HW3

Follow up questions for R Intro Part 2

Download and work through the file posted on our moodle page called R Intro Part 2. When you're finished, complete the following exercises using R. Upload a pdf document that contains the r commands you executed and the results they produced.

1. Indexing

Generate 1000 random standard normal values. Using indexing, construct a vector that omits all the values between 0 and 1.5. Create a histogram of your observations. Include your code and a copy of your histogram in your write-up.

2. Sequencing

Generate a vector contains the following 199 numbers: [1 2 3 …. 99 100 99 98 … 3 2 1]. Verify that the length of the vector is 199 using the length command in R. Include your code and a copy of your vector in your write-up.

3. Sorting

Examine the USArrests dataset from the 'datasets' package. Sort the data by murder rate using the order command. Provide a copy of the dataset sorted from highest murder rate to lowest murder rate (hint: you'll have to set the decreasing option in the order function to TRUE, see the help file for order by typing ?order). Also include the command that you used to generate the sorted dataset.

4. Factors

Again, using the USArrests dataset create a new variable called Urban.Cat that categorizes the percent of the population that is urban into Low, Medium, and High categories. To do this, you'll first need to use the 'cut' command. Try entering the following command that divides the range of UrbanPop into three equal parts:

```
USArrests$Urban.Cat <- cut(USArrests$UrbanPop, 3)
```

You'll see that a new factor variable was created in the USArrests dataframe. It won't have very useful level names though. Change the level names appropriately. Once you've done that, include your commands and a copy of your new (and improved) dataset in your write-up.

5. Boxplots

Is Percent Urban associated with the murder rate? To examine this, create side-by-side boxplots of Murder rate based on Urban.Cat. Include both your code and your plot in your write-up.

6. Loading a dataset / performing a t-test

The Heart and Estrogen / Progestin Study (HERS) was a clinical trial of hormone therapy for prevention of recurrent heart attacks and death among post-menopausal women with existing coronary heart disease. Data from this study are posted on our moodle page (under the R Intro Part 2 heading). Download and save these data to your computer. They're stored as a comma-separated-value file (a .csv file). Make sure you remember where you save the data. Once you've downloaded the data, load it into R. If you're using the webserved version, load the file into your directory by clicking on the 'Upload' button after you download it. Then you can load the file into R-Studio with the following command:

```
hersdata <- read.csv(file=file.choose())
```

The above command should open a window on your machine from which you can navigate to and select the file.

Once you've loaded the data, you can type:

```
head(hersdata)
```

to see the first few records of the dataframe. To see all the variables contained in the dataframe type:

```
names(hersdata)
```

a) Produce side-by-side boxplots of BMI for white and non-white women (use the variable nonwhite). Include your code and your plot in your write-up.

b) Is BMI associated with white/non-white status? Conduct a hypothesis test to find out. First, create histograms of BMI for each level (both white and non-white) to check assumptions. Do you think the distributions are approximately normal? (Hint: you should probably transform the BMI variable using the natural log transform). Include histograms of ln(BMI) in your write-up.

You can perform a two-sample t-test using the following syntax (assuming you've created a variable called lnBMI in the hersdata dataframe):

```
t.test(lnBMI ~ nonwhite, hersdata)
```

Report your code, output, and final conclusion.

7. Functions

Construct a function called ci.prop that calculates a confidence interval for a proportion. You should "pass" the numerator and the denominator to the function. Your function should return the proportion and the upper and lower bounds for a 95% confidence interval. Recall that the standard error for a proportion can be estimated using:

$$\text{SE}(p) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Test out your function. Suppose your sample proportion is 46/200. Use your function to calculate $\hat{p}$ and a 95% C.I.

You should include the code for your function and its results in your write-up.

BONUS: for extra credit, code your function so that the level of confidence can be specified as well. In other words, level of confidence is an additional value that is passed to your function. Use your improved function to calculate 80%, 90%, and 99% confidence intervals for the sample proportion 46/200. (Hint: you'll probably want to utilize the function 'qnorm'). Include your function code and its results in your write-up.

More with prevalence and incidence

8. Latex Lament

A sharp increase in recognition of latex sensitization occurred after the implementation of universal precautions in response to 1987 recommendations by the Centers for Disease Control and Prevention. Since then, latex sensitization has become an important occupational health concern, especially for occupations where latex glove use is common. Several studies have been conducted to determine whether latex allergy is more common among health care workers than in the general population.

A study conducted by Valks, et al. (Contact Dermatitis, 2004 April;50(4):222-4) assessed latex allergy among persons seeking care at a clinic for occupation-related skin diseases. They compared latex allergy among workers in the health care and non-health care professions.

a) Calculate the frequency (per 1000) of latex sensitization for each the following groups. In this study, which occupation has the highest frequency of latex allergy?

Occupation	# With Latex Allergy	Total # in Group	Frequency per 1,000 people
Health Care Workers	13	58	
Non-Health Care Workers	25	1093	
Painters	2	32	
Hairdressers	3	59	
Food Handlers	7	41	
Cleaners	3	78	

b) What is the name of the measurement of disease occurrence that you calculated?

A hypothetical researcher was interested in measuring the number of new cases of latex sensitization occurring in health care workers. The researcher invited 2062 employees of local hospitals to participate in the study. At the first study visit, all 2030 participants (32 decided not to participate) were given a skin prick test with latex reagents. One hundred and sixty four participants tested positive at baseline. The remaining participants were followed for two years and tested every month for latex allergy. During the follow-up period, 47 participants were newly diagnosed with latex allergy.

c) What is the cumulative incidence (per 1,000) of new latex allergy in the follow-up cohort over the two years?

Now consider the possibility that not all of the participants completed each of the 24 monthly study visits (because they either dropped out or were diagnosed with latex allergy). The table below summarizes the length of follow-up for the participants.

# of Participants	# Monthly Visits Completed
1532	24
17	22
19	19
41	16
232	12
25	7

d) Calculate the total person time for which the participants were followed.

e) Using the information from the previous questions, calculate the incidence density of latex allergy per 1,000 person-months.

f) In one particular hospital, approximately 12.1% of the hospital workers tested positive for a latex allergy during the first week of June.

Epidemiologically speaking, this number is best described as a:

RATE RATIO PROPORTION

You could also describe it as:

INCIDENCE PROPORTION INCIDENCE DENSITY

PERIOD PREVALENCE POINT PREVALENCE