Stat 641 Fall 2012 - Solutions for Assignment 6

I. (28 points) Lifetime of 25 batteries:

1. The exact distribution of $\bar{Y}$ is determined using the fact that Y_i's are independent exponential r.v.s with parameter β, therefore $T = \sum_{i=1}^{n} Y_i$ is the sum of iid Exponential and hence has a Gamma distribution with shape parameter $\alpha = n$ and scale parameter β. The distribution of $\bar{Y}$ is obtained from $\bar{Y} = \frac{1}{n}T$ which is a Gamma distribuion with shape parameter $\alpha = n$ and scale parameter β/n.

   a. We can verify this result as follows:
   Let $\bar{Y}$ have cdf H, T have cdf G and pdf given by : $g(t) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} t^{\alpha-1} e^{-t/\beta}$(Gamma with parameters α and β). Then we have

$$H(y) = P(\bar{Y} \leq y) = P\left(\frac{1}{n}T \leq y\right) = G(ny) \Rightarrow h(y) = \frac{d}{dy}H(y) = \frac{d}{dy}G(ny) = ng(ny)$$

$$\Rightarrow h(y) = ng(ny) = \frac{n}{\Gamma(\alpha)\beta^{\alpha}}(ny)^{\alpha-1}e^{-ny/\beta} = \frac{1}{\Gamma(\alpha)(\beta/n)^{\alpha}} y^{\alpha-1} e^{-y/(\beta/n)}$$

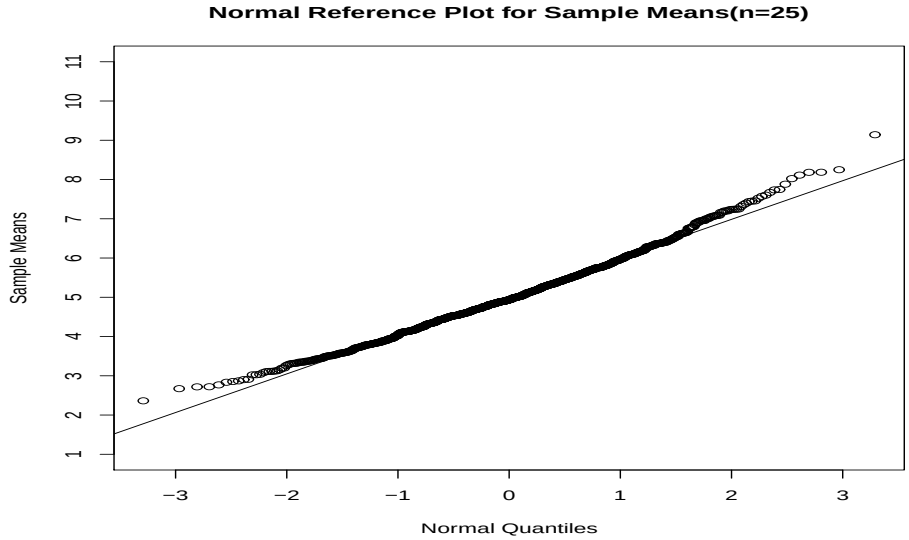
   From the above we can identify that $h(y)$ is the pdf of a Gamma distribution with shape parameter α and scale β/n. In particular, $\bar{Y}$ has a Gamma(25,5/25) = Gamma(25,.2) distribution.

   b. $\mu_{\bar{Y}} = \mu_Y = \beta = 5$

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{\beta}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$$

2. From the normal reference distribution plot on the next page, the sample means deviate from a normal distribution, with the right tail somewhat longer and the left tail shorter than a normal distribution. The Shapiro-Wilk test has p-value = 3.229e-06 indicating that the sampling distribution of the sample means has a very poor fit to a normal distribution. If the sample size was increased from 25 to 50 we would observe a better fit of the normal distribution, although the p-value from the Shapiro-Wilk test will remain relatively small due to the very large number of sample means, 1000, which results in a test that is very sensitive to small deviations. Note that you might get a little bit different normal reference plot from your simulation. The following R code was used:

```
M=1000
means = numeric(M)
for(i in 1:M){
  sample = rexp(25,1/5)
  means[i] = mean(sample)
}
means = sort(means)
k = seq(1,M,1)
u = (k-.5)/M
Z = qnorm(u)
plot(Z,means,main="Normal Reference Plot for Sample Means(n=25)",
  xlab="Normal Quantiles",ylab="Sample Means",lab=c(5,11,7),ylim=c(1,11))
abline(lm(means~Z))
shapiro.test(means)
```



3. Compute $P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right)$:

   a. We have from Part 1.a that $\bar{Y}$ has a Gamma(25,.2) distribution. Therefore,
   $P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right) = P(4.8 \leq \bar{Y} \leq 5.2) = 0.1580747$ obtained using the following R command (recall in R, the gamma distribution uses $1/\beta$ not β

   ```
   pgamma(5.2,25,1/.2)-pgamma(4.8,25,1/.2)
   ```

   b. Since $Y_1, \ldots, Y_{25} \overset{\text{iid}}{\sim} Exp(5)$, we have by the central limit theorem that approximately (for large n), $\bar{Y} \sim N(\mu, (\sigma/\sqrt{n})^2) = N(5, (5/\sqrt{25})^2) = N(5, 1)$. Thus

   $$\begin{aligned} P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right) &= P\left(\frac{-0.2}{1} \leq \frac{\bar{Y} - 5}{1} \leq \frac{0.2}{1}\right) \\ &\approx P\left(-0.2 \leq Z \leq 0.2\right) = 0.1585194 \end{aligned}$$

   obtained using the R command:

   pnorm(.2) - pnorm(-.2)

   c. In my generated sample, there were 156 out of 1000 values of $\bar{Y}$ that satisfied $-0.2 \leq \bar{Y} - 5 \leq 0.2$. Thus, the estimated value of the probability is 156/1000=0.156. Again, this value changes from simulation to simulation, and can vary considerably. (I ran the simulation four times, and obtained the following results 0.151, 0.156, 0.163, 0.159). These results can be obtained by including the following R command along with the code on the previous page:

   ```
   sum((means-5>=-0.2 & means-5<=0.2)*1)/1000
   ```

4. The three values for computing $P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right)$ are quite close.

   Exact=.1580747; CLT=.1585194; Simulation=.156

II. (24 points) CPUE problem.

The following R code is used to obtain the solutions to parts 1. and 2.:

```
CPUE=c(0.6 ,   0.7 ,   1.1   ,   1.3    , 1.8  ,   2.0 ,    2.3 ,    2.7    ,
2.9    ,  3.1  , 3.9   ,  4.3  ,   4.4  ,    4.9 ,   5.2 ,    5.4 ,    6.1 ,    6.8 ,
7.1 ,    8.0 ,  9.4   , 10.3  ,  12.9  ,   15.9 ,  16.0,    22.0 ,   22.2 ,   22.5 ,
23.0 ,   23.1 ,  23.9  , 26.5  ,  26.7  ,   28.4 ,  28.5 ,   32.2 ,   40.2 ,   42.5 ,
47.2 ,   48.3 ,  55.8  , 57.0  ,  57.2  ,   64.9 ,  67.6  ,   71.3 ,   79.5 ,   114.5 ,
128.6 ,  293.5)
y=log(CPUE)
mean.hat = mean(y)
se.hat = sd(y)/sqrt(50)
M =  5000
est = matrix(0,M,4)

for(i in 1:M){
  sample.boot = sample(y,replace=T)
  a = mean(sample.boot)
  b = median(sample.boot)
  c = sd(sample.boot)
  d = mad(sample.boot)
  est[i,] = c(a,b,c,d)
}

se.boot = sd(est[,1])
med.m.boot = mean(est[,2])
med.se.boot = sd(est[,2])
S.m.boot = mean(est[,3])
S.se.boot = sd(est[,3])
MAD.m.boot = mean(est[,4])
MAD.se.boot = sd(est[,4])
output=c(se.boot,med.m.boot,med.se.boot,S.m.boot,S.se.boot,MAD.m.boot,MAD.se.boot)
output
```

1. The standard error of $\bar{Y}$ by bootstrapping is se.boot = 0.2027. From the 50 original log(CPUE) values compute $S_Y = 1.453$. The estimated standard error of $\bar{Y}$ using just the original data is equal to $S_Y/\sqrt{n} = 1.453/\sqrt{50} = 0.2055$ (se.hat in the above R code). The two estimates of the standard error of $\bar{Y}$ are relatively close.

2. The result using the bootstrap R code is summarized in the following table.

Statistic	$\hat{Q}(0.5)$	S	$\widehat{MAD}$
Estimated Mean	2.8087	1.4334	1.5733
Estimated Standard Deviation	0.3729	0.1167	0.2255

3. Using the results from Handout 10, when sampling from a $N(3,(1.5)^2)$ distribution,

- The sample median, $\hat{Q}(.5)$, asymptotically, has mean and standard deviation

  $\mu_A = Q(0.5) = \mu = 3$

  $\sigma_A = \sqrt{0.5(1-0.5)}/\left[f(Q(0.5))\sqrt{50}\right] = \sqrt{0.5(1-0.5)}/\left[(.265962)\sqrt{50}\right] = 0.26587$

  where $f(Q(0.5) = f(3) = 1/(\sigma\sqrt{2\pi}) = 1/(1.5\sqrt{2\pi}) = .265962$ with f the $N(3,(1.5)^2)$ pdf.

- The sample standard deviation, S, asymptotically, has mean and standard deviation

  $\mu_A = \sigma = 1.5$

  $\sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{2(1.5)^4}}{2(1.5)\sqrt{50}} = .15$

Note that the sample standard deviation has sampling distribution $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Therefore, we can compute the exact values of $\mu_S = E[S]$ and $\sigma_S = \sqrt{Var[S]}$:

$$E(S) = c_n\sigma = \left[\sqrt{\frac{2}{n-1}}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)\right]\sigma = (.9949113)(1.5) = 1.492367$$

$$\sigma_S = \sigma\sqrt{(1-c_n^2)} = \sigma\sqrt{(1-(.9949113)^2)} = 0.151132$$

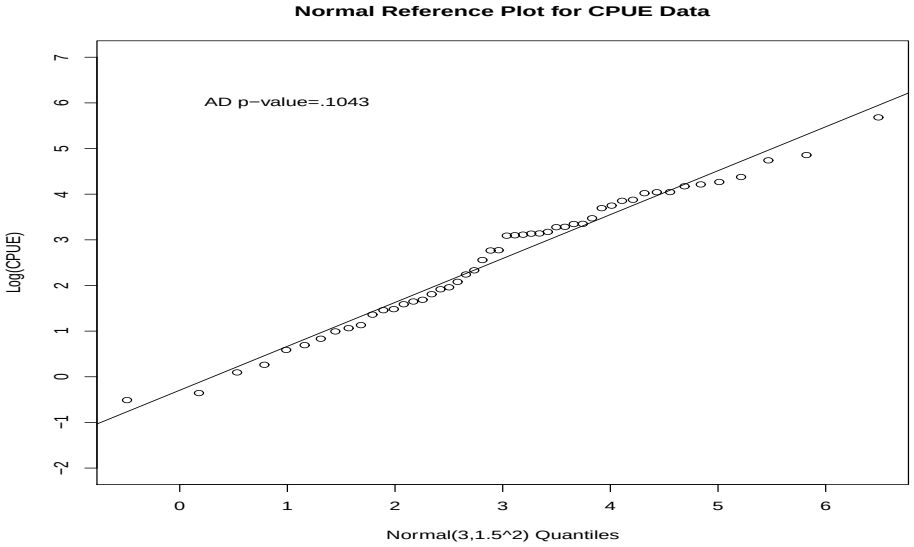
where the following R functions are used to compute $c_n\sigma = \left[\sqrt{\frac{2}{n-1}}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)\right]\sigma$

```
c_n = sqrt(2/(50-1))*gamma(50/2)/gamma((50-1)/2)
```

The result is summarized in the following table.

Statistic	AsympMean	AsympStDev	BootMean	BootStDev	Exact Mean	Exact StDev
$\hat{Q}(0.5)$	3.0	0.2659	2.8087	.3729	N/A	N/A
S	1.4924	0.1511	1.4334	.1167	1.5	.15

Comparing the two sets of values, there is a general agreement between the bootstrap values and the theoretical asymptotic values. For the median, $Q(.5)$, the asymptotic mean is about 7% larger than the bootstrap mean, whereas, for the sample standard deviation, S, the asymptotic mean is about 4% larger than the bootstrap mean. The reflects the fact that the sample size is n=50, not infinity and that the 50 values of CPUE may not have exactly a log-normal distribution. In fact, if we fit a $N(3,(1.5)^2)$ distribution to the data, we obtain a p-value of .1043 from the Anderson-Darling test and the corresponding normal reference distribution plot is presented here.



Note that there were a few points that deviate from the straight line in the normal reference distribution plot, which indicates a good but not perfect fit. Also, note that the bootstrap mean for S is about 4% smaller that the exact mean of S but the bootstrap standard deviation for S is 22% smaller than the true standard deviation of S. This lends more evidence that the distribution of log(CPUE) does not exactly fit a $N(3,(1.5)^2)$ distribution.

III. (24 points)

III-1. Y is the number of correct answers out of 100 questions. Y has a binomial distribution with n=100, p=0.2

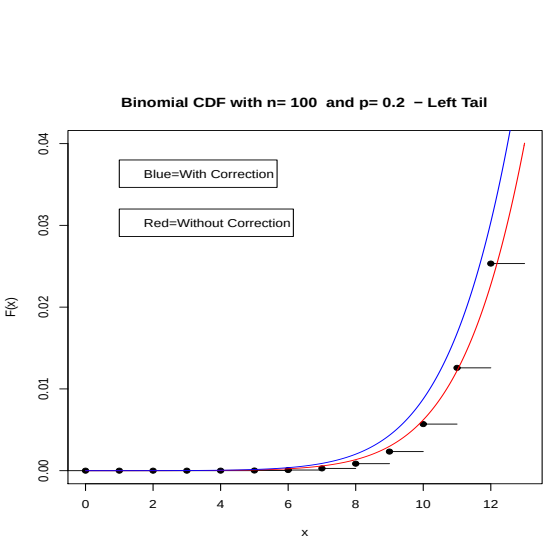
a. $\mu_Y = E[Y] = np = 100(.2) = 20$ and $\sigma_Y = \sqrt{Var(Y)} = \sqrt{np(1-p)} = \sqrt{100(.2)(.8)} = 4$

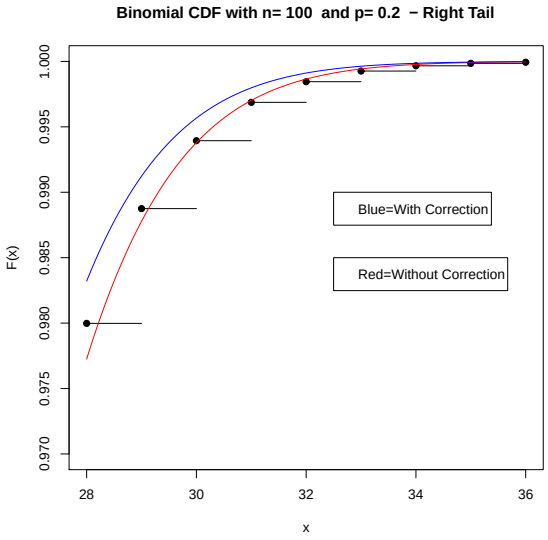
b. $\hat{p} = \frac{Y}{100} \Rightarrow \mu_{\hat{p}} = E\left(\frac{Y}{n}\right) = \frac{E(Y)}{n} = \frac{np}{n} = p = .2$ and

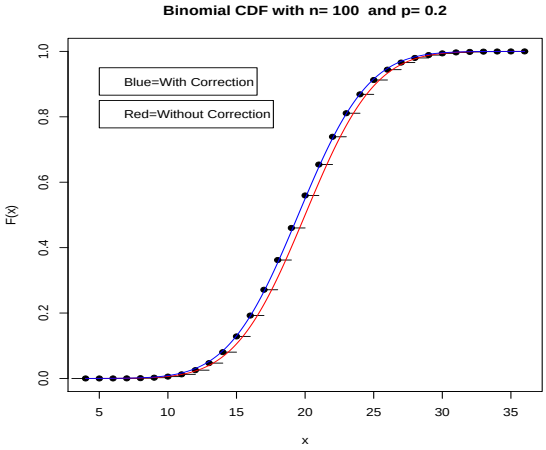
$sd(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{Var(\frac{Y}{n})} = \sqrt{\frac{Var(Y)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.2)(.8)}{100}} = .04$

c. $P(Y \geq 30) = 1 - P(Y \leq 29)$

- Exact calculation using B(100,.2) distribution:
  $P(Y \geq 30) = 1 - P(Y \leq 29) = 1 - pbinom(29, 100, .2) = 0.01125$
- Normal Approx to Binomial With Correction:
  $P(Y \geq 30) = 1 - P(Y \leq 29) \approx 1 - P\left(Z \leq \frac{29-20+.5}{\sqrt{100(.2)(.8)}}\right) = 1 - pnorm\left(\frac{29-20+.5}{\sqrt{100(.2)(.8)}}\right) = .008771$
- Normal Approx to Binomial Without Correction:
  $P(Y \geq 30) = 1 - P(Y \leq 29) \approx 1 - P\left(Z \leq \frac{29-20}{\sqrt{100(.2)(.8)}}\right) = .01222$
- The following graphs illustrates that for the values of Y in the middle of the binomial distribution the normal approximation with the correction is more accurate than the approximation without the correction. However, in the tails of the distribution, there are regions where the opposite is true.







III-2. Let $U = (n-1)S^2/\sigma^2$ where S^2 is the sample variance and σ^2 is the population variance.

a. Generate the 1000 values of U based on data from the normal distribution using the R code:

```
r = 1000
yn = rep(0,10)
Un = rep(0,r)
for (i in 1:r){
yn[i] = rnorm(10,20,5)
Un[i] = (10-1)*var(yn[i])/25
}
p = c(.1,.25,.5,.9,.95,.99)
samquantn = quantile(Un,p)
```

b. Generate the 1000 values of U based on data from a transformed gamma distribution using the R code:

```
r = 1000
yg = rep(0,10)
Ug = rep(0,r)
for (i in 1:r){
yg[i] = 18.4 + 15.8*rgamma(10,.1,1)
Ug[i] = (10-1)*var(yg[i])/25
}
p = c(.1,.25,.5,.9,.95,.99)
samquantg = quantile(Ug,p)
```

c. Generate the 1000 values of U based on data from a transformed t distribution using the R code:

```
r = 1000
yt = rep(0,10)
Ut = rep(0,r)
for (i in 1:r){
yt[i] = 20+(5/sqrt(3))*rt(10,3)
Ut[i] = (10-1)*var(yt[i])/25
}
p = c(.1,.25,.5,.9,.95,.99)
samquantt = quantile(Ut,p)
theoryquant = qchisq(p,9)
out = cbind(p,theoryquant,samquantn,samquantg,samquantt)
out
```

d. From the above we obtain the following four sets of values for Q(p):

2nd column - Q(p)=qchisq(.1,9); columns 3-5 are the estimated values of Q(p) from the simulations

p	Chisquare Distribuion	Normal Simulation	Gamma Simulation	t Simulation
0.10	4.168159	4.077350	0.03323729	1.398460
0.25	5.898826	5.931916	0.33862411	1.981606
0.50	8.342833	8.352569	1.90705478	2.804554
0.90	14.683657	14.449155	19.91701559	4.754567
0.95	16.918978	16.492388	41.63029146	5.621921
0.99	21.665994	20.157174	96.07975651	7.019786

The sample quantiles of U when sampling from a Normal distribution are very close to the theoretical chi-square quantiles. When the samples are drawn from a Gamma distribution, the quantiles of U would indicate the distribution of U is much more highly skewed in comparison to the chi-square distribution. When the samples are drawn from a t-distribution with df=3, the quantiles of U would indicate that the distribution of U has a more shorter tail than the distribution of U when sampling from a Normal distribution.

III-3. Let p=the proportion of people in the population belonging to the sensitive group. Thus, p is the probability of Yes response to Question 1 and $1-p$ is the probability of Yes response to Question 2 (assuming that people answer truthfully).

(a)

$$\begin{aligned}\pi &= P(\text{Yes}) \\ &= P(\text{Yes }|\text{Question 1 is Answered})P(\text{Question 1 is Answered}) + \\ &\quad P(\text{Yes }|\text{Question 2 is Answered})P(\text{Question 2 is Answered}) \\ &= p\theta + (1-p)(1-\theta) = p(2\theta-1) + (1-\theta)\end{aligned}$$

(b) Let X be the number of "Yes" responses in a simple random sample of n people.
Then $\hat{\pi} = \frac{X}{n}$ with $X \sim Binomial(n, \pi)$. Therefore, $E[\hat{\pi}] = E[X]/n = n\pi/n = \pi$, we have the following:

$$E[\hat{p}] = \frac{E[\hat{\pi}] - (1-\theta)}{2\theta - 1} = \frac{\pi - (1-\theta)}{2\theta - 1} = \frac{p(2\theta-1) + (1-\theta) - (1-\theta)}{2\theta - 1} = p$$

by plugging the value of $\pi = p(2\theta-1) + (1-\theta)$ from (a) into this expression. Thus, $\hat{p}$ is an unbiased estimator of p.

(c) $\hat{\pi} \sim Binomial(n, \pi), \Rightarrow Var(\hat{\pi}) = \pi(1-\pi)/n$, thus yielding

$$Var(\hat{p}) = \frac{Var[\hat{\pi}]}{(2\theta-1)^2} = \frac{\pi(1-\pi)}{n(2\theta-1)^2} = \frac{(p(2\theta-1) + (1-\theta))(\theta - p(2\theta-1)))}{n(2\theta-1)^2} = \frac{p(1-p)}{n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2}$$

By plugging the value of $\pi = p(2\theta-1) + (1-\theta)$ from (a) into this expression and simplifying, we can obtain the desired result.

(d) The second term in $Var(\hat{p})$ goes to infinite as θ goes to $\frac{1}{2}$ and goes to 0 as θ goes to 0 or 1. Thus, the variability in the estimation of p due to randomization is getting smaller as θ moves away from $\frac{1}{2}$ toward 0 or 1 but would be extremely large if θ is too close to 1/2.

IV. (24 points -3 pts each) Multiple Choice Questions.

1. **C -** The MSE takes into account both the variance and bias of the estimator
2. **B or C** The bootstrap is often used when the form of θ is very complex and/or the asymptotic results cannot be applied.
3. **C -** The accuracy of the approximation is controlled by both the sample size n and how close a match there is between $\hat{F}$ and F.
4. **E -** The asymptotic efficiency of the median to the mean varies from greater than 1 to less than 1 depending of the population cdf (see page 10 in Handout 10)
5. **E -** The CLThm does not hold for all estimators;
   (see page 13 in Handout 10 concerning the asymptotic distributions of min and max.
6. **D -** In most cases we will want the estimator having smallest MSE.
7. **E -** The CLThm does not apply to minimum and maximum;
   Exact derivation requires knowing the population cdf F;
   Simulation requires knowing the population cdf F;
   Bootstrapping would not work because all bootstrap values for R would be smaller than the value of R in the original data.
8. **D -** If $\hat{\theta}$ is an unbiased estimator of θ then $E[\hat{\theta}] = \theta$ which implies that the mean of the sampling distribution of $\hat{\theta}$ is θ