Project 4 Solutions

Statistics 401: Fall 2016

Tuesday, October 18

1. (Exercise 3.4 on p. 158; 8 pts: 2 pts for parts (a) and (b), 1 pt each for other parts) In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

   - The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds.
   - The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds.
   - The distributions of finishing times for both groups are approximately Normal.

   (a) Let M denote the finishing times for men, ages 30-34. Then the normal distribution of finishing times is $M \sim N(4313, 583)$. Let W denote the finishing times for women, ages 25-29. Then the normal distribution of finishing times is $W \sim N(5261, 807)$.

   (b) Leo's z-score is $Z = \frac{M-\mu_M}{\sigma_M} = \frac{4948-4313}{583} = 1.09$. Mary's z-score is $Z = \frac{W-\mu_W}{\sigma_W} = \frac{5513-5261}{807} = 0.31$ This means that Mary was 0.31 SD's above the mean for her group, and Leo was 1.09 SD's above the mean for his group.

   (c) Since Mary was only 0.31 SD's above the mean for her group, and Leo was 1.09 SD's above the mean for his group, then Mary ranked better in here group than Leo did in his group.

   (d) Because $P(Z > 1.09) = 0.138$, then Leo finished faster than 13.8% of the runners in his group. See R code and output in the Appendix.

   (e) Because $P(Z > 0.32) = 0.378$, then Mary finished faster than 37.8% of the runners in her group. See R code and output in the Appendix.

   (f) If the data are non-normal, then Leo's z-score $= \frac{M-\mu_M}{\sigma_M} = \frac{4948-4313}{583} = 1.09$ and Mary's z-score $= \frac{W-\mu_W}{\sigma_W} = \frac{5513-5261}{807} = 0.31$ do not change. However, the probabilities in (d) and (e) associated with these z-scores will be different!

2. (Exercise 3.12 on p. 160; 6 pts: 2 pts for part (e), 1 pt each for other parts) The distribution of passenger vehicle speeds, X, traveling on the Interstate 5 Freeway (I-5) in California is nearly normal with a mean of $\mu = 72.6$ miles/hour and a standard deviation of $\sigma = 4.78$ miles/hour.

   (a) The percent of passenger vehicles that travel slower than 80 miles/hour is calculated by

$$P(X < 80) = P\left(Z < \frac{80 - 72.6}{4.78}\right) = P(Z < 1.55) = 0.939.$$

   In other words, 93.9% of travelers on I-5 go less than 80mph.

(b) The percent of passenger vehicles that travel between 60 and 80 miles/hour is calculated by

$$P(60 < X < 80) = P\left(\frac{60 - 72.6}{4.78} < Z < \frac{80 - 72.6}{4.78}\right) = P(-2.64 < Z < 1.55)$$

$$= P(Z < 1.55) - P(Z < -2.64) = 0.939 - 0.004 = 0.935.$$

In other words, 93.5% of travelers on I-5 go between 60mph and 80mph.

(c) The fastest 5% of passenger vehicles travel is found by first noticing that to satisfy $P(Z > z^*) = 0.05$, $z^* = 1.645$. Because $Z = \frac{X-\mu}{\sigma}$, then to get X we must solve $Z = 1.645 = \frac{X-72.6}{4.78}$, which shows that $X = 80.46$ mph. In other words, the fastest 5% of drivers go 80.46mph or more.

(d) The speed limit on this stretch of the I-5 is 70 miles/hour. The percentage of passenger vehicles that travel above the speed limit on this stretch of the I-5 is found by

$$P(X > 70) = P\left(Z > \frac{70 - 72.6}{4.78}\right) = P(Z < -0.544) = 0.707$$

In other words, on I-5, 70.7% of passenger vehicles speed.

(e) If 10 motorists are clocked by the California Highway Patrol driving down I-5, the probability that at least 1 is speeding is:

$$P(\text{at least 1 is speeding}) = 1 - P(\text{none are speeding}) = 1 - (1 - 0.707)^{10} \approx 1.$$

In other words, it's a sure thing that at least one motorist is speeding.

3. (12 pts; 2 pts each for parts (b), (d), (e) and (h), 1 pt each for the other parts) A random sample of 100 calls made to the customer service center of a small bank in a month was collected.

(a) Figure 1 contains a density plot, a boxplot, and a normal probability plot of the service calls data.

(b) Table 1 gives the sample mean, sample standard deviation, and the five number summary for the service calls data.

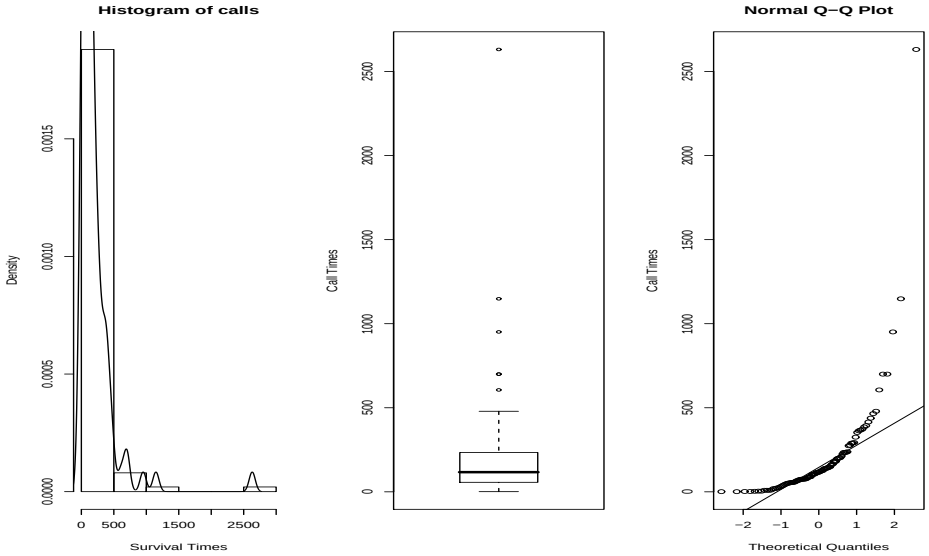
Table 1: Statistics for Service Calls data

$\bar{x}$	s	min	Q_1	$\tilde{X}$	Q_3	max
200.79	313.03	1.00	56.75	117.00	234.00	2631.00

(c) The correlation between the normal scores and the calls data shown in Figure 1 is 0.712 (see the Appendix for details). Since $.712 < r_{\text{critical}} = .98$, then the evidence suggests that there is a deviation from normality.

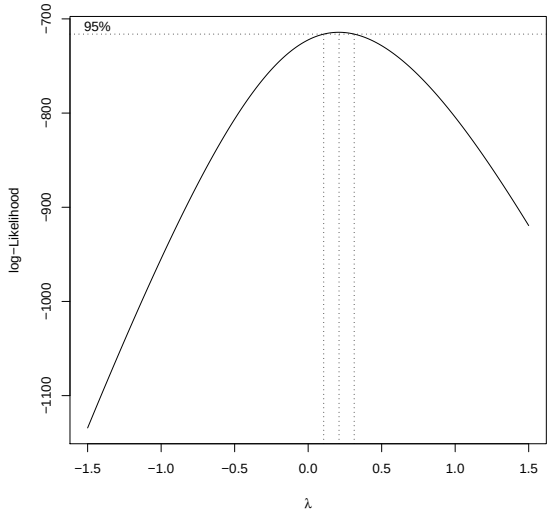
(d) The severe right skew in the histogram and boxplot, the severe deviation from linearity in the normal probability plot, as well as the test of correlation all indicate that the service calls data are not normal.

Figure 1: Checking for Normality



(e) Figure 2 shows the confidence interval for the appropriate λ for the Box-Cox transform.

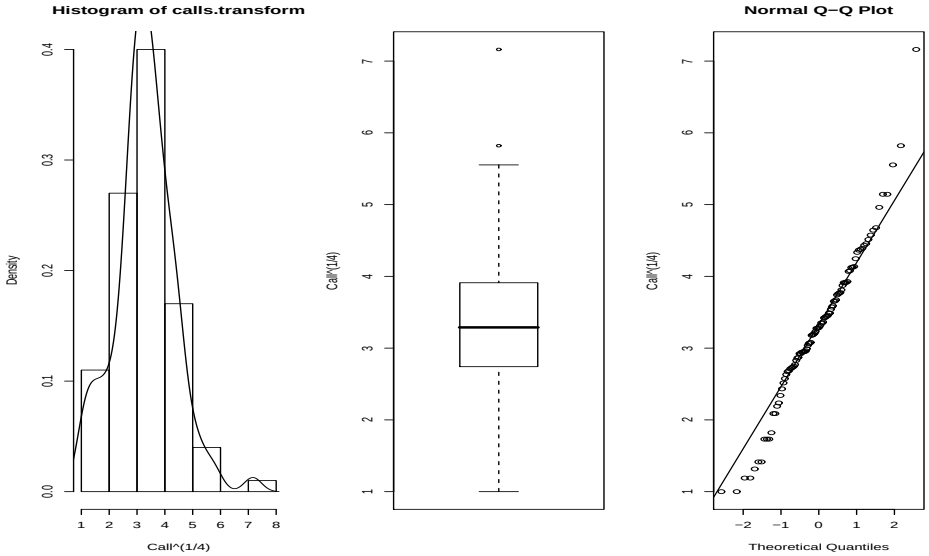
Figure 2: Confidence interval for λ for the Box-Cox transform



(f) From the confidence interval, it appears that the transform $Y = X^{\frac{1}{4}}$ is appropriate for the service calls data.

(g) Figure 3 contains a density plot, a boxplot, and a normal probability plot of the transformed calls data.

Figure 3: Checking for normality of the transformed data



(h) Table 2 gives the sample mean, sample standard deviation, and and five number summary for the transformed service calls data.

Table 2: Statistics for (Service Calls)$^{\frac{1}{4}}$

$\bar{x}$	s	min	Q_1	$\tilde{X}$	Q_3	max
3.3016	1.0573	1.000	2.745	3.289	3.911	7.162

(i) The correlation between the normal scores and the transformed calls is 0.9865378 (see the Appendix for details). Since this value is larger than the critical r value of .98 then the evidence fails to suggest that the transformed data deviates from normality.

(j) The histogram of the transformed data is less skewed than the original service calls data. The boxplot looks more symmetric and the normal probability plot looks better. Also, considering the test of correlation, we conclude that the evidence fails to suggest that the transformed data is not normal. It appears that our transformation to normality was effective.

Appendix

```
> #PROBLEM 1

# Leo Z:
> (4948 - 4313)/583
[1] 1.089194

# Mary's Z:
> (5513-5261)/807
[1] 0.3122677

# Percentiles for Leo and Mary
```

```
> 1-pnorm(c(1.09,.31))
[1] 0.1378566 0.3782805

> # PROBLEM 2

# 2(a)
> pnorm(80,mean=72.6,sd=4.78)
[1] 0.939203

# 2(b)
> pnorm(80,mean=72.6,sd=4.78)-pnorm(60,mean=72.6,sd=4.78)
[1] 0.9350083

#2(c)
> qnorm(.95,mean=72.6,sd=4.78)
[1] 80.4624

#2(d)
> 1-pnorm(70,72.6,4.78)
[1] 0.7067562


#2(e)
> 1-pnorm(70,mean=72.6,sd=4.78)^10
[1] 0.9999953

>
> # PROBLEM 3
> service=read.table("project4Calls.txt",header=T)
> attach(service)
>
> #3(a)
> par(mfrow=c(1,3))
> hist(calls,freq=F,xlab="Survival Times")
> lines(density(calls))
> boxplot(calls,ylab="Call Times")
> qqnorm(calls,ylab="Call Times")
> qqline(calls)
>
>
> #3(b)
> mean(calls);sd(calls);summary(calls)
[1] 200.79
[1] 313.0268
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   56.75  117.00  200.80  234.00 2631.00
>
>
```

```
> # 3(c)
> xy=qqnorm(calls)
> cor(xy$y,xy$x)
[1] 0.710987
>
>
> #3(e)
> library(MASS)
> par(mfrow=c(1,1))
> boxcox(calls~1,plotit=T,lambda=seq(-1.5,1.5,.01))
>
> #3(f)
> calls.transform=calls^(1/4)
>
>
> #3(g)
> par(mfrow=c(1,3))
> hist(calls.transform,freq=F,xlab="Call^(1/4)")
> lines(density(calls.transform))
> boxplot(calls.transform,ylab="Call^(1/4)")
> qqnorm(calls.transform,ylab="Call^(1/4)")
> qqline(calls.transform)
>
> # 3(h)
> mean(calls.transform);sd(calls.transform);summary(calls.transform)
[1] 3.301633
[1] 1.057268
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.745   3.289   3.302   3.911   7.162
>
>
> # 3(i)
> xy=qqnorm(calls.transform)
> cor(xy$y,xy$x)
[1] 0.9865378
```