Arif Ali
ANLY-511 Prob. Modeling & Stat. Computing
Dec 21, 2015

Take Home Examination

```
load("~/Dropbox/School/Georgetown/Analytics 511 Fall 2015/final.rdata")
```

Problem 1

Given that X_1 and X_2 are independent of one another, we are able to use the properties of Expectation and variance in the case of indepedence which is $E(Y) = E(X_1) + E(X_2)$ and $Var(Y) = Var(X_1) + Var(X_2)$. Since, we know that both σ and λ are whole numbers, the exact values should be rounded values of the the sample values.

```
mean(problem1)
## [1] 6.238424
var(problem1)
## [1] 14.77133
```

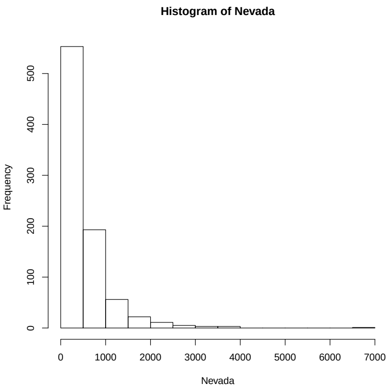
$$E(Y) = E(X_1) + E(X_2) = 0 + \hat{\lambda} = 6.238424 \implies \lambda = 6 \tag{1}$$

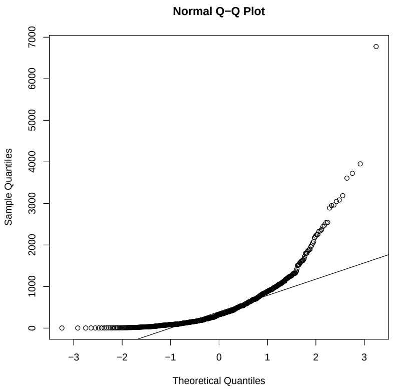
$$Var(Y) = Var(X_1) + Var(X_2) = \hat{\sigma}^2 + \hat{\lambda} = 14.77 \implies \hat{\sigma}^2 = 14.77 - 6.24 = 8.53 \implies \sigma = 3 \tag{2}$$

Problem 2

Part A

```
Nevada = handle.times[handle.times$program=="Nevada","time"]
hist(Nevada)
qqnorm(Nevada)
qqline(Nevada)
bootstrap.ne = replicate(10000, mean(sample(Nevada,
                                            length(Nevada),
                                            replace = T)))
# 95% Confidence Interval:
quantile(bootstrap.ne, c(0.025,0.975))
##     2.5%    97.5%
## 453.8555 531.3701
```





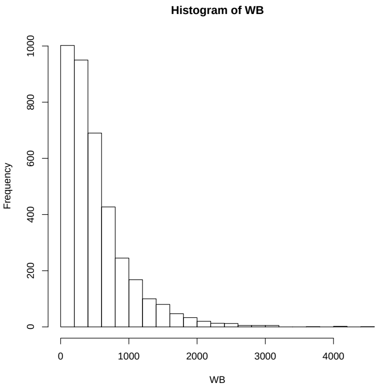
Using the z-statistic method, requires knowing the mean and variance of the entire population relative to a sample. The t-statistic confidence interval method assumes that μ is the sample mean and that we can the sample standard deviation. However, this method also assumes that the sample should follow close the normal distribution. Based on the histogram and qqnorm plot, it is clear that this is not the case. Given out skewed the data is, the bootstrap method for the confidence interval is better compared the the t-statistic or z-statistic (normal method).

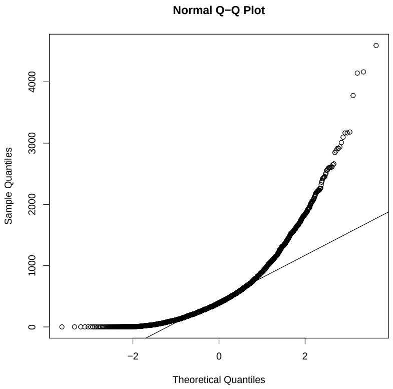
Part B

$H_0 : \mu = 540$ $H_a : \mu < 540$

```
WB = handle.times[handle.times$program=="West Billing","time"]
hist(WB, breaks = 20)
qqnorm(WB); qqline(WB);
E2b = replicate(10000, mean(sample(WB, length(WB), replace = T)))

#P-value
mean(E2b>=540)
## [1] 0.0021
```





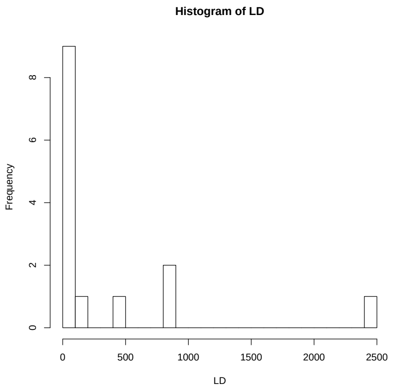
The t test assumes that the data is somewhat normal; however, based on the histogram and the qqnorm plots, it is clear that the data is not normal and skewed to the left. Therefore, the best idea is too see permutations of WB by sampling them with replacement and determining the mean. Based on the p-value, we reject the null hypothesis that mean of handle times for West Billing is 540 in favor of the alternative hypothesis, that handle times are less than 540 seconds.

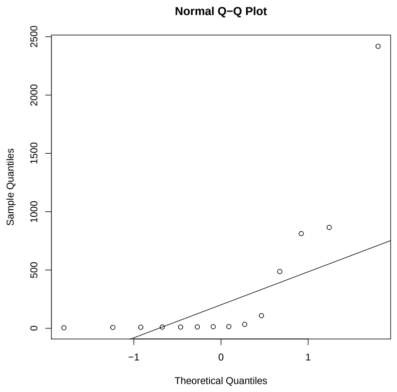
Part C

```
LD = handle.times[handle.times$program=="LD Care","time"]
hist(LD, breaks = 20)
qqnorm(LD); qqline(LD);

LD.WB = function(){
  l.sample = mean(sample(LD, length(LD), replace = T))
  w.sample = mean(sample(WB, length(WB), replace = T))
  return(l.sample-w.sample)
}
lw.bootstrap = replicate(10000, LD.WB())

# 90% Confidence Interval:
quantile(lw.bootstrap, c(0.05,0.95))
##        5%       95%
## -423.9365  139.2426
```



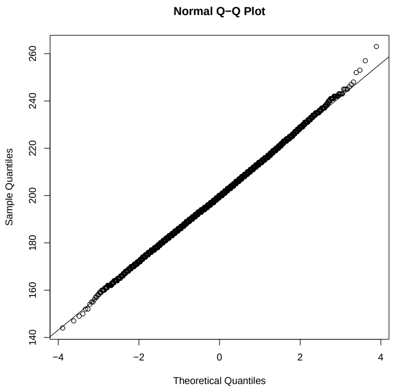


Neither of the samples follow close to a normal distribution, so both the the z and t statistics methods for building a confidence interval is not appropriate. Based on the histogram, both samples are very skewed, so it seems to be best to apply the bootstrap method for the confidence interval.

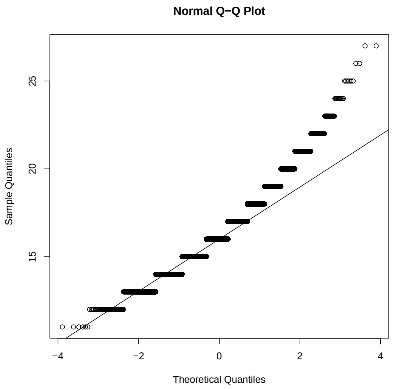
Problem 3

Part A

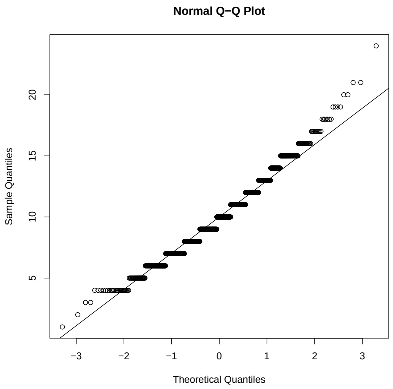
```
Part3ai = replicate(10000,sum(rpois(n = 20, lambda = 10)))
alpha = 0.05
qqnorm(Part3ai)
qqline(Part3ai)
```



```
c1 = qnorm(1-alpha, mean(Part3ai), sd(Part3ai))
c1
## [1] 222.9913
Part3aii = replicate(10000,max(rpois(n = 20, lambda = 10)))
qqnorm(Part3aii)
qqline(Part3aii)
```



```
sample.poisson = rpois(n = 1000, lambda = 10)
qqnorm(sample.poisson)
qqline(sample.poisson)
```



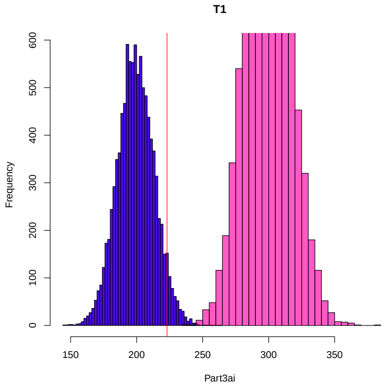
```
c2 = qpois(1-alpha, mean(Part3aii))
c2
## [1] 23
```

The replications for both T_1 and T_2 was done 10000 times in order to simulate the various values. In order to properly T_1 and T_2, I replicated both methods using rpois to generate the values of $x_1, x_2, ..., x_{20}$. For T_2, the assumption of normality cannot hold. This is confirmed by the qqnorm plot which shows a step like structure. This structure is similar to a qqnorm of numbers generated from a poisson distribution. Therefore, for c_2, the best approach to estimating is using the qpois approach.

Part B

With the T_1 Test:

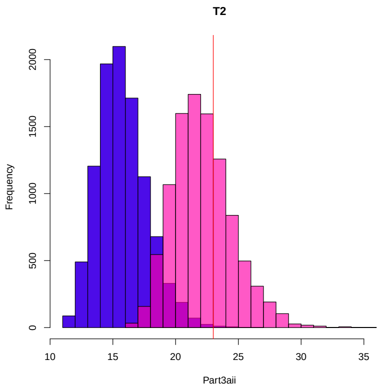
```
Part3bi = replicate(10000,sum(rpois(n = 20, lambda = 15)))
hist(Part3ai, xlim = c(min(Part3ai), max(Part3bi)), breaks = 50, col = "#4C0CE8", main = "T1")
hist(Part3bi, breaks = 50, add = T, col = "#FF00A8A6")
abline(v = c1, col = 'red')
```



```
# Type II error:
pnorm(c1, mean(Part3bi), sd(Part3bi))
## [1] 3.612225e-06
```

The Histogram shows the data from the two simulations next to one another. The second histogram is transparent to show where the graphs overlap. The line represents where c_1 is in relation to both simulations. It's interesting to see that c_1 appears to not occur at any point of the histogram based on $\lambda = 15$. Running the pnorm gives the probability of this ever occuring. The value is so small, so the Type II error is very small, which means that the Power of the test is large. With the T_2 Test:

```
Part3bii = replicate(10000,max(rpois(n = 20, lambda = 15)))
hist(Part3aii, xlim = c(min(Part3aii), max(Part3bii)), breaks = 20, col = "#4C0CE8", main = "T2")
hist(Part3bii, breaks = 20, add = T, col = "#FF00A8A6")
abline(v = c2, col = 'red')
```



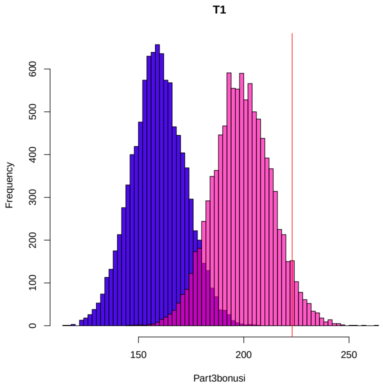
```
# Type II error:
ppois(c2, mean(Part3bii))
## [1] 0.5843319
```

The same logic behind the histogram from T_1 was the same that was applied to T_2 using ppois instead of pnorm, but the line represents where c_2 is capturing more of the distribution generated from $\lambda = 15$. Unlike the histograms for T_1, there is noticable overlap between the two histograms. The value of the type II is larger for T_2 as opposed to T_1.

Bonus

Using the T_1 Test:

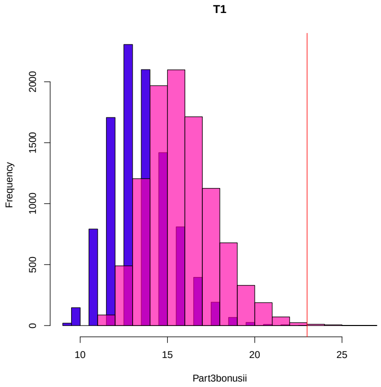
```
Part3bonusi = replicate(10000,sum(rpois(n = 20, lambda = 8)))
hist(Part3bonusi, xlim = c(min(Part3bonusi), max(Part3ai)), breaks = 50, col = "#4C0CE8", main = "T1")
hist(Part3ai, breaks = 50, add = T, col = "#FF00A8A6")
abline(v = c1, col = 'red')
```



```
# Probability of rejecting H0:
pnorm(c1, mean(Part3bonusi), sd(Part3bonusi), lower.tail = F)
## [1] 3.985083e-07
```

Using the T_2 Test:

```
Part3bonusii = replicate(10000,max(rpois(n = 20, lambda = 8)))
hist(Part3bonusii, xlim = c(min(Part3bonusii), max(Part3aii)), breaks = 20, col = "#4C0CE8", main = "T1
hist(Part3aii, breaks = 20, add = T, col = "#FF00A8A6")
abline(v = c2, col = 'red')
```



```
# Probability of rejecting H0:
ppois(c2, mean(Part3bonusii), lower.tail = F)
## [1] 0.007217956
```

The second test (T_2) has a higher probability of rejecting H_0. In order to determine this, I applied the same logic as in part b; however, I had to look at $P(X > c_i)$ so the lower.tail needed to be set to False in pnorm. Since, the alternative is that $\lambda > 10$, it should be expected that the Type I error would be small. Since (T_1) involves summing up all the values, each value is lower which helps keep the overwhelming majority of simulated sums under the value of c_1.

Problem 4

Part A

```
EX = (mean(problem4)); MX = median(problem4); SDX = (sd(problem4))
3*(EX-MX)/SDX
## [1] 1.080882
```

For the plug in estimate, I used the the sample mean, median, and standard deviation. I computed the sample mean, median, and standard deviation, denoting them as EX, MX, and SDX repectively, then applied the formula for Pearson 2 skewness given in problem 4.

Part B

```
bootstrap.p2s = function(){
  p4 = sample(problem4, length(problem4), replace = T)
  return(3*(mean(p4)-median(p4))/(sd(p4)))
}
bootstrap.p2sb = replicate(10000, bootstrap.p2s())
quantile(bootstrap.p2sb, c(0.05, 0.95))
##       5%      95%
## 0.763949 1.274624
```

In order to run the bootstrap method, the first step was to create a no input function that samples the Problem4 data with replacement and then applies the Pearson 2 skewness formula. I replicated the process 10,000 times and constructed a 90% confidence interval by using the quantile function at 5% and 95%.

Bonus

```
mean(bootstrap.p2sb)-3*(EX-MX)/SDX
## [1] -0.05069033
```

There is a slightly negative bias using the bootstrap method. To calculate the Bias I did the following

$$Bias_{boot}[\hat{\theta}^*] = E[\hat{\theta}^*] - \hat{\theta} \tag{3}$$

where $E[\hat{\theta}^*]$ is the mean of the bootstrap distribution and $\hat{\theta}$ is the estimate from the orginal data.

Problem 5

Part A

```
mean(problem5$A==0)
## [1] 0.624
mean(problem5$A==1)
## [1] 0.376
```

These are just straight calculations from the data set with respect to column A. The mean of boolean vectors returns the percentage of the vector that is defined as TRUE by the boolean structure.

Part B

```
mean((problem5$B[problem5$A==1])==1)
## [1] 0.5718085
mean((problem5$B[problem5$A==0])==1)
## [1] 0.2035256
```

Subsetted Column by column A to represent the subset of space, then I applied the same logic as Part A.

Part C

```
partC = table(problem5$B, problem5$C)
chisq.test(partC)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  partC
## X-squared = 2.1224, df = 1, p-value = 0.1452
```

The cell counts for the oberved values are are greater than 5, so using the chisq.test with result in no warnings. Since the P-value of the Chi-Square for B and C is 0.1452, the null hypothesis that B and C are independent at an $\alpha = 0.05$ will hold since $0.1452 > 0.05$.

Part D

```
partD = table(problem5$A, problem5$D)
chisq.test(partD)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  partD
## X-squared = 1.5225, df = 1, p-value = 0.2172
```

The cell counts for the oberved values are are greater than 5, so using the chisq.test with result in no warnings. Since the P-value of the Chi-Square for A and D is 0.2172, the null hypothesis that A and D are independent at an $\alpha = 0.05$ will hold since $0.2172 > 0.05$

Part E

```
mean((problem5$D[problem5$B==1 & problem5$C==0])==1)
## [1] 0.2132701
```

The And statement corresponds to using And in the probability statement $P(D = 1|B = 1, C = 0)$.