

R Lab 2: What Impacts Strength of Evidence?

Do Olympic athletes in certain uniform colors have an advantage over their competitors?

Team Captain: NAME HERE Facilitator: NAME HERE
Recorder: NAME HERE Resource Manager: NAME HERE

Invalid Date

```
# Load in the packages necessary for our analysis: tidyverse
library(tidyverse)
library(infer)
```

Officials noticed that competitors in the combat sports of boxing, tae kwon do, Greco-Roman wrestling, and freestyle wrestling are randomly assigned red or blue uniforms. For each match in the 2004 Olympics, they recorded the uniform color of the athlete who won.

```
# load the data
athletes <- read_csv("data/athletes.csv")
```

1. Click on the data set to open it up in the data previewer. What is the observational unit for this study?

Your answer goes here.

Visualizing One Categorical Variable

Let's start by exploring how many matches in our data set were won by each uniform color

2. Create a *frequency bar plot* of the variable color.

Step 1: Set-up the data and the variable(s) of interest

Step 2: Define what geometric object should be plotted on the plot – now we're using bars!

Step 3: Declare nice x- and y-axis labels

Let's carry out these steps in the code chunk below!

```
ggplot(data = athletes,
       mapping = aes(x = ____))
) +
geom_bar(stat = "count") +
labs(title = "____",
     x = "____",
     y = "____")
```

Error: <text>:2:27: unexpected input

```
1: ggplot(data = athletes,
2:       mapping = aes(x = __
                        ^
```

We could also choose to display the data as a proportion in a *relative frequency bar plot*. To find the relative frequency, we divide the count in each level by the total sample size. The resulting values are sample proportions.

To get *proportions* instead of *counts* we need to override the default statistic `geom_bar(stat = "count")` wants to use. To do this, we tell it we want proportions on the y-axis. We do this using the following code:

```
geom_bar(stat = "count",
        aes(y = ..prop.., group = 1)
)
```

Notice, there are two periods (..) before **and** after prop. These periods are necessary to define a new statistic that should be used for the y-axis.

3. Use this new code to modify the code you had previously. Copy-and-paste your code from the frequency bar plot above and change the y-axis to plot proportions instead of counts. Make sure to change your y-axis label to match proportions rather than counts!

```
# Copy-and-paste your code from #2
# Add the aes() input to geom_bar() so your plot has proportions instead of counts!
```

4. Try removing the `group = 1` input from `geom_bar()`. What happens? What do your bars look like? Once you've answered the question, make sure to put the `group = 1` back in to your code!

Your answer goes here.

Summary statistics for categorical variables

The other part of Exploratory Data Analysis (EDA) is making summary statistics. For a categorical variable, this means making frequency tables and relative frequency tables. Let's explore how to do that.

For this piece you are going to learn about a tool – “the pipe.” This is a special operator written as `|>`. The pipe is used as a connecting piece to keep your data flowing through a pipeline. We will use the pipe to string together two R functions:

- `count()`, which counts the number of observations (rows) for each group of a categorical variable
- `mutate()`, which modifies our dataset by adding new variables or changing existing variables

Let's start with a short data pipeline.

```
athletes |>
  count()
```

```
# A tibble: 1 x 1
      n
<int>
1  457
```

5. What does the variable `n` tell us?

Your answer goes here.

We can add another piece to the pipeline to make it a bit more interesting. Let's add the `won` variable into the `count()` function and see what we get!

```
athletes |>
  count(_____)
```

```
Error: <text>:2:10: unexpected input
1: athletes |>
2:   count(_
  ^
```

6. How many matches resulted in the red team winning? How many resulted in the blue team winning?

Your answer goes here.

Finally, we might want to add one more piece to the pipeline to create a `proportion` variable of the relative frequencies of each level of `won`.

```
athletes |>
  count(won) |>
  mutate(proportion = n / sum(n))
```

```
# A tibble: 2 x 3
  won      n proportion
<chr> <int>     <dbl>
1 blue   209     0.457
2 red    248     0.543
```

7. What proportion of matches were won by athletes wearing red uniforms? Is this a statistic or a parameter?

Your answer goes here.

Scenario One: Original

Research Question: Do competitors that wear red uniforms win a *majority* of the time?

8. State the null and alternative hypotheses in words.

- Null:
- Alternative:

9. Check the conditions for using the normal approximation method to test our hypotheses.

10. Run the code below. Based on the output, what decision do you make about your hypotheses? Justify your answer using the p-value.

```
binom.test(x = 248, n = 457, p = 0.5, alternative = "greater")
```

Exact binomial test

data: 248 and 457

```
number of successes = 248, number of trials = 457, p-value = 0.03768
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5031206 1.0000000
sample estimates:
probability of success
      0.5426696
```

11. Write a conclusion in context of the scenario.

Scenario Two: One Sided vs. Two Sided

Research Question: Does one color win more often than the other?

12. State the null and alternative hypotheses in words.
 - Null:
 - Alternative:
13. Copy and change the code from question 10 to conduct a two-sided test.
14. Compare to scenario one (original). How is the p-value different? What is similar and what is different about the hypothesis? Are the conclusions the same?

Key Idea

The p-value for a two-sided test is “about” twice as large as that for a one-sided test. Thus the strength of evidence **decreases**. (i.e. reject the null less often).

Scenario Three: Distance Between the Statistic and the Null

Research Question: Do competitors that wear red uniforms win a majority of the time?

- Recall: Going back to scenario one (one-sided hypothesis), we had the following information: hypothesized null = 0.5 and observed statistic = 0.543.
 - Now suppose 261 matches were won by competitors wearing red (i.e., an observe a statistic of 0.571).
15. How does the distance between our observed statistic and hypothesized null change? Might be helpful to draw a picture.

16. Run the code below. What is our p-value?

```
binom.test(x = ____, n = 457, p = 0.5, alternative = "greater")
```

Error: <text>:1:17: unexpected input

```
1: binom.test(x = __  
               ^
```

17. Is your p-value larger or smaller than your original one? Explain why this makes sense.

Key Idea

As the observed sample statistic moves farther away from the hypothesized value, the strength of evidence increases. (i.e. reject the null more often.)

Scenario 4: Effect of Sample Size

Research Question: Do competitors that wear red uniforms win a majority of the time?

- Recall: In scenario one (original) we had a sample size of 457 and a statistic of 0.543.
- In boxing, the researchers found that out of the 272 boxing matches 150 of them were won by competitors wearing red.

(Note: the sample statistic of $150/272 = 0.551$ is similar to the original sample statistic from scenario one - original of 0.543.)

Run the code below to filter out a *subset* containing only the boxing matches.

```
# filter the boxing data  
boxing <- athletes |>  
  filter(sport == "boxing")  
head(boxing)
```

```
# A tibble: 6 x 2  
  sport won  
  <chr> <chr>  
1 boxing blue  
2 boxing blue  
3 boxing blue
```

```
4 boxing blue
5 boxing red
6 boxing red
```

19. Run the code below to test our hypothesis.

```
binom.test(x = _____, n = _____, p = 0.5, alternative = "greater")
```

Error: <text>:1:17: unexpected input

```
1: binom.test(x = __
  ^
```

20. Compare to the original scenario. Note the probability of success. Is the p-value larger or smaller?

Key Idea

As the **sample size increases** (and the value of the observed sample statistic stays the same) the strength of evidence **increases**. (i.e. reject the null more).

As the **sample size decreases** (and the value of the observed sample statistic stays the same) the strength of evidence **decreases**. (i.e. reject the null less).