Homework 1

BSTA 513/613

Your name here - update this!!!!

2024-04-11

Purpose

This homework is designed to help you practice the following important skills and knowledge that we covered in Classes 1-2:

- Practicing and outlining your decision process to analyze the relationship between two categorical variables
- Interpreting research aims into questions/tests that can be answered with statistics
- Using R to calculate sample proportions
- Using R to calculate test statistic values for inference tests
- Interpreting new phrasing of questions that were introduced in class

Directions

- Download the `.qmd` file here.
- You will need to download the datasets. Use this link to download the HW1 datasets needed in this assignment. If you do not want to make changes to the paths set in this document, then make sure the files are stored in a folder named "data" that is housed in the same location as your HW1 `.qmd` file.
- Please upload your homework to Sakai. **Upload both your `.qmd` code file and the rendered `.html` file**
- For each question, make sure to include all code and resulting output in the html file to support your answers
- Show the work of your calculations using R code within a code chunk. Make sure that both your code and output are visible in the rendered html file. This is the default setting.
- Write all answers in complete sentences as if communicating the results to a collaborator.

Tip

It is a good idea to try rendering your document from time to time as you go along! Note that rendering automatically saves your `.qmd` file and rendering frequently helps you catch your errors more quickly.

Questions

PART 1

The following questions are intended to give you practice in understanding concepts and completing calculations.

Question 1

If the probability that one white blood cell is a lymphocyte is 0.3, compute the probability of 2 lymphocytes out of 10 white blood cells. Also, compute the probability that at least 3 lymphocytes out of 10 white blood cells. You may calculate by hand, using a web app, or using R.

Question 2

Consider a 2 x 2 table from a prospective cohort study:

```
Warning: 'tidy.numeric' is deprecated.
See help("Deprecated")
```

```
# A tibble: 2 x 1
  x[,"Favorable"] [,"Unfavorable"]
            <dbl>            <dbl>
1              30               20
2              10               60
```

Part a

Estimate the probability of having favorable results for subjects in the treatment group. Report with the 95% confidence interval.

Part b

Repeat part a for the placebo group.

Part c

Conduct a statistical test to evaluate whether there is an association between group and outcome. What is the name of the test? Write down the null and alternative hypotheses. Compute the test statistic (by hand or by a software). What distribution does the test statistic follow under the null hypothesis? Give the p-value and interpret your result.

Question 3

Consider a cohort study with results shown as in following table:

```
Warning: 'tidy.numeric' is deprecated.
See help("Deprecated")
```

```
# A tibble: 2 x 1
  x[,"Favorable"] [,"Unfavorable"]
            <dbl>            <dbl>
1               6                1
2               2                5
```

Conduct a statistical test to evaluate whether there is an association between group and outcome. Write down the null and alternative hypotheses. Compute the expected cell counts under null hypothesis. What is the name of the test? Give the p-value and interpret your result.

Question 4

Table 4 shows the information of a selected group of adolescents on whether they use smokeless tobacco and their perception of risk for using smokeless tobacco.

Table 4:

```
Warning: 'tidy.numeric' is deprecated.
See help("Deprecated")
```

```
# A tibble: 3 x 1
  x[,"YES"] [,"NO"]
      <dbl>   <dbl>
1        25      60
2        35     172
3        10     200
```

Part a

Conduct a statistical test to examine **general** association between adolescent smokeless tobacco users and risk perception. What is the name of the test? Write down the null and alternative hypotheses. Compute the test statistic (use software). What distribution does the test statistic follow under the null hypothesis? Give the p-value and interpret your result.

Part b

Is there a trend of increased risk perception for smokeless tobacco users? What test would you use? State the test, state any assumptions, conduct the inference test, and state your conclusions.

PART 2

The following questions are intended to give you practice in connecting concepts that will help you make decisions in real world applications.

Question 5

Start making a comprehensive table or outline for the inference tests that we have covered. Here is a list of the tests we have covered:

- Single proportion
- Chi-squared test for general association
- Likelihood ratio test for general association
- Fisher's Exact test for general association
- Cochran-Armitage test for trend
- Mantel-Haenszel test for linear trend

And here is a list of attributes to include:

- Number of variables testing

- Types of variables
- Criteria (if any)
- Hypothesis test
- Test statistic (if we went over it)
- R code for test
- Sample size / Power calculation (**optional**, not discussed in class)
- Special notes (**optional**)

For example, I could make a table with different rows corresponding to different tests and different columns for each attribute.

Question 6

I want you to gain experience exploring a package and function. This is an important skill in coding that can help you grow as an applied statistician.

In your previous course, the function lm() was introduced to perform linear regression. In this class, we will heavily use the function glm(). By typing "?glm" in the R console, we can open the Help page for glm(). The following questions ask about the glm() function. You can Google or use R documentation to answer the questions.

Feel free to read more about the differences between lm() and glm().

Part a

What does the input "family" mean? If I wanted to perform regression using a Poisson distribution, what would I input into family?

Part b

What is the default action for the "na.action" input?

Part c

How does the glm() function fit our model? (Hint: see "method")

Part d

Do you think the output of summary() will be the same for lm() and glm()?

Question 7

OPTIONAL

Let's make a decision tree on the different tests we learned! I would like you to make a flow chart for the different tests we learned in Classes 1 and 2. You'll need to include characteristics for:

- Number of variables (1, 2, or 3 - we will go over 3 variables in Class 4)
- Number of categories in each variable
- Sample size is small
- Ordinal/nominal independent variable
- Ordinal/nominal response variable(s)

For example, if I make a decision tree that includes end nodes for different animals (cat, dog, snake, turtle, and hawk) using yes/no characteristics (has a shell, woof/meows, has fur, or flies), then my flow chart would look like: See my example under Sakai Resources. You are welcome to draw this chart. I used SmartArt under the Insert tab in Word to create mine.