Homework 5 Key

Estimation

Instructions

Please answer the following questions in the order in which they are posed. Add a few empty lines below each and write your answers there. **Focus on answering in complete sentences and show work whether we ask for it or not**. You will also need scratch paper/pen to work out the answers before typing it.

For help with formatting documents in RMarkdown, please consult R Markdown: The Definitive Guide. Another option is to search using Google.

Exercises

1. (MIAA basketball) The `MIAA05` dataset from the **fastR2** package contains statistics on each of the 134 players from the 2004-2005 season . In this problem we will consider modeling the players' free throw shooting percentage. The variable is called `FTPct`. It is the ratio of free throws made (`FT`) to free throws attempted (`FTA`). Please type `?MIAA05` in the Console for a description of the variables in the dataset.

   Since `FTPct` is a proportion, we will consider the Beta distribution:

$$f(x) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)}\, x^{\alpha_0 - 1}\,(1-x)^{\beta_0 - 1} \quad 0 < x < 1.$$

   a. Show that the method of moments estimators of α_0 and β_0 are:

$$\hat{\alpha}_0^{mom} = \bar{x}\left[\frac{\bar{x} - s}{s - \bar{x}^2}\right],$$
$$\hat{\beta}_0^{mom} = \hat{\alpha}_0^{mom}\frac{1-\bar{x}}{\bar{x}}$$

   where $s = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

   Hint: You can use without proof from homework 1 key:

$$E\left[X\right] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$
$$E\left[X^2\right] = \frac{\alpha_0(\alpha_0+1)}{(\alpha_0+\beta_0)\,(\alpha_0+\beta_0+1)}.$$

First we may proceed to solve our Method of Moments estimators by replacing $E[X]$ and $E[X^2]$ with their empirical (data drawn) estimates $\bar{x}$ and s, respectively.

From here we have:

$$\bar{x} = \frac{\alpha_0}{\alpha_0 + \beta_0} \tag{1}$$
$$s = \frac{\alpha_0(\alpha_0+1)}{(\alpha_0+\beta_0)\,(\alpha_0+\beta_0+1)}. \tag{2}$$

Subbing equation (1) into equation (2) we see we have:

$$s = \frac{\bar{x}(\alpha_0 + 1)}{(\alpha_0 + \beta_0 + 1)}$$

We also notice that we have that we can solve for $\alpha_0 + \beta_0$ from Eq (1), resulting in:

$$\alpha_0 + \beta_0 = \frac{\alpha_0}{\bar{x}}$$

Thus lets now sub in this expression into our modified (2) and we have:

$$s = \frac{\bar{x}(\alpha_0 + 1)}{(\frac{\alpha_0}{\bar{x}} + 1)} = \frac{\bar{x}^2(\alpha_0 + 1)}{\alpha_0 + \bar{x}}$$

From here multiplying the denominator over and combining like terms yields:

$$s(\alpha_0) - \bar{x}^2\alpha_0 = \bar{x}^2 - s\bar{x} \Rightarrow \hat{\alpha_0}^{MoM} = \frac{\bar{x}^2 - s\bar{x}}{s - \bar{x}^2}$$

This is exactly the form we wish, thus we have that: $\hat{\alpha_0}^{MoM} = \frac{\bar{x}(\bar{x}-s)}{s-\bar{x}^2}$

Now plugging this equation into (1) we have:

$$\bar{x} = \frac{\hat{\alpha_0}^{MoM}}{\hat{\alpha_0}^{MoM} + \beta_0}$$

From here by multiplying the denominator and doing some algebra:

$$\hat{\beta_0}^{MoM} = \frac{\hat{\alpha_0}^{MoM} - \bar{x}\hat{\alpha_0}^{MoM}}{\bar{x}} = \hat{\alpha_0}^{MoM}\frac{1-\bar{x}}{\bar{x}}$$

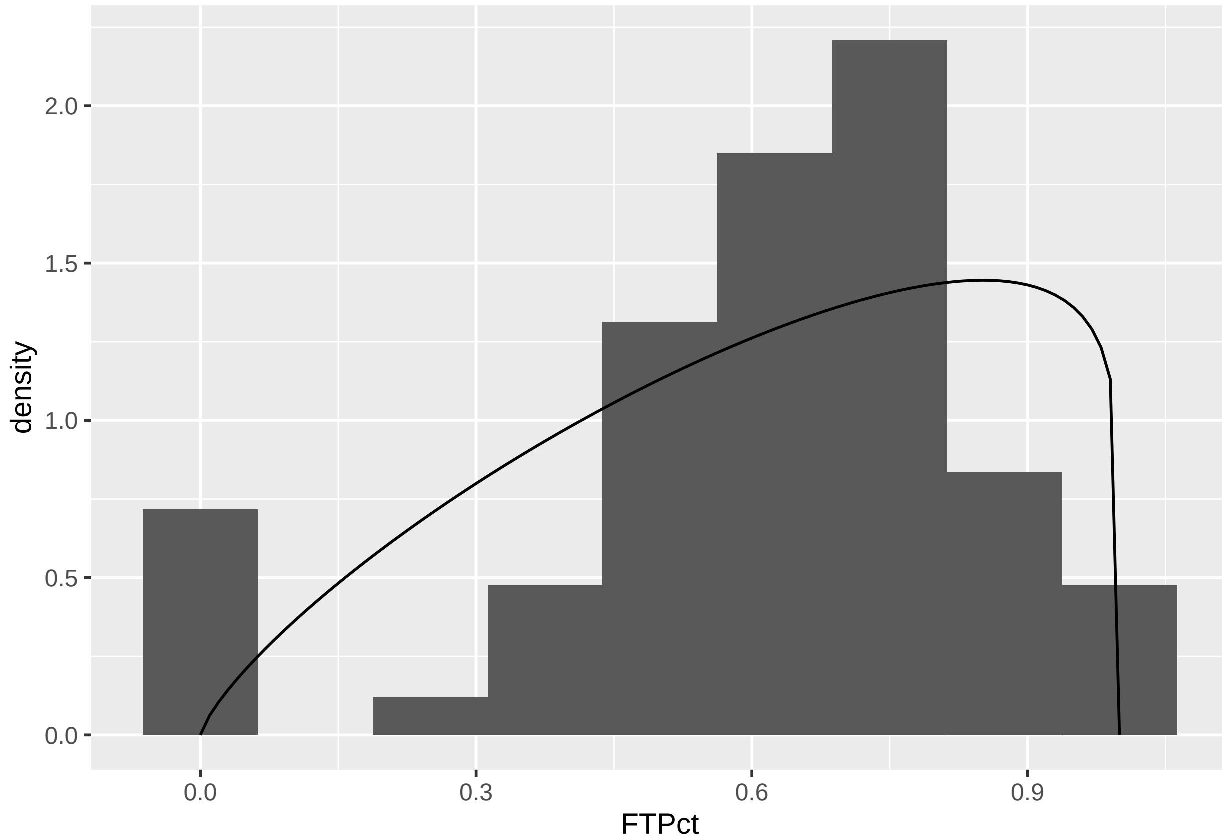
Thus we have found the forms given in the problem.

The remaining two parts involve coding. Be sure to show the code and output, however, suppress warnings and messages.

b. Make a histogram and a QQplot to assess the goodness of fit.

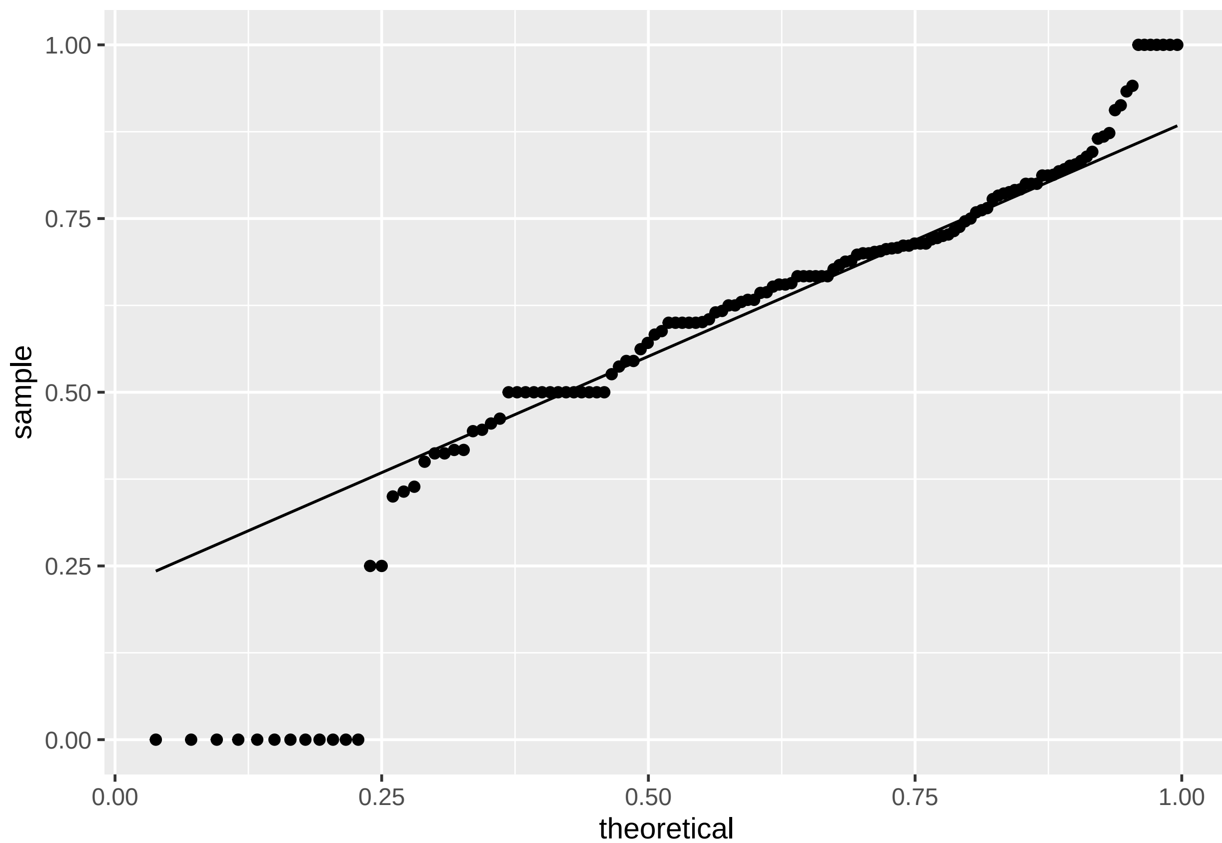
For our histogram we have:

```
#sturges rule, num_bins = 1+log_2(n)
bb_dat <- MIAA05
bin_num <- ceiling(1+log(nrow(bb_dat),2))
x_bar <- mean(MIAA05$FTPct)
s <- mean(MIAA05$FTPct^2)
alpha_mom <- x_bar*(x_bar-s)/(s-x_bar^2)
beta_mom <- alpha_mom*(1-x_bar)/x_bar
ggplot(data=bb_dat,aes(x=FTPct))+
  geom_histogram(bins = bin_num, aes(y=..density..))+
  stat_function(fun=dbeta,
                args = list(shape1 = alpha_mom,
                            shape2 = beta_mom))
```



We see we have an ok fit, but some outliers on the left side of the plot are preventing a better fit. We also have the qqplot as:

```
ggplot(data=bb_dat,aes(sample=FTPct))+
  stat_qq(distribution=qbeta,
               dparams = list(shape1 = alpha_mom,
                              shape2 = beta_mom))+
   stat_qqline(distribution=qbeta,
               dparams = list(shape1 = alpha_mom,
                              shape2 = beta_mom))
```



We see a similar story here, where we have a decent fit aside from the tail values.

c. Are there any players you should remove from the data before attempting the analysis? Decide on an elimination rule, apply it, and repeat the analysis. Do you like this fit better?

We see that we are trying to fit a beta distirbution to the free throw percentage of players. However, we notice that for many of players in the data frame we have a very low number of attempts. This means that the free throw percentage in the DF is not a good estimate of their true FTPct. Removing the values with low numbers of FTA, thus would clean our data and possibly provide a better fit. We can experiment with different cutoff values for this FTA value.

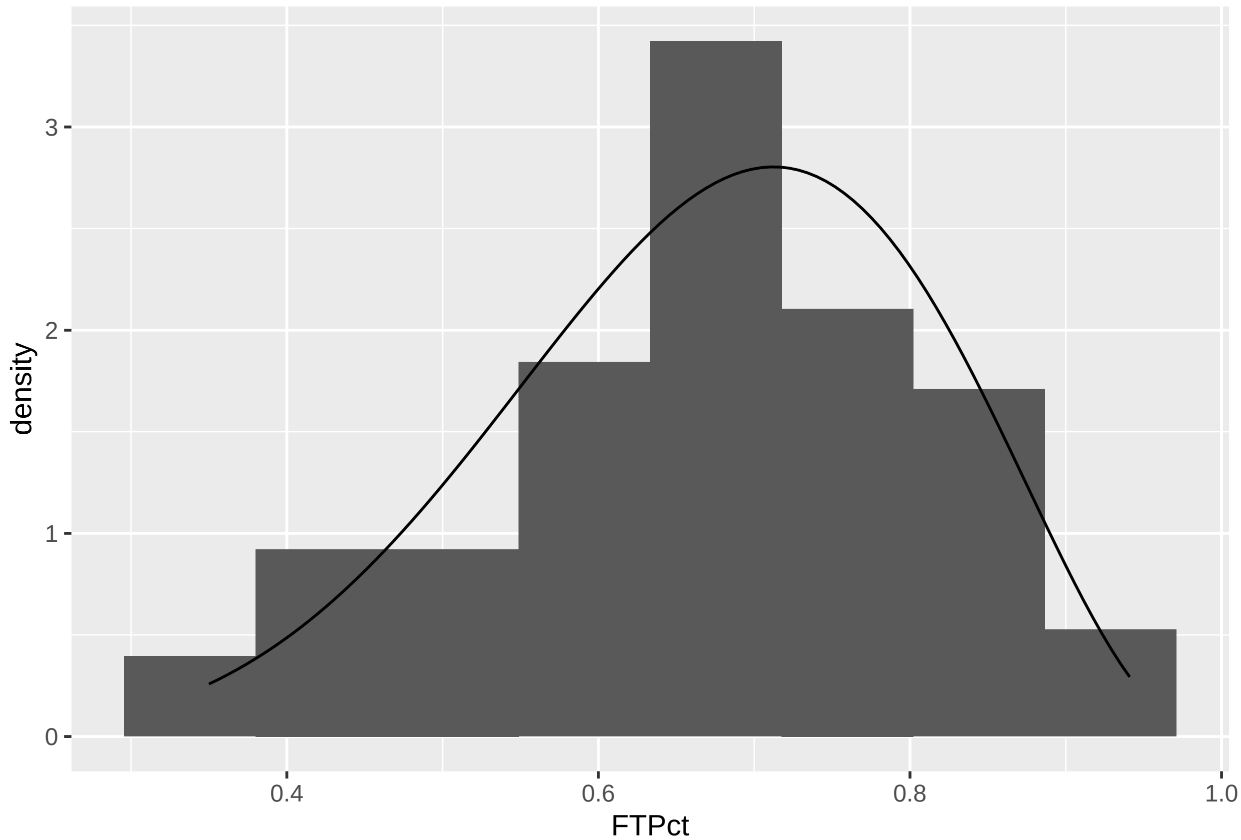
```
#filter out low FTA players
bb_dat_new <- MIAA05 %>% filter(FTA>10)
nrow(bb_dat_new)
```

```
## [1] 90
```

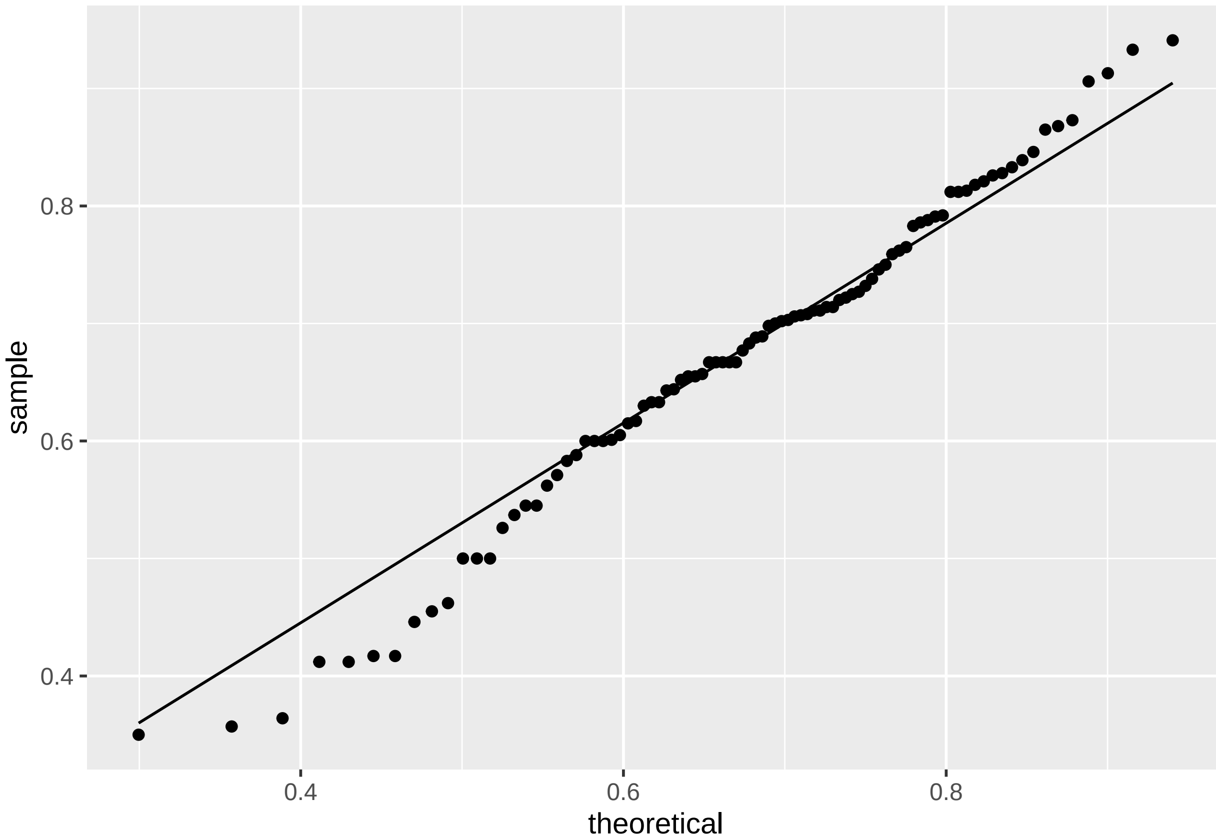
We see if we make our filter to be greater than 10 FTA, we still have 90 data points which is a relatively decent sample size still. We thus have the following fits on our histogram and qq plot:

```
#sturges rule, num_bins = 1+log_2(n)
bin_num <- ceiling(1+log(nrow(bb_dat_new),2))
x_bar <- mean(bb_dat_new$FTPct)
s <- mean(bb_dat_new$FTPct^2)
alpha_mom <- x_bar*(x_bar-s)/(s-x_bar^2)
beta_mom <- alpha_mom*(1-x_bar)/x_bar
ggplot(data=bb_dat_new,aes(x=FTPct))+
  geom_histogram(bins = bin_num, aes(y=..density..))+
  stat_function(fun=dbeta,
```

```
                args = list(shape1 = alpha_mom,
                            shape2 = beta_mom))
```




```
ggplot(data=bb_dat_new,aes(sample=FTPct))+
  stat_qq(distribution=qbeta,
                dparams = list(shape1 = alpha_mom,
                            shape2 = beta_mom))+
   stat_qqline(distribution=qbeta,
                dparams = list(shape1 = alpha_mom,
                            shape2 = beta_mom))
```



We see that both the qq-plot and histogram are much better fits than before.

2. (Unbias your estimator) Let $X \sim Binom(n, \pi_0)$.

a. Show, with justification, that

$$E\left[\frac{X}{n}\left(1-\frac{X}{n}\right)\right] = \frac{(n-1)\,\pi_0\,(1-\pi_0)}{n}$$

Simplifying this expression, we are looking for $E[\frac{X}{n} - \frac{X^2}{n^2}]$

We have for a Binomial RV, X that $E[X] = n\pi_0$ and $Var[X] = n(\pi_0)(1-\pi_0)$. From here we see that:

$$E[X^2] = n(\pi_0)(1-\pi_0) + n^2\pi_0^2 = n\pi_0 - n\pi_0^2 + n^2\pi_0^2 = n\pi_0(1-\pi_0+n\pi_0)$$

Thus we have that:

$$
\begin{aligned}
E[\frac{X}{n} - \frac{X^2}{n^2}] &= \frac{n\pi_0}{n} - \frac{n\pi_0(1-\pi_0+n\pi_0)}{n^2} \\
&= \frac{n^2\pi_0 - n\pi_0 + n\pi_0^2 - n^2\pi_0^2}{n^2} \\
&= \frac{n\pi_0 - \pi_0 + \pi_0^2 - n\pi_0^2}{n} \\
&= \frac{\pi_0(n-1+\pi_0-n\pi_0)}{n} \\
&= \frac{\pi_0((n-1)+(n-1)(\pi_0))}{n} \\
&= \frac{\pi_0(n-1)(1-\pi_0)}{n}
\end{aligned}
$$

b. Suppose we want an unbiased estimator for $\pi_0\,(1-\pi_0)$. Use your answer from (a) to construct such an estimator.

We have:

$$E\left[\frac{X}{n}\left(1-\frac{X}{n}\right)\right] = \frac{(n-1)\,\pi_0\,(1-\pi_0)}{n}$$

Thus we have that:

$$E\left[\left(\frac{n}{n-1}\right)\frac{X}{n}\left(1-\frac{X}{n}\right)\right] = \pi_0\,(1-\pi_0)$$

So we have the unbiased estimator $\left(\frac{n}{n-1}\right)\frac{X}{n}\left(1-\frac{X}{n}\right)$

3. (Bayes estimator) Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables drawn from the PMF:

$$f(x) = \begin{cases} (1-\pi_0) & x=0 \\ \pi_0 & x=1 \end{cases}$$

Consider the Bayesian estimator of π_0[1]:

$$\hat{\pi}_0^{bayes} = \frac{X+1}{n+2}$$

where $X = X_1 + X_2 + \cdots + X_n$ is $Binom(n, \pi_0)$.

a. Is $\hat{\pi}_0^{bayes}$ an unbiased estimator of π_0? If not, is it asymptotically unbiased?

We know that $E[X] = n\pi_0$

We have:

$$E[\hat{\pi}_0^{bayes}] = E[\frac{X+1}{n+2}] = \frac{n\pi_0+1}{n+2}$$

Since this is not equal to π_0, the bayes estimator is biased. We see that the exact bias is:

$$Bias = E[\frac{X+1}{n+2}] - \pi_0 = \frac{n\pi_0+1}{n+2} - \pi_0 = \frac{1-2\pi_0}{n+2} \to 0 \text{ as n goes to } \infty$$

Since this bias goes to zero as n goes to infinity, this estimator is asymptotically unbiased.

[1] don't worry about how to derive this estimator

b. Is $\hat{\pi}_0^{bayes}$ a consistent estimator?

We have $Var(\hat{\pi}_0^{bayes}) = Var(\frac{X+1}{n+2}) = \frac{1}{(n+2)^2}Var(X) = \frac{n(\pi_0)(1-\pi_0)}{(n+2)^2}$. We see as n goes to infinity this variance goes to 0. Since this estimator is also asymptotically unbiased, it is also consistent.

c. Based on the sample 1, 0, 1, 0, 1, calculate the value of $\hat{\pi}_0^{bayes}$ for this sample. Also calculate its estimated standard error.

We have in this case that $x = x_1 + \cdots + x_5 = 3$.

Thus we have that $\hat{\pi}_0^{bayes} = \frac{3+1}{5+2} = 4/7$.

We have $Var(\hat{\pi}_0^{bayes}) = \frac{n(\pi_0)(1-\pi_0)}{(n+2)^2}$. Since we do not know π_0, we must estimate it with our Bayes estimator. Thus we have that $Var(\hat{\pi}_0^{bayes}) = \frac{n(\hat{\pi}_0^{bayes})(1-\hat{\pi}_0^{bayes})}{(n+2)^2}$. Plugging in our values we get that $Var(\hat{\pi}_0^{bayes}) = 0.0249896$. So we have an estimated standard error of 0.158081.

4. (Bias variance trade-off) In this problem we will continue working with the model described in problem 3. Our focus will be on comparing the mean square error (MSE) of $\hat{\pi}_0^{bayes}$ with the MSE of

$$\hat{\pi}_0^{mom} = \frac{X}{n}.$$

a. Give expressions for the MSE of each estimator. Show your work clearly.

We have that $E[\hat{\pi}_0^{mom}] = E[\frac{X}{n}] = \frac{n\pi_0}{n} = \pi_0$. Thus this estimator is unbiased, so the MSE will simply equal the variance of the estimator. We have that $Var(\hat{\pi}_0^{mom}) = Var(\frac{X}{n}) = \frac{Var(X)}{n^2} = \frac{\pi_0(1-\pi_0)}{n}$.

Thus we have that $MSE(\hat{\pi}_0^{mom}) = \frac{\pi_0(1-\pi_0)}{n}$.

For the Bayes estimator we found that the Bias was equal to $\frac{1-2\pi_0}{n+2}$. We also saw that $Var(\hat{\pi}_0^{bayes}) = \frac{n(\pi_0)(1-\pi_0)}{(n+2)^2}$.

Thus we have that $MSE(\hat{\pi}_0^{bayes}) = [\frac{1-2\pi_0}{n+2}]^2 + \frac{n(\pi_0)(1-\pi_0)}{(n+2)^2} = \frac{(1-2\pi_0)^2 + n\pi_0(1-\pi_0)}{(n+2)^2}$

b. For $n = 10$, plot the MSE of both estimators on the same graph as a function of π_0. Describe (just visually) the π_0 values for which the MSE is smaller for the Bayes estimator. Repeat for $n = 1,000$. Don't forget to the label the plot, and make it easy for the reader to know which estimator is being represented by which curve.

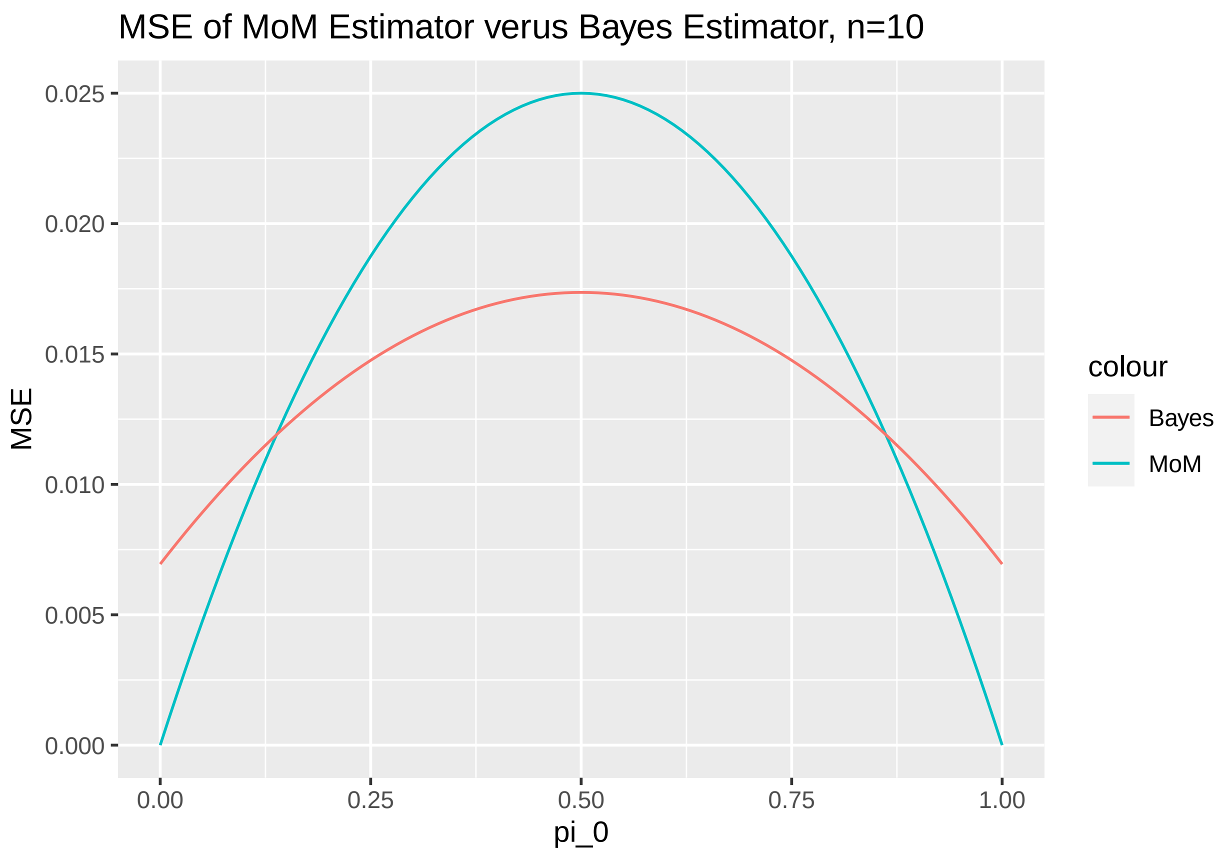
This example illustrates the bias variance tradeoff. Sometimes a biased estimator does better than an unbiased one, if it has a smaller variance!

For n=10 we have the following plot:

```
n1 <- 10
n2 <- 1000
pi_0 <- seq(0,1,.01)
mse_mom <- function(n){
  return(pi_0*(1-pi_0)/n)
}
mse_bayes <- function(n){
  return( ((1-2*pi_0)^2+n*pi_0*(1-pi_0))/(n+2)^2)
}

ggplot(data=NULL, mapping = aes(x=pi_0)) +
  geom_line(aes(y=mse_mom(n1),color="MoM"))+
  geom_line(aes(y=mse_bayes(n1),color="Bayes"))+
  labs(title="MSE of MoM Estimator verus Bayes Estimator, n=10",
```

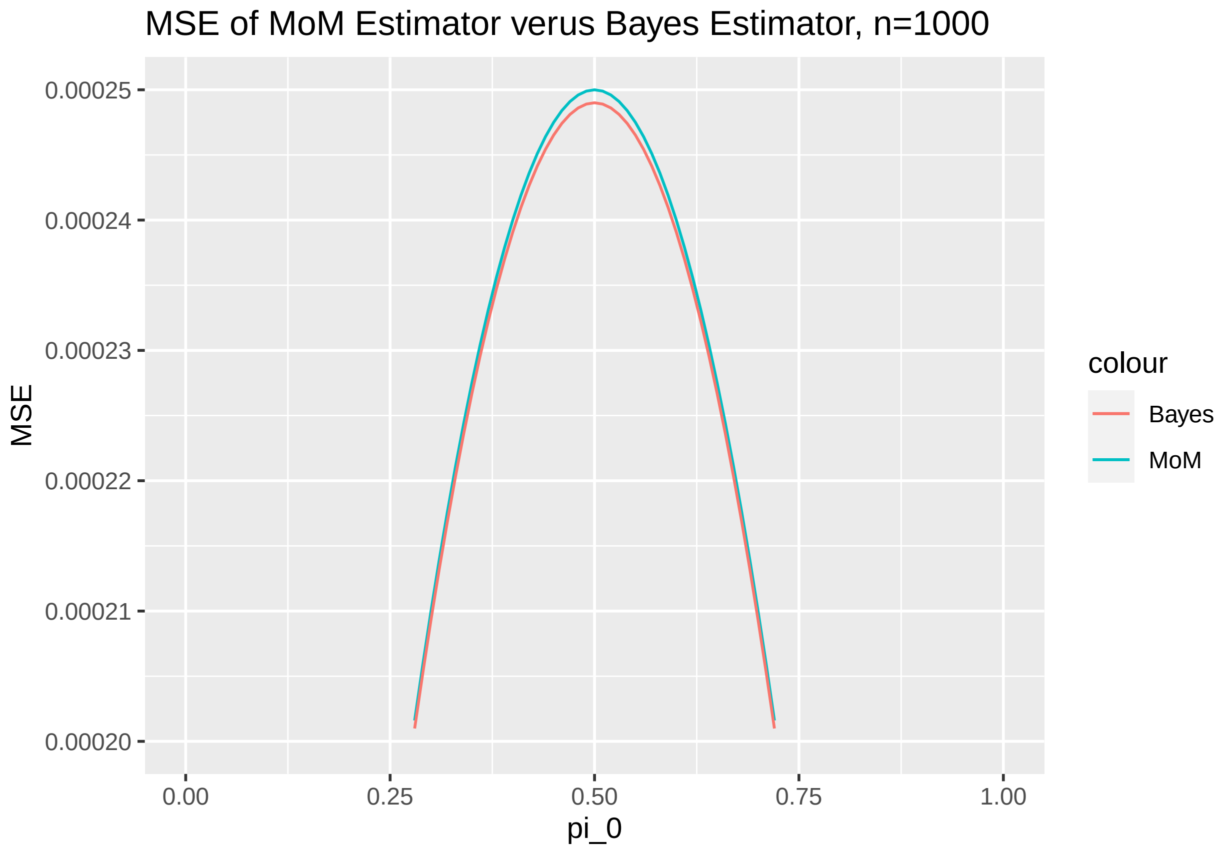
```
       x = "pi_0",
       y = "MSE")
```



We see that for small n, the bias added by the bayes estimator, decreases the variance enough such that the MSE will be lower for values of π_0 away from the tails (the boundary). For extreme values of π_0 however, the MoM estimator performs better.

For n=1000 we have:

```
ggplot(data=NULL, mapping = aes(x=pi_0)) +
  geom_line(aes(y=mse_mom(n2),color="MoM"))+
  geom_line(aes(y=mse_bayes(n2),color="Bayes"))+
  labs(title="MSE of MoM Estimator verus Bayes Estimator, n=1000",
       x = "pi_0",
       y = "MSE")+
  ylim(c(0.0002,.00025))
```



For large values of n we see a similar pattern, however in this case the difference between the MSE's is negligble. This is a common pattern that appears among consistent biased or unbiased estimators, where asymptotically they will perform very similarly.