STAT 515 hw 4

Binomial, Normal probabilities/quantiles, Q-Q plots

Include the R plots and R code with your homework. You may write out your other answers by hand if you want. Just make sure you upload your homework as a single pdf (you can merge things).

1. Each visitor entering a museum must pass through one of five turnstiles: turnstiles A, B, C, D, and E. Suppose 10 visitors enter the museum and that each chooses a turnstile at random, independently of the others.
   (a) Find P(None pass through turnstile A).
   (b) Find P(Exactly three pass through turnstile A)
   (c) Find P(All pass through turnstiles C, D, and E)
   (d) Find P(Fewer than half of the visitors enter through turnstiles C,D, and E)

2. Your company makes bars of soap sold with labels claiming that they weigh 100 grams. A consumer advocacy group has said they will sue you if more than 5% of your bars weigh less than the stated 100 grams.
   (a) If your current production process makes bars with weights having the Normal($\mu = 105, \sigma^2 = 5^2$) distribution, find the percentage of bars which will weigh less than 100 grams.
   (b) You decide you must alter your production process: If your process makes bars with weights having the Normal($\mu = 105, \sigma^2$) distribution, find the largest value of σ such that no more than 5% of your bars will weigh less than the stated 100 grams.
   (c) Your process engineer claims it is impossible to make σ any smaller than 4. If your process makes bars with weights having the Normal($\mu, \sigma^2 = 4^2$) distribution, find the smallest μ such that no more than 5% of your bars will weigh less than the stated 100 grams.

3. Suppose a population has values that are Normally distributed. Then:
   (a) How many standard deviations should you go above and below the mean to in order to capture 1/3 of the values?
   (b) The top 10% of the values are all at least how many standard deviations above the mean?
   (c) What proportion of the values lie within 3 standard deviations of the mean?
   (d) What is the distance between the median and the 75th percentile in terms of the number of standard deviations?
   (e) At what percentile is a value that is 2 standard deviations above the mean?
   (f) The bottom 5% of values lie at least how many standard deviations below the mean?

4. Open R and enter `data()` into the console. A window opens showing you a bunch of data sets. You may close this window. This command just imported these standard R datasets into the workspace so you can play with them. Now type `iris` and press enter. You will see a data set printed out. Find a description of this data set at https://en.wikipedia.org/wiki/Iris_flower_data_set. Now type the command

```
hist(iris$Sepal.Length[iris$Species=="virginica"])
```

You can increase the number of bars in the histogram with

```
hist(iris$Sepal.Length[iris$Species=="virginica"],breaks=20)
```

(a) Choose one of these plots to hand in. It doesn't matter which one.

(b) Do you think that the sepal length of virginica irises follows a Normal distribution? Say why or why not.

(c) Compute the mean $\bar{X}$ and the sample standard deviation s for the this species of iris. I suggest storing the sepal lengths in a variable:

```
x <- iris$Sepal.Length[iris$Species=="virginica"]
```

Then to compute the mean $\bar{X}$, you can simply type `mean(x)` and for the standard deviation, you can type `sd(x)`.

(d) What proportion of the sepal lengths of virginica irises are less than 6.25? Hint: You can type `mean(x < 6.25)`. Note that if you type `x < 6.25`, you will get a vector of 0s and 1s, a 1 whenever $x < 6.25$ is true. Taking the mean of this vector is like counting the number of 1s and dividing by n, which is the same as computing a proportion.

(e) If the sepal lengths of virginica irises followed a Normal distribution with a mean μ equal to `mean(x)` and a standard deviation σ equal to `sd(x)`, then what proportion of sepal lengths would you expect to fall below 6.25?

(f) What is the 0.90 quantile of the 50 sepal lengths of the virginica irises? Hint: Sort the values using `sort(x)` and take the 45th value. This is the value exceeded by only 10% of the sample values.

(g) What is the 0.90 quantile of a Normal distribution with mean μ equal to `mean(x)` and a standard deviation σ equal to `sd(x)`?

(h) Would the answers from parts (d)–(g) support or not support the claim that the distribution of sepal lengths of virginica irises is Normal?

(i) Imagine doing parts (f) and (g) for many different quantiles, not just the 0.90 quantile. If the population distribution of sepal lengths of virginica irises is Normal, then the quantiles of the sample should agree with the quantiles of a Normal distribution. We can perform a visual test of this by producing a Normal Quantile-Quantile or Normal QQ plot. This plots many quantiles of the sample against the corresponding quantiles of the standard Normal distribution. *If the points in the Normal QQ plot fall more or less on a straight line*, then we say that the distribution is Normal. If the points do not fall on a straight line, we say that the distribution is not Normal.

Make a Normal QQ plot of the 50 sepal lengths of the virginica iris by using the command `qqnorm(x)`. Turn in this plot.

(j) Do you think that the distribution of sepal lengths of the virginica iris follow a Normal distribution? Why or why not?

5. The purpose of this question is to make you familiar with boxplots. Enter the following commands to store the petal lengths of the three species of irises in the objects `vir.pl`, `set.pl`, and `ver.pl`:

```
vir.pl <- iris$Petal.Length[iris$Species == "virginica"]
set.pl <- iris$Petal.Length[iris$Species == "setosa"]
ver.pl <- iris$Petal.Length[iris$Species == "versicolor"]
```

Now use the following command to create a plot which has three boxplots side by side which are labeled with the names of the species.

```
boxplot(vir.pl,set.pl,ver.pl, names=c("virginica","setosa","versicolor"))
```

(a) For which species does the sample contain an outlying petal length value?

(b) Which species appears to have the greatest average petal length?

(c) Do the petal lengths look symmetrically distributed or skewed?

(d) Order the species according to what you believe the sample variances s^2 are: which has the smallest, the second-smallest, and the largest sample variance (do this by just looking at the plot)?

(e) Now compute the sample variances s^2 of the petal lengths for each species using the commmands `var(vir.pl)`, `var(set.pl)`, and `var(ver.pl)`.

(f) If you have two boxplots, how do you judge which one (probably) has the larger sample variance s^2?