Stat 111:
Introduction to Statistical Inference, 2017-18
Susan A. Murphy & Neil Shephard
Assignment 1

Credit will be given to the best grades from 3 individually marked questions (each question gets equal credit). Answer as many questions are you would like.

1. Consider the CEO Golf and Stock Data at Dartmouth, `https://www.dartmouth.edu/~chance/teaching_aids/data.html`. The direct data link, which relates the golf handicap of the CEO to the rank of the company stock performance, is
`https://www.dartmouth.edu/~chance/teaching_aids/data/golf.txt`
Based on the description of the data, pose a descriptive question, a prediction question and a causal question. Use the data to address your descriptive question. Hint: you can read data into `R` from a text file using, in this case,

```
mydata <- read.table("http://www.dartmouth.edu/
~chance/teaching_aids/data/golf.txt", header=TRUE)
```

2. A popular descriptive measure is the "histogram". An example of this is the histogram of the S&P500 arithmetic returns given in Figure 1. This is generated by the command

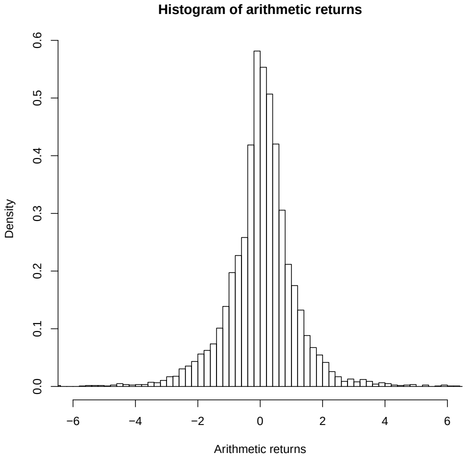

Figure 1: Histogram of arithmetic returns of S&P500 returns.

```
hist(xRet,xlim=c(-6,6),breaks=100,xlab="Arithmetic returns",
freq=FALSE,main="Histogram of arithmetic returns")
```

in R, where we have asked the computer to plot the histogram over the range of ±6 with 100 equally spaced bins. Histograms can be reported in a number of different ways: in Figure 1 we have been using the version which records the number of data points in the bin a, b scaled by the sample size and the length of the bin,

$$H_{a,b} = \frac{1}{n(b-a)} \sum_{i=1}^{n} 1_{y_i \in (a,b]}, \quad b > a,$$

where the data is $y = (y_1, ..., y_n)'$.
(a) Compute the histogram for the durations of birth data, reporting the code.
(b) Derive the expectation of

$$H_{a,b} = \frac{1}{n(b-a)} \sum_{i=1}^{n} 1_{Y_i \in (a,b]}, \quad b > a,$$

assuming that the scalar random variable $Y_i \overset{iid}{\sim}$ from some distribution function F.
(c) Derive the variance of $H_{a,b}$, assuming the data is i.i.d. from some distribution function F.

3. Consider the Auto MPG Data Set,
`https://archive.ics.uci.edu/ml/datasets/Auto+MPG`
Use the data in auto-mpg.data to plot $\bar{y}_x$ versus x for x the weight of the auto (in lbs) and y the MPG. Do this for bandwidths 60 and 160 (not 1 as in this chapter). Provide your R code as well as the graph. Do you see anything interesting in terms of prediction? Note: this data was collected by David Donoho and Ernesto Ramos in 1982.

4. Consider the Census Income data set
`https://archive.ics.uci.edu/ml/datasets/Census+Income`.
Suppose you want to address the causal question:“Does divorce impact income?” What would the purpose be of asking this question? Discuss why or why not this data might be useful in answering this question.

5. Consider the univariate density (or probability mass function for discrete random variables) of a random variable Y,

$$f(y|\eta) = h(y)\exp\{\eta T(y) - A(\eta)\}, \quad \text{where} \quad h(y) > 0,$$

where η is a scalar parameter. This is called the “canonical exponential family” of probability density functions. This family contains many famous models as special cases.
(a) Let $Y|\eta \sim N(\eta, c^2)$ where c is a known constant. What are $h(y)$, $T(y)$ and $A(\eta)$?
(b) Let $Y|\eta \sim Binomial(n, p)$ where $\eta = \log\{p/(1-p)\}$. What are $h(y)$, $T(y)$ and $A(\eta)$?
(c) Use the property of a density (or probability mass function) that it integrates (or sums) to one to show that, for any non-random δ,

$$\mathrm{E}\left\{e^{\delta T(Y)}|\eta\right\} = e^{A(\eta+\delta)-A(\eta)},$$

and use that result to show that

$$\mathrm{E}\{T(Y)|\eta\} = \frac{\partial A(\eta)}{\partial \eta}.$$

24th January 2018
To be handed in by 1.05pm, Tuesday, 6th February (week 3).