Stat 641 Fall 2021

Solutions for Assignment 5

Prob. 1. (5 points) The process distribution appears to be right skewed with the left tail much shorter (lighter) than a normal distribution the right tail much longer (heavier) than a normal distribution.

Prob. 2. (10 points) Let X be the number of breaks on a given bar. Then, the estimated probability of a break at a given location is

$$\hat{p} = \frac{(0)(140) + (1)(197) + (2)(115) + (3)(41) + (4)(5) + (5)(2)}{5 \times 500} = .232$$

Next evaluate if the distribution of X is Binomial(500, p).
Under the Binomial model, $P[X = i] = p_i = \binom{5}{i} p^i (1-p)^i$, for $i = 0, 1, 2, 3, 4, 5$
Because p is unknown use $\hat{p} = .232$.
An initial calculation of the $E_i = 500 p_i$s yields $E_5 = .336$ which is less than 1
Therefore, combine the last two cells and then compute the following using the R-function $p_i = dbinom(i, 5, .232)$ for $i = 0, 1, 2, 3$ and $p_4 = P[X \geq 4] = 1 - P[X \leq 3] = 1 - pbinom(3, 5, .232)$
then compute $E_i = 500 * p_i$ yielding the following values

$$\begin{aligned}
\widehat{p}_0 = \text{dbinom}(0, 5, .232) = 0.2672 \quad &\Rightarrow \quad \widehat{E}_0 = (500)(\widehat{p}_0) = 133.59 \\
\widehat{p}_1 = \text{dbinom}(0, 5, .232) = 0.4036 \quad &\Rightarrow \quad \widehat{E}_1 = (500)(\widehat{p}_1) = 201.78 \\
\widehat{p}_2 = \text{dbinom}(0, 5, .232) = 0.2438 \quad &\Rightarrow \quad \widehat{E}_2 = (500)(\widehat{p}_2) = 121.91 \\
\widehat{p}_3 = \text{dbinom}(0, 5, .232) = 0.0737 \quad &\Rightarrow \quad \widehat{E}_3 = (500)(\widehat{p}_3) = 36.82 \\
\widehat{p}_4 = P[X \geq 4] = 1 - P[X \leq 3] = 1 - \text{pbinom}(3, 5, .232) = 0.0118 \quad &\Rightarrow \quad \widehat{E}_4 = (500)(\widehat{p}_4) = 5.90
\end{aligned}$$

i	pi	Ei	Oi	(Oi-Ei)^2/Ei
0	0.26718	133.590663	140	0.3075036
1	0.40355	201.777564	197	0.1131202
2	0.24381	121.907278	115	0.3913670
3	0.07365	36.826157	41	0.4730596
4	0.01179	5.898339	7	0.2057626
Total	1.0	500	500	1.4908

The evaluation statistic is

$$Q^* = \sum_{i=0}^{4} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} = 1.491$$

and Q^* has approximately a chi-squared distribution with df=$5 - 1 - 1 = 3$.
Using Chi-square distribution: p-value= $Pr(\chi^2_3 \geq 1.491) = 1 - pchisq(3.316, 3) = 0.6843$, thus conclude that a Binomial model provides an excellent fit to the data.
Notice that the expected counts for the five cells under the Binomial model are very close to the observed counts.

Prob. 3. (20 points)For the 4 plots we have

Plot 1: K - because Plot 1 is a mixture distribution with $Q(.5) \approx 5$ which implies that half of the data values are less than 5.

Plot 2: D - because Plot 2 is a right skewed distribution with left tail lighter (shorter) than a normal distribution and the right tail heavier (longer) than a normal distribution. Also, all its mass on (0,1.0)

Plot 3: H - because Plot 3 displays a right skewed distribution with $Q(.5) \approx 20$ and
Gamma$(\alpha = 2, \beta = 25)$ has $Q(.5) = qgamma(.5, 2, 1/25) = 41.96 >> 20$
Weibull$(\gamma = 1.2, \alpha = 25)$ has $Q(.5) = 25[-log(.5)]^{1/1.2} = 18.42 \approx 20$
Exponential$(\beta = 100)$ has $Q(.5) = -100log(.5) = 69.3 >> 20$

Plot 4: G because Plot 4 displays a right skewed distribution with
$\hat{Q}(.159) \approx 20,\ \hat{Q}(.5) \approx 70,\ \hat{Q}(.75) \approx 145,\ \hat{Q}(.84) \approx 200,\ \hat{Q}(.977) \approx 400$

• Gamma$(\alpha = 2, \beta = 25)$ has $Q(.5) = qgamma(.5, 2, 1/25) = 41.96 < \mu = \alpha\beta = 50$

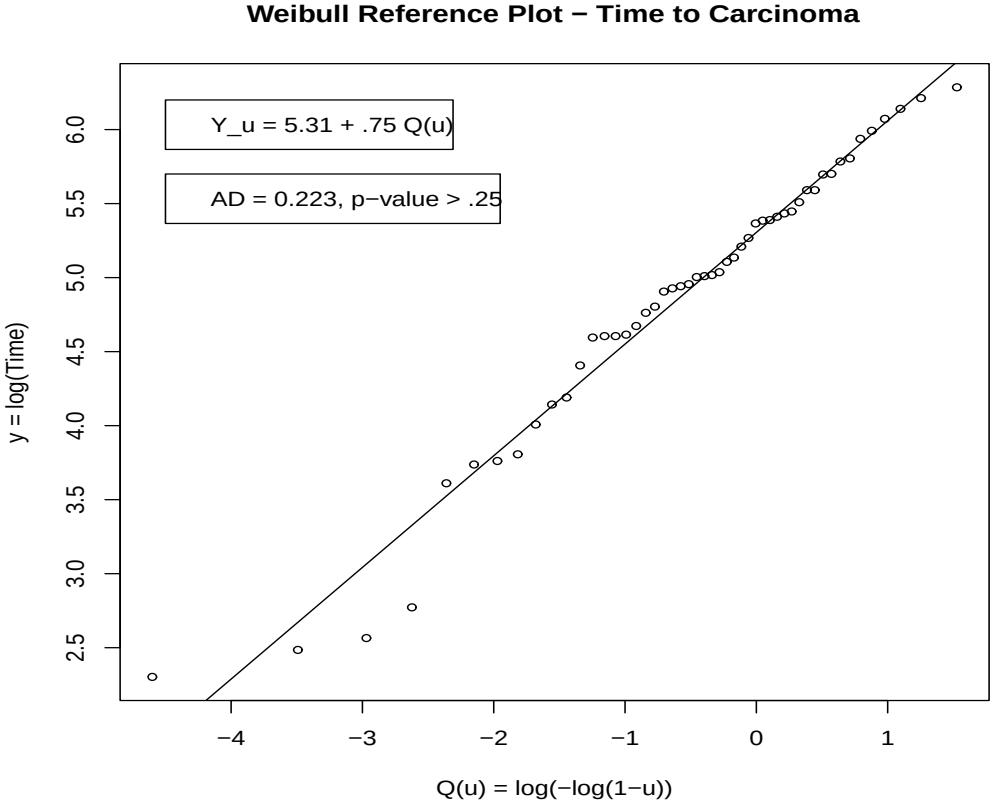
• Weibull$(\gamma = 1.2, \alpha = 25)$ has $QW(u) = 25[-log(1-u)]^{1/1.2} \Rightarrow$
$QW(.159) = 6,\ QW(.5) = 18.4,\ QW(.75) = 32.8,\ QW(.84) = 41.4,\ QW(.977) = 75.6$

• Exponential$(\beta = 100)$ has $QE(u) = -100log(1-u) \Rightarrow$
$QE(.159) = 17,\ QE(.5) = 69.3,\ QE(.75) = 138.6,\ QE(.84) = 1183.3,\ QE(.977) = 377.2$

Therefore, the Exp$(\beta = 100)$ provides the best match to the sample quantiles.

Prob. 4. (15 points) Using the times to the appearance of carcinoma for the 50 rats, we obtain the following:

Weibull Reference distribution plot:



The plotted points are still relatively close to the fitted line with $R^2 = .98$ hence the Weibull model appears to be appropriate.

To calculate AD-GOF statistics, we first need the MLE's of γ and α:

```
#mle's for Weibull
library(MASS)
mle_weibull=fitdistr(x,"weibull")
mle_weibull
# shape         scale
#    1.3861041   201.1786476
# (  0.1560356) ( 21.5933886)
```

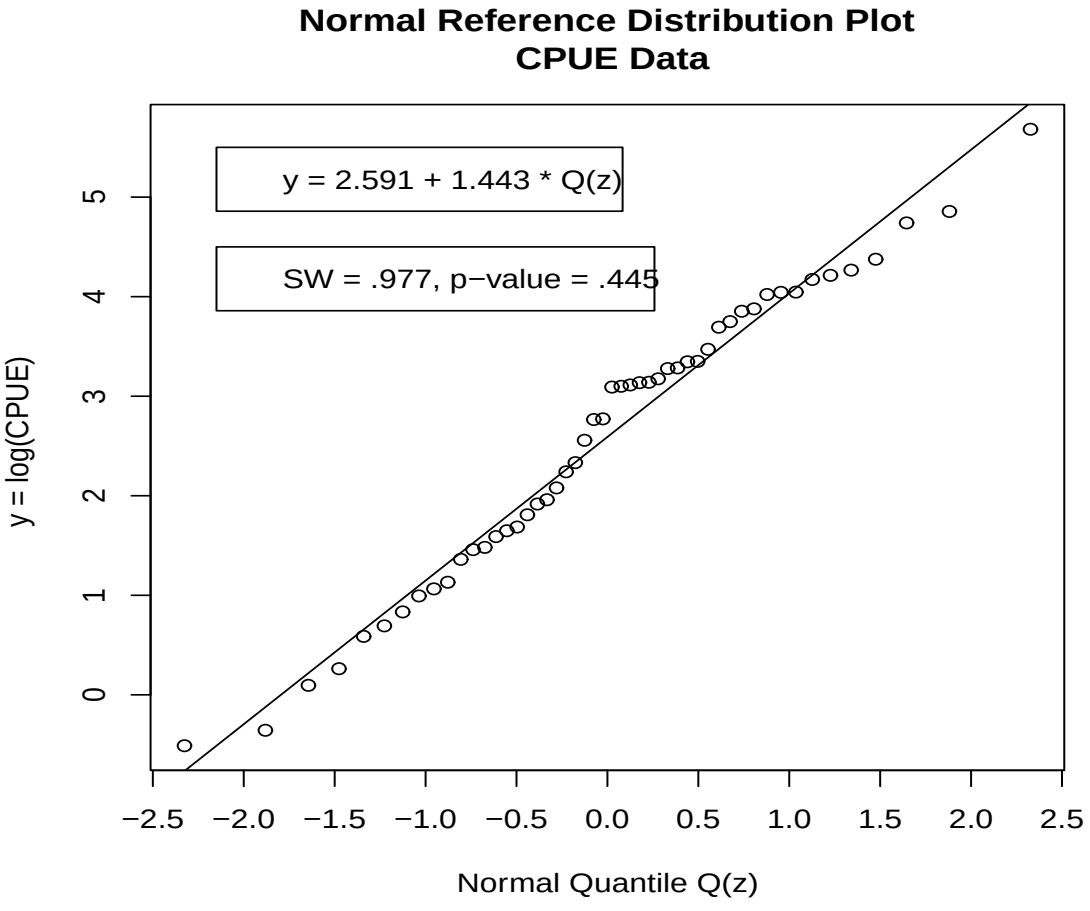
The MLE estimates using the above code are

$$\hat{\gamma} = 1.3861 \qquad\qquad \hat{\alpha} = 201.1786$$

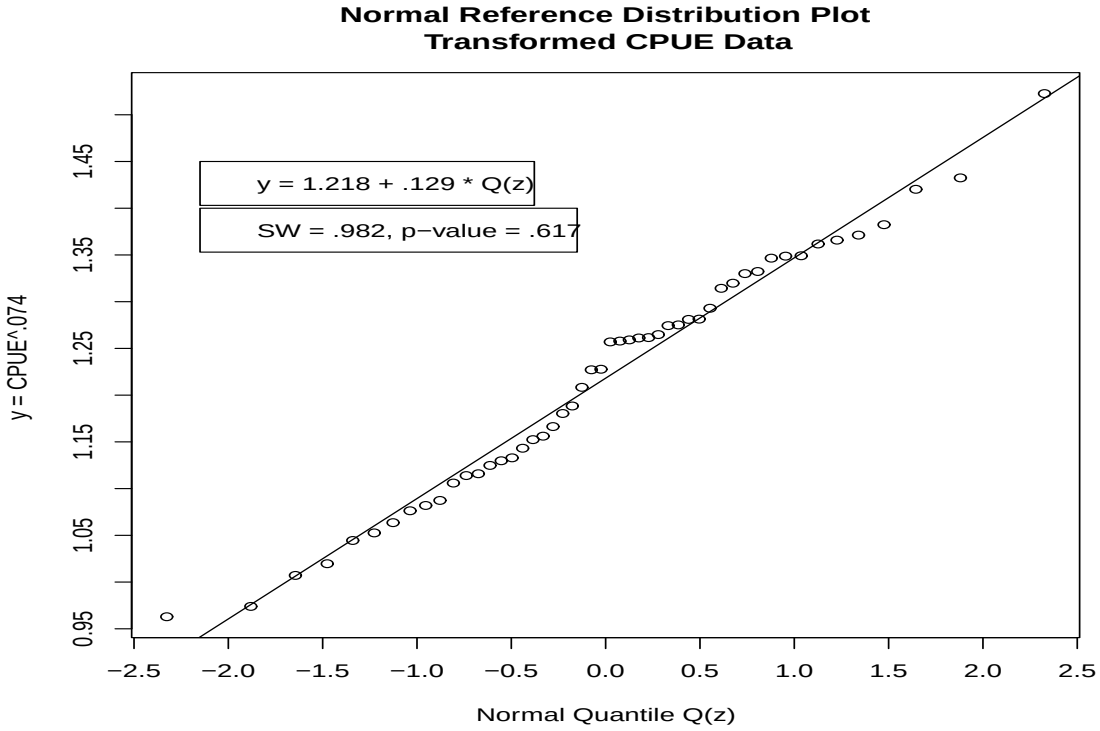
The AD GOF test for the Weibull model using the mle's:
AD=.217 which yields Modified ADM = AD$*\left(1.0 + \frac{0.2}{\sqrt{50}}\right) = .223$. Using Table 5 with ADM = .223 yields p-value $> .25$. This would lend strong evidence that a Weibull provides an excellent fit to the data which is consistent with the conclusion from the reference distribution plot.

Prob. 5. (25 points) CPUE problem.

1. The Normal Reference Distribution plot has nearly all the 50 close to a straight line with $R^2 = 0.98$.
   The Shapiro-Wilk test has the following values:
   SW=.9773, p-value = .445.
   Thus, we would conclude there is an excellent fit of the log normal model to the observed CPUE data.



2. Because the Log-Normal fit the CPUE data so well it is not necessary to perform a power transformation using the Box-Cox transformation. However, if you conduct the Box-Cox process on the CPUE data, we obtain $\widehat{\theta}_{max} = .074$ with a 95% CI of (-.112, 0.259) which contains 0, the log transformation. The Shapiro-Wilk test on $(CPUE)^{.074}$ yields SW=.9815 and p-value=.617 which is a slight improvement over the fit for the Log(CPUE) data.



3. The standard error of $\bar{Y}$ is se.boot = 0.2040616 by bootstrap, and is approximately equal to $S_Y/\sqrt{n} = 1.453/\sqrt{50} = 0.2054852$ the estimated standand error using the sample standard deviation.
4. The result using the bootstrap R code is summarized in the following table.

	$\widehat{Q}(0.5)$	S	$\widehat{MAD}$
Boot mean	2.814	1.434	1.568
Boot sd	0.374	0.117	0.223

Note that bootstrap mean for S and MAD are nearly equal which is empirical evidence that they are both estimating the same parameter, σ. Recall, this would be true only if the is normally distributions. Also, the boot strap estimates of their standard deviations has the value for S smaller than the value for MAD. Thus, indicating that S is a more precise estimator of σ than MAD when the data is from a normal distribution.

When the data is from a non-normal distribution, do S and MAD estimate the same parameter? The answer is no. We can observe this using bootstrap estimators of the mean of S and MAD from the untransformed values of CPUE:

	S	$\widehat{MAD}$
Boot mean	45.285	21.378
Boot sd	14.391	6.800

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

- The sample median, 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 has sampling distribution $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Therefore,

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

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

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

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

Alternatively, using the asymptotic results, μ_A for $S = \sigma = 1.5$ and σ_A for $S = \frac{\sqrt{\mu_4-\sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{2\sigma^4}}{2\sigma\sqrt{n}} = \frac{\sigma}{\sqrt{2n}} = \frac{1.5}{\sqrt{100}} = .15$ The result is summarized in the following table

	Boot mean	Boot sd	Asymp. mean	Asympt. sd
$\hat{Q}(0.5)$	2.814	0.374	3	0.266
S	1.434	0.117	1.492	0.151

Comparing the two sets of values, the asymptotic mean for $\widehat{Q}(0.5)$ is larger than the bootstrap value with the bootstrap standard deviation somewhat larger than the asymptotic standard deviation. However, the differences are not too large.

For S, there is very strong agreement between the bootstrap mean and the asymptotic mean. The bootstrap standard deviation is slightly smaller than the asymptotic value.

Prob. 6. **(25 points)** Lifetime of 25 batteries:

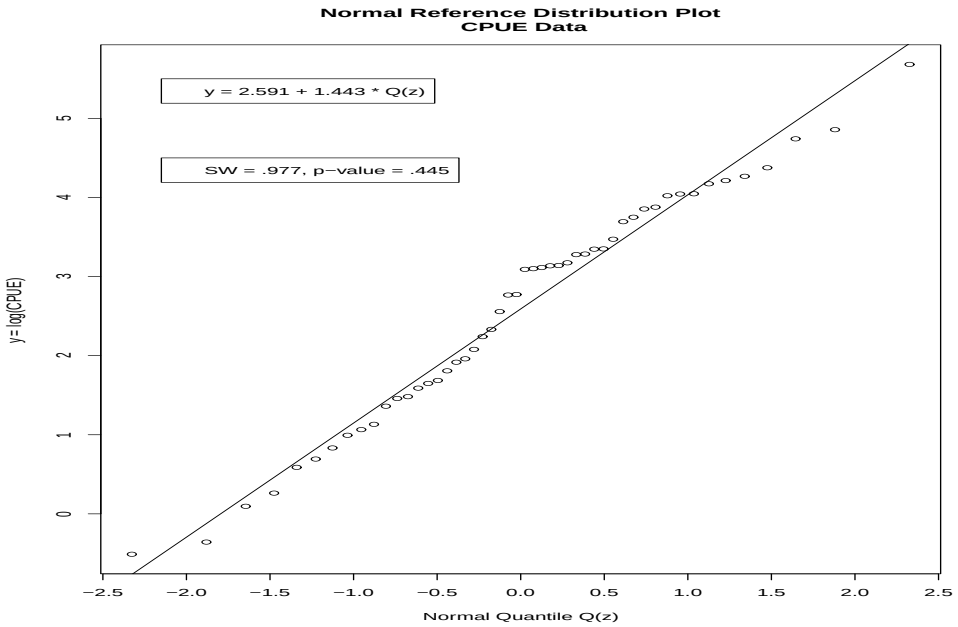
1. The exact distribution of $\bar{Y}$ is determined using the fact that Y_i's are independent exponential r.v.s with parameter β, therefore $S = \Sigma_{i=1}^n Y_i$ is the sum of iid Exponential and hence has a Gamma distribution with shape parameter $\alpha = n$ and scale parameter β. The distribution of $\bar{Y}$ is obtained from $\bar{Y} = \frac{1}{n}S$ which is a Gamma distribution with shape parameter $\alpha = n$ and scale parameter β/n.
   a. We can verify this result as follows:
   Let $\bar{Y}$ have cdf H, S have cdf G and pdf given by : $g(s) = \frac{1}{\Gamma(\alpha)\beta^\alpha}s^{\alpha-1}e^{-s/\beta}$(Gamma with parameters α and β). Then we have

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

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

   From the above we can identify that $h(y)$ is the pdf of a Gamma distribution with shape parameter $\alpha = n$ and scale parameter β/n.
   In particular, $\bar{Y}$ has a Gamma$(\alpha = 25, \beta = 5/25)$ distribution.
   b. $\mu_{\bar{Y}} = \mu_Y = E[Y] = \beta = 5$
   $\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{\beta}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$ because $Var(Y) = \beta^2 \quad \Rightarrow \quad \sigma_Y = \beta = 5$.

2. From the normal reference distribution plot, the sample means are close to normal distribution, with the right tail slightly heavier and the left tail slightly shorter than a normal distribution. This is an indication that $n = 25$ is not a large enough sample size to use the results from the Central Limit Theorem to infer that the sampling distribution of $\bar{Y}$ is nearly a normal distribution when the data is from a highly right skewed distribution.



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(25,5/25) distribution. Therefore,

$$P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right) = P(4.8 \leq \bar{Y} \leq 5.2) = pgamma(5.2, 25, 25/5) - pgamma(4.8, 25, 25/5) = 0.1580747$$

   b. Since $Y_1, \ldots, Y_{25} \overset{\text{iid}}{\sim} Exp(5)$, the Central Limit Theorem yields: approximately (for large n), $\bar{Y} \sim N(\mu, (\sigma/\sqrt{n})^2) = N(5, (5/\sqrt{25})^2) = N(5, 1)$.

$$\begin{aligned} P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right) &= P\left(\frac{-0.2}{1} \leq \frac{\bar{Y} - 5}{1} \leq \frac{0.2}{1}\right) \\ &\approx P\left(-0.2 \leq Z \leq 0.2\right) = pnorm(.2) - pnorm(-.2) \ = 0.1585194 \end{aligned}$$

   c. The R function

```
sum(1*(abs(means-5)<.2))
```

   counts the number of $\bar{Y}$'s out of the 10,000 values that satisfy

$$P\left(-0.2 \leq \bar{Y} - 5 \leq 0.2\right)$$

   In my generated sample, there were 1579 out of 10000 values of $\bar{Y}$ that satisfied $-0.2 \leq \bar{Y} - 5 \leq 0.2$. Thus, the estimated value of the probability is 1579/10000=0.1579. Again, this value changes from simulation to simulation, and can vary considerably. (I ran the simulation four times, and obtained the following results 0.1589, 0.1624, 0.1615, 0.1549).

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=.1579

```
####
#### (2)
####

y <- c(140, 197, 115, 41, 5, 2)
ii <- 0:5
n <- sum(y)
m <- 5

## MLE
p_hat <- (y %*% ii) / (n * m)

## GOF statistic
p_i <- dbinom(ii, 5, p_hat)
E_i <- n * p_i

## Combine last two categories
p_i <- c(p_i[1:4], 1 - pbinom(3, 5, p_hat))
E_i <- n * p_i
O_i <- c(y[1:4], y[5] + y[6])

Q <- sum((O_i - E_i) ^ 2 / E_i)
p_val <- 1 - pchisq(Q, 5 - 2)

####
#### (3)
####

##
## Gamma(alpha = 2, beta = 25)
##

x_seq <- seq(0, 350, length = 1000)
y_seq <- dgamma(x_seq, 2, 1 / 25)
plot(x_seq, y_seq, type = "l")

y_sim <- rgamma(10000, 2, 1 / 25)
mean(y_sim)
median(y_sim)

##
## Weibull(gamma = 1.2, alpha = 25)
##

x_seq <- seq(0, 200, length = 1000)
y_seq <- dweibull(x_seq, 1.2, 25)
plot(x_seq, y_seq, type = "l")

y_sim <- rweibull(10000, 1.2, 25)
mean(y_sim)
median(y_sim)

##
## Exponential(beta = 100)
##

x_seq <- seq(0, 950, length = 1000)
```

```
y_seq <- dexp(x_seq, 1 / 100)
plot(x_seq, y_seq, type = "l")

y_sim <- rexp(10000, 1 / 100)
mean(y_sim)
median(y_sim)

####
#### (4)
####

x <- c(10, 12, 13, 16, 37, 42, 43, 45, 55, 63, 66, 82, 99, 100, 100, 101, 107, 117,
  122, 135, 138, 140, 142, 149, 150, 151, 154, 165, 170, 183, 194, 214, 218, 219, 224,
  229, 232, 247, 268, 268, 298, 299, 325, 332, 379, 400, 434, 464, 499, 537)
n <- length(x)

y <- -log(x)

##
## Reference distribution plot
##

i <- 1:n
u <- (i - 0.5) / n

Q_Z <- log(-log(1 - u))
plot(Q_Z, log(x), xlab = "Q(u) = log(-log(1 - u))", ylab = "y = log(time)")

## Linear fit
fit <- lm(log(x) ~ Q_Z)
abline(fit)
text(-3, 6, labels = "y(u) = 5.31 + 0.75 Q(u)")

##
## Anderson-Darling GOF test
##

## MLEs of Weibull distribution
library(MASS)
mles <- fitdistr(x, "weibull")
gamma_hat <- mles$est[1]
alpha_hat <- mles$est[2]

## Weibull CDF
u <- 1 - exp(-(x / alpha_hat) ^ gamma_hat)
u_sort <- sort(u)

## AD statistics
AD <- -n - (1 / n) * sum((2 * i - 1) * log(u_sort) +
  (2 * n + 1 - 2 * i) * log(1 - u_sort))
ADm <- AD * (1 + 0.2 / sqrt(n))

####
#### (5)
####

x <- 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)
n <- length(x)

##
## (1)
##

y <- log(x)
mu_hat <- mean(y)
sigma_hat <- sqrt(var(x) * (n - 1) / n)

## Reference distribution plot
i <- 1:n
u <- (i - 0.5) / n

Q_Z <- qnorm(u)
plot(Q_Z, y, xlab = "Q(u) = qnorm(u)", ylab = "log(x)")

fit <- lm(y ~ Q_Z)
abline(fit)
text(-1, 5, labels = "y = 2.59 + 1.47 x")

## Shapiro-Wilks test
shapiro.test(y)

##
## (2)
##

## Determine theta that maximizes log likelihood
l <- 0
theta_seq <- seq(-3, 3, by = 0.001)

s_0 <- sum(y)
v_0 <- var(y)
for(i in 1:length(theta_seq)) {
  if(abs(theta_seq[i]) < 1e-10) {
    l[i] <- -s_0 - (n / 2) * (log(2 * pi * v_0) + 1)
    theta_seq[i] <- 0
  } else {
    x_theta <- (x ^ theta_seq[i] - 1) / theta_seq[i]
    v_1 <- var(x_theta)
    l[i] <- (theta_seq[i] - 1) * s_0 - (n / 2) * (log(2 * pi * v_1) + 1)
  }
}

i_max <- which.max(l)
theta_hat <- theta_seq[i_max]

plot(theta_seq, l, xlab = "theta", ylab = "L(theta)", type = "l")

## 95% confidence interval for theta
which_ci <- (1:length(theta_seq))[(l[i_max] - l) <= 0.5 * qchisq(0.95, 1)]
theta_ci <- theta_seq[c(min(which_ci), max(which_ci))]
```

```
## Box-Cox transformed data
x_tr <- (x ^ theta_hat - 1) / theta_hat

## Shapiro-Wilks test on transformed data
shapiro.test(x_tr)

##
## (3)
##

B <- 10000
y_b <- matrix(NA, nrow = n, ncol = B)
ybar_b <- Q_b <- sd_b <- mad_b <- numeric(B)

for(b in 1:B) {
  y_b[, b] <- sample(y, replace = TRUE)
  ybar_b[b] <- mean(y_b[, b])
  Q_b[b] <- median(y_b[, b])
  sd_b[b] <- sd(y_b[, b])
  mad_b[b] <- mad(y_b[, b])
}

sd(ybar_b)
sd(y) / sqrt(n)

##
## (4)
##

mean(Q_b); sd(Q_b)
mean(sd_b); sd(sd_b)
mean(mad_b); sd(mad_b)

##
## (5)
##

mu <- 3
sg <- 1.5

## Asymptotic sd of median
0.5 / dnorm(3, mu, sg) / sqrt(n)

## Mean of S
c_n <- sqrt(2 / (n - 1)) * gamma(n / 2) / gamma((n - 1) / 2)
c_n * sg

## SD of s
sg * sqrt(1 - c_n ^ 2)

####
#### (6)
####

##
## (2)
##
```

```
n <- 25
N <- 10000
y_N <- matrix(rexp(n * N, 1 / 5), nrow = n, ncol = N)

ybar_N <- colMeans(y_N)
ybar_N <- sort(ybar_N)
u <- (1:N - 0.5) / N
Q_z <- qnorm(u)
plot(Q_z, ybar_N, xlab = "Normal Quantiles", ylab = "Sample Quantile Function")
abline(lm(ybar_N ~ Q_z))

##
## (3)
##

## (a)
pgamma(5.2, 25, 5) - pgamma(4.8, 25, 5)

## (b)
pnorm(5.2, 5, 1) - pnorm(4.8, 5, 1)

## (c)
mean(ybar_N <= 5.2) - mean(ybar_N <= 4.8)
```