Carson Slater STA 5380 Homework #4

Questions 1 to 4 are from Chapter 5.

1. Derive the probability mass function of the sample maximum for the experiment in Example 5.2 (and Table 5.2) of the class notes.

   The PMF for the sample maximum for the experiment in Example 5.2 is:

Sample Maximum	10	20	30	40	50
Probability	$\frac{1}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{7}{25}$	$\frac{9}{25}$

Table 1: PMF for sample maximum.

2. A soft drink company uses a filling machine to fill cans. Each 12 oz. can is to contain 355 milliliters of beverage. In fact, the amount varies according to a normal distribution with mean $\mu = 355.2$ ml and standard deviation $\sigma = 0.5$ ml.

   (a) What is the probability that an individual can contains less than 355 ml?

   ```r
   pnorm(355, 355.2, 0.5)
   ```

   ```
   ## [1] 0.3445783
   ```

   Let X be the random variable where X is the volume of soda filling a can. From R, we observe $P(X < 355) = 0.345$.

   (b) What is the distribution of mean content of a six-pack of cans?

   We know that the distribution for one can is $\mathcal{N}(355, 0.5)$, which mean the sampling distribution for mean volume of soda in six cans would be $\mathcal{N}(355, \frac{0.5}{\sqrt{6}})$.

   (c) Does your answer in (b) require the use of the Central Limit Theorem? Why or why not?

   This answer does not require the Central Limit Theorem, because any normally distributed variable also possesses a sampling distribution whose mean sampling distribution is also normally distributed, regardless of the sample size.

   (d) What is the probability that the mean content of a six-pack of cans is less than 355 ml?

   ```r
   pnorm(355, 355.2, 0.5/sqrt(6))
   ```

   ```
   ## [1] 0.1635934
   ```

   From R, we observe $P(\bar{x} < 355) = 0.164$.

3. In an extrasensory perception (ESP) experiment, five choices are offered for each question. Assume that a person without ESP guesses randomly and thus correctly answers with probability 1/5. Further assume that the responses are independent. Suppose that 100 questions are asked.

   (a) What are the mean and standard deviation of the number of correct answers?
   Let random variable $X = \sum_{i=1}^{n} X_i$ such that

$$X_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ trial is a success} \\ 0, & \text{if the } i^{\text{th}} \text{ trail is a failure.} \end{cases}$$

   We know that success for each X_i occurs with probability $p = \frac{1}{5}$. We can approximate the random variable X, the number of correct answers by $\mathcal{N}\left(np, np(1-p)\right)$, if $np > 10$ and $n(1-p) > 10$. We have that

$$np = 100 \times \frac{1}{5} = 20 > 10 \qquad \text{and} \qquad n(1-p) = 100 \times \frac{4}{5} = 80 > 10.$$

   So then we can approximate X with a normal distribution, $X \sim \mathcal{N}\left(np = 20, \sqrt{np(1-p)} = 4\right)$. Hence, the mean and standard deviation for the number of correct answers is 20 and 4 respectively.

   (b) What are the mean and standard deviation of the proportion of correct answers?
   Let random variable $\frac{X}{n} = p =$ proportion of correct answers a person scores while doing the experiment. We can also approximate the distribution of $\hat{p}$ with a normal distribution, $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ when we apply the Central Limit Theorem. This would imply that the mean and standard deviation of $\hat{p} \sim \mathcal{N}\left(p = \frac{1}{5}, \sqrt{\frac{p(1-p)}{n}} = 0.04\right)$ is 0.2 and 0.04.

   (c) What is the probability that a person without ESP will correctly answer at least 30 of the 100 questions? Check if the normal approximation will give accurate results, and if so, use it.

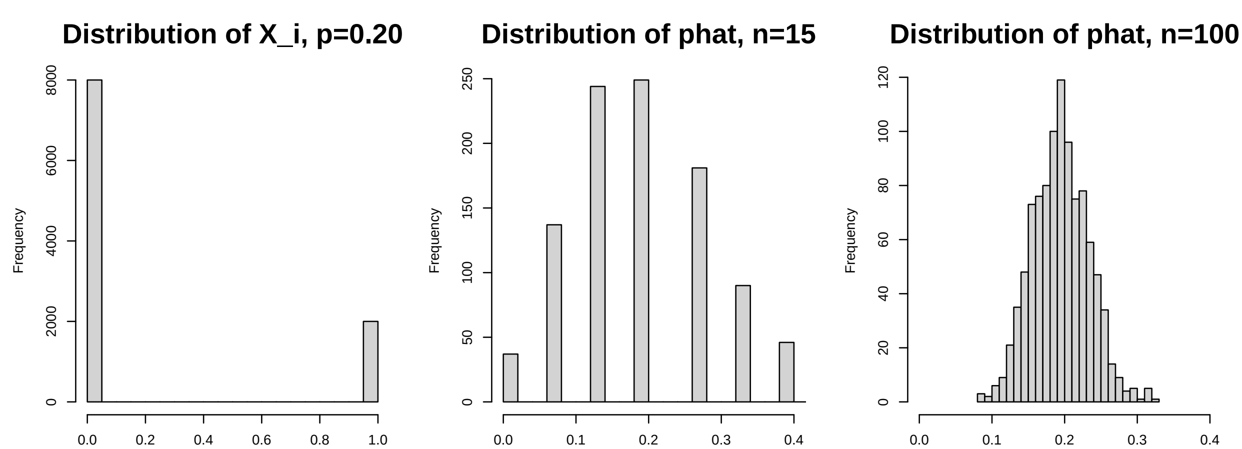


Figure 1: Distribution of X_i and $\hat{p}$.

We are interested in $P(X \geq 30)$. Figure 1 indicated that the number of observations in each experiment is sufficient to approximate this probability with a normal distribution. So then $P(X \geq 30) = P(Z \geq 2.5)$. We can calculate this in `R` with the following script.

```
pnorm(2.5, 0 , 1, lower.tail = FALSE)

## [1] 0.006209665
```

So $P(X \geq 30) = 0.006$.

4. An engineer maintains the temperature inside an oven to within ±10°F, which translates into roughly $\pm 2\sigma$ (assuming the temperatures are normally distributed), and $\sigma^2 = 25$°F. If he takes 20 measurements of the temperature in the oven, what is the probability that the sample variance exceeds 30°F?

```r
pchisq((19*30)/25, 19, lower.tail = FALSE)

## [1] 0.2462659
```

We are looking for $P(S^2 > 30) = P(\chi^2_{19} > \frac{(19)(30)}{25})$. By R, we can find $P(S^2 > 30) = 0.246$.

The remaining questions are from Chapter 6.

5. Let $X_1, \dots, X_n$ be a random sample where each X_i has the following pdf:

$$f(x_i|\theta) = \frac{1}{\theta}, \quad 0 \leq x_i \leq \theta.$$

(a) Find the method of moments estimator for θ.

We know that $X_i \sim \text{Unif}(0, \theta)$. To find the method of moments estimator, we can find the sample moment:

$$\begin{aligned} E[X|\theta] = \bar{x} &= \frac{\theta}{2} \\ \implies \hat{\theta}_{MOM} &= 2\bar{x}. \end{aligned}$$

(b) Is the method of moments estimator unbiased? Justify your answer.

$$\begin{aligned} E[\hat{\theta}_{MOM}] &= E[2\bar{x}] \\ &= 2E[\bar{x}] \\ &= 2\mu \\ &= \theta \qquad \left(\text{because } \mu = \frac{\theta}{2}\right). \end{aligned}$$

This method of moments estimator is unbiased because the expected value of the estimator is equal to the true parameter.

6. Let X denote the proportion of allotted time that a randomly selected student spends working on a certain aptitude test. Suppose the pdf of X is

$$f(x|\theta) = \begin{cases} (\theta+1)x^{\theta}, & \text{for } 0 \leq x \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > -1$. A random sample of ten students yields the following data

0.92, 0.79, 0.90, 0.65, 0.86, 0.47, 0.73, 0.97, 0.94, 0.77

(a) Use the method of moments to obtain an estimator of θ, and then compute the estimate for this data. We have it that:

$$\begin{aligned} E[X|\theta] &= \int_0^1 (\theta+1)x^{\theta+1}dx \\ &= \frac{(\theta+1)x^{\theta+2}}{(\theta+2)}\Big|_0^1 \\ \implies \bar{x} &= \frac{(\theta+1)}{(\theta+2)}. \end{aligned}$$

Solving for θ, we have

$$\begin{aligned} \bar{x} &= \frac{(\theta+1)}{(\theta+2)} \\ \implies (\theta+2)\bar{x} &= (\theta+1) \\ \implies 2\bar{x}+1 &= \theta(1-\bar{x}) \\ \implies \hat{\theta}_{MOM} &= \frac{2\bar{x}+1}{1-\bar{x}} \end{aligned}$$

This would mean that the estimate for this data is $\hat{\theta}_{MOM} = 13$.

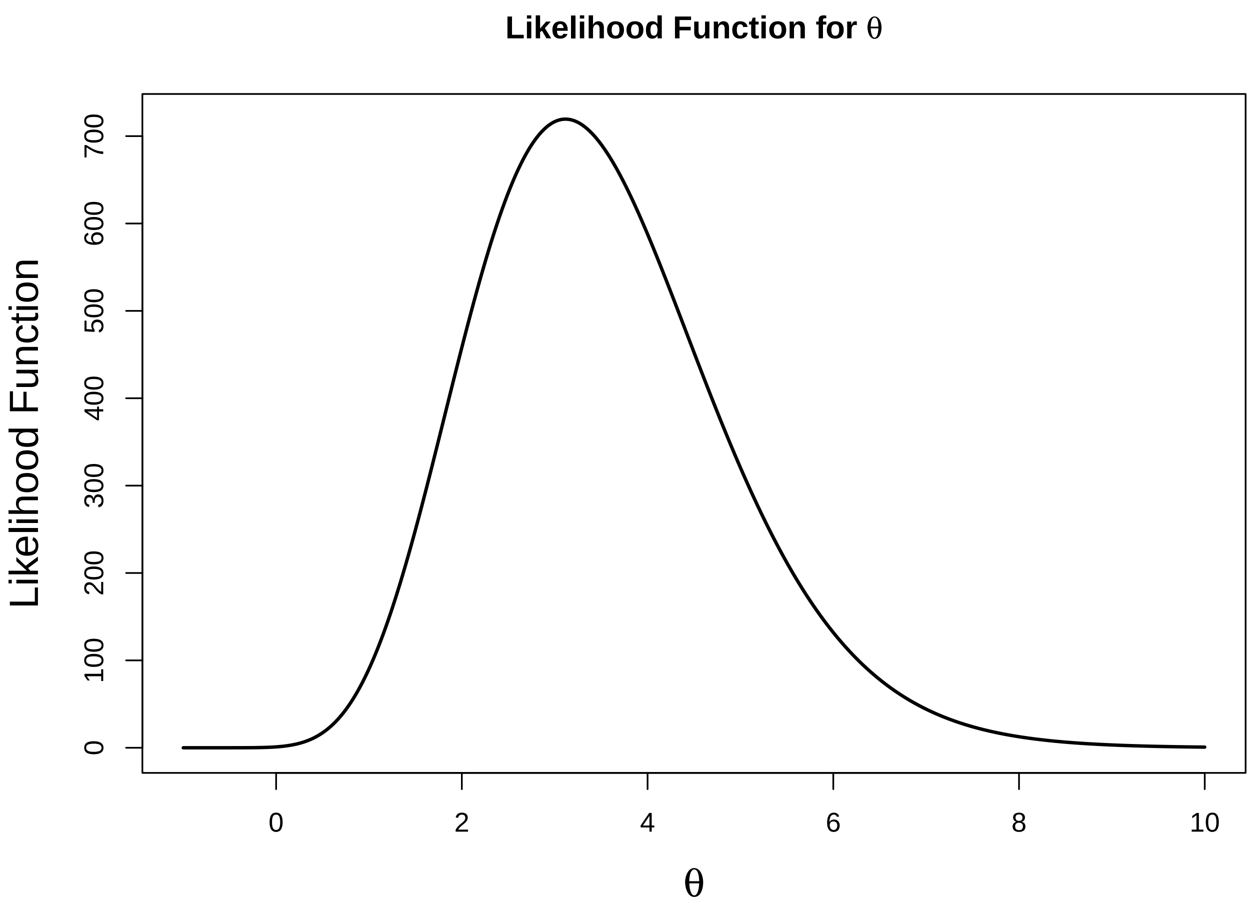
(b) Obtain the maximum likelihood estimator of θ. Plot the likelihood function for this data. Compute the estimate of θ for the given data.

$$\begin{aligned} f(x_1, x_2, ..., x_{10}|\theta) &= \prod_{i=1}^{10}(\theta+1)x_i^{\theta} \\ \implies L(\theta|x_1, x_2, ..., x_{10}) &= (\theta+1)^{10}\left(\prod_{i=1}^{10} x_i\right)^{\theta} \\ &= (\theta+1)^{10}(0.0880806)^{\theta}\left(\text{since } \prod_{i=1}^{10} x_i = 0.0880806\right) \\ \implies \mathcal{L}(\theta) &= 10\ln(\theta+1) + \theta\ln(0.0880806) \end{aligned}$$

Maximizing the log-likelihood function, we solve,

$$0 = \frac{10}{(\theta+1)} + \frac{\theta}{0.0880806}$$

where the root of this function is approximately $\hat{\theta}_{MLE} \approx 3.11607$ (via WolframAlpha).


Likelihood Function for θ
Likelihood Function
0
100
200
300
400
500
600
700
0
2
4
6
8
10
θ

7. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ,

$$f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, 3, \ldots$$

(a) Write down the log-likelihood function for λ.

We can calculate the likelihood function as:

$$\begin{aligned} L(\lambda|x_1, x_2, ..., x_n) &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}. \\ \implies \mathcal{L}(\lambda|x_1, x_2, ..., x_n) &= -n\lambda \ln(e) + \sum_{i=1}^{n} x_i \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right) \end{aligned}$$

(b) Find the maximum likelihood estimator for λ.

Maximizing the log-likelihood function, we find:

$$\begin{aligned} 0 &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \\ \implies \hat{\lambda}_{MLE} &= \bar{x}. \end{aligned}$$

(c) Find the MLE to estimate $P(X_1 = k)$, where k can be any number in the set $\{0, 1, 2, 3, \ldots\}$.

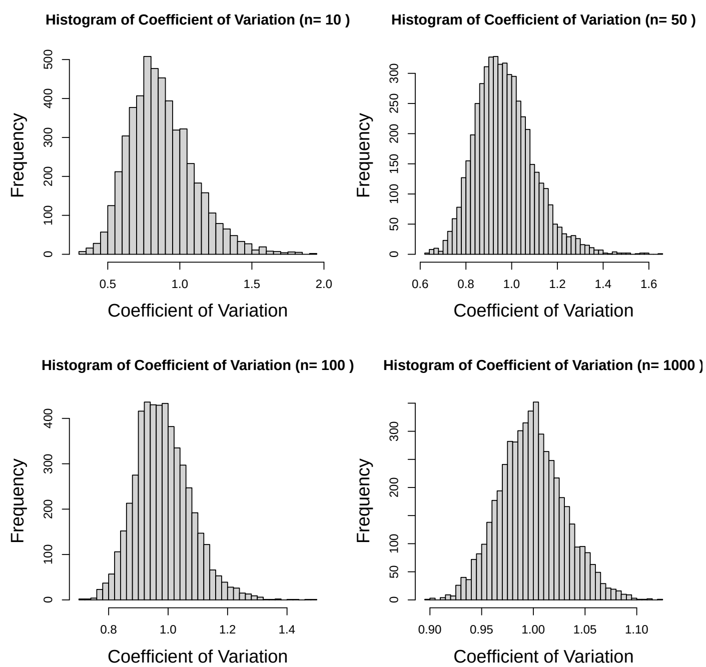
Due to the invariance property of the MLE, we can say that if $\theta = P(X_1 = k)$, then

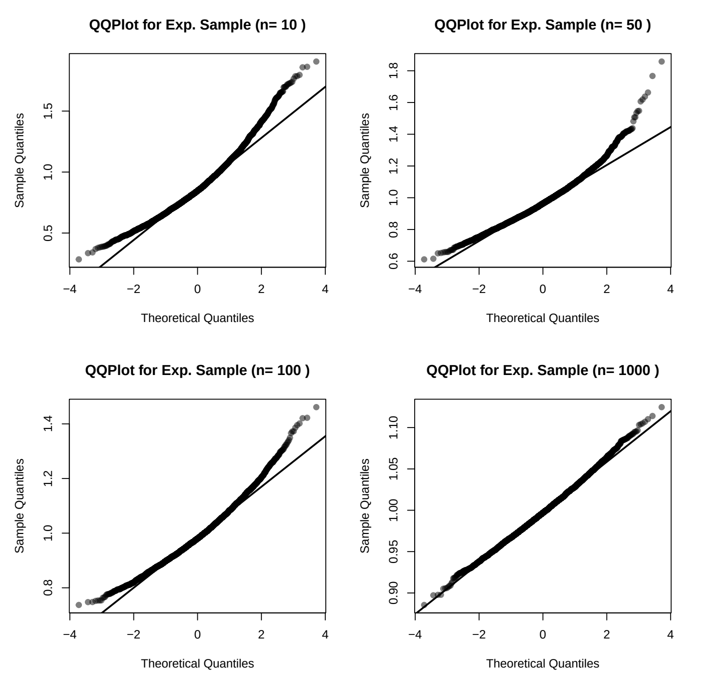
$$\hat{\theta}_{MLE} = \frac{\bar{x}^k e^{-\bar{x}}}{k!}.$$

8. From Example 6.14 of the class notes, the MLE of the coefficient of variation is

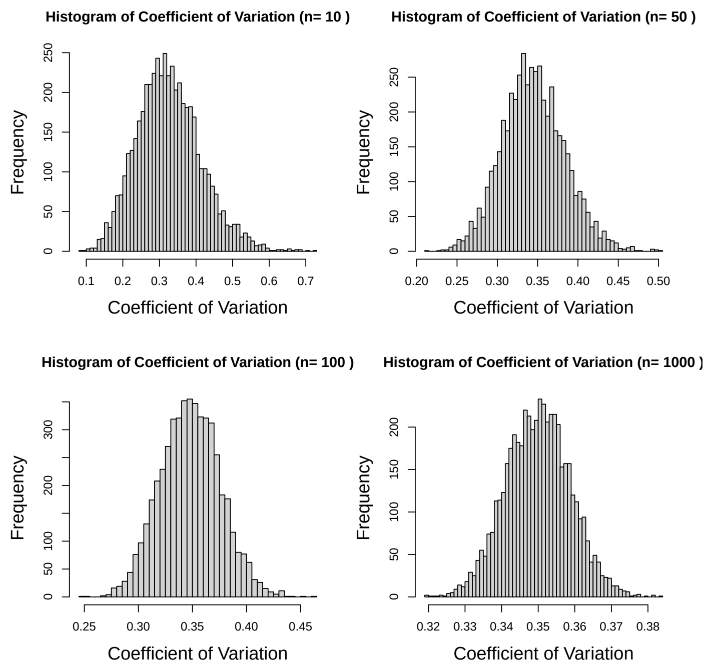
$$\hat{C}_v = \frac{\sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}}{|\bar{X}|}.$$

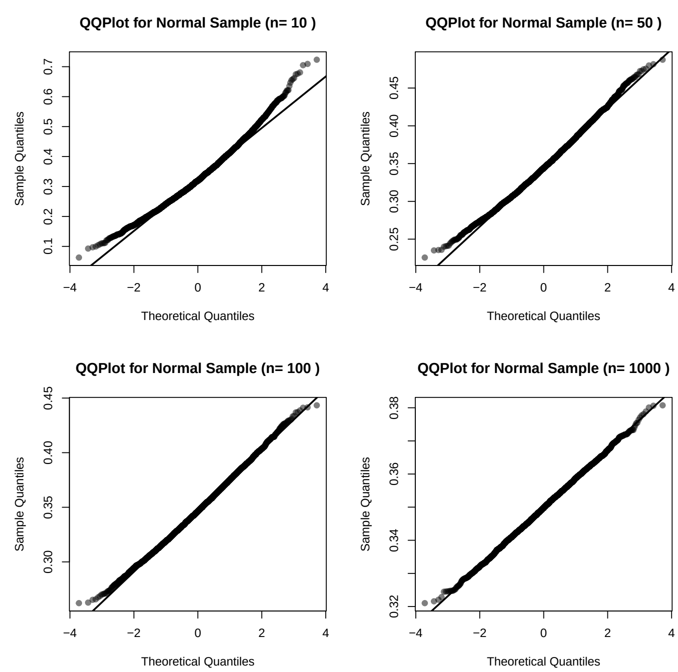
(a) Simulate 5,000 samples of size $n = 10$ from an exponential with parameter $\alpha = 1$. Compute $\hat{C}_v$ for each sample, and then plot a histogram of all 5,000 estimated $\hat{C}_v$'s. Repeat for $n = 50$, 100, and 1000. For which sample size, if any, does the distribution of the $\hat{C}_v$'s appear to become normal? See part (c) for answer as to when the distribution appears to become normal. (For formatting reasons.)





(b) Repeat part (a) but simulate from a normal distribution with mean 10 and standard deviation 3.5. See part (c) for answer as to when the distribution appears to become normal. (For formatting reasons.)





(c) Compare your results from parts (a) and (b). Explain what is causing the difference in your answers between parts (a) and (b).

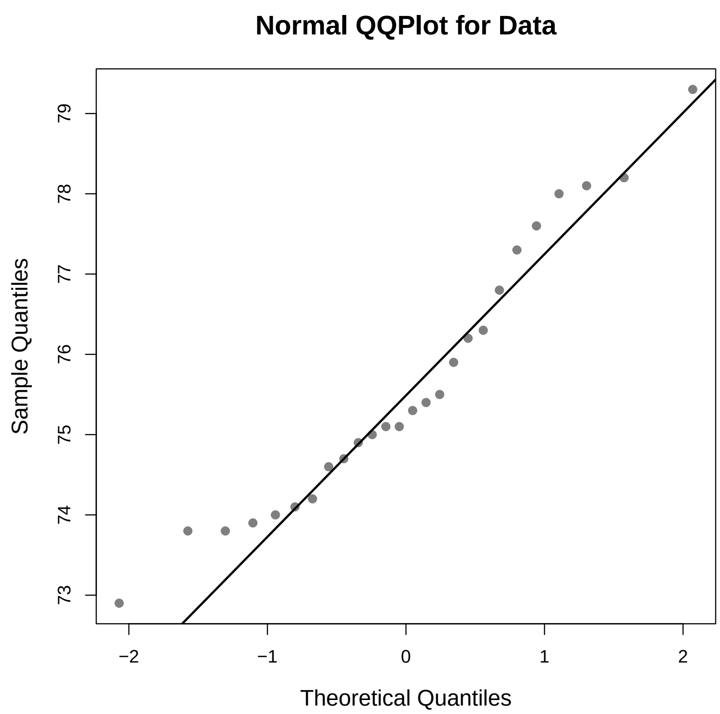
In part (a), the distribution truly starts to look normal after the sample size is increased to 1000. Before then, the distribution for $\hat{C}_v$ is fairly right-skewed. Intuitively, it appears that when sampling from an exponential distribution, a larger sample size is required to overcome right-skewness in the parent distribution. As $n \to \infty$, the central limit theorem kicks in and the mean and variance of the sampling distribution converge to a normal distribution. Since $\hat{C}_v$ is a function of those two things, it is natural that it would follow suit and converge to be normally distributed if its parameters are normally distributed.

In part (b), the distribution for $\hat{C}_v$ appears normal when the sample size reaches 50. It converges much faster than the distribution from part (a) because the parent population is normal already. Hence, the sampling distribution does not need to overcome as much (if any) skew in the parent population to look normal.

9. Data on the pull-off force (in pounds) for connectors used in an automobile engine application are as follows:

$$79.3, 75.1, 78.2, 74.1, 73.9, 75.0, 77.6, 77.3, 73.8, 74.6, 75.5, 74.0, 74.7,$$
$$75.9, 72.9, 73.8, 74.2, 78.1, 75.4, 76.3, 75.3, 76.2, 74.9, 78.0, 75.1, 76.8$$

(a) What type of distribution, if any, that we have covered in class does it appear that this data may come from?



The shape of the distribution slightly resembles a bell-shape like a normal distribution, even with a smaller sample size. There does, however, appear to be slight right skew.

(b) Assuming that the data comes from a normal population, use the parametric bootstrap method ($B = 5000$) to generate an estimate of the standard error of the standard deviation S.

```
set.seed(613)
B <- 5000
pbs <- replicate(B, sd(rnorm(length(data), mean(data), sd(data))))

mean(pbs) # mean of sd estimate for pbs

## [1] 1.635599

sd(pbs) # MY ANSWER: se of sd estimate for pbs

## [1] 0.2336573
```

The estimate of the standard error of S is 0.2336.

(c) Not assuming anything about the distribution that the data is sampled from, use the nonparametric bootstrap method ($B = 5000$) to generate an estimate of the standard error of the standard deviation S. How does this value compare to the estimate you obtained in (b)?

```
B <- 5000
npbs <- replicate(B, sd(sample(data, replace = TRUE)))

mean(npbs) # mean of sd estimate for npbs

## [1] 1.61225

sd(npbs) # MY ANSWER: se of sd estimate for npbs

## [1] 0.1887268
```

The estimate of the standard error of S is 0.1980.

10. A researcher has performed an experiment on the flow of a liquid through a pipe that is 2 inches in diameter. He varies the flow of the liquid through the pipe (flow is measured in gallons per minute) and records the resulting velocity that the liquid travels in feet per second.

   (a) Identify the explanatory and the response variables.

   In this experiment, the explanatory variable is the flow of liquid through the pipe in gallons per minute, and the response variable is the velocity the liquid travels in feet per second.

   (b) The researcher wants to compute the correlation between the two variables, but in his sample of size 60, there are some outliers. He plans to use the the 10% trimmed means, $\bar{X}_{tr(10)}$ and $\bar{Y}_{tr(10)}$ in place of $\bar{X}$ and $\bar{Y}$ in the sample correlation coefficient. He calls this new estimator of the population correlation, $R_{tr(10)}$. Outline the steps he would need to use to get a nonparametric bootstrap estimate of the standard error of $R_{tr(10)}$.

      i. He would need to generate n samples of size 60 from the data with replacement.

      ii. He would need to calculate the correlation, ρ, between the two variables X and Y using the 10% trimmed means, using the following formula:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_{tr(10)})(y_i - \bar{x}_{tr(10)})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_{tr(10)})^2 \sum_{i=1}^{n}(y_i - \bar{y}_{tr(10)})^2}}$$

      for each sample.

      iii. He would need to estimate the standard error of the simulated correlations by finding the square root of the variance of the generated correlation estimates.

11. Suppose that X is a normal random variable with unknown mean and variance $\sigma^2 = 9$. The prior distribution for μ is normal with $\mu_0 = 4$ and $\sigma_0^2 = 1$. A random sample of 25 observations is taken, and the sample mean is $\bar{x} = 4.85$.

   (a) Find the Bayes estimate of μ.

   We assume that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 = 9)$. Additionally, the mean $\mu_{\bar{x}} \sim \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 1)$. The Bayes estimator for μ would be computed as

$$\begin{aligned}\tilde{\mu} &= \frac{\left(\frac{9}{25}\right)(4) + (1)(4.85))}{(1) + \left(\frac{9}{25}\right)} \\ &= 4.625.\end{aligned}$$

   (b) Compare the Bayes estimate with the MLE.

   We know that the $\hat{\mu}_{MLE}$ is $\bar{x} = 4.85$, which is greater than the Bayes estimate $\tilde{\mu} = 4.625$. This intuitively makes sense because the Bayes estimator incorporates the fact that prior knowledge supposes the true mean is far less than $\bar{x}$.

12. Let $X_1, X_2, \ldots, X_n$ represent a random sample from a distribution with pdf

$$f(x|\theta) = \frac{x}{\theta} e^{-x^2/(2\theta)}, \quad x > 0.$$

(a) It can be shown that $E(X^2) = 2\theta$. Use this fact to construct an unbiased estimator of θ based on $\sum_{i=1}^n X_i^2$, and use the rules of expected value to show that it is unbiased.

Given the second sample moment, $E(X^2) = 2\theta$, we can construct an unbiased estimator for θ,

$$\hat{\theta}_{MOM} = \frac{1}{2n} \sum_{i=1}^n X_i^2.$$

Showing it is unbiased,

$$\begin{aligned}
E[\hat{\theta}_{MOM}] &= E\left[\frac{1}{2n}\sum_{i=1}^n X_i^2\right] \\
&= \frac{1}{2n} E\left[\sum_{i=1}^n X_i^2\right] \\
&= \frac{1}{2n} \sum_{i=1}^n E\left[X_i^2\right] \\
&= \frac{1}{2n} \sum_{i=1}^n 2\theta \\
&= \frac{1}{2n} 2n\theta \\
&= \theta.
\end{aligned}$$

(b) Estimate θ from the following $n = 10$ observations on vibratory stress of a turbine blade under specified conditions:

16.88, 10.23, 4.59, 6.66, 13.68, 14.23, 19.87, 9.40, 6.51, 10.95

```
data <- c(16.88, 10.23, 4.59, 6.66, 13.68,
          14.23, 19.87, 9.40, 6.51, 10.95)^2
(thetahat_mom <- sum(data)/(2*length(data)))

## [1] 74.50529
```

Our estimate for θ is 74.505.

13. Let $\hat{\lambda}$ be the MLE estimator for λ when $X \sim \text{Pois}(\lambda)$. To make large sample inferences on λ, we need the variance of $\hat{\lambda}$, which is $\text{Var}(\hat{\lambda}) = 1/(nI(\lambda))$ where $I(\lambda)$ is estimated by

$$I(\hat{\lambda}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d^2 \ln f(x_i|\lambda)}{d\lambda^2}\right]_{\lambda=\hat{\lambda}}.$$

An approximate large sample $(1-\alpha)$-level confidence interval for λ is given by

$$\hat{\lambda} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\lambda})}}.$$

Find this interval for λ.

We can find $I(\hat{\lambda})$ by first finding the second derivative of the log-Poission pdf, with as the parameter.

$$\begin{aligned}
\frac{d^2 f(x_i|\lambda)}{d\lambda^2} &= \frac{d^2}{d\lambda^2}\left(x_i \ln(\lambda) - \lambda - \ln(x_i!)\right) \\
&= \frac{d}{d\lambda}\left(\frac{x_i}{\lambda} - 1\right) \\
&= -\frac{x_i}{\lambda^2}.
\end{aligned}$$

Substituting our result from above into $I(\hat{\lambda})$, we would find

$$\begin{aligned}
I(\hat{\lambda}) &= -\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{x_i}{\lambda^2}\right]_{\lambda=\hat{\lambda}} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \left[\frac{1}{\hat{\lambda}^2}\right] \\
&= \frac{\bar{x}}{\hat{\lambda}^2} \\
&= \frac{1}{\bar{x}} \text{ (since } \hat{\lambda} = \bar{x}).
\end{aligned}$$

Since an approximate large sample $(1-\alpha)$-level confidence interval for λ is given by $\hat{\lambda} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\lambda})}}$, we can find the confidence interval by substituting the prior result for $I(\hat{\lambda})$.

$$\hat{\lambda} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\lambda})}} = \hat{\lambda} \pm z_{\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}}.$$