S2S Lab 6

Exercise Solutions

Exercise 1

Suppose $X \sim \text{Bi}(16, 0.5)$. Investigate whether sample sizes of $n = 2$, $n = 5$, $n = 10$, $n = 50$ and $n = 100$ are sufficiently large for the sampling distribution of $\bar{X}$ to follow the normal distribution, as stated by the Central Limit Theorem.

For each different sample size, you should,

- Draw 10,000 random samples from $X \sim \text{Bi}(16, 0.5)$.
- Calculate the mean of each of these samples.
- Plot the kernel density estimate of the sampling distribution for $\bar{X}$.
- Superimpose the distribution that the Central Limit Theorem states the sampling distribution for $\bar{X}$ should follow above the kernel density estimate - use the mean and standard deviation of X to find which distribution this should be.

Hint You can draw a random sample from the Binomial distribution using the function `rbinom()`. Use the `help()` function to find out which arguments need to be provided to `rbinom()`.

Solution

Start by setting up empty vectors, of length 10,000, to store the sample means.

```
m <- 10000

means_2 <- numeric(m)
means_5 <- numeric(m)
means_10 <- numeric(m)
means_50 <- numeric(m)
means_100 <- numeric(m)
```

Use for loops draw the 10,000 samples and calculate the mean of each. Save these means in the vectors created above.

```
for(i in 1:m){
  means_2[i] <- mean(rbinom(n = 2, size = 16, prob = 0.5))
}

for(i in 1:m){
  means_5[i] <- mean(rbinom(n = 5, size = 16, prob = 0.5))
}

for(i in 1:m){
  means_10[i] <- mean(rbinom(n = 10, size = 16, prob = 0.5))
}

for(i in 1:m){
  means_50[i] <- mean(rbinom(n = 50, size = 16, prob = 0.5))
```

```
}

for(i in 1:m){
  means_100[i] <- mean(rbinom(n = 100, size = 16, prob = 0.5))
}
```

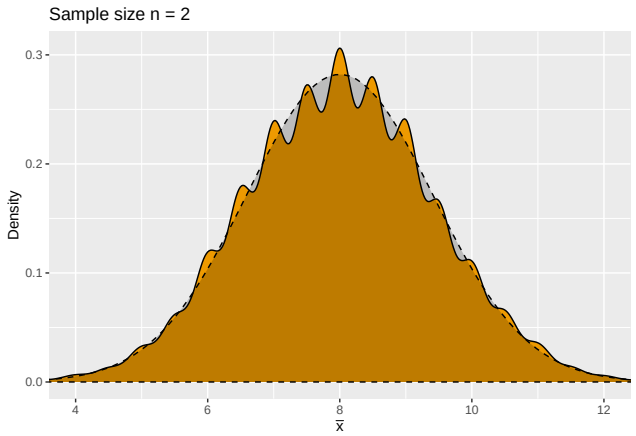
Store all the sample means, and the sample size they are calculated from, in a data frame so it can be used for plotting in `ggplot2`.

```
means <- data.frame(mean = c(means_2, means_5, means_10,
                             means_50, means_100),
                    size = rep(c("n = 2", "n = 5", "n = 10", "n = 50", "n = 100"), each = m))

means$size <- factor(means$size, levels = c("n = 2", "n = 5", "n = 10", "n = 50", "n = 100"))
```

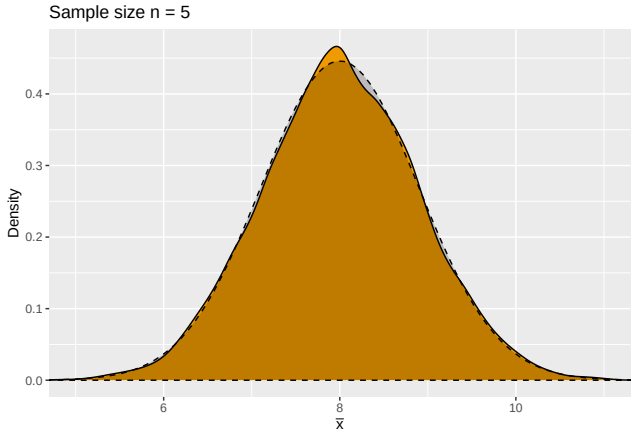
Plot the sampling distribution for $\bar{X}$ when the samples are of size $n = 2$. Since $E[X] = 8 = \mu$ and $sd(X) = 2 = \sigma$, superimpose the $N\left(8, \frac{2}{\sqrt{2}}\right)$.

```
ggplot(data = subset(means, subset = (size == "n = 2"))) +
  geom_density(aes(x = mean), fill = "orange2") +
  coord_cartesian(xlim = c(4, 12)) +
  labs(title = "Sample size n = 2",
       x = expression(bar(x)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 8, sd = 2/sqrt(2)), geom = "area",
                col = "black", linetype = 2, fill = "black", alpha = 0.2)
```



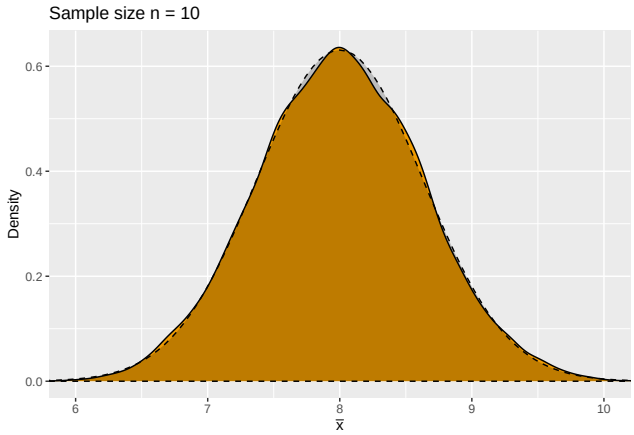
Plot the sampling distribution for $\bar{X}$ when the samples are of size $n = 5$. Superimpose the $N\left(8, \frac{2}{\sqrt{5}}\right)$.

```
ggplot(data = subset(means, subset = (size == "n = 5"))) +
  geom_density(aes(x = mean), fill = "orange2") +
  coord_cartesian(xlim = c(5, 11)) +
  labs(title = "Sample size n = 5",
       x = expression(bar(x)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 8, sd = 2/sqrt(5)), geom = "area",
                col = "black", linetype = 2, fill = "black", alpha = 0.2)
```



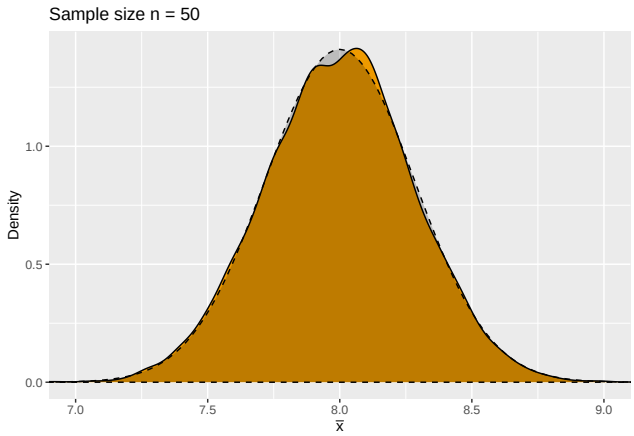
Plot the sampling distribution for $\bar{X}$ when the samples are of size $n = 10$. Superimpose the $N\left(8, \frac{2}{\sqrt{10}}\right)$.

```
ggplot(data = subset(means, subset = (size == "n = 10"))) +
  geom_density(aes(x = mean), fill = "orange2") +
  coord_cartesian(xlim = c(6, 10)) +
  labs(title = "Sample size n = 10",
       x = expression(bar(x)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 8, sd = 2/sqrt(10)), geom = "area",
                col = "black", linetype = 2, fill = "black", alpha = 0.2)
```



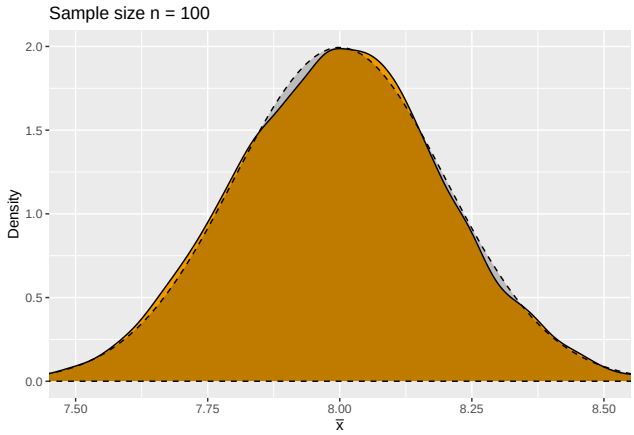
Plot the sampling distribution for $\bar{X}$ when the samples are of size $n = 50$. Superimpose the $N\left(8, \frac{2}{\sqrt{50}}\right)$.

```
ggplot(data = subset(means, subset = (size == "n = 50"))) +
  geom_density(aes(x = mean), fill = "orange2") +
  coord_cartesian(xlim = c(7, 9)) +
  labs(title = "Sample size n = 50",
       x = expression(bar(x)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 8, sd = 2/sqrt(50)), geom = "area",
                col = "black", linetype = 2, fill = "black", alpha = 0.2)
```



Plot the sampling distribution for $\bar{X}$ when the samples are of size $n = 100$. Superimpose the $N\left(8, \frac{2}{\sqrt{100}}\right)$.

```
ggplot(data = subset(means, subset = (size == "n = 100"))) +
  geom_density(aes(x = mean), fill = "orange2") +
  coord_cartesian(xlim = c(7.5, 8.5)) +
  labs(title = "Sample size n = 100",
       x = expression(bar(x)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 8, sd = 2/sqrt(100)), geom = "area",
                col = "black", linetype = 2, fill = "black", alpha = 0.2)
```



Exercise 2

The number of honeybees found in commercial beehives made by company X has a mean of 35,000 and a standard deviation of 12,000. The number of honeybees found in beehives made by company Y also has a mean of 35,000, but a standard deviation of 20,000. Both these hive sizes follow a normal distribution.

a) Suppose 40 beehives have been randomly sampled from company X and 22 beehives have been randomly sampled from company Y. Calculate the probability that the mean number of honeybees found in the beehives sampled from company X is at least 2,000 bees more than the mean number found in the sample from company Y's beehives.

b) Simulate taking the samples of beehives from companies X and Y by drawing 15,000 versions of the samples from the respective distributions. Calculate the sample mean in each case and use these to estimate the probability found in part (a). Is this empirical probability similar to the theoretical probability?

It might help to show the kernel density estimate of the sampling distribution of $\bar{X} - \bar{Y}$ in a plot, with the normal distribution it is expected to follow superimposed on top.

Solution

(a)

We know that the number of bees in beehives from both companies follow the distributions,

$$X \sim N(35\,000,\, 12\,000^2) \qquad Y \sim N(35\,000,\, 20\,000^2)$$

Because the samples drawn are of sizes $n_X = 40$ and $n_Y = 22$, then we also know that the difference between the mean of X and Y has the distribution,

$$\bar{X} - \bar{Y} \sim N\left(35\,000 - 35\,000 = 0,\, \sqrt{\frac{12\,000^2}{40} + \frac{20\,000^2}{22}}\right)$$

We can use this distribution to calculate $P\big(\bar{X} - \bar{Y} > 20\,000\big)$.

```
1-pnorm(2000, mean = 0, sd = sqrt((12000^2/40)+(20000^2/22)))
```

```
[1] 0.334132
```

This shows that $P\big(\bar{X} - \bar{Y} > 20\,000\big) = 0.334132$

(b)

Start by setting up the size of the samples to be drawn.

```
m <- 15000
nx <- 40
ny <- 22

meansX <- numeric(m)
meansY <- numeric(m)
```

Draw each sample 15,000 times and calculate the sample mean.

```
for(i in 1:m){
  meansX[i] <- mean(rnorm(n = nx, mean = 35000, sd = 12000))
}
```

```r
for(i in 1:m){
  meansY[i] <- mean(rnorm(n = ny, mean = 35000, sd = 20000))
}
```

The empirical estimate of $P\left(\bar{X} - \bar{Y} > 20\,000\right)$ is 0.3360667, as shown below.

```r
XY <- data.frame(diff = meansX - meansY)

mean(XY$diff > 2000)
```

```
[1] 0.3360667
```

Plot the kernel density estimate of the sampling distribution of $\bar{X} - \bar{Y}$ and superimpose the $N\left(0, \sqrt{\frac{12\,000^2}{40} + \frac{20\,000^2}{22}}\right)$ distribution.

```r
ggplot(data = XY) +
  geom_density(aes(x = diff), fill = "hotpink") +
  labs(x = expression(bar(X)-bar(Y)), y = "Density") +
  stat_function(fun = dnorm, args = list(mean = 0, sd = sqrt((12000^2/40)+(20000^2/22))),
                geom = "area", col = "black", linetype = 2, fill = "black", alpha = 0.2)
```