Assessment: Confounding (Verified Learners only)

For this set of exercises, we examine the data from a <u>2014 PNAS paper that analyzed success rates from funding agencies in the Netherlands</u> and concluded:

> *"our results reveal gender bias favoring male applicants over female applicants in the prioritization of their "quality of researcher" (but not "quality of proposal") evaluations and success rates, as well as in the language used in instructional and evaluation materials."*

A response was published a few months later titled <u>No evidence that gender contributes to personal research funding success in The Netherlands: A reaction to Van der Lee and Ellemers</u>, which concluded:

> *However, the overall gender effect borders on statistical significance, despite the large sample. Moreover, their conclusion could be a prime example of Simpson's paradox; if a higher percentage of women apply for grants in more competitive scientific disciplines (i.e., with low application success rates for both men and women), then an analysis across all disciplines could incorrectly show "evidence" of gender inequality.*

Who is right here: the original paper or the response? Here, you will examine the data and come to your own conclusion.

The main evidence for the conclusion of the original paper comes down to a comparison of the percentages. The information we need was originally in Table S1 in the paper, which we include in **dslabs**:

```
library(dslabs)
data("research_funding_rates")
research_funding_rates
```

Question 1

2.0/2.0 points (graded)

Construct a two-by-two table of gender (men/women) by award status (awarded/not) using the total numbers across all disciplines.

What is the number of men not awarded?

1345

What is the number of women not awarded?

1011

✔ **Answer:** 1011

1011

Explanation
The two-by-two table can be constructed using the following code:

```
two_by_two <- research_funding_rates %>%
    select(-discipline) %>%
    summarize_all(funs(sum)) %>%
    summarize(yes_men = awards_men,
              no_men = applications_men - awards_men,
              yes_women = awards_women,
              no_women = applications_women - awards_women) %>%
    gather %>%
    separate(key, c("awarded", "gender")) %>%
    spread(gender, value)
two_by_two
```

Submit   You have used 2 of 10 attempts

ⓘ   Answers are displayed within the problem

Question 2

2.0/2.0 points (graded)
Use the two-by-two table from Question 1 to compute the percentages of men awarded versus women awarded.

What is the percentage of men awarded?
Report a percentage between 0 and 100.

17.737

✔ **Answer:** 17.7

17.737

Explanation
The percentage awarded can be calculated using the following code:

```
two_by_two %>%
    mutate(men = round(men/sum(men)*100, 1), women = round(women/sum(women)*100, 1)) %>%
    filter(awarded == "yes") %>%
    pull(men)
```

What is the percentage of women awarded?
Report a percentage between 0 and 100.

14.898 ✔ **Answer:** 14.9

14.898

Explanation
The percentage awarded can be calculated using the following code:

```
two_by_two %>%
    mutate(men = round(men/sum(men)*100, 1), women = round(women/sum(women)*100, 1)) %>%
    filter(awarded == "yes") %>%
    pull(women)
```

Submit You have used 1 of 10 attempts

ⓘ   Answers are displayed within the problem

Question 3

1.0/1.0 point (graded)
Run a chi-squared test on the two-by-two table to determine whether the difference in the two success rates is significant. (You can use `tidy` to turn the output of `chisq.test` into a data frame as well.)

What is the p-value of the difference in funding rate?

0.0509 ✔ **Answer:** 0.0509

0.0509

Explanation
The p-value can be calculated using the following code:

```
two_by_two %>% select(-awarded) %>% chisq.test() %>% tidy() %>% pull(p.value)
```

Submit You have used 1 of 10 attempts

Question 4

3.0/3.0 points (graded)

There may be an association between gender and funding. But can we infer causation here? Is gender bias causing this observed difference? The response to the original paper claims that what we see here is similar to the UC Berkeley admissions example. Specifically they state that this "could be a prime example of Simpson's paradox; if a higher percentage of women apply for grants in more competitive scientific disciplines, then an analysis across all disciplines could incorrectly show 'evidence' of gender inequality."

To settle this dispute, use this dataset with number of applications, awards, and success rate for each gender:

```
dat <- research_funding_rates %>%
    mutate(discipline = reorder(discipline, success_rates_total)) %>%
    rename(success_total = success_rates_total,
           success_men = success_rates_men,
           success_women = success_rates_women) %>%
    gather(key, value, -discipline) %>%
    separate(key, c("type", "gender")) %>%
    spread(type, value) %>%
    filter(gender != "total")
dat
```

To check if this is a case of Simpson's paradox, plot the success rates versus disciplines, which have been ordered by overall success, with colors to denote the genders and size to denote the number of applications.

In which fields do men have a higher success rate than women?
Select ALL that apply.

- ☑ Chemical sciences
- ☑ Earth/life sciences
- ☐ Humanities
- ☐ Interdisciplinary
- ☑ Medical sciences
- ☐ Physical sciences

- ☑ Physics
- ☑ Social sciences
- ☐ Technical sciences

✔

Which two fields have the most applications from women?
Select TWO.

- ☐ Chemical sciences
- ☐ Earth/life sciences
- ☐ Humanities
- ☐ Interdisciplinary
- ☑ Medical sciences
- ☐ Physical sciences
- ☐ Physics
- ☑ Social sciences
- ☐ Technical sciences

✔

Which two fields have the lowest overall funding rates?
Select TWO.

- ☐ Chemical sciences
- ☐ Earth/life sciences
- ☐ Humanities

- ☐ Interdisciplinary
- ☑ Medical sciences
- ☐ Physical sciences
- ☐ Physics
- ☑ Social sciences
- ☐ Technical sciences

✔

Explanation

The plot can be generated using the following code:

```
dat %>%
    ggplot(aes(discipline, success, size = applications, color = gender)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    geom_point()
```

Submit   You have used 2 of 3 attempts

ⓘ Answers are displayed within the problem