

Homework 1

BSTA 513/613

Your name here - update this!!!!

Caution

This homework is ready to go! (Nicky 4/1/25)

Purpose

This homework is designed to help you practice the following important skills and knowledge that we covered in Lessons 2-5:

- 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
- Calculate and interpret the estimated risk difference, relative risk, and odds ratios, and their confidence intervals
- Expand work on contingency tables to evaluate the agreement or reproducibility using Cohen's Kappa
- Understand important differences between linear regression and logistic regression
- Construct a simple logistic regression model

Directions

- [Download the .qmd file here.](#)
- You will not need to download datasets for this homework.
- Please upload your homework to Sakai. Upload both your .qmd code file and the rendered .html file

- Please rename your homework as Lastname_Firstinitial_HW1.qmd. This will help organize the homeworks when the TAs grade them.
- Please also add the following line under subtitle: “BSTA 513/613”: author: First-name Last-name with your first and last name so it is attached to the viewable document.
- 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:

	Favorable	Unfavorable
Treatment	30	20
Placebo	10	60

Part a

Estimate the probability of having favorable results for subjects in the treatment group. Include an interpretation and 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? Make sure to follow the entire test process demonstrated in the slides.

Question 3

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

	Favorable	Unfavorable
Treatment	6	1
Placebo	2	5

Conduct a statistical test to evaluate whether there is an association between group and outcome. What is the name of the test? Make sure to follow the entire test process demonstrated in the slides.

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:

	YES	NO
Minimal	25	60
Moderate	35	172
Substantial	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? Make sure to follow the entire test process demonstrated in the slides.

Part b

Is there a trend of increased risk perception for smokeless tobacco users? What test would you use? Make sure to follow the entire test process demonstrated in the slides.

Question 5

A study looked at the effects of oral contraceptive (OC) use on heart disease in women 40 to 44 years of age. The researchers prospectively tracked whether or not the women developed a myocardial infarction (MI) over a 3-year period. The table below summarizes their results with columns indicating whether or not women developed MI and rows indicating their OC use.

	Yes	No
OC users	13	4987
Non-OC users	7	9993

Part a

Compute the estimated risk difference comparing OC users to non-OC users. Include a 95% CI for the estimate and interpretation of the estimated value.

Part b

Compute the estimated relative risk comparing OC users to non-OC users. Include a 95% CI for the estimate and interpretation of the estimated value.

Part c

Compute the estimated odds ratio comparing OC users to non-OC users. Include a 95% CI for the estimate and interpretation of the estimated value.

Part d

Is the OR a good approximation of the RR? Explain why or why not.

Question 6

One important aspect of medical diagnosis is its reproducibility. Suppose that two different doctors examine 100 patients for dyspnea in a respiratory-disease clinic, and that doctor A diagnosed 15 patients as having dyspnea (while doctor B did not), doctor B diagnosed 10 patients as having dyspnea (while doctor A did not), and both doctor A and doctor B diagnosed 7 patients as having dyspnea.

Part a

Construct a two-way contingency table to summarize the dyspnea diagnoses from doctor A and B.

Part b

Compute the Cohen's kappa, 95% confidence interval, and interpret the results. What level of agreement does your kappa indicate?

Questions 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 7

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
- 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 8

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 9

This question is meant to emphasize the differences between linear regression and logistic regression. Each part will ask about different aspects of the two regression models. If the question has multiple choice answers, then you must write 1-2 sentences justifying your answer.

Part a

In **linear** regression, what type of variable is our response/outcome variable?

- a. binary
- b. continuous
- c. count
- d. ordinal

Part b

In **logistic** regression, what type of variable is our response/outcome variable?

- a. binary
- b. continuous
- c. count
- d. normal

Part c

What assumptions in linear regression? Please state the assumption name and characteristics of that assumption.

Part d

Which assumptions of linear regression are violated if we try to fit a binary response using linear regression? You may choose more than one answer.

- a. Independence
- b. Linearity
- c. Normality
- d. Homoscedasticity

Part e

Please use our notes on generalized linear models (GLMs) to answer this question. What is the random component used in linear regression? What is the random component used in logistic regression? Name the specific variable type and distribution for it.

Part f

Please use our notes on generalized linear models (GLMs) to answer this question. What link function do we use in linear regression? What link function do we use in logistic regression? Name the link and write out the function.

Part g

Please use our notes on generalized linear models (GLMs) to answer this question. What is the systematic component used in simple linear regression? What is the systematic component used in simple logistic regression? Please write out the function for the systematic component for a single covariate.

Part h

How do we determine our coefficient estimates (estimates of the parameter values) in linear regression? You may choose more than one answer.

- a. Ordinary least squares
- b. Maximum likelihood estimation

Part i

How do we determine our coefficient estimates (estimates of the parameter values) in logistic regression? You may choose more than one answer.

- a. Ordinary least squares
- b. Maximum likelihood estimation

Question 10

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.