R Assignment #2:

Summary Statistics and Graphs with R

(Due Oct 4 – by email)

BIOL 4090 - 6090

David Hyrenbach

Tasks - Assignment #2

1. Install ggplot2 function

2. Learn how to access data from a package

3. Learn how to make simple plots with qplot
4. Install and explore ggplot2 function

5. Practice making five types of graphs using the ggplot function

6. Learn how to make more advanced plots with ggplot

Learning Objectives – Assignment #2

Explore the RStudio windows

Create and manipulate datasets

Install and run other R packages

Summarize data graphically / visually

Practice using qplot and ggplot2

Access the R help

Instructions

This assignment is worth 5 points. Paste your answers into a word file named "BIOL6090_Assignment1_YOURNAME".

Email your word file and any other required files to me (khyrenba@gmail.com) using a message entitled "BIOL6090 Assignment #2") by the end of Sept. 13.

(PENALTIES: 10% per partial/full day late)
(5% for not using right email address/title)

Questions / Answers

1a. Create a vector named "age", with 18 age estimates:
34 36 37 37 38 38 38 38 39
40 40 41 41 42 42 42 41 48

Report the command you used to create the "age" vector (+0.125):

Questions / Answers

1b. Report the command that calculates skewness in R (make sure you include and explain all of the arguments used with this command) (+0.125):

Questions / Answers

1c. Use the R help to find out how the three skewness types (1, 2 and 3) are calculated in R. Which one is the default type used, unless specified? Copy and paste the meta-data for each type below (+0.125):

Questions / Answers

1d. Report the command that calculates kurtosis in R (make sure you include and explain all of the arguments used with this command) (+0.125):

Questions / Answers

1e. Use the R help to find out how the three kurtosis types (1, 2 and 3) are calculated in R. Which one is the default type used, unless specified? Copy and paste the meta-data for each type below (+0.125):

Questions / Answers

1f. Calculate the following summary statistics for "age" using R (with the commands you found) and RCmdr

Summary Statistic (calculated using R)	Value
Skewness (type 1)	
Skewness (type 2)	
Skewness (type 3)	
Kurtosis (type 1)	
Kurtosis (type 2)	
Kurtosis (type 3)	

Summary Statistic (calculated with Rcmdr)	Value
Skew	
Kurtosis	

Report which Skewness type does RCmdr use: _____ (+0.125)

Report which Kurtosis type does RCmdr use: _____ (+0.125)

Questions / Answers

2a. Using RCmdr, calculate the following summary statistics for the "age" dataset: Mean, SD, SE, IQR, CV, and 5 Quantiles (0, 25, 50, 75, 100) (+0.125).

Questions / Answers

2b. Using RCmdr make a Histogram and a Density plot of "age" (use gaussian kernel) and default settings. Paste both figures below (+0.125).

Questions / Answers

2c. Do the Histogram and Density plot of the "age" data set look like normal distributions? Explain Why / Why not? (+0.125).

Questions / Answers

3a. Distributions of random variables.
In this lab you will investigate the normal distribution. First, you will use the graphical tools of R to assess the normality of your data. Then, you will learn how to generate random numbers from a normal distribution.

You will work with a dataset of body measurements from 247 men and 260 women, considered healthy young adults.

```
> download.file
("http://www.openintro.org/stat/data/bdims.RData" ,
destfile = "bdims.RData")
load ( "bdims.RData" )
```

How large is the bdims dataset? (+0.125)

Questions / Answers

3b. Since males and females tend to have different body dimensions, it will be useful to create two separate datasets: one with only men and another with only women.

Next, make two separate plots of men's heights and women's heights that illustrate the frequency distributions of these two datasets. HINT: you can create two stacked plots using a single command in Rcmdr.

Explain how the male and female distributions differ, by describing the summary statistics you can observe in the distribution plots. Be as specific as you can (+0.125).

Questions / Answers

3b. In your description of the distributions, did you use words like "bell-shaped" or "normal"? It is tempting to say so when describing unimodal and symmetrical distributions.

However, to determine how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution.

This normal curve should have the same mean and standard deviation as the data. Let's do this for the women's data.

```
> fhgtmean <- mean(fdims$hgt)
> fhgtsd <- sd( fdims$hgt)
```

Questions / Answers

3b. Next, we use these parameters to overlay a normal probability curve over the density histogram. We use two functions: dnorm creates the normal distribution and lines plots the continuous distribution. You can copy and paste the four lines in one command or line by line:

```
> hist (fdims$hgt, probability = TRUE )
x <- 140 : 190
y <- dnorm (x = x, mean = fhgtmean, sd = fhgtsd)
lines (x = x, y = y, col = "blue")
```

NOTICE:
To compare the histogram to the normal distribution, we need to plot a density histogram, where the areas of the bars add up to 1. The area of the normal curve also adds to 1.

Questions / Answers

3c. Use the same instructions to make a plot of the male heights, and overlay the corresponding normal distribution.

Report the commands you used to generate the plot (+0.125).

Paste the plot (+0.125).

Questions / Answers

3d. Report the summary statistics for the male and the female weight datasets (+0.125 for each). To receive full credit, briefly explain why you selected each of statistic.

Questions / Answers

4a. Eyeballing the shape of the histogram is one way to determine if the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve.

An alternative approach involves constructing a normal probability plot, also called a Q-Q plot for "quantile-quantile". For the female data:

```
> qqnorm (fdims$hgt)
> qqline (fdims$hgt)
```

A data set that is nearly normal will result in a probability plot where the points closely follow the theoretical line. Any deviations from normality leads to deviations of these points from the line.

Questions / Answers

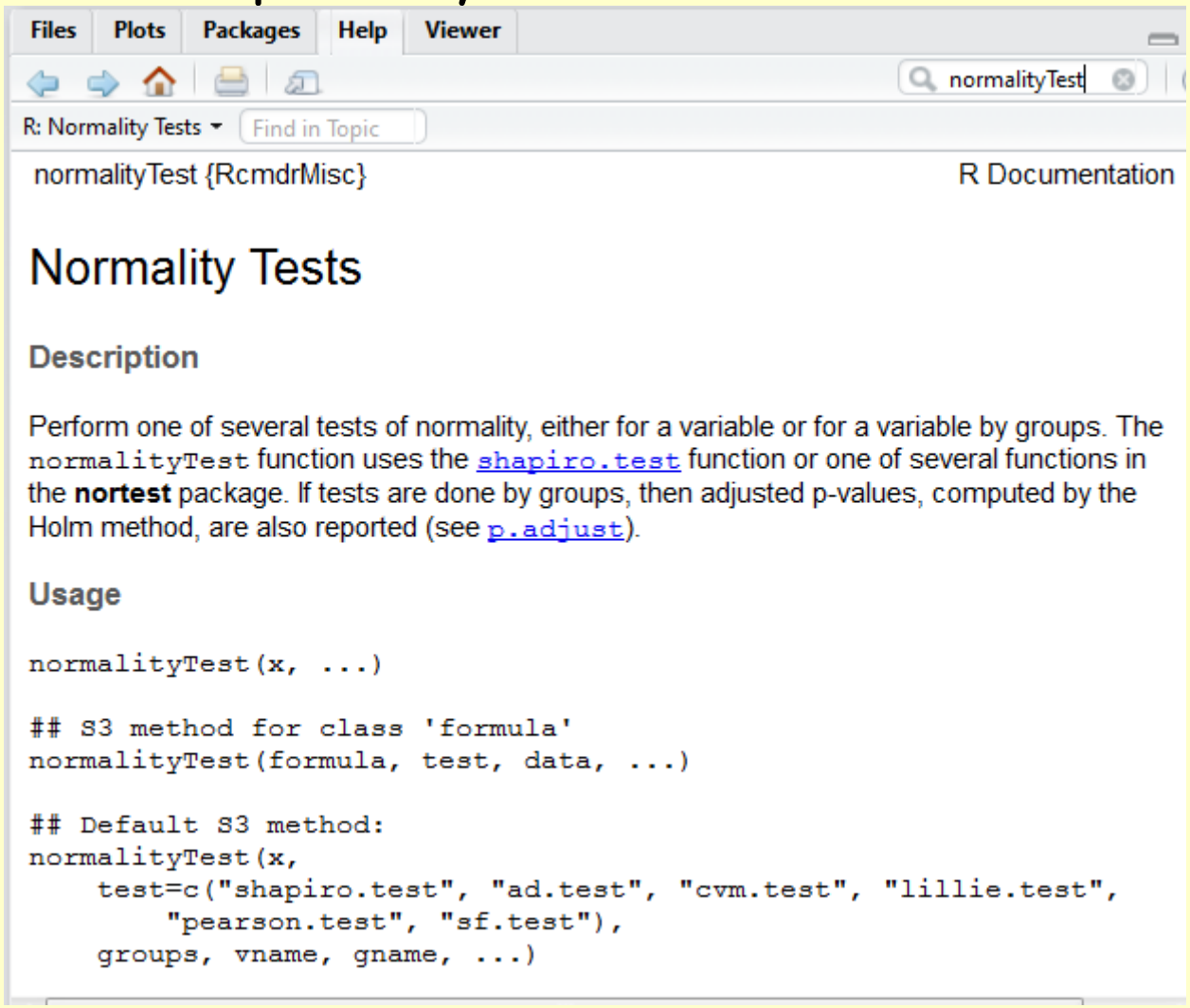
4a. Create and paste a Q-Q plot for "quantile-quantile". for the male data (+0.125).

Questions / Answers

4b. Let's perform a test of normality (S-W) for the male and the female datasets separately.

We will use the normalityTest function in the RcmdrMisc package.

Start by activating the RcmdrMisc package



Questions / Answers

4b. For the male data:

- Create a new variable containing the male height data

```
> mheight <- mdims$hgt
```

- Perform the normality analysis of the male data

```
> normalityTest (mheight, test = "shapiro.test")
```

Result:
Shapiro-Wilk normality test data:
mheight W = 0.99358, p-value = 0.3716

Interpret this result (+0.125):
Are the data normal? Why / Why not?

Questions / Answers

4e. Do the same analysis for the female data:

Interpret this result (+0.125):
Are the data normal? Why / Why not?

Questions / Answers

4d. It turns out that statisticians know a lot about the normal distribution. Once we decide that a random variable is normally distributed, we can answer all sorts of questions about that variable, by calculating various probabilities.

For example: "What is the probability that a randomly chosen male or female is taller than 6 feet (182 cm of more) ?"

We can find this probability by calculating a Z score and consulting a Z table (also called a normal probability table). In R, this is done in one step with the function pnorm.

```
> 1 - pnorm (q = 182 , mean = fhgtmean , sd = fhgtsd)
> 0.004434387
```

(NOTE: this calculation is for the female distribution)

Questions / Answers

4d. The function pnorm calculates the area under the normal curve below a given value, q , with a given mean and S.D. Since we are interested in the probability that a woman is taller than 182 cm, we use one minus that probability.

Assuming a normal distribution has allowed us to calculate a theoretical probability. If we want to calculate the probability empirically, we determine how many observations fall above 182, and divide this number by the sample size.

```
> sum (fheight > 182 ) / length (fheight)
> 0.003846154
```

Although the probabilities are not exactly the same, they are very close. The closer the distribution is to a normal, the more accurate the theoretical probabilities will be.

Questions / Answers

4d. Calculate the probability that a randomly chosen male is taller than 6 feet (182 cm of more) (+0.125).

```
> 1 - pnorm (q = 182 , mean = mhgtmean , sd = mhgtsd)
> 0.2768345
```

Questions / Answers

5a. Extra Credit: Calculate the actual Z scores for the male and the female data. Briefly discuss how the values you calculated relate to the probabilities R computed (+0.1 each):

5b. Extra Credit: Finally, find and install an R function that calculates Z scores from a dataset. Report the following information (+0.1 each):

- name of function and the package containing the function:
- usage of function (copy and paste from meta-data)
- Z values for male and female data you calculated above

Questions / Answers

6a. Install package ggplot2. List all the data sets attached to the ggplot2 package. Paste a screen capture of your console, showing the list of available data sets (+0.125).

Questions / Answers

6b. View the data sets attached to package ggplot2. Read the data from data set "diamonds". Report how large is the "diamond" data set (how many rows and columns) (+0.125).

Questions / Answers

6c. Create a simple scatterplot of the diamond data set using the quickplot command (qplot). First of all, use qplot to create a scatterplot showing the relationship between the price and carats (weight) of a diamond, with this command:

```
> qplot(carat, price, data = diamonds)
```

Paste the scatterplot into your answer file. (+0.125).

Questions / Answers

6d. The plot shows a strong correlation with notable outliers. The relationship looks exponential, though, so the first thing we would like to do is to transform the variables. Because qplot() accepts functions of variables as arguments, plot log(price) vs. log(carat) with the command:

```
> qplot(log(carat), log(price), data = diamonds)
```

Paste the scatterplot below. (+0.125).

Questions / Answers

6e. Note that arguments can also be combinations of existing variables. If we are curious about the relationship between the volume of the diamond (approximated by $x \times y \times z$) and its weight, we could use the following command:

```
> qplot(carat, x * y * z, data = diamonds)
```

Paste the scatterplot below. (+0.125).

Questions / Answers

6f. Lets add some aesthetic attributes to the plot, like specific colors and symbols. Note that qplot can do this and will automatically provide a legend that maps the displayed attributes to the data values. This makes it easy to include additional data on the plot.

Next, lets augment the plot of carat and price with new information about diamond color and cut, from the data set "dsmall", using these commands:

```
> qplot(carat, price, data = diamonds, colour = color)
> qplot(carat, price, data = diamonds, shape = cut)
```

Questions / Answers

6f. Paste the two scatterplots into your file (+0.125).

Questions / Answers

6g. Report the command that you would use to plot the diamond shape data by symbol colour, rather than symbol shape, and paste the revised scatterplot, into your file (+0.125).

Questions / Answers

6h. You can also manually set the aesthetics attributes using I(). For example, to make a semi-transparent colour you can use the *alpha* aesthetic, which takes a value between 0 (completely transparent) and 1 (complete opaque). It's often useful to specify the transparency as a fraction, e.g., 1/10 or 1/20, as the denominator specifies the number of points that must overplot to get a completely opaque colour.

To start, make a default plot, using this command:

```
> qplot(carat, price, data = diamonds, alpha = I(1))
```

Questions / Answers

6h. Try these two commands:

```
> qplot(carat, price, data = diamonds, alpha = I(1/10))
> qplot(carat, price, data = diamonds, alpha = I(1/100))
```

Paste both scatterplots into your file and explain what happens as you change I from 1/10 to 1/100 (+0.125).

Questions / Answers

6i. You can also manually set other aesthetics attributes using I(), e.g., colour = I("red") or size = I(2).

To start, make a default plot, using this command:

```
> qplot(carat, price, data = diamonds)
```

Then, create the same scatterplot with large red symbols. Report the command you used (+0.125):

Questions / Answers

6j. Paste the resulting scatterplot into your file (+0.125).

Questions / Answers

6k. qplot is not limited to scatterplots, but can produce almost any kind of plot by varying the geometric object (or geom), which describes the type of object being used to display the data. For example, the geom "smooth" enables you to investigate two-dimensional relationships:

geom "smooth" fits a smoother to the data and displays the smooth and its standard error

Notice that you can combined multiple geoms by creating a vector of geom names created with command c(). The geoms will be overlaid in the order in which they are listed.

Questions / Answers

6k. Try this command:

```
> qplot(carat, price, data = diamonds, geom = c("point", "smooth")
```

Report the smoothing method used, as reported by R:

Paste the resulting plot into your file (+0.125):

Questions / Answers

6l. Paste the help documentation for the geom_smooth, command. Report how many smoothing methods are available, and list their names (+0.125).

Questions / Answers

7a. Import the dataset "ExamAnxiety.xlsx" and create a simple scatterplot of Exam and Anxiety. Show the points, and add a straight line to the plot. For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command (NOTE: the intro to graphing ppt shows the steps for making this figure) (+0.10):

Questions / Answers

7b. Import the dataset "ExamAnxiety.xlsx" and create a grouped scatterplot of Exam and Anxiety, with the male / female data plotted separately. Show the points, and add a straight line to the plot. For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command (NOTE: the intro to graphing ppt shows the steps for making this figure) (+0.10):

Questions / Answers

8a. Import the dataset "MusicFestival.xlsx" and create a histogram of the day 1 data using a binwidth of 1 (NOTE: the intro to graphing ppt shows the steps for making this figure) (+0.10):

Questions / Answers

8b. Create a histogram of the day 2 data using a binwidth of 1. Make sure you change the figure labels accordingly. For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command (+0.10):

Questions / Answers

8c. Create a boxplot for the day1 MusicFestival dataset, that separates the data for males / females. For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command (+0.10):

Questions / Answers

8d. Create a density plot of the day1 MusicFestival dataset. For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command (+0.10):

Questions / Answers

9a. Load the data "ChickFlick.xlsx" and create a bar graph for one independent variable (Film). For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command. Do the means of the two films seem significantly different? State why / why not? (+0.10):

Questions / Answers

9b. Load the data "ChickFlick.xlsx" and create a bar graph for one independent variable (Gender). For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command. Do the means of the two genders seem significantly different? State why / why not? (+0.10):

Questions / Answers

9c. Using the dataset "ChickFlick.xlsx", create a bar graph for two independent variables (Gender X Film), so the bar color is applied to "gender". For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command. Do the means of the two films seem significantly different? State why / why not? (+0.10):

Questions / Answers

9d. Using the dataset "ChickFlick.xlsx", create a bar graph for two independent variables (Film X Gender), so the bar color is applied to "film". For full credit, paste the figure you created, report the command you used to create this plot and explain all of the arguments used with this command. Do the means of the two films seem significantly different? State why / why not? (+0.10):