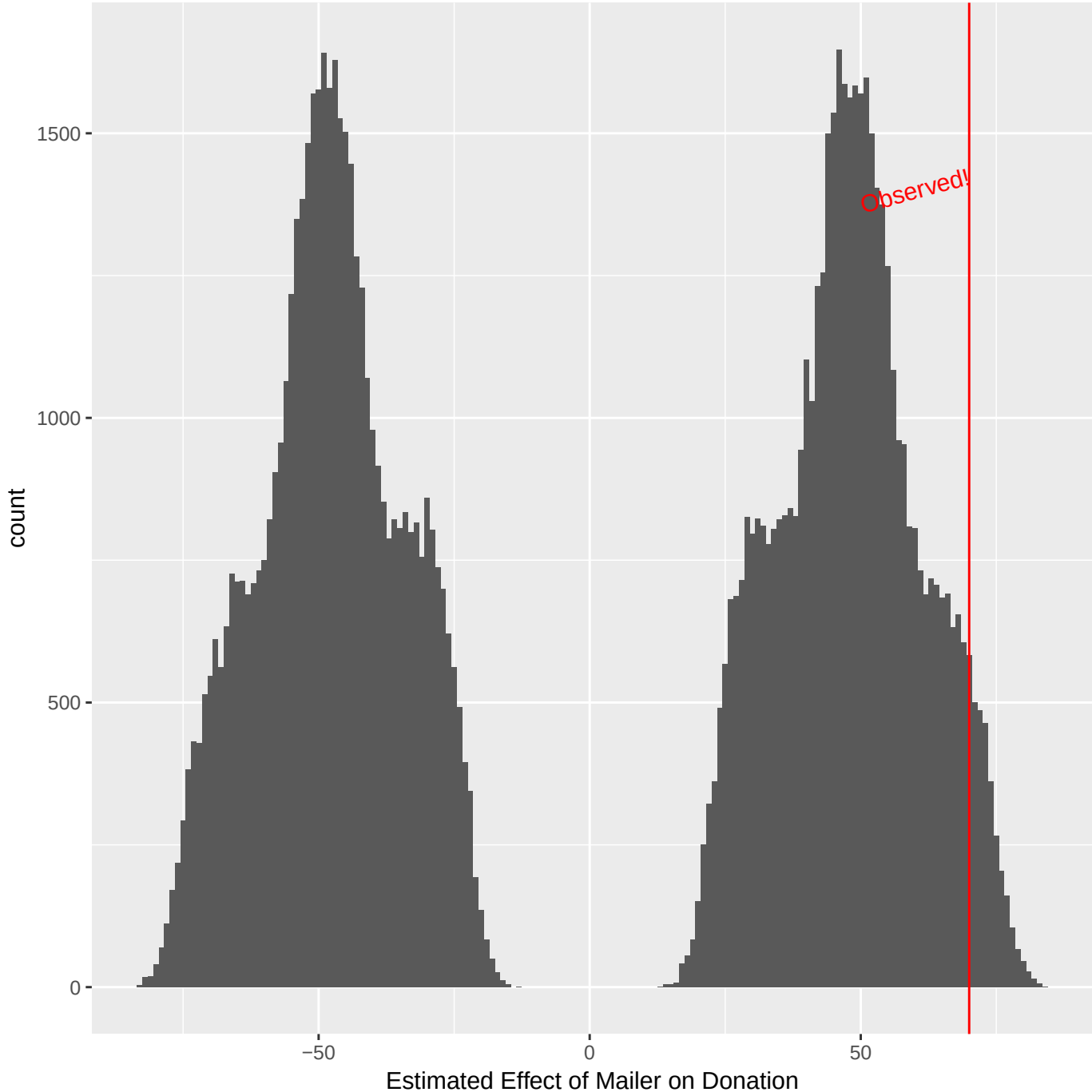

4000
3000
2000
1000
0
count
−0.02
0.00
0.02
Estimated Effect of Black Name on CV


Observed!
4000
3000
2000
1000
0
count
−0.02
0.00
0.02
Estimated Effect of Black Name on CV

Randomization Inference

- Gerber & Green donations example, p. 65
- Possible values $\tau_i \in (-\infty, \infty)$
- Y_1, Y_0, τ likely very skewed
- See `02-ri-resume-donate.R`


1500
1000
500
0
count
−50
0
50
Estimated Effect of Mailer on Donation


1500
1000
500
0
count
Observed!
−50
0
50
Estimated Effect of Mailer on Donation

The RI Confidence Interval

Recall that

"reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

The RI Confidence Interval

Recall that

> "reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

Create RI confidence intervals

- Posit $H_0 : \tau = \tau^* \in \{\ldots, -2, 1, 0, 1, 2, \ldots\}$

The RI Confidence Interval

Recall that

> "reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

Create RI confidence intervals

- Posit $H_0 : \tau = \tau^* \in \{\ldots, -2, 1, 0, 1, 2, \ldots\}$
- RI test whether to reject H_0

The RI Confidence Interval

Recall that

> "reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

Create RI confidence intervals

- Posit $H_0 : \tau = \tau^* \in \{\ldots, -2, 1, 0, 1, 2, \ldots\}$
- RI test whether to reject H_0
- If not, then τ^* is in CI

The RI Confidence Interval

Recall that

> "reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

Create RI confidence intervals

- Posit $H_0 : \tau = \tau^* \in \{\ldots, -2, 1, 0, 1, 2, \ldots\}$
- RI test whether to reject H_0
- If not, then τ^* is in CI
- CI consists of set of τ^* not unusual, given data

The RI Confidence Interval

Recall that

> "reject H_0 at $\alpha = 0.05$" $\equiv$ "H_0 falls outside 95% CI"

Create RI confidence intervals

- Posit $H_0 : \tau = \tau^* \in \{\ldots, -2, 1, 0, 1, 2, \ldots\}$
- RI test whether to reject H_0
- If not, then τ^* is in CI
- CI consists of set of τ^* not unusual, given data
- See `03-ri-resume-donate.R`

References

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Imai, Kosuke. 2017. *Quantitative Social Science: An Introduction.* Princeton, NJ: Princeton University Press.