PROJECT 3: SUMMARIZING DATA

Statistics 401: Fall 2016

Due: Tuesday at 1:40pm, September 27

To complete this project, download two data files from the Stat 401 web site: crime.txt and jellyfish.txt. Assemble your answers, including relevant tables and/or plots into a coherent well-organized write-up. Remember to label all of your graphs, and reference them from the body of your report. Put ALL R code in an Appendix. You will need the "pastecs" package in R to complete the last problem. You should have downloaded it when you installed R.

1. The National Institute of Drug Abuse (part of National Institutes of Health) conducted a survey in 2013 (if you are interested, results of the survey are at: https://www.drugabuse.gov/publications/drugfacts/nationwide-trends). Each of the people surveyed were asked what the first illicit drug was that they used. The responses are provided in Table 1.

Table 1: Frequencies of First Specific Drug Associated with Initiation of Illicit Drug Use 2013

Drug	Frequency
Marijuana	47,663
Pain Relievers	8475
Inhalants	4271
Tranquilizers	3526
Hallucinogens	1763
Sedatives	136
Cocaine	68

   (a) From Table 1, report how many people responded to the survey (i.e., give the sample size).

   (b) Convert the frequencies in Table 1 to percentages using the **prop.table( )** command.

   (c) Using your plot of choice, graphically display the results in Table 1 using percentages.

   (d) Do these data support the claim that the majority of illicit drug users in America used marijuana first? Explain.

   (e) Give the value of the statistic $\hat{p}$, defined as the proportion of survey responders who reported using marijuana first. Is $\hat{p}$ a good estimate for the true proportion p of all Americans in 2013 who used marijuana as the first illicit drug? Explain.

   (f) Regardless of your explanation to #1e, assume that $\hat{p}$ is a good estimate of p for all of the categories of Table 1. If an illicit drug user is randomly chosen, what is the probability that one of marijuana or pain relievers were used first?

   (g) Regardless of your explanation to #1e, assume that $\hat{p}$ is a good estimate of p for all of the categories of Table 1. If two illicit drug users are randomly chosen, what is the probability that neither used marijuana first?

2. The numbers of hikers at Bear Trap Canyon trail-head was observed on ten different afternoons in the month of August 2016: 64, 48, 42, 41, 57, 32, 34, 35, 42, 58.

   (a) Graphically display the distribution of these data using both a stem-plot and a histogram. Insert these graphs into your report.

(b) Describe the shape of the distribution (i.e., number of modes, symmetry, skew, and outliers).

(c) Use the sample mean $\bar{x}$ and sample median $\tilde{x}$ to give two different measures of the center of the distribution. Why are they different?

(d) Give the Greek symbol that represents the true mean number of hikers at Bear Trap over all August afternoons. Is $\bar{x}$ that you calculated in #2c a good estimate of the true mean? Explain.

(e) Give the values of two measures of the spread of the data. Which is resistant to outliers?

(f) Give your best guess of the probability that the number of hikers on an afternoon next August will be larger than 32.

3. In July, 2016, *The Washington Post* ran the article "Yes, violence in America has suddenly increased. But thats far from the whole story" (the full story is available from https://www.washingtonpost.com/news/wonk/wp/2016/07/08/why-america-feels-so-violent-right-now/). The article reports that:

   "Its hard not to feel like we're experiencing a surge of gun violence in the United States. And we are - but in the grand scheme of history, not as much ... focusing on this one-year uptick ignores the larger trend of steadily declining violence in the United States. Between 1993 and 2013, total gun homicides were nearly cut in half, primarily during the 1990s."

   The data file "crime.txt" on the STAT401 web site shows the incidences of gun homicides (as number of homicides per 100,000 Americans) from 1993 to 2013 (data available from Pew Research at http://www.pewresearch.org/fact-tank/2015/10/21/gun-homicides-steady-after-decline-in-90s-suicide-rate-edges-up/). Download crime.txt to perform the following analysis.

   (a) Construct a time series plot with the variable Year on the x-axis and gun homicides (per 100,000 people) on the y-axis.

   (b) Based on the plot, give a brief description of the relationship between the Year and gun homicides between 1993 and 2013. Be certain to describe the form, association, and strength of the relationship.

4. 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. Use 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? Explain.

   (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 4b? Explain.