# Statistics for Data Science
# Winter Institute in Data Science

Ryan T. Moore

2023-01-05

Warm Up Quiz

Descriptive Statistics

Probability and Bayes Rule

Distributions, Expectation, Variance, the LLN and CLT

Uncertainty: The Confidence Interval

Uncertainty: Null-Hypothesis Significance Testing

Randomization (Design-based) Inference

# Descriptive Statistics

Descriptive statistics: summarize observed features of data

- *Univariate* statistics: describe single variable
- *Bivariate* statistics describe relationship betwn two vars ("are higher values of $X$ assoc'ed w/ higher values of $Y$?")
- *Multivariate* statistics summarise several relatnshps 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

```r
max(times_voted)
```

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

```r
min(times_voted)
```

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

```r
range(times_voted)
```

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

```r
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
```

# Summary Statistics with R

```
IQR(times_voted)
```

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

# Summary Statistics with R

```
IQR(times_voted)
```

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

```
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 average is a typical value of the var?"

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 "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 $\sqrt{\phantom{x}}$.

# 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 linear transformation of $x_i$ with $y_i = ax_i + b$, then

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

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.

$z$-scores can compare variables on different scales, since we divide each variables' values by its *own* SD. Let $X$ be household income, $Y$ be left-right ideology on $[0, 10]$ scale:

$z$-scores can compare variables on different scales, since we divide each variables' values by its *own* SD. Let $X$ be household income, $Y$ be left-right ideology on $[0, 10]$ scale:

| 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 | | |

$z$-scores can compare variables on different scales, since we divide each variables' values by its *own* SD. Let $X$ be household income, $Y$ be left-right ideology on $[0, 10]$ scale:

| 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 | | |

Process of calculating $z$-scores: *standardizing* the variable.

# Correlation

Are larger values of $X$ assoc'd with larger (or smaller?) values of $Y$?

This is question of *correlation* between $X$ and $Y$. When $X$ and $Y$ are *positively correlated*, larger values of $X$ are assoc'd with larger values of $Y$.

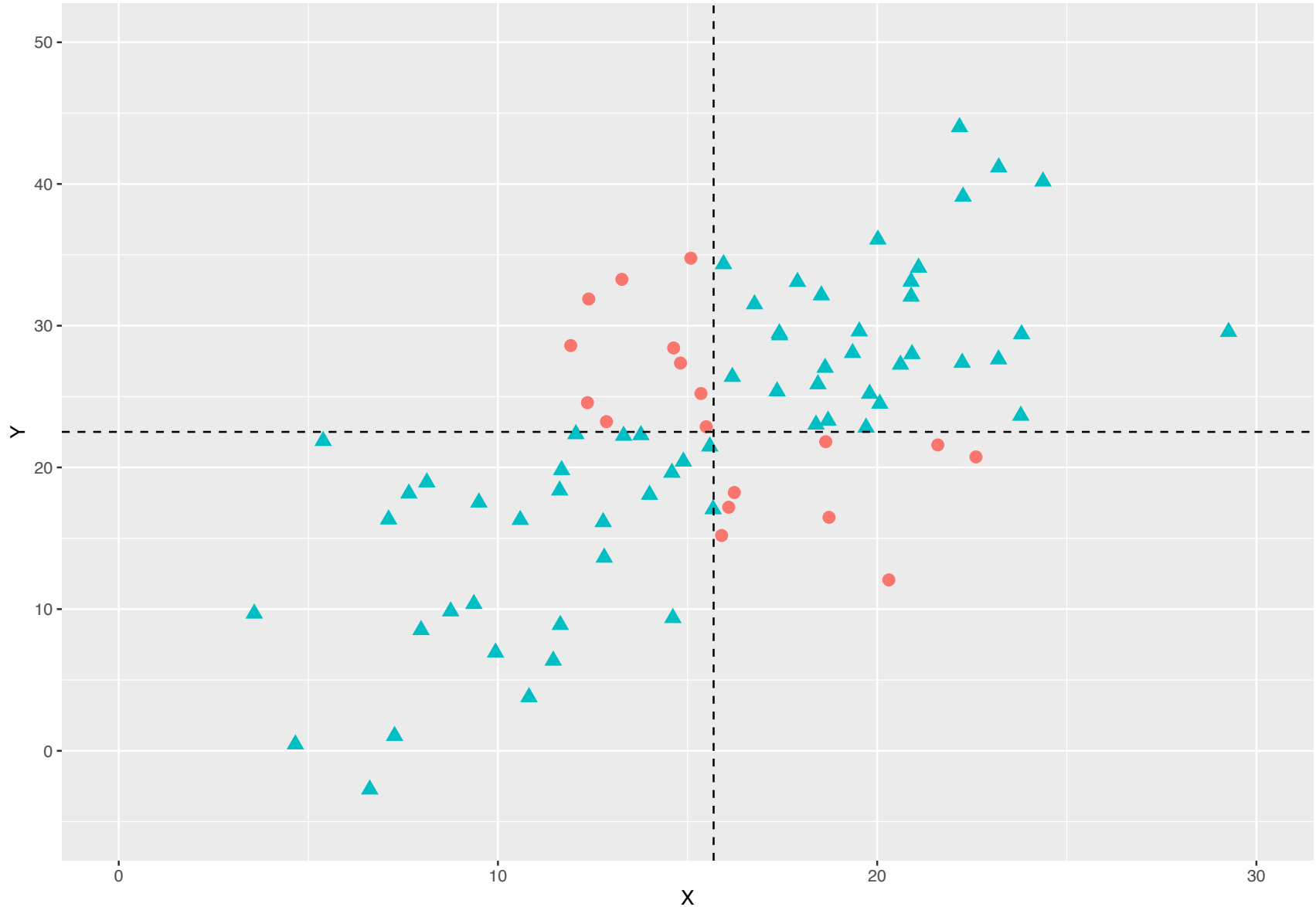



Figure 1: Positive Correlation. Blue triangles outweigh red discs.

On the other hand, when $X$ and $Y$ are negatively correlated, larger values of $X$ tend to be assoc'd with *smaller* values of $Y$.

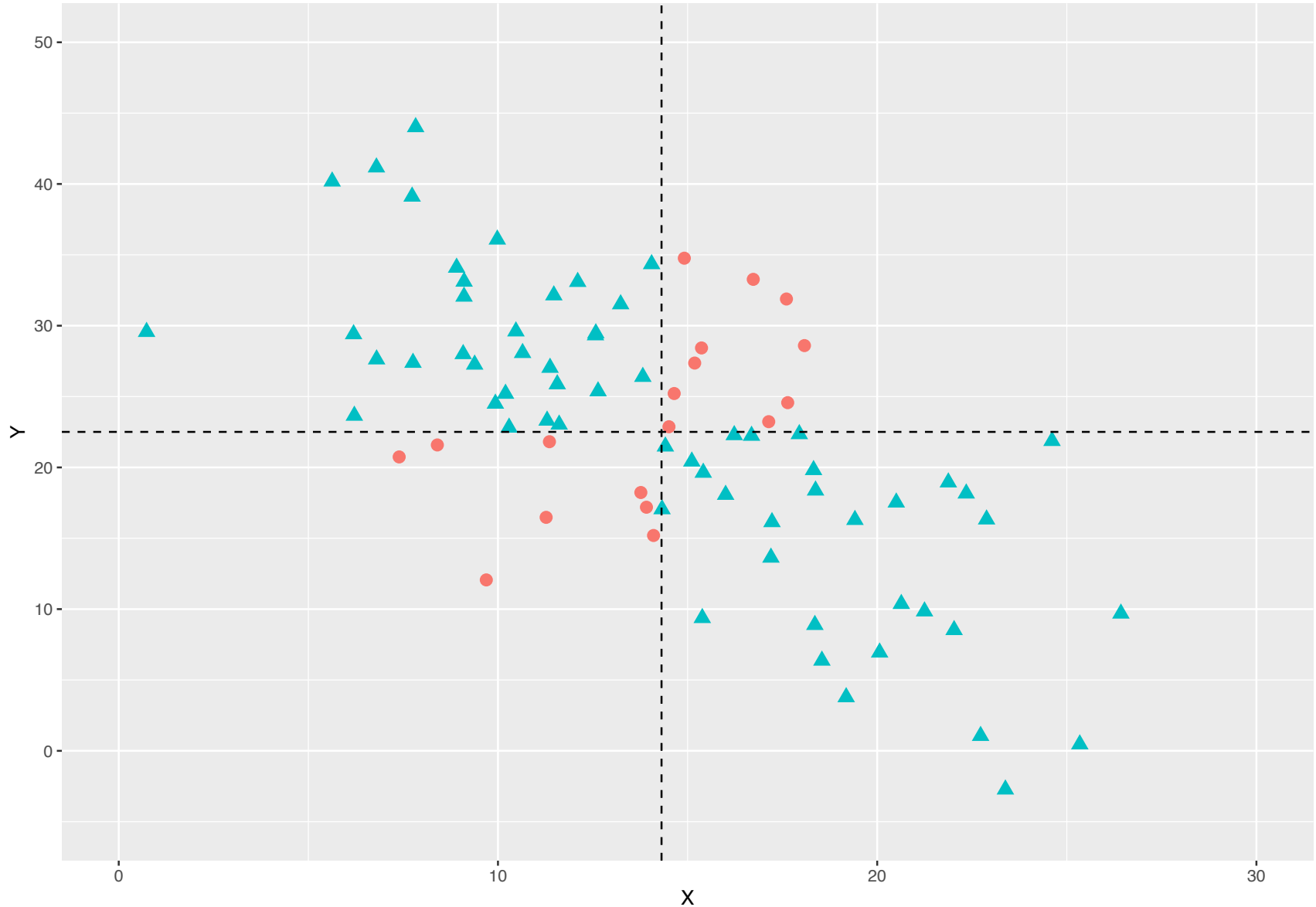



Figure 2: Negative Correlation. Blue triangles outweigh red discs.

Formally, correlation is average of products of $z$-scores:

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

Formally, correlation is average of products of $z$-scores:

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

Positive correlation: on average, $[z_i(x_i) \times z_i(y_i)] > 0$ – only true if both scores positive or both scores negative.

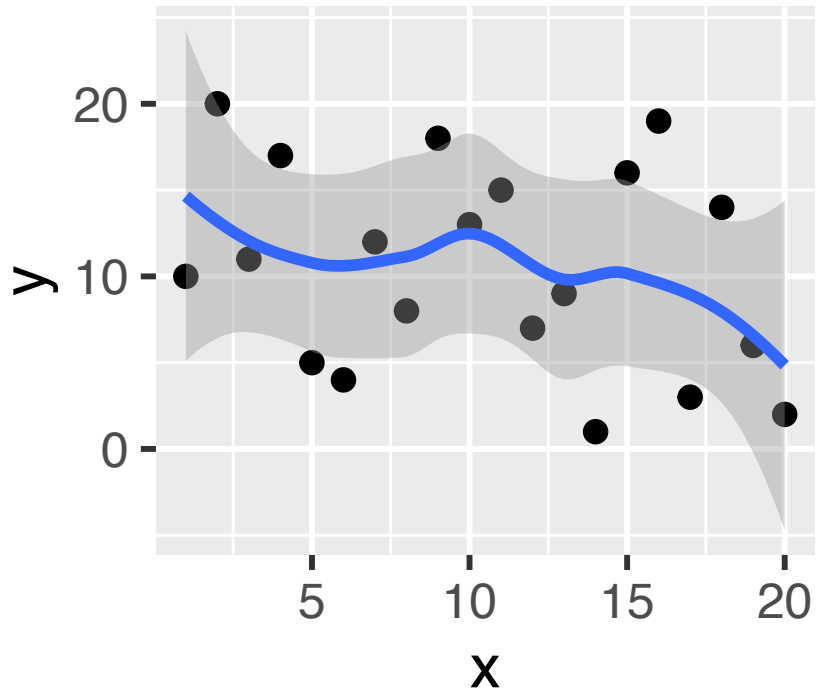
Whether $z_i$ is positive or negative depends on whether unit $i$ is above or below mean on that variable. Its magnitude determined by *how far* above or below average unit $i$ is.

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

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

```
df <- tibble(x = sample(20),
             y = sample(20))

ggplot(df, aes(x, y)) + geom_point() + geom_smooth()
```




```
cor(df$x, df$y)
```

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

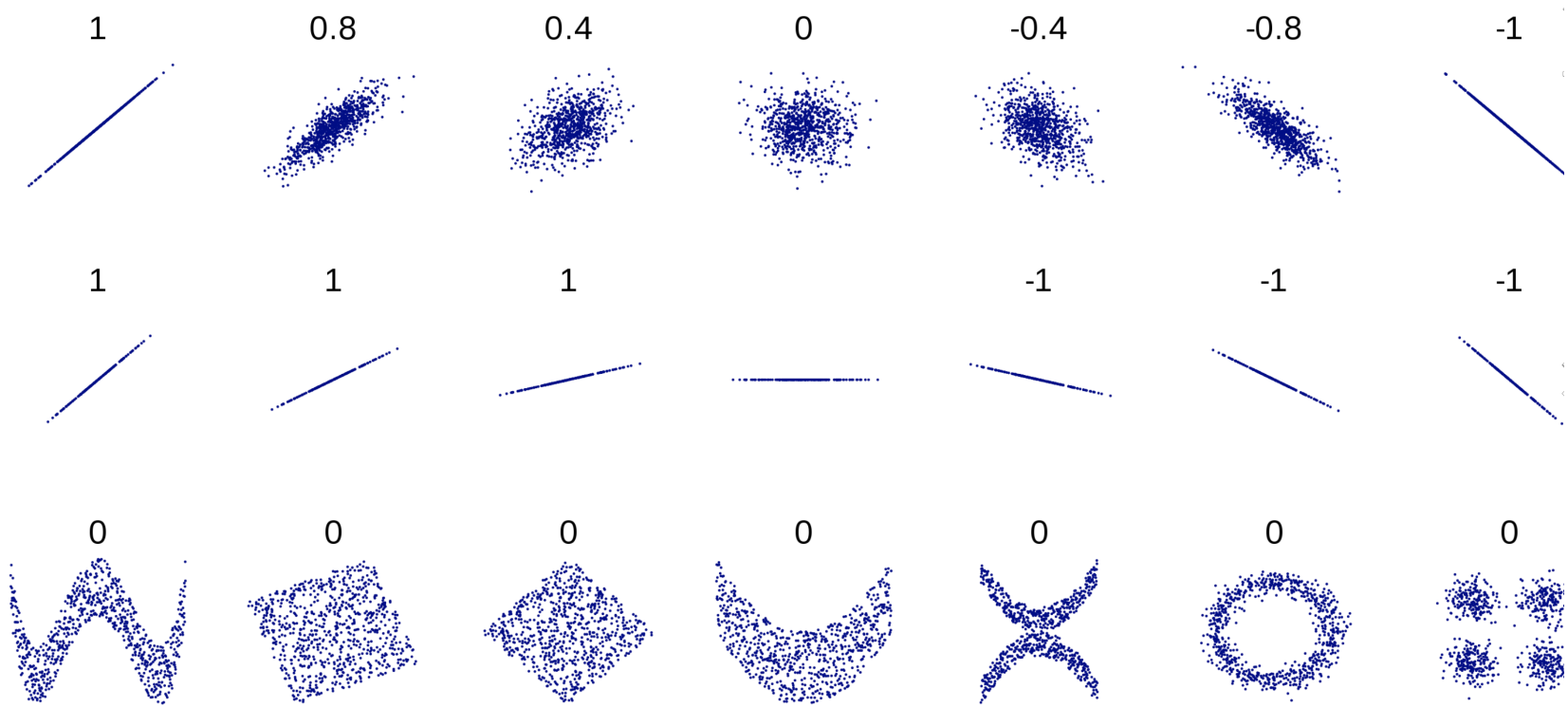



Figure 3: Correlations (Wikipedia)

# Uncertainty: The Confidence Interval

# The Standard Error

*Standard error* (SE) of an estimator: 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.

# The Standard Error

$$SE\left(\hat{p}\right) = \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

$$SE\left(\bar{X}\right) = \sqrt{\frac{Var(X)}{n}}$$

$$SE\left(\bar{X}-\bar{Y}\right) = \sqrt{\frac{Var(X)}{n_X} + \frac{Var(Y)}{n_Y}}$$

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.

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?"

| $\alpha$ | 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 uncertainty with "alpha level" ($\alpha$) instead of confidence level. These are mathematically identical – just different ways to express the same level of uncertainty. $\alpha = 0.05$ corresponds to 95% confidence.

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

# An Example

For a Congress, suppose that probability $\pi$ of bills passing is 0.35. Calculate an 80% confidence interval for a sample of bills. After you create each object, 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
   ```

When you're finished, post your CI to the chat.

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

Note that 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 $t$ Instead of Normal

Often, $t$-distribution will improve upon CI's from Normal distribution. The $t$-distribution is somewhat fatter-tailed than the Normal, implying more variation in data than Normal involves. The $t$ is a family of distributions that are wider when we have less data, narrower when we have more. When we have $n$ observations, we select $t$-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

Data from randomized experiment creating village council seats for women in Indian villages.

```
w_res <- read.csv("https://bit.ly/2KYOiHE")
head(w_res)
```

```
##   GP village reserved female irrigation water
## 1  1       2        1      1          0    10
## 2  1       1        1      1          5     0
## 3  2       2        1      1          2     2
## 4  2       1        1      1          4    31
## 5  3       2        0      0          0     0
## 6  3       1        0      0          0     0
```

Calculate 95% CI for number of `water` projects that happen in villages with reservations for women. Get the estimate, the SE, the critical value (from a $t$), and form 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

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

# Example 2: CI for Difference Betwn Two Sample Means

Calculate the 90% CI for *difference* between water projects in villages with and without reservations for women. Stat 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. Use `t.test()` to calculate the confidence interval directly. `water` is outcome; `reserved` is treatment.

```
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 interval R provides. Note: `t.test()` always takes "first group − second group", which here means "`reserved == 0` − `reserved == 1`".

# 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,

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

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

Find minimum sample size for certain level of precision in a survey. Suppose have survey asking whether Scottish voters support Brexit, and want it to be precise to within 0.03 (three percentage points), with 95% confidence. The largest SE for 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

# Uncertainty: Null-Hypothesis Significance Testing

SEs and CIs quantify uncertainty around estimates. When you report estimates in your final paper, you should provide a measure of the uncertainty around them.

*Null hypothesis significance testing* (NHST)

Idea: *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: "how strange would the data be, if, in fact, there is no relationship between $X$ and $Y$?"

A bit odd, since usually think there might be a relationship – that's what caused us to investigate this phenomenon in the first place.

Caveat: causal interpretations can be valid when design is good, but **no test result by itself indicates causality**. Linear regression coefficients, $R^2$, $p$-values below are not inherently imbued with causal meaning. That only comes 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 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 $\alpha$. 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.

6. Standardize the test statistic. Subtract off the null hypothesis value and divide by the SE.

For four tests,

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

7. 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.
8. Compare the $p$-value to $\alpha$. 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**.

# 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
```

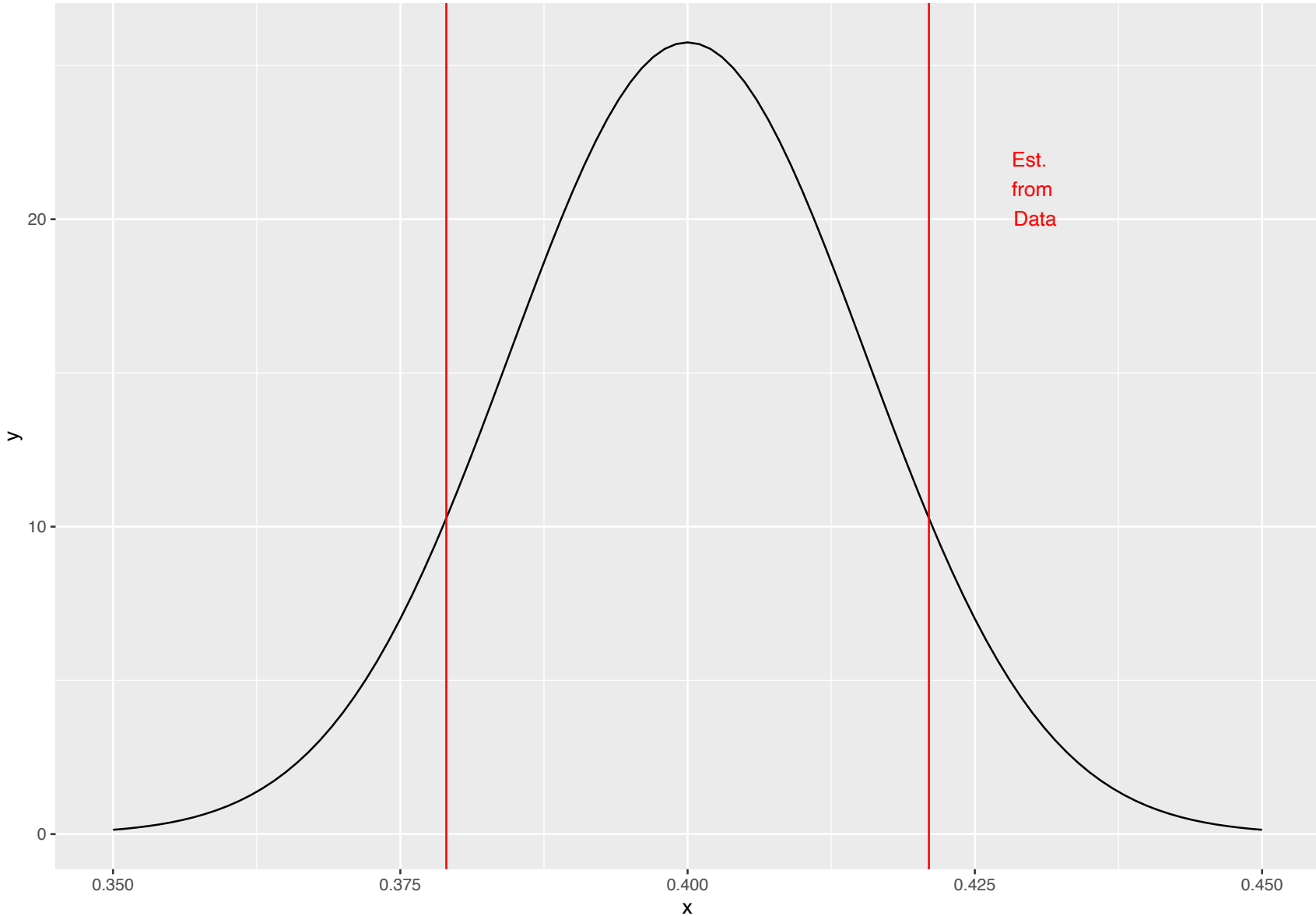
8. Here $0.1752442 > \alpha$, so we **do not reject** $H_0$. We do **not** have evidence that the true $\pi$ is different from 0.4.

# More Detail

How to figure out if 1.4 SE's is "too far" away? 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 $\alpha$. If $p < \alpha$, it's too far, and we reject the null hypothesis. If $p > \alpha$, we do not reject it.

Since testing proportions, reference dist'n is normal dist'n.

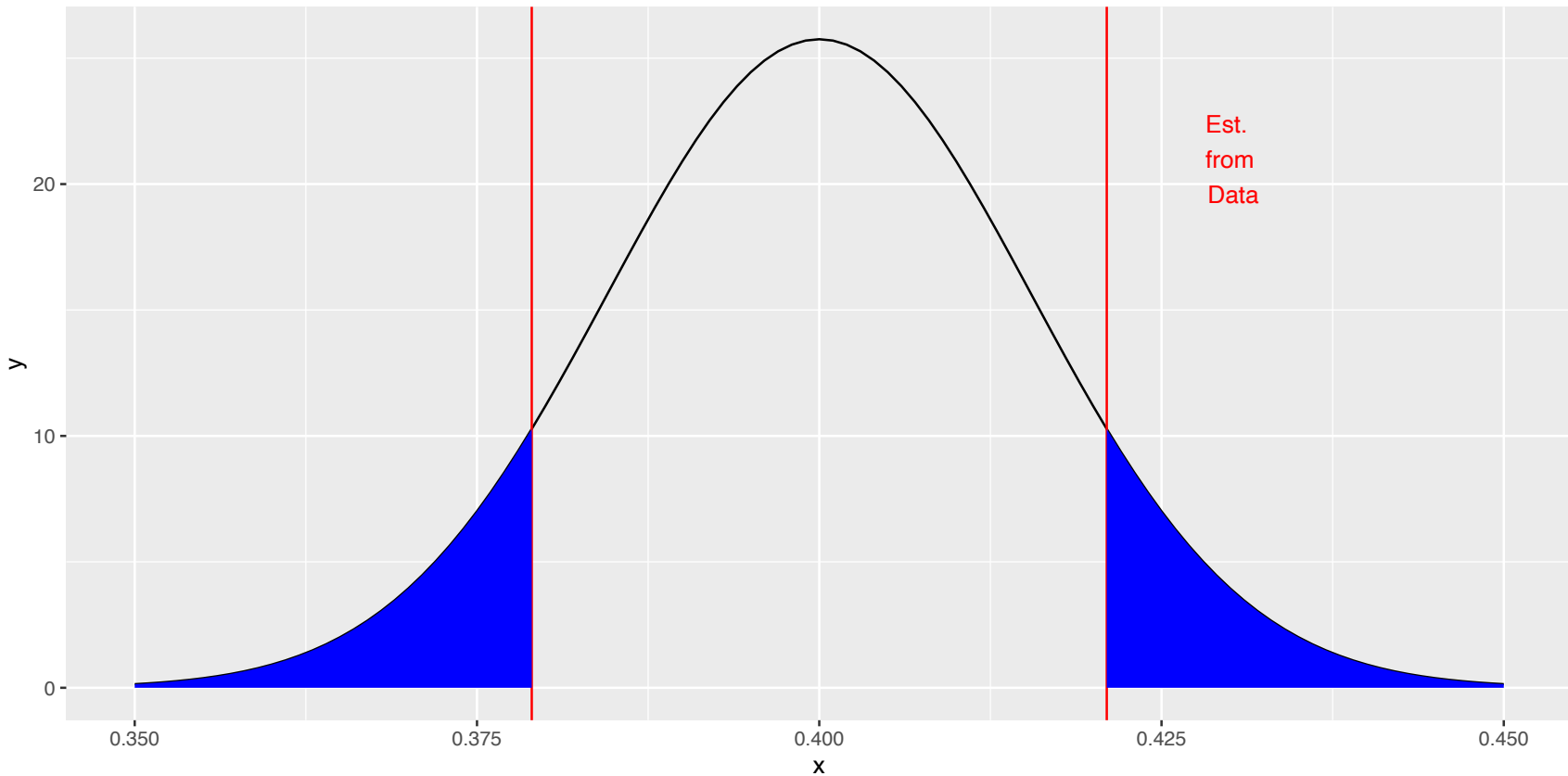
Normal dist'n, mark $-1.4$, $+1.4$ SE's from hypo ($40\% \pm 2.1\%$):




Set threshold for rejecting $H_0$, denoted $\alpha$. (Like inverse of confidence level, on different scale. E.g., for a 90% confidence

Set threshold for rejecting $H_0$; say, $\alpha = 0.1$.

Calculate $p$-value of est, assuming $H_0$ true. $p$-value = area under ref dist'n *as extreme or more extreme* than data we saw. What fraction as or more extreme than $40 \pm 2.1\%$? $p$-value area:



To calculate exact proportion of 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,
                            lower.tail = FALSE)
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,
                   lower.tail = TRUE)
prop_left
```

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

Finally, we compare the $p$-value to our threshold $\alpha$. 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

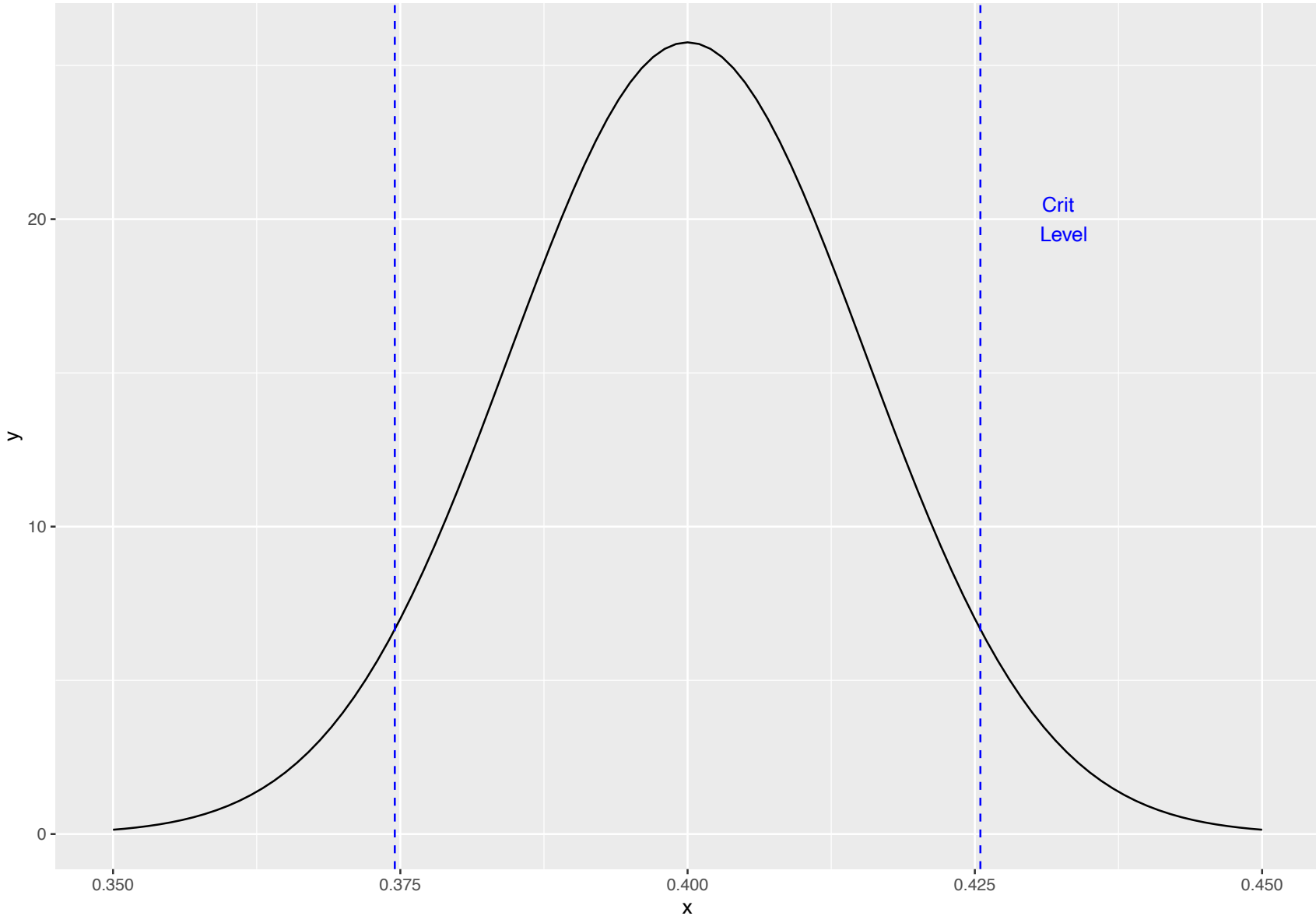
Another way of getting the same result is to ask, "what is the largest number of SE's away from the null we could be to not be considered 'too far'?".

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:

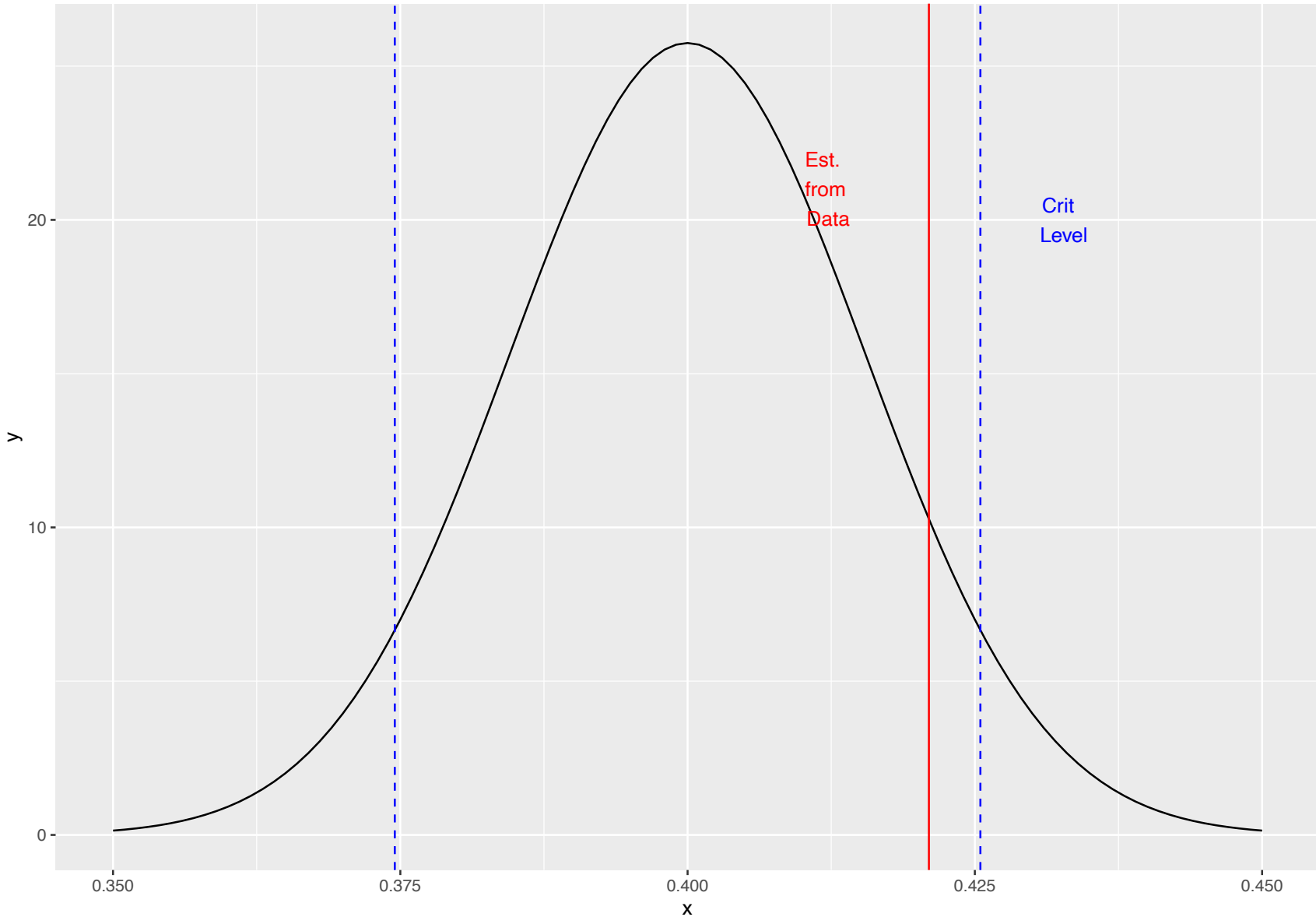
```
alpha <- .1
qnorm(alpha / 2)  ## divide by 2: half prob on each side
```

```
## [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

Data from randomized experiment creating village council seats for women in Indian villages.

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

Test whether mean number of water projects for villages w/o reservations is stat. 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

```
n <- length(water_proj_no_res)
se <- sqrt(var(water_proj_no_res) / n)
se
```

```
## [1] 1.298276
```

6. Standardize. How many SE's away is this?

```
mu <- 15
how_many_ses <- (x.bar - mu) / se
how_many_ses  # the t-statistic
```

```
## [1] -0.2015614
```

7. Calculate $p$-value. For test of one mean, use `df = n - 1`.

```
prop_left_of_data <- pt(how_many_ses, df = n - 1)
2 * prop_left_of_data
```

```
## [1] 0.840452
```

(If `x.bar` were to the right of $H_0$, use `lower.tail = FALSE`.)

8. Since `2 * prop_left_of_data` $\approx 0.84$, which is much greater than $\alpha = .1$, we **do not reject** the $H_0$.

# 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()`

Use `t.test()` to test whether 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 betwee
## 90 percent confidence interval:
##  -17.7058080  -0.7990379
## sample estimates:
```

# 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.)

# 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.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* New York, NY: WW Norton.

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