

PROJECT 3: DESCRIPTIVE STATISTICS

Statistics 401: Fall 2006

Due: Friday, September 22

In this project, you will use R to summarize different data sets using numerical and graphical techniques. Two of these data sets need to be downloaded from the Stat 401 web site: BAC.txt and jellyfish.txt. You will need to assemble the tables, plots, and answers into a coherent write-up. Your write-up must be typed. Please number your answers as the questions were numbered. Your grade will be determined by how well you answer the questions and by the professionalism and clarity of our write-up. Note, you will need the “pastecs” library in R to complete the last problem. You should have downloaded it when you installed R. If you did not, then check out the Chapter 1 Handout for download instructions.

1. Define the term “distribution”.
2. Email spam is the curse of the internet. Here is a compilation of the most common types of spam (from Greenspan, Robyn, “The deadly duo:spam and viruses, October 2003” at cyberatlas.internet.com):

Table 1: Distribution of Spam Type

Type of Spam	Relative Frequency
Adult	.145
Financial	.162
Health	.073
Leisure	.078
Products	.21
Scams	.142

Graphically display this distribution. NOTE: In R, when implementing

```
> barplot(percent, names=c("Adult", "Finance", "Health", "Leisure", "Product", "Scams"),  
          xlab="Spam Type", ylab="Percent")
```

the labels of each category on the horizontal axis can not exceed 7 characters. For example, trying to set

```
> barplot(percent, names=c("Adult", "Finance", "Health", "Leisure", "Products", "Scams"),  
          xlab="Spam Type", ylab="Percent")
```

will leave the 5th category with no label.

3. The numbers of hikers at Bear Trap Canyon trail-head was observed on ten different afternoons in the month of August 2006: 64, 48, 42, 41, 57, 32, 34, 35, 42, 58.
 - (a) Graphically display this distribution using a stem-plot.
 - (b) Use the mean and median to give two different measures of the center of the distribution. Why are they different?
 - (c) Give two measures of the spread of the data.

4. Consider the article *Shiny Happy People* in the August 2006 issue of Discover, seen at <http://www.discover.com/issues/aug-06/features/shinyhappy/>
- How did Seligman collect his data. That is, give the sampling design. How does this limit the inferences that can be made based on results from Seligman’s data?
 - Consider Seligman’s comment: “It’s a random-assignment, placebo controlled study, the best kind of study there is.” Is this comment accurate? Why or why not?
 - Describe the placebo that Seligman is referring to. What is the purpose of the placebo in this experiment?
 - Consider the statement made in the article: “The tricky bit, skeptics say, is that they (the responses) are self reported.” Give the name of the bias (which we covered in class) that these skeptics are worried about.
5. Cocaine Addiction is hard to break. Addicts need cocaine to feel any pleasure, so perhaps giving them an antidepressant will help. A 3 year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium (the standard treatment) and a placebo. The subjects were each randomly assigned to one of the three treatments (Barnes, D.M. “Breaking the cycle of addiction.” *Science* p1029-1030, 1988). The data is summarized in Table 2:

Table 2: Cocaine relapse for three different antidepressant drugs

Cocaine Relapse?	Treatment:		
	Desipramine	Lithium	Placebo
Yes	10	18	20
No	14	6	4

- What type of experimental design is being used?
 - Construct a segmented bar chart to compare the distributions for those addicts who relapse versus those addicts who do not relapse. Be sure to put the categories of Treatment across the horizontal axis. Include the segmented bar chart in your report.
 - Based on the segmented bar chart, does there appear to be a relationship between drug type and cocaine relapse?
6. Eight college aged males took part in a study to determine the relationship between weight and number of drinks (called drinks) on Blood Alcohol Content (BAC). Download the data file “BAC.txt” from the Stat 401 web site to perform the following analysis.
- Construct a scatterplot with the variable drinks on the x -axis and BAC on the y -axis. One way to insert the plot into your report document: have the plot window in R selected and hit **Ctrl C** to copy. Go to your report document and hit **Ctrl V** to paste.
 - Based on the scatterplot, give a brief description of the relationship between the number of drinks consumed and BAC. Be certain to describe the form, direction, and strength of the relationship.

7. Recall the Jellyfish data from Project #1, where the length and breadth (in mm) of jellyfish were measured from two different locations, Dangar Island and Salamander Bay. Download the data file “jellyfish.txt” from the Stat 401 web site to perform the following analysis.
- (a) Construct a scatterplot with breadth on the x -axis and length on the y axis while using different symbols for the data from Dangar Island versus Salamander Bay. Include the scatterplot in your report.
 - (b) Does this scatterplot provide evidence that the jellyfish at one of the two locations is larger than the other?
 - (c) Give the 5 number summary of the lengths for each location.
 - (d) Construct comparative boxplots of length for each of the two locations. Include the comparative boxplot in your report.
 - (e) Discuss the similarities and differences in the boxplots. Do the boxplots support your answer to 7b? Explain.