Homework 1 Answers

BSTA 513/613

Nicky Wakim

Questons 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.

Answer:

0.233 and 0.617

Question 2

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

```
t1 <- matrix(c(30,20,10,60),ncol=2,byrow=TRUE)
colnames(t1) <- c("Favorable","Unfavorable")
rownames(t1) <- c("Treatment","Placebo")
t1 <- as.table(t1)
t1 %>% kable(table.attr = 'data-quarto-disable-processing="true"') %>%
  kable_styling(full_width = F, bootstrap_options = c("striped", "hover"))
```

	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.

Answer:

0.6 (95% CI: 0.452, 0.733)

Part b

Repeat part a for the placebo group.

Answer:

0.143 (95% CI: 0.074, 0.252)

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.

Answer:

p-value approx 4.63×10^{-7}

Question 3

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

```r
t2 <- matrix(c(6,1,2,5),ncol=2,byrow=TRUE)
colnames(t2) <- c("Favorable","Unfavorable")
rownames(t2) <- c("Treatment","Placebo")
t2 <- as.table(t2)
t2 %>% kable(table.attr = 'data-quarto-disable-processing="true"') %>%
  kable_styling(full_width = F, bootstrap_options = c("striped", "hover"))
```

	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.

Answer:

p-value= 0.103

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:

```r
t4 <- matrix(c(25,60,35,172,10,200),ncol=2,byrow=TRUE)
colnames(t4) <- c("YES","NO")
rownames(t4) <- c("Minimal","Moderate","Substantial")
t4 <- as.table(t4)
t4 %>% kable(table.attr = 'data-quarto-disable-processing="true"') %>%
  kable_styling(full_width = F, bootstrap_options = c("striped", "hover"))
```

	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.

Answer:

```r
chisq.test(t4)
```

```

    Pearson's Chi-squared test

data:  t4
X-squared = 33.218, df = 2, p-value = 6.122e-08
```

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.

Answer:

```r
library(DescTools)
CochranArmitageTest(t4)
```

```

    Cochran-Armitage test for trend

data:  t4
Z = -5.7632, dim = 3, p-value = 8.253e-09
alternative hypothesis: two.sided
```

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.

Answer:

0.0019 (95% CI: 0.0004, 0.0034)

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.

Answer:

3.71 (95% CI: 1.48, 9.30)

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.

Answer:

3.68 (95% CI: 1.49, 9.95)

Part d

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

Answer:

Yes

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.

Answer:

Not given

Part b

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

Answer:

0.207 (95% CI: -0.062, 0.476)

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
- 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.

Tests	Number of Variables	Types of Variables	Criteria	Hypothesis Test	R Function
Single proportion	1	Nominal, binary	None	$H_0 : \hat{p} = p, H_1 : \hat{p} \neq p$	`prop.test()`
Chi-squared test for general association					
Fisher's Exact test for general association					
Cochran-Armitage test for trend					
Mantel-Haenszel test for linear trend					

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?

Not given

Part b

What is the default action for the “na.action” input?

Not given

Part c

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

Not given

Part d

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

Not given

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

Answer:

b. continuous

Part b

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

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

Answer:

a. binary

Part c

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

Answer:

LINE!

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

Answer:

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.

Answer:

Not given

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.

Answer:

Not given

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.

Answer:

Not given

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

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

Answer:

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.