ST 516 - Homework 8 Paul ReFalo 11/17/17

1. (5 points) Suppose you are embarking on a baking business. A key ingredient for you is flour, and of most importance is that the protein content has minimal variation. That is, for consistent baked products you need the protein levels to be consistent across bags. You sample and test 60 bags from your current favorite flour brand for their protein content (measured in %), a histogram of the results is shown below.

- (a) Get this protein flour data into R

```
> protein <- c(12.06, 11.16, 11.35, 11.89, 12.49, 12.19, 11.89, 12.47, 12.42,
+              11.57, 12.2, 11.04, 12.17, 12.82, 11.81, 11.86, 11.75, 11.82,
+              12.17, 11.63, 11.54, 12.76, 12.2, 12.13, 12.08, 12.56, 12.77,
+              13.12, 12.15, 12.07, 11.48, 11.61, 12.28, 12.38, 11.67, 11.67,
+              11.55, 12.16, 12.92, 11.85, 12.53, 12.29, 12.06, 12.06, 12.01,
+              12.81, 11.78, 11.66, 11.4, 12.33, 12.21, 11.93, 12.71, 11.65,
+              12.32, 12.52, 11.84, 12.56, 13.72, 11.29)
> protein
 [1] 12.06 11.16 11.35 11.89 12.49 12.19 11.89 12.47 12.42 11.57 12.20 11.04 12.17 12.82
11.81 11.86
[17] 11.75 11.82 12.17 11.63 11.54 12.76 12.20 12.13 12.08 12.56 12.77 13.12 12.15 12.07
11.48 11.61
[33] 12.28 12.38 11.67 11.67 11.55 12.16 12.92 11.85 12.53 12.29 12.06 12.06 12.01 12.81
11.78 11.66
[49] 11.40 12.33 12.21 11.93 12.71 11.65 12.32 12.52 11.84 12.56 13.72 11.29
```

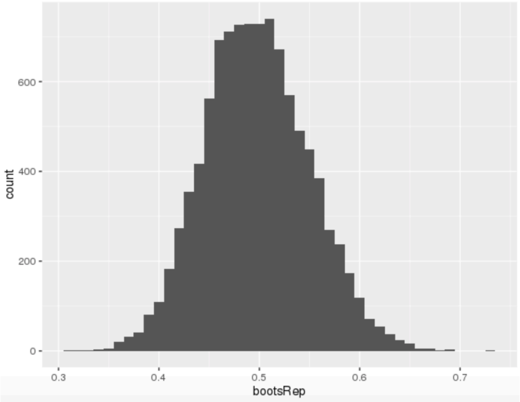
- (b) What is the point estimate for the standard deviation of protein content for the population of bags of this brand of flour?

```
> proteinSD <- sd(protein)
> proteinSD
[1] 0.504584
```

- (c) Write a function called boots() that takes a bootstrap sample of size 60 from the flour protein data; and calculates and returns the standard deviation of the bootstrap sample.

```
> boots <- function(bootStrapData){ # Arguments are data set
+   b <- sample(bootStrapData, replace = TRUE)
+   sd(b)
+}
```

. (d) Now use your function and replicate() to take and calculate the standard deviation of 10,000 bootstrap samples. Make a histogram of the results with qplot().



```
> bootsRep <- replicate(10000,
boots(protein))
> qplot(bootsRep, binwidth = .01)
```

. (e) Find the 95% bootstrap confidence interval using the percentile method and quantile(). One of the arguments for quantile() is probs, which allows you to specify the empirical quantiles you want returned. See Lecture 6 (Module 8) to review the percentile method.

```
> boots95 <- quantile(bootsRep, probs = c(.025, 0.975))
> boots95
    2.5%    97.5%
0.401219 0.602238
```

. (f) Suppose we want to use bootstrapping to test the following null and alternative hypotheses at the 5% level. In two sentences at most, what do we conclude and why?

There is strong evidence that at the 95% Confidence Interval that the standard deviation of the protein content of flour from this supplier is 0.5 (Percentile Method with quantiles 0.025 = 0.40, 0.975 = 0.60, n = 60). This is because the null hypothesis of

H_0: $\sigma = 0.5$ H_a: $\sigma \neq 0.5$

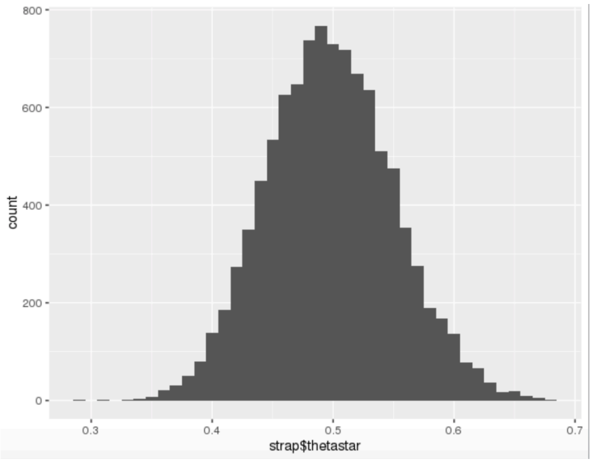
is contained in the confidence interval of 0.401 and 0.602.

2. (3 points) Not surprisingly, there is an R package called bootstrap, with a function bootstrap() that can expedite the bootstrap procedure.

- (a) Load the bootstrap package (you may need to install it first). Take a look at the help and then use bootstrap() to replicate your work in 1(d), i.e. take and calculate the standard deviation of 10,000 bootstrap samples. (Hint: the argument theta should take the value sd) Store the result in strap

```
> strap <- bootstrap(protein, 10000, sd)
```

- (b) Examine the contents strap using str(). Can you see where the 10,000 bootstrap values are stored? Pull them out and draw a histogram of them.



```
> str(strap)
List of 5
 $ thetastar     : num [1:10000] 0.464
0.526 0.522 0.474 0.503 ...
 $ func.thetastar: NULL
 $ jack.boot.val : NULL
 $ jack.boot.se  : NULL
 $ call          : language bootstrap(x =
protein, nboot = 10000, theta = sd)
> qplot(strap$thetastar, binwidth = .01)
```

- (c) Use quantile() to find the 95% bootstrap confidence interval from bootstrap(), and then repeat Problem 1, part (f).

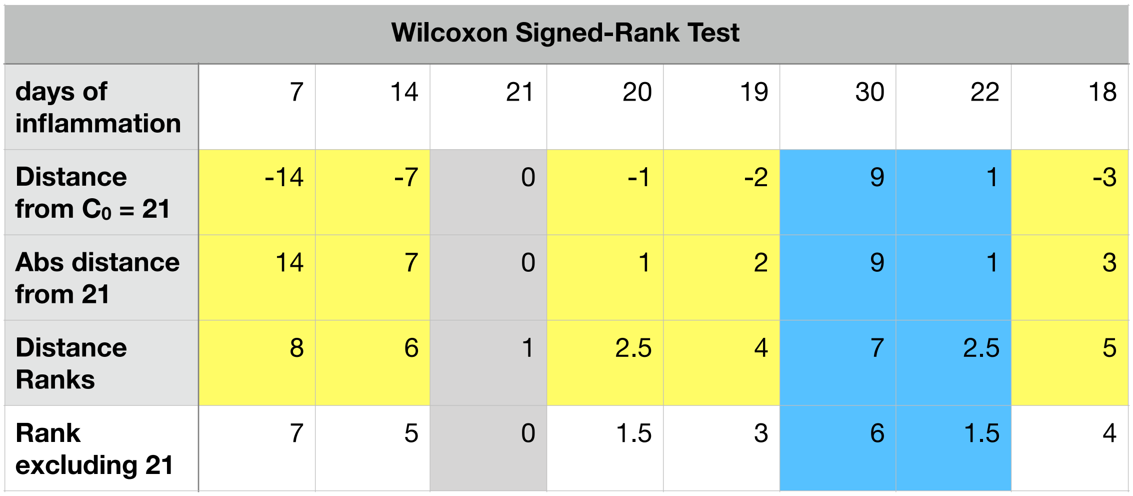
```
> strap95 <- quantile(strap$thetastar, probs = c(.025, 0.975))
> strap95
    2.5%    97.5%
0.399320 0.603723
```

There is strong evidence that at the 95% Confidence Interval that the standard deviation of the protein content of flour from this supplier is 0.5 (Percentile Method with quantiles 0.025 = 0.40, 0.975 = 0.60, n = 60). This is because the null hypothesis of

H_0: σ = 0.5 H_a: σ ≠ 0.5

is contained in the confidence interval of 0.339 and 0.604.

4. (1 point) A pathologist wishes to study the length of sinus inflammation due to rhino-sinusitis. She asks a random sample of people who have suffered from rhino-sinusitis how long they had the inflammation (in days). She wishes to test the hypothesis that the center of the distribution of inflammation times is 21 days. Calculate a signed-rank statistic for this hypothesis by hand.



Wilcoxon Signed-Rank Test								
days of inflammation	7	14	21	20	19	30	22	18
Distance from C_0 = 21	-14	-7	0	-1	-2	9	1	-3
Abs distance from 21	14	7	0	1	2	9	1	3
Distance Ranks	8	6	1	2.5	4	7	2.5	5
Rank excluding 21	7	5	0	1.5	3	6	1.5	4

The Wilcoxon signed-rank statistic is: $S_w = 6 + 1.5 = 7.5$

Data point = 21 removed from consideration (http://vassarstats.net/textbook/ch12a.html)

$\mu = n (n + 1) / 4 = (60) (61) / 4 = 915$

$\sigma^2 = n (n + 1) (2n + 1) / 24 = (60) (61) (121) / 24 = 18452.5$

$\sigma = 135.8$

$z = (\mu - S) / \sigma = (7.5 - 915) / 135.8 = -6.7$

5. (1 point) A state park ranger would like to test the hypothesis that visitors spend about a typical time of 3 hours at the state park at which he works. He records the time at which cars arrive and depart, and calculates the time spent in the park in hours. Use the signed-rank test to test the hypothesis using R and state a conclusion in the context of the problem.

```
> times <- c(3.6, 3.3, 1.2, 3.6, 2.3, 5.6, 3.4, 3.5, 3.1, 2.6)
> mean(times) # one measure of center
[1] 3.22
> median(times) # another measure of center
[1] 3.35
> wilcox.test(times, mu = 3, alternative = "two.sided")
# p-value = 0.507
```

```
      Wilcoxon signed rank test with continuity correction

data:  times
V = 34.5, p-value = 0.507
alternative hypothesis: true location is not equal to 3
```

There is strong evidence to support the ranger's hypothesis (Wilcoxon signed-rank test, V = 34.5, p-value = 0.507). With 95% confidence the typical time spent at the state park is 3 hours.