Extension Problems

Econ 103

Spring 2018

About This Document

Extension problems are designed to give you a deeper understanding of the lecture material and challenge you to apply what you have learned in new settings. Extension problems should only be attempted *after* you have completed the corresponding review problems. As an extra incentive to keep up with the course material, each exam of the semester will contain at least one problem taken *verbatim* from the extension problems. We will circulate solutions to the relevant extension problems the weekend before each exam. You are also welcome to discuss them with the instructor, your RI, and your fellow students at any point.

Lecture #1 – Introduction

1. A long time ago, the graduate school at a famous university admitted 4000 of their 8000 male applicants versus 1500 of their 4500 female applicants.

   (a) Calculate the difference in admission rates between men and women. What does your calculation suggest?

   > **Solution:** The rate for men is $4000/8000 = 50\%$ while that for women is $1500/4500 \approx 33\%$ so the difference is 17%. It appears that women are less likely to be accepted to the graduate school.

   (b) To get a better sense of the situation, some researchers broke these data down by area of study. Here is what they found:

	Men		Women	
	# Applicants	# Admitted	# Applicants	# Admitted
Arts	2000	400	3600	900
Sciences	6000	3600	900	600
Totals	8000	4000	4500	1500

Calculate the difference in admissions rates for men and women studying Arts. Do the same for Sciences.

Solution: For Arts, the admission rate is 400/2000 = 20% for men versus 900/3600 = 25% for women. For Sciences 3600/6000 = 60% for men versus 600/900 ≈ 67% for women. In summary:

	Men	Women	Difference
Arts	20%	25%	-5%
Sciences	60%	67%	-7%
Overall	50%	33%	17%

(c) Compare your results from part (a) to part (b). Explain the discrepancy using what you know about observational studies.

Solution: When we compare overall rates, women are less likely to be admitted than men. In each field of study, however, women are *more* likely to be admitted. In this example, field of study is a *confounder*: women are disproportionately applying to study Arts and Arts have much lower admissions rates than Sciences.

Lecture #2 – Summary Statistics I

2. The *mean deviation* is a measure of dispersion that we did not cover in class. It is defined as follows:

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

(a) Explain why this formula averages the absolute value of deviations from the mean rather than the deviations themselves.

Solution: As we showed in class, the average deviation from the sample mean is zero regardless of the dataset. Taking the absolute value is similar to squaring the deviations: it makes sure that the positive ones don't cancel out the negative ones.

(b) Which would you expect to be more sensitive to outliers: the mean deviation or the variance? Explain.

Solution: The variance is calculated from squared deviations. When x is far from zero, x^2 is much larger than $|x|$ so large deviations "count more" when calculating the variance. Thus, the variance will be more sensitive to outliers.

3. Let m be a constant and $x_1, \ldots, x_n$ be an observed dataset.

   (a) Show that $\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2$.

Solution:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - m)^2 &= \sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2) \\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2mx_i + \sum_{i=1}^{n} m^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2\end{aligned}$$

   (b) Using the preceding part, show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Solution: Solving this requires two observations. First, note that $\bar{x}$ is a *constant*, i.e. that it does not have an index of summation. Second, note that $\sum_{i=1}^{n} x_i = n\bar{x}$. Hence, taking $m = \bar{x}$ in the formula from the preceding part,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

Lecture #3 – Summary Statistics II

4. Consider a dataset $x_1, \ldots, x_n$. Suppose I multiply each observation by a constant d and then add another constant c, so that x_i is replaced by $c + dx_i$.

   (a) How does this change the sample mean? Prove your answer.

   Solution:

   $$\frac{1}{n}\sum_{i=1}^{n}(c + dx_i) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} c + d\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = c + d\bar{x}$$

   (b) How does this change the sample variance? Prove your answer.

   Solution:

   $$\frac{1}{n-1}\sum_{i=1}^{n}[(c + dx_i) - (c + d\bar{x})]^2 = \frac{1}{n-1}\sum_{i=1}^{n}[d(x_i - \bar{x})]^2 = d^2 s_x^2$$

   (c) How does this change the sample standard deviation? Prove your answer.

   Solution: The new standard deviation is $|d|s_x$, the positive square root of the variance.

   (d) How does this change the sample z-scores? Prove your answer.

   Solution: They are unchanged as long as d is positive, but the sign will flip if d is negative:

   $$\frac{(c + dx_i) - (c + d\bar{x})}{ds_x} = \frac{d(x_i - \bar{x})}{ds_x} = \frac{x_i - \bar{x}}{s_x}$$

Lecture #4 – Regression I

5. Define the z-scores

   $$w_i = \frac{x_i - \bar{x}}{s_x}, \quad \text{and} \quad z_i = \frac{y_i - \bar{y}}{s_y}.$$

Show that if we carry out a regression with z_i in place of y_i and w_i in place of x_i, the intercept a^* will be zero while the slope b^* will be r_{xy}, the correlation between x and y.

Solution: All we need to do is replace x_i with w_i and y_i with z_i in the formulas we already derived for the regression slope and intercept:

$$a = \bar{y} - b\bar{x}, \quad b = \frac{s_{xy}}{s_x^2}$$

And use the properties of z-scores from class. Let a^* be the intercept for the regression with z-scores, and b^* be the corresponding slope. We have:

$$a^* = \bar{z} - b^*\bar{w} = 0$$

since the mean of the z-scores is zero, as we showed in class. To find the slope, we need to covariance between the z-scores, and the variance of the z-scores for x:

$$b^* = \frac{s_{wz}}{s_w^2}$$

But since sample variance of z-scores is always one, $b^* = s_{wz}$. Now, by the definition of the sample covariance, the fact that the mean of z-scores is zero, and the definition of a z-score:

$$\begin{aligned} s_{wz} &= \frac{1}{n-1}\sum_{i=1}^{n}(w-\bar{w})(z-\bar{z}) = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ &= r_{xy} \end{aligned}$$

6. This question concerns a phenomenon called *regression to the mean*. Before attempting this problem, read Chapter 17 of *Thinking Fast and Slow* by Kahneman.

   (a) Lothario, an unscrupulous economics major, runs the following scam. After the first midterm of Econ 103 he seeks out the students who did extremely poorly and offers to sell them "statistics pills." He promises that if they take the pills before the second midterm, their scores will improve. The pills are, in fact, M&Ms and don't actually improve one's performance on statistics exams. The overwhelming majority of Lothario's former customers, however, swear that the pills really work: their scores improved on the second midterm. What's your explanation?

Solution: This is an example of regression to the mean. The students Lothario seeks out were both unprepared for the midterm *and* got unlucky: the correlation between exam scores is less than one. It is very unlikely that they will be unlucky twice in a row, so their performance on the second exam will almost certainly be higher. Our best guess of their second score is closer to the mean than their first score.

(b) Let $\hat{y}$ denote our prediction of y from a linear regression model: $\hat{y} = a + bx$ and let r be the correlation coefficient between x and y. Show that

$$\frac{\hat{y} - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right)$$

Solution:

$$\begin{aligned}
\hat{y} &= a + bx \\
\hat{y} &= (\bar{y} - b\bar{x}) + bx \\
\hat{y} - \bar{y} &= b(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x^2}(x - \bar{x}) \\
\hat{y} - \bar{y} &= \frac{s_{xy}}{s_x}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= \frac{s_{xy}}{s_x s_y}\left(\frac{x - \bar{x}}{s_x}\right) \\
\frac{\hat{y} - \bar{y}}{s_y} &= r\left(\frac{x - \bar{x}}{s_x}\right)
\end{aligned}$$

(c) Using the equation derived in (b), briefly explain"regression to the mean."

Solution: The formula shows that unless r is one or negative one, perfect positive or negative correlation, our best linear prediction of y based on knowledge given x is closer to the mean of the y-observations (relative to the standard deviation of the y-observations) than x is to mean of the x-observations (relative to the standard deviation of the x-observations). If x is very large, for example, we would predict that y will be large too, but not as large.

No extension problems for Lecture #5

Lecture #6 – Basic Probability II

7. You have been entered into a very strange tennis tournament. To get the $10,000 Grand Prize you must win at least two sets *in a row* in a three-set series to be played against your Econ 103 professor and Venus Williams alternately: professor-Venus-professor or Venus-professor-Venus according to your choice. Let p be the probability that you win a set against your professor and v be the probability that you win a set against Venus. Naturally $p > v$ since Venus is much better than your professor! Assume that each set is independent.

   (a) Let W indicate win and L indicate lose, so that the sequence WWW means you win all three sets, WLW means you win the first and third set but lose the middle one, and so on. Which sequences of wins and losses land you the Grand Prize?

   Solution: To get the prize, you have to win the middle set. Thus, the only possibilities are WWW, WWL, and LWW.

   (b) If you elect to play the middle set against Venus, what is the probability that you win the Grand Prize?

   Solution: The probabilities of mutually exclusive events sum. Thus,

$$\begin{aligned} P(WWW) + P(LWW) + P(WWL) &= pvp + (1-p)vp + pv(1-p) \\ &= p^2v + pv - p^2v + pv - p^2v \\ &= 2pv - p^2v \\ &= pv(2-p) \end{aligned}$$

   (c) If you elect to play the middle set against your professor, what is the probability that you win the Grand prize?

   Solution: Again, the probabilities of mutually exclusive events sum. Thus,

$$\begin{aligned} P(WWW) + P(LWW) + P(WWL) &= vpv + (1-v)pv + vp(1-v) \\ &= v^2p + vp - v^2p + vp - v^2p \\ &= 2pv - v^2p \\ &= pv(2-v) \end{aligned}$$

(d) To maximize your chance of winning the prize, should you choose to play the middle set against Venus or your professor?

Solution: Manipulating the inequality,

$$\begin{aligned} p &> v \\ -p &< -v \\ 2-p &< 2-v \\ pv(2-p) &< pv(2-v) \end{aligned}$$

You can't get the prize without winning the middle set, so it turns out that it's better to face Venus twice rather than face her in the middle set. You should elect to play the middle set against your professor.

8. Rossa and Rodrigo are playing their favorite game: matching pennies. The game proceeds as follows. In each round, both players flip a penny. If the flips match (TT or HH) Rossa gets one point; if the flips do not match (TH or HT) Rodrigo gets one point. The game is best of three rounds: as soon as one of the players reaches two points, the game ends and that player is declared the winner. Since there's a lot of money on the line and graduate students aren't paid particularly well, Rossa secretly alters each of the pennies so that the probability of heads is 2/3 rather than 1/2. In spite of Rossa's cheating, the individual coin flips remain independent.

(a) (6 points) Calculate the probability that Rossa will win the first round of this game.

Solution: Rossa wins a given round if either of the two mutually exclusive outcomes HH or TT occurs. Thus:

$$P(\text{Rossa Wins}) = P(HH) + P(TT) = (2/3)^2 + (1/3)^2 = 5/9$$

(b) Calculate the probability that the game will last for a full three rounds.

Solution: We need to calculate the probability of a tie after two rounds. There are two ways that a tie could occur: either Rossa wins the first round while Rodrigo wins the second, or Rodrigo wins the first round while Rossa wins the second. These two events are mutually exclusive and the probability of each is $5/9 \times 4/9 = 20/81$ since successive coin flips are independent. Thus, the desired probability is 40/81.

(c) Calculate the probability that Rodrigo will win the game.

Solution: Rodrigo needs to win two rounds to win the game. There are three ways this can happen. First, Rodrigo could win both rounds 1 and 2, in which case no third round is played. The probability of this event is $4/9 \times 4/9 = 16/81$. Second Rodrigo could lose round 1 but win rounds 2 and 3. The probability of this event is $5/9 \times 4/9 \times 4/9 = 80/729$. Finally, Rodrigo could lose round 2 but win rounds 1 and 3. The probability of this event is $4/9 \times 5/9 \times 4/9 = 80/729$. Summing these probabilities, since their corresponding events are mutually exclusive, the probability that Rodrigo wins the game is $304/729 \approx 0.417$.

(d) Yiwen is walking down the hallway and sees Rodrigo doing his victory dance: clearly Rossa has lost in spite of rigging the game. Given that Rodrigo won, calculate the probability that the game lasted for three rounds.

Solution: By the definition of conditional probability,

$$P(\text{3 Rounds}|\text{Rodrigo Won}) = \frac{P(\text{3 Rounds} \cap \text{Rodrigo Won})}{P(\text{Rodrigo Won})}$$

We already calculated the denominator in the preceding part: it equals 304/729. To calculate the numerator we simply add up the probabilities of the two mutually exclusive ways in which Rodrigo could win in three rounds: (Win, Lose, Win) and (Lose, Win, Win). We calculated these probabilities in the preceding part: both were 80/729 so the numerator is 160/729. Taking the ratio of these gives $160/304 \approx 0.526$. Given that Rodrigo won, it is slightly more likely than not that the game lasted for a full three rounds.

Lecture #7 – Basic Probability III / Discrete RVs I

9. A plane has crashed in one of three possible locations: the mountains (M), the desert (D), or the sea (S). Based on its flight path, experts have calculated the following prior probabilities that the plane is in each location: $P(M) = 0.5$, $P(D) = 0.3$ and $P(S) = 0.2$. If we search the mountains then, given that the plane is actually there, we have a 30% chance of *failing* to find it. If we search the desert then, given that the plane is actually there, we have a 20% chance of *failing* to find it. Finally, if we search the sea then, given that the plane is actually there, we have a 90% chance of *failing* to find it. Naturally if the plane is *not* in a particular location but we search for it there, we will not find it. You may assume that searches in each location are independent. Let F_M be the event that we *fail* to find the plane in the mountains. Define F_D and F_S analogously.

(a) We started by searching the mountains. We did not find the plane. What is the conditional probability that the plane is nevertheless in the mountains? Explain.

Solution: By Bayes' Rule: $P(M|F_M) = P(F_M|M)P(M)/P(F_M)$. We first calculate the denominator using the Law of Total Probability:

$$\begin{aligned} P(F_M) &= P(F_M|M)P(M) + P(F_M|M^C)P(M^C) \\ &= 0.3 \times 0.5 + 1 \times 0.5 = 0.15 + 0.5 = 0.65 \end{aligned}$$

Hence, the desired probability is $15/65 = 3/13 \approx 0.23$.

(b) After failing to find the plane in the mountains, we searched the desert, and the sea. We did not find the plane in either location. After this more exhaustive search what is the conditional probability that the plane is in the mountains? Explain.

Solution: We are asked to calculate $P(M|F_M \cap F_D \cap F_S)$. By Bayes' rule,

$$P(M|F_M \cap F_D \cap F_S) = \frac{P(F_M \cap F_D \cap F_S|M)P(M)}{P(F_M \cap F_D \cap F_S)}$$

Define the shorthand $A = F_M \cap F_D \cap F_S$. By the Law of Total Probability

$$\begin{aligned} P(A) &= P(A|M)P(M) + P(A|D)P(D) + P(A|S)P(S) \\ &= (0.3 \times 1 \times 1) \times 0.5 + (1 \times 0.2 \times 1) \times 0.3 + (1 \times 1 \times 0.9) \times 0.2 \\ &= 0.15 + 0.06 + 0.18 = 0.39 \end{aligned}$$

using independence. Hence, the desired probability is $15/39 \approx 0.38$.

Lecture #8 – Discrete RVs II

10. I have an urn that contains two red balls and three blue balls. I offer you the chance to play the following game. You draw one ball at a time from the urn. Draws are made at random and *without replacement*. You win \$1 for each red ball that you draw, but lose \$1 for each blue ball that you draw. You are allowed to stop the game at any point. Find a strategy that ensures your expected value from playing this game is *positive*.

Solution: There are $\binom{5}{2} = 10$ possible sequences of two red and blue balls, each of which has probability $1/10$ of occurring. The following table enumerates all of them:

WBBBW	BWBWB
WBBWB	BBWWB
WBWBB	BBBWW
WWBBB	BBWBW
BWWBB	BWBBW

I have found two strategies that yield a positive expected value. (There may be more.) The first to *keep playing if and only if your cumulative winnings are negative*. For example, if your first draw is W, your cumulative winnings are +1 so you should stop. On the other hand, if your first draw is B then you cumulative winnings are −1 so you should keep playing. If your first draw is a B and your second draw is a W, then your cumulative winnings are zero, so you should stop playing. The following table uses parentheses to show how this rule applies to each possible sequence of draws. You should *stop playing* as soon as you hit the first parenthesis. The value to the right of the arrow denotes your winnings for a given sequence when following the strategy.

W(BBBW) $\to +1$	BW(BWB) $\to 0$
W(BBWB) $\to +1$	BBWW(B) $\to 0$
W(BWBB) $\to +1$	BBBWW() $\to -1$
W(WBBB) $\to +1$	BBWBW() $\to -1$
BW(WBB) $\to 0$	BW(BBW) $\to 0$

The expected value of this strategy is $(1 + 1 + 1 + 1 - 1 - 1)/10 = 1/5$ since each sequence has a probability of 1/10.

A slightly different strategy that also gives a positive expected value is as follows:

> If your first draw is W, stop; if your first draw is B, keep drawing until both white balls are removed.

Following the same notational convention as above, we can summarize the results of this strategy as follows:

W(BBBW) $\to +1$	BWBW(B) $\to 0$
W(BBWB) $\to +1$	BBWW(B) $\to 0$
W(BWBB) $\to +1$	BBBWW() $\to -1$
W(WBBB) $\to +1$	BBWBW() $\to -1$
BWW(BB) $\to +1$	BWBBW() $\to -1$

so the expected value is $(1 + 1 + 1 + 1 + 1 - 1 - 1 - 1)/10 = 1/5$. Although these two strategies have the same expected value, they different. Let p denote the pmf of

your winnings under the first strategy and q denote the pmf of your winnings under the second. Then we see that:

$$p(-1) = 2/10, \quad p(0) = 4/10, \quad p(1) = 4/10$$

while

$$q(-1) = 3/10, \quad q(0) = 2/10, \quad p(1) = 5/10$$

Hence, you have a larger chance of winning a positive amount using the second strategy, but also a larger chance of *losing* a positive amount. A helpful review problem would be to calculate the variance of your winnings under each strategy.

11. An ancient artifact worth \$100,000 fell out of Indiana Jones's airplane and landed in the Florida Everglades. Unless he finds it within a day, it will sink to the bottom and be lost forever. Dr. Jones can hire one or more helicopters to search the Everglades. Each helicopter charges \$1,000 per day and has a probability of 0.9 of finding the artifact. If Dr. Jones wants to maximize his *expected value*, how many helicopters should he hire?

Solution: Let $p(n)$ be the probability of finding the artifact if Indiana Jones hires n helicopters. By the complement rule, $p(n) = 1 - 1/10^n$ since the probability of *not* finding the artifact when n helicopters are search for it is $(1 - 0.9)^n = 1/10^2$. Now let $E(n)$ denote Indiana Jones's expected value if he hires n helicopters. If he does *not* find the artifact, Jones *loses* $n\times$\$1,000. If he *does* find the artifact, Jones *gains* \$100,000$-n\times$\$1,000. Therefore,

$$\begin{aligned} E(n) &= [1 - p(n)] \times (-1000 \times n) + p(n) \times [100000 - 1000 \times n] \\ &= 100000 \times p(n) - 1000 \times n \\ &= 1000 \times [100 \times p(n) - n] \end{aligned}$$

The question of whether Jones should hire an *additional* – i.e. *marginal* – helicopter comes down to whether the marginal expected benefit, $100000 \times [p(n+1) - p(n)]$, exceeds the marginal expected cost, 1000. From the factorization above, we see that this will be the case whenever $100 \times [p(n+1) - p(n)]$ is larger than one. Notice that, because you cannot hire a fraction of a helicopter, it may not be possible to exactly equate marginal expected cost and benefit as we would do in a continuous optimization problem. Instead we'll make a table of values. It's easy enough to do this by hand, but we could also use R:

```
n <- 1:4
p <- 1 - 1/10^n
EV <- 100000 * p - 1000 * n
cbind(n,p,EV)

##      n      p    EV
## [1,] 1 0.9000 89000
## [2,] 2 0.9900 97000
## [3,] 3 0.9990 96900
## [4,] 4 0.9999 95990
```

From the table is appears that the optimal number of helicopters is 2. But how can we be sure when we have only examined five possible values for n? Notice that the marginal expected cost is a constant \$1000 regardless of n. In contrast, the marginal expected benefit is

$$\begin{aligned} p(n+1) - p(n) &= \left(1 - \frac{1}{10^{n+1}}\right) - \left(1 - \frac{1}{10^n}\right) \\ &= \frac{1}{10^n} - \frac{1}{10^{n+1}} \\ &= \frac{1}{10^n}\left(1 - \frac{1}{10}\right) \\ &= \frac{1}{10^n} \times 0.9 \end{aligned}$$

This is *decreasing* with n, so we know that it is it unnecessary to examine larger values of n. An alternative way to solve this problem is to "pretend" that n is continuous, derive the first and second order conditions to characterize the unique global optimum, and then look at the whole number values of n on each side of the (infeasible) optimum from the continuous problem.

No extension problems for Lecture #9

Lecture #10 – Discrete RVs IV

12. Let X and Y be discrete random variables. Prove that if X and Y are independent, then $Cov(X, Y) = 0$. Hint: write out the definition of covariance for two discrete RVs in terms of a double sum, and use independence to substitute $p_{XY}(x, y) = p_X(x)p_Y(y)$.

Solution:

$$\begin{aligned}
Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x,y) \\
&= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(x)p(y) \\
&= \sum_x (x-\mu_X)p(x)\left[\sum_y (y-\mu_Y)p(y)\right] \\
&= E[Y-\mu_Y]\sum_x (x-\mu_X)p(x) \\
&= E[Y-\mu_Y]E[X-\mu_X] \\
&= 0
\end{aligned}$$

13. Let a, b, c be constants and X, Y be RVs. Show that:

$$Var(aX+bY+c) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y).$$

Hint: the steps are similar our proof that $Var(aX+b) = a^2Var(X)$ from lecture.

Solution:

$$\begin{aligned}
Var(aX+bY) &= E\left[\{(aX+bY) - E[aX+bY]\}^2\right] \\
&= E\left[\{a(X-\mu_X) + b(Y-\mu_Y)\}^2\right] \\
&= E\left[a^2(X-\mu_X)^2 + b^2(Y-\mu_Y)^2 + 2ab(X-\mu_X)(Y-\mu_Y)\right] \\
&= a^2E[(X-\mu_X)^2] + b^2E[(Y-\mu_Y)^2] + 2abE[(X-\mu_X)(Y-\mu_Y)] \\
&= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)
\end{aligned}$$

14. Let X_1 be a random variable denoting the returns of stock 1, and X_2 be a random variable denoting the returns of stock 2. Accordingly let $\mu_1 = E[X_1]$, $\mu_2 = E[X_2]$, $\sigma_1^2 = Var(X_1)$, $\sigma_2^2 = Var(X_2)$ and $\rho = Corr(X_1, X_2)$. A *portfolio*, Π, is a linear combination of X_1 and X_2 with weights that sum to one, that is $\Pi(\omega) = \omega X_1 + (1-\omega)X_2$, indicating the proportions of stock 1 and stock 2 that an investor holds. In this example, we require $\omega \in [0,1]$, so that *negative* weights are not allowed. (This rules out short-selling.)

(a) Express $Var[\Pi(\omega)]$ in terms of ρ, σ_1^2 and σ_2^2.

Solution:

$$\begin{aligned} Var[\Pi(\omega)] &= Var[\omega X_1 + (1-\omega)X_2] \\ &= \omega^2 Var(X_1) + (1-\omega)^2 Var(X_2) + 2\omega(1-\omega)Cov(X_1, X_2) \\ &= \omega^2\sigma_1^2 + (1-\omega)^2\sigma_2^2 + 2\omega(1-\omega)\rho\sigma_1\sigma_2 \end{aligned}$$

(b) Find the value of ω^* that minimizes $Var[\Pi(\omega)]$. You do not have to check the second order condition. In finance, $\Pi(\omega^*)$ is called the *minimum variance portfolio*.

Solution: The First Order Condition is:

$$2\omega\sigma_1^2 - 2(1-\omega)\sigma_2^2 + (2-4\omega)\rho\sigma_1\sigma_2 = 0$$

Dividing both sides by two and rearranging:

$$\begin{aligned} \omega\sigma_1^2 - (1-\omega)\sigma_2^2 + (1-2\omega)\rho\sigma_1\sigma_2 &= 0 \\ \omega\sigma_1^2 - \sigma_2^2 + \omega\sigma_2^2 + \rho\sigma_1\sigma_2 - 2\omega\rho\sigma_1\sigma_2 &= 0 \\ \omega(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) &= \sigma_2^2 - \rho\sigma_1\sigma_2 \end{aligned}$$

So we have

$$\omega^* = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

(c) I have posted five years of daily closing stock prices for Apple (AAPL) and Google (GOOG) on my website: `http://ditraglia.com/econ103/closing_prices.csv`. Write code to read this data into R and store it in a dataframe called `prices`.

Solution:

```
data_url <- 'http://ditraglia.com/econ103/closing_prices.csv'
prices <- read.csv(data_url)
```

(d) If P_t is today's closing price and P_{t-1} is yesterday's closing price, then we define the *log daily return* as $\log(P_t) - \log(P_{t-1})$ where log denotes the natural logarithm. Write R code that uses `prices` to create two vectors `apple_returns` and `google_returns` containing *log returns* for Apple and Google respectively. The R function for natural logarithms is `log` while `diff` takes successive *differences* of a vector.

Solution:

```
google_returns <- diff(log(prices$GOOG))
apple_returns <- diff(log(prices$APPL))
```

(e) Using `apple_returns` and `google_returns`, write R code to compute the sample standard deviation of Apple and Google daily log returns, along with the sample correlation between them.

Solution:

```
s1 <- sd(google_returns)
s2 <- sd(apple_returns)
r <- cor(google_returns, apple_returns)
```

(f) Suppose we make the assumption that *future* Apple and Google returns are random variables with the standard deviations and correlation equal to the values you computed in the preceding part. If we have \$100 and wish to invest in these two companies, how much money should we allocate to Apple and Google stock to minimize the variance of our portfolio? (If we take variance as a measure of risk, this problem is asking you to calculate the "safest" portfolio.)

Solution: The formula from part (b) gives

```
wstar <- (s2^2 - r * s1 * s2) / (s1^2 + s2^2 - 2 * r * s1 * s2)
wstar
## [1] 0.5422114
```

so we should invest \$54 in Google and \$46 in Apple.

No extension problems for Lecture #11

Lecture #12 – Continuous RVs II

15. Suppose that X is continuous RV with PDF $f(x) = 3x^2$ for $x \in [0, 1]$, zero otherwise. Calculate the median of X.

Solution: To answer this question, we need to derive the quantile function of X. Since $Q(p)$ is simply the *inverse* of the CDF, we first need to derive $F(x_0)$. We have:

$$F(x_0) \equiv \int_{-\infty}^{x_0} f(x)\,dx = \int_0^{x_0} 3x^2\,dx = x_0^3$$

for $x_0 \in [0,1]$, $F(x_0) = 0$ for $x_0 < 0$ and $F(x_0) = 1$ for $x_0 > 1$. Hence the quantile function of X is $Q(p) = F^{-1}(p) = \sqrt[3]{p}$. To find the median of X, we simply evaluate this at $p = 0.5$, yielding median $= \sqrt[3]{0.5} \approx 0.79$.

16. Students applying for a job at a consulting firm are required to take two tests. For a given applicant, scores on the two tests can be viewed as random variables: X and Y. From information on past applicants, we know that both tests have means equal to 50, and standard deviations equal to 10. The correlation between scores on the two tests is 0.62. Let $Z = X + Y$ denote an applicant's combined score.

    (a) Calculate $E[Z]$

    Solution: $E[Z] = E[X+Y] = E[X] + E[Y] = 50 + 50 = 100$

    (b) Calculate $Var(Z)$.

    Solution:

$$\begin{aligned}\text{Var}(Z) &= Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \\ &= 10^2 + 10^2 + 2\sigma_X\sigma_Y\rho_{XY} \\ &= 200 + 2 \times 10 \times 10 \times 0.62 = 324\end{aligned}$$

    (c) Only students whose combined score is at least 136 are hired. If Z is normally distributed, approximately what percentage of applicants applicants will be hired?

    Solution: From the preceding parts $\mu_Z = 100$ and $\sigma_Z = \sqrt{324} = 18$. Hence,

$$\begin{aligned}P(Z > 136) &= 1 - P(Z \le 136) = 1 - P\left(\frac{Z-100}{18} \le \frac{136-100}{18}\right) \\ &= 1 - P\left(\frac{Z-100}{18} \le 2\right) = 1 - \texttt{pnorm(2)} \approx 0.025\end{aligned}$$

    So about 2.5% of applicants will be hired.

(d) Continuing under the assumption that Z is normally distributed, as in the preceding part, suppose the firm wanted to hire the top 16% of applicants. Approximately what cutoff should they use for each applicant's combined score?

Solution: We need to find c such that $P(Z > c) = 0.16$. Using the same reasoning as in the preceding solution,

$$P(Z > c) = 1 - P\left(\frac{Z - 100}{18} \leq \frac{c - 100}{18}\right) = 0.16$$

and hence

$$1 - \texttt{pnorm}\big((c - 100)/18\big) = 0.16$$

Rearranging,

$$\texttt{pnorm}\big((c - 100)/18\big) = 0.84$$

Since `pnorm` and `qnorm` are inverses of each other, if we apply them to both sides, we get

$$(c - 100)/18 = \texttt{qnorm}(0.84)$$

Here's another way of saying the same thing: we need to choose c so that $\texttt{pnorm}\big((c-100)/18\big) = 0.84$ which means that $(c-100)/18$ needs to be the 0.84 quantile of the standard normal distribution. Since $\texttt{qnorm}(0.84) \approx 1$, we have

$$\begin{aligned} (c - 100)/18 &\approx 1 \\ c - 100 &\approx 18 \\ c &\approx 118 \end{aligned}$$

So the firm should lower their cutoff from 136 to about 118 if they want to hire the top 16% of applicants.

Lecture #13 – Sampling and Estimation I

17. Garth wants to learn how much taller NBA players are than Penn Undergraduates, on average. To answer this question, he's recruited volunteers to make up two independent random samples. The first sample contains 10 NBA players: $X_1, \ldots, X_{10} \sim$ iid with mean μ_X and variance σ^2. The second sample is independent of the first and contains 15 Penn Undergrads: $Y_1, \ldots, Y_{15} \sim$ iid with mean μ_Y and variance σ^2. Just to be completely clear, in this question we are assuming that the variance is identical for Penn Students and NBA players to make the calculations simpler.

(a) To answer his question, Garth needs to estimate $\mu_X - \mu_Y$. There is an obvious

unbiased estimator of this quantity. What is it? Prove that it is unbiased.

Solution: The obvious choice is the difference of sample means: $\bar{X} - \bar{Y}$. We know that this is unbiased because $\bar{X}$ is an unbiased estimator of μ_X, $\bar{Y}$ is an unbiased estimator of μ_Y and by the linearity of expectation

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_X - \mu_Y$$

(b) Calculate the variance of the estimator you proposed in part (a).

Solution:

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma^2}{10} + \frac{\sigma^2}{15} = \frac{\sigma^2}{6}$$

(c) When measuring the players and students Garth makes a mistake: although he accurately records each of the 25 heights, he forgets to note which correspond to Penn students and which correspond to NBA players. Fortunately Garth remembers that, among the first 10 heights on his list, there were exactly 5 students and 5 NBA players. In other words, the first 10 heights on his list are $X_1, \ldots, X_5$ and $Y_1, \ldots, Y_5$ *in some unknown order* and the last 15 are $X_6, \ldots, X_{10}$ and $Y_6, \ldots, Y_{15}$ *in some unknown order*. Let $\bar{Z}_1$ be the sample mean of the first 10 heights on Garth's list and $\bar{Z}_2$ be the sample mean of the last 15. Prove that

$$E[\bar{Z}_1] = \frac{\mu_X + \mu_Y}{2} \qquad \text{and} \qquad E[\bar{Z}_2] = \frac{1}{3}\mu_X + \frac{2}{3}\mu_Y$$

Solution: We have

$$\bar{Z}_1 = \frac{1}{10}\left(X_1 + X_2 + \cdots + X_5 + Y_1 + Y_2 + \cdots + Y_5\right)$$

Hence, by the linearity of expectation,

$$E[\bar{Z}_1] = \frac{1}{10}\left(5\mu_X + 5\mu_Y\right) = \frac{\mu_X + \mu_Y}{2}$$

Similarly,

$$\bar{Z}_2 = \frac{1}{15}\left(X_6 + X_7 + \cdots + X_{10} + Y_6 + Y_7 + \cdots + Y_{15}\right)$$

and, again, by the linearity of expectation

$$E[\bar{Z}_2] = \frac{1}{15}\left(5\mu_X + 10\mu_Y\right) = \frac{1}{3}\mu_X + \frac{2}{3}\mu_Y$$

(d) Let α and β be two arbitrary constants. Using part (c), calculate $E[\alpha\bar{Z}_1 + \beta\bar{Z}_2]$. Simplify your answer so that it takes the form $c_1\mu_X + c_2\mu_Y$ where c_1 and c_2 are two constants that will depend on α and β.

Solution: By the linearity of expectation:

$$\begin{aligned} E[\alpha\bar{Z}_1 + \beta\bar{Z}_2] &= \alpha E[\bar{Z}_1] + \beta E[\bar{Z}_2] \\ &= \alpha\left(\frac{\mu_X + \mu_Y}{2}\right) + \beta\left(\frac{1}{3}\mu_X + \frac{2}{3}\mu_Y\right) \\ &= \mu_X\left(\frac{\alpha}{2} + \frac{\beta}{3}\right) + \mu_Y\left(\frac{\alpha}{2} + \frac{2\beta}{3}\right) \end{aligned}$$

(e) Back to Garth's problem: *amazingly* it's still possible to construct an unbiased estimator of $\mu_X - \mu_Y$ *without knowing* which observations corresponded to NBA players and which corresponded to Penn students. The trick is to take a particular linear combination of $\bar{Z}_1$ and $\bar{Z}_2$. Use your answer to the previous part to find the values of α and β that give $E[\alpha\bar{Z}_1 + \beta\bar{Z}_2] = \mu_X - \mu_Y$.

Solution: We have two linearly independent equations in two unknowns:

$$\begin{aligned} \alpha/2 + \beta/3 &= 1 \\ \alpha/2 + 2\beta/3 &= -1 \end{aligned}$$

Solving the first equation gives $\alpha/2 = 1 - \beta/3$. Substituting this into the second:

$$\begin{aligned} (1 - \beta/3) + 2\beta/3 &= -1 \\ \beta &= -6 \end{aligned}$$

Hence $\alpha = 2(1 - \beta/3) = 6$. Plugging these values into the result of the previous part, we can verify that these values indeed give an unbiased estimator:

$$E[6\bar{Z}_1 - 6\bar{Z}_2] = 3(\mu_X + \mu_Y) - (2\mu_X + 4\mu_Y) = \mu_X - \mu_Y$$

(f) Although it's unbiased, there's a clear downside to the estimator from the previous part: it has a very high variance. Calculate its variance and compare it to that of the estimator from part (a). Explain the intuition behind the difference.

Solution: We have:

$$Var(6\bar{Z}_1 - 6\bar{Z}_2) = 36\left[Var(\bar{Z}_1) + Var(\bar{Z}_2)\right] = 36(\sigma^2/10 + \sigma^2/15) = 6\sigma^2$$

which is 36 times as large as $Var(\bar{X} - \bar{Y})$! Although $Var(\bar{X} - \bar{Y}) = Var(\bar{Z}_1 - \bar{Z}_2)$, when we multiplied $\bar{Z}_1$ and $\bar{Z}_2$ by six to get an unbiased estimator, this multiplied the variance by $6^2 = 36$.

Lecture #14 – Sampling and Estimation II

18. Problem 7-13 parts (a)–(c) from Wonnacott & Wonnacott.

Solution: The point is that S, the number of successes in n trials each with probability π of success, is a Binomial(n, π) random variable. We calculated the mean and variance of such a RV in class (see the slides) and we will use this information to find the MSE of $P = S/n$ as well as that of

$$P^* = \frac{nP + 1}{n + 2} = \left(\frac{n}{n+2}\right) P + \left(\frac{1}{n+2}\right)$$

The reason the book gives you the above expression is to give you a hint: namely that once you've solved for the MSE of P you can use this to get the MSE of P^* fairly easily by writing it as above.

$$\begin{aligned} E[P] &= E[S/n] = E[S]/n = n\pi/n = \pi \\ \text{Bias}(P) &= E[P] - \pi = \pi - \pi = 0 \\ Var(P) &= Var(S/n) = Var(S)/n^2 = n\pi(1-\pi)/n^2 = \pi(1-\pi)/n \\ MSE(P) &= \text{Bias}(P)^2 + Var(P) = 0^2 + \pi(1-\pi)/n = \pi(1-\pi)/n \end{aligned}$$

where we have used our rules for manipulating expectation and variance, as well as

the expressions for the mean and variance of a Binomial random variable. Now:

$$\begin{aligned}
E[P^*] &= E\left[\left(\frac{n}{n+2}\right)P + \left(\frac{1}{n+2}\right)\right] = \left(\frac{n}{n+2}\right)E[P] + \left(\frac{1}{n+2}\right) \\
&= \left(\frac{n}{n+2}\right)\pi + \left(\frac{1}{n+2}\right) \\
\text{Bias}(P^*) &= E[P^*] - \pi = \left(\frac{n}{n+2}\right)\pi + \left(\frac{1}{n+2}\right) - \pi \\
&= \left(\frac{n}{n+2} - 1\right)\pi + \left(\frac{1}{n+2}\right) = \frac{1-2\pi}{n+2} \\
Var(P^*) &= Var\left[\left(\frac{n}{n+2}\right)P + \left(\frac{1}{n+2}\right)\right] = \left(\frac{n}{n+2}\right)^2 Var(P) \\
&= \frac{n^2}{(n+2)^2}\frac{\pi(1-\pi)}{n} = \frac{n\pi(1-\pi)}{(n+2)^2} \\
MSE(P^*) &= \text{Bias}(P^*)^2 + Var(P^*) = \left(\frac{1-2\pi}{n+2}\right)^2 + \frac{n\pi(1-\pi)}{(n+2)^2} \\
&= \frac{(1-2\pi)^2 + n\pi(1-\pi)}{(n+2)^2} = \frac{1-4\pi+4\pi^2+n\pi-n\pi^2}{(n+2)^2} \\
&= \frac{1+(n-4)\pi-(n-4)\pi^2}{(n+2)^2} = \frac{1+\pi(1-\pi)(n-4)}{(n+2)^2}
\end{aligned}$$

If we take limits, we'll see that both P and P^* are consistent, since their mean-squared errors go to zero as $n \to \infty$. For different values of π and n, however, the two estimators will have different MSE.

19. Problem 7-18 from Wonnacott & Wonnacott.

Solution: 7-18

The point of this question is non-response bias: the people who respond are not representative of the population as a whole. Note that P and P^* as defined in this question *do not correspond* to question 7-13. Our goal is to estimate the population proportion who will buy a computer. Using the table, we calculate the total number of people who will buy a computer as:

$$0.02 \times 40 + 0.04 \times 5 + 0.1 \times 3 + 0.2 \times 2 = 1.7 \text{ million}$$

which corresponds to a fraction $\pi^* = 1.7/50 = 0.034$. Again using the table, the number of people who will buy a computer *among the sub-population who would*

respond can be calculated as:

$$0.02 \times 7 + 0.04 \times 1 + 0.1 \times 1 + 0.2 \times 1 = 0.48 \text{ million}$$

which corresponds to a fraction $\pi = 0.48/10 = 0.048$. The point is that $\pi \neq \pi^*$. In other words, the proportion of people who would buy a computer *differs* across people who would and would not respond to the phone survey. The estimator P is based on calling 1000 people chosen at random and recording responses for *only those who reply*. The proportion of people who will reply is $10/50 = 1/5$. Thus, P will end up with a sample size of approximately $n = 200$ individuals. These individuals correspond to the sub-population for which the proportion who would buy a computer is $\pi^* = 0.048$. In contrast, the estimator P^* is based on calling $n^* = 100$ people chosen at random and then following up with these people repeatedly until *all of them respond*. Thus, P^* draws from the *full population*, in which a proportion $\pi^* = 0.034$ of people will buy a computer. The *true parameter* is π^* since we want to estimate the *overall* fraction of people who will buy a computer, *not* the fraction of people who would buy a computer among those who are likely to respond to a telephone survey. Hence bias is calculated *relative to* π^*. Variance is calculated *relative to the mean of each sampling distribution*. For P this mean is π while for P^* it is π^*. That is:

$$\begin{aligned}
MSE(P) &= \text{Bias}(P)^2 + Var(P) = (E[P] - \pi^*)^2 + E[(P-\pi)^2] \\
&= (\pi - \pi^*)^2 + E[(P-\pi)^2] \\
&= (\pi - \pi^*)^2 + \pi(1-\pi)/n \\
MSE(P^*) &= \text{Bias}(P^*)^2 + Var(P^*) = (\pi^* - \pi^*)^2 + E[(P^* - \pi^*)^2] \\
&= E[(P^* - \pi^*)^2] = \pi^*(1-\pi^*)/n^*
\end{aligned}$$

The estimator P^* does not have any bias because of the follow-ups to ensure that everyone in the original random sample responds. However, since it is based on a smaller sample, we would expect it to have a higher variance. The question is how this trade-off comes out in the expressions for MSE. To find out, we simply plug in the values $\pi^* = 0.034$, $\pi = 0.048$, $n = 200$ and $n^* = 100$. We find $MSE(P^*) \approx 0.000328$ and $MSE(P) \approx 0.000424$.

20. Problem 7-19 from Wonnacott & Wonnacott. (Note that the answer in the back of the book is *incorrect*.)

Solution: The book's solution erroneously takes $n = 100$ rather than $n = 200$ when calculating the variance of P. This is wrong because 1000 is the number of peo-

ple *called* not the number of people who *reply*. The question statement specifically states that P should be calculated for those who respond. The correct simply requires us to take the square root of the answers from the previous question. We find that $RMSE(P^*) \approx 0.018$ and $RMSE(P) \approx 0.021$. These are the root mean squared errors for estimators of the population *proportion*. To answer the question for estimators of *market size*, i.e. the population proportion multiplied by the size of the market (50 million), we simply multiply each of the RMSE values by 50 million yielding values of approximately 900,000 and 1,000,000 respectively.

Lecture #15 – Confidence Intervals I

21. In this question you will carry out a simulation exercise similar to the one I used to make the plot of twenty confidence intervals from lecture 16.

    (a) Write a function called `get_CI` that calculates a confidence interval for the mean of a normal population when the population standard deviation is known. It should take three arguments: `data` is a vector containing the observed data from which we will calculate the sample mean, `pop_sd` is the population standard deviation, and `alpha` controls the confidence level (e.g. `alpha = 0.1` for a 90% confidence interval). Your function should return a vector whose first element is the lower confidence limit and whose second element is the upper confidence limit. Test out your function on a simple example to make sure it's working properly.

    Solution:

```r
get_CI <- function(data, pop_sd, alpha){

  x_bar <- mean(data)
  n <- length(data)

  SE <- pop_sd / sqrt(n)
  ME <- qnorm(1 - alpha/2) * SE

  lower <- x_bar - ME
  upper <- x_bar + ME

  out <- c(lower, upper)
  return(out)
}
```

Testing this out on fake data containing twenty-five zeros and assuming a population variance of one, we have

```r
fake_data <- rep(0, 25)
get_CI(fake_data, pop_sd = 1, alpha = 0.05)
## [1] -0.3919928  0.3919928
```

If we calculated the corresponding interval by hand, assuming a population standard deviation of one, we'd get

$$0 \pm 2 \times 1 \times 1/5 = (-0.4, 0.4)$$

Which is almost exactly the same. The reason for the slight discrepancy is that when working by hand we use the approximation `qnorm(0.975)`≈ 2 whereas the exact value, which R provides, is more like 1.96.

(b) Write a function called `CI_sim` that takes a single argument `sample_size`. Your function should carry out the following steps. First generate `sample_size` draws from a standard normal distribution. Second, pass your sample of standard normals to `get_CI` with `alpha` set to 0.05 and `pop_sd` set to 1. Third, return the resulting confidence interval. Test your function on a sample of size 10. (What we're doing here is constructing a 95% confidence interval for the mean of a normal population using simulated data. The population mean is in fact zero, but we want to see how our confidence interval procedure works. To do this we "pretend" that we don't know the population mean and only know the population variance. Think about this carefully and make sure you understand the intuition.)

Solution:

```
CI_sim <- function(sample_size){
    sims <- rnorm(sample_size)
    CI <- get_CI(sims, pop_sd = 1, alpha = 0.05)
    return(CI)
}
CI_sim(10)
## [1] -0.8160878  0.4235023
```

(c) Use `replicate` to construct 10000 confidence intervals based on simulated data using the function `CI_sim` with `sample_size` equal to 10. (Note that `replicate` will, in this case, return a matrix with *2* rows and *10000* columns. Each column corresponds to one of the simulated confidence intervals. The first row contains the lower confidence limit while the second row contains the upper confidence limit.) Calculate the proportion of the resulting confidence intervals contain the true population mean. Did you get the answer you were expecting?

Solution:

```
simCIs <- replicate(10000, CI_sim(10))
simCIs[,1:5]
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] -0.6429955 -0.6474871 -0.3554235 -0.1951084 -0.3309972
## [2,]  0.5965946  0.5921029  0.8841666  1.0444817  0.9085928
lower <- simCIs[1,]
upper <- simCIs[2,]
covers_truth <- (lower < 0) & (upper > 0)
sum(covers_truth) / length(covers_truth)
## [1] 0.9478
```

The answer is pretty much dead on: almost exactly 95% of the intervals contain the true population mean (zero).

(d) Repeat the preceding but rather than using `CI_sim` write a new function called `CI_sim2`. This new function should be identical to `CI_sim` except that, when calling `get_CI`, it sets `pop_sd = 1/2` rather than 1. How do your results change? Try to provide some intuition for any differences you find.

Solution:

```
CI_sim2 <- function(sample_size){
    sims <- rnorm(sample_size)
    CI <- get_CI(sims, pop_sd = 1/2, alpha = 0.05)
    return(CI)
}
simCIs <- replicate(10000, CI_sim2(10))
simCIs[,1:5]
##           [,1]      [,2]       [,3]        [,4]      [,5]
## [1,] 0.4284204 -0.232400 -0.4313465 -0.64034061 -1.102286
## [2,] 1.0482155  0.387395  0.1884485 -0.02054558 -0.482491
lower <- simCIs[1,]
upper <- simCIs[2,]
covers_truth <- (lower < 0) & (upper > 0)
sum(covers_truth) / length(covers_truth)
## [1] 0.6701
```

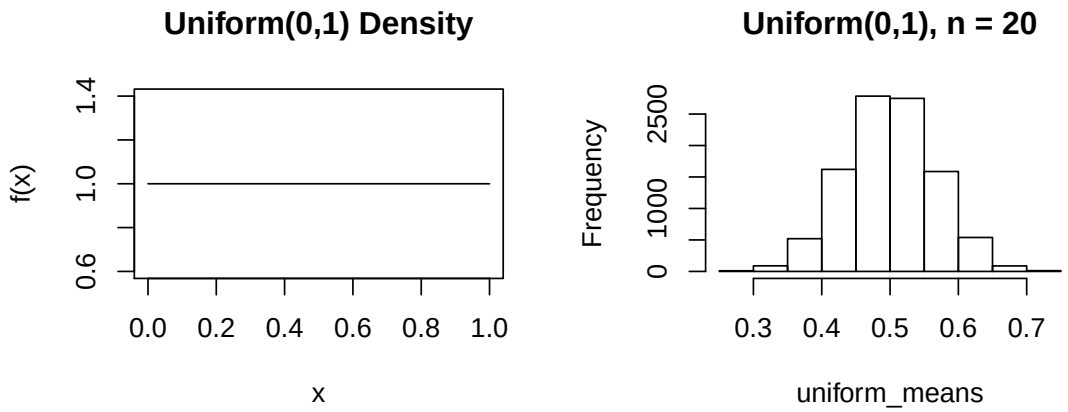
In this case the procedure didn't work: many fewer than 95% of the intervals contain the true population mean. The problem is that `CI.sim2` constructs a confidence interval using the *wrong* population standard deviation! Since it uses 1/2 rather than 1, the resulting intervals are too short, so too few of them contain the true population mean.

Lecture #16 – Confidence Intervals II

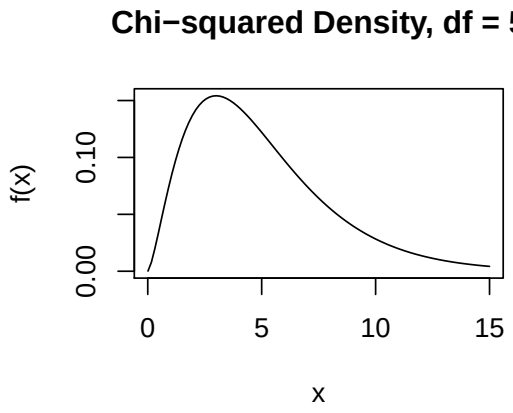
22. Write R code to carry out the simulation experiments presented on slides 13–15 of Lecture 16 illustrating the central limit theorem. In each case, plot population density (or mass function) and compare it to the histograms of the sample mean. Use a sample size of 20 and 10,000 simulation replications. The R command for making n draws from a $\chi^2(5)$ distribution is `rchisq(n, df = 5)` and the density is `dchisq(x, df = 5)`. You can use `curve` to plot both the uniform and $\chi^2(5)$ densities, but you'll need another approach to plot the Bernoulli pmf. I suggest that you create a vector `x` to represent the support set, and another called `p` to represent the pmf. You can then use the `plot` command with the option `type = 'h'` to plot vertical bars as in my slides. I also suggest setting `ylim = c(0,1)` so that the y-axis in this plot starts at zero and ends at one. You can also try setting `xlim` to make it easier to see the vertical bars.

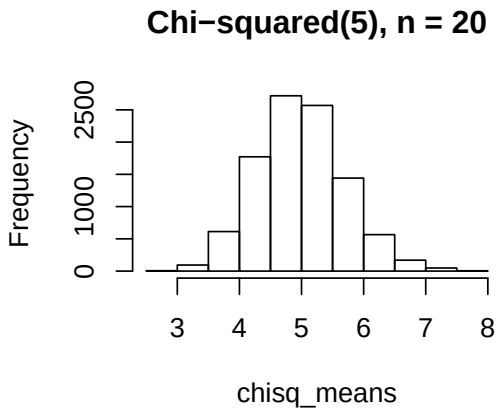
Solution:

```r
# Uniform
uniform_means <- replicate(10000, mean(runif(20)))
par(mfrow = c(1, 2))
curve(dunif(x), 0, 1, main = "Uniform(0,1) Density", ylab = 'f(x)')
hist(uniform_means, main = "Uniform(0,1), n = 20")
```

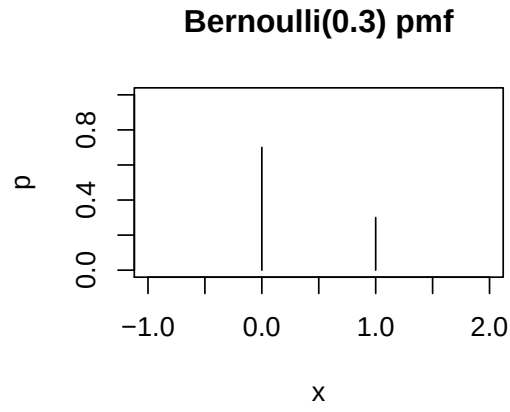


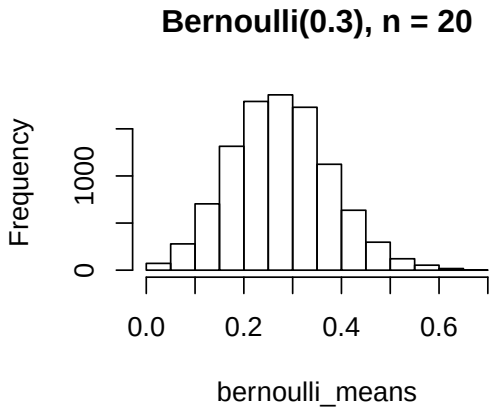

```r
par(mfrow = c(1, 1))

# Chi-squared(5)
chisq_means <- replicate(10000, mean(rchisq(20, df = 5)))
par(mfrow = c(1,2))
curve(dchisq(x, 5), 0.01, 15,
      main = "Chi-squared Density, df = 5", ylab = 'f(x)')
hist(chisq_means, main = "Chi-squared(5), n = 20")
```






```r
par(mfrow = c(1, 1))

# Bernoulli(0.3)
par(mfrow = c(1, 2))
x <- c(0,1)
p <- c(0.7, 0.3)
plot(x, p, type = 'h', main = "Bernoulli(0.3) pmf", ylim = c(0, 1),
     xlim = c(-1, 2))
bernoulli_means <- replicate(10000, mean(rbinom(20, 1, 0.3)))
hist(bernoulli_means, main = "Bernoulli(0.3), n = 20")
```






```r
par(mfrow = c(1, 1))
```

Lecture #17 – Confidence Intervals III

23. In April of 2013, Public Policy Polling carried out a survey of 1247 registered voters to determine whether Republicans and Democrats differ in their beliefs about various conspiracy theories. To answer this question, you'll need to download the full results of their survey which I've posted on my website for convenience: `http://www.ditraglia.com/econ103/conspiracy.pdf`. Note that this is a *pdf file* so you can't import it into R. You'll need to go read through the document to find the relevant data from the poll.

    (a) Construct a 99% confidence interval for the proportion of registered voters who belive that a UFO crashed at Roswell, New Mexico in 1947 and the US Government covered it up.

    Solution: Overall percentages appear on page 2 of the report, and this question refers to Q3. The sample size is 1247 and $\widehat{p} = 0.21$. We can carry out the calculations in R as follows:

```
p <- 0.21
n <- 1247
SE <- sqrt(p * (1 - p)/n)
ME <- qnorm(1 - 0.01/2) * SE
LCL <- p - ME
UCL <- p + ME
c(LCL, UCL)
## [1] 0.1802897 0.2397103
```

    (b) Is there evidence that male and female voters differ in their beliefs about Roswell and UFOs?

    Solution: Percentages broken down by sex appear on page 15, while overall percentages of men and women appear on page 3. Of the 1247 registered voters in the poll, about 50% were women and 50% were men. We'll call that $n = 623$ for each. The sample proportions are $\widehat{p}_W = 0.19$ for women versus $\widehat{p}_M = 0.24$ for men. Using R, we find:

```
n <- 623
p_M <- 0.24
p_W <- 0.19
SE <- sqrt(p_M * (1 - p_M)/n + p_W * (1 - p_W) / n)
ME <- qnorm(1 - 0.01/2) * SE
LCL <- (p_M - p_W) - ME
UCL <- (p_M - p_W) + ME
c(LCL, UCL)
## [1] -0.009846188  0.109846188
```

This 99% CI just barely includes zero. A 95% wouldn't (try this out for yourself). We have found evidence suggesting that a higher proportion of men believe in the Roswell conspiracy compared to women.

(c) Is there evidence that Romney voters differ from Obama voters in their beliefs about Roswell and UFOs?

Solution: Percentages broken down by 2012 vote appear in page 5. Overall percentages of Romney and Obama voters in the sample appear on page 3. Of the 1247 registered voters in the sample, 50% voted for Obama and 44% voted for Romney. We'll call this $n_O = 623$ and $n_R = 547$. The sample proportions are $\widehat{p}_O = 0.16$ for Obama voters versus $\widehat{p}_R = 0.27$ for Romney voters. Using R, we find:

```
n_R <- 547
p_R <- 0.27
n_O <- 623
p_O <- 0.16
SE <- sqrt(p_R * (1 - p_R) / n_R + p_O * (1 - p_O) / n_O)
ME <- qnorm(1 - 0.01/2) * SE
LCL <- (p_R - p_O) - ME
UCL <- (p_R - p_O) + ME
c(LCL, UCL)
## [1] 0.04817691 0.17182309
```

We have found strong evidence that a substantially greater proportion of Romney voters believe in the Roswell conspiracy.

(d) How should we interpret the results of the preceding two parts?

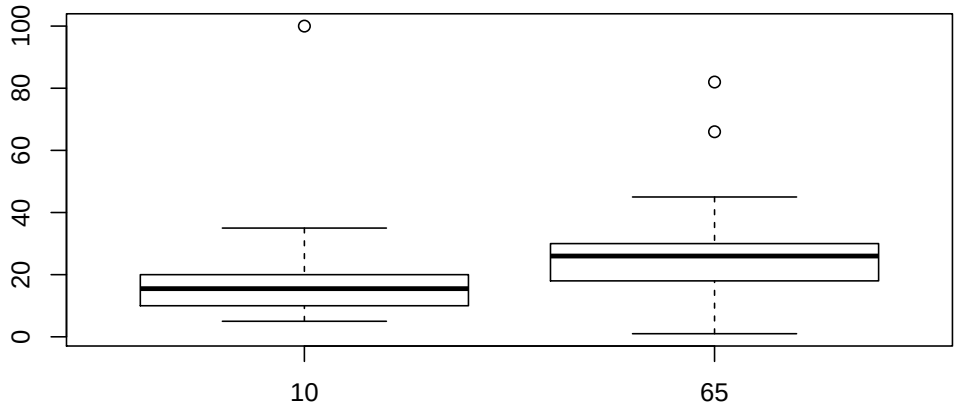
Solution: Since we know the men are more likely to vote for Republican candidates than women, it's difficult to tell whether the effect has to do with sex or political affiliation. To learn more, we'd need to compare *female* Romney voters to *female* Obama voters and then *separately* compare male Obama voters to male Romney voters.

24. In this question you will analyze data from the Spring 2019 anchoring experiment in Econ 103, contained in the columns `rand.num` and `africa.percent` of our class survey: `http://ditraglia.com/econ103/survey-spring-2019.csv`.

    (a) Make a boxplot of the results from the anchoring experiment and discuss your findings.

Solution:

```
data.url <- "http://ditraglia.com/econ103/survey-spring-2019.csv"
survey <- read.csv(data.url)
anchoring <- survey[,c("rand.num", "africa.percent")]
boxplot(africa.percent ~ rand.num, anchoring)
```



    (b) Construct an approximate 95% confidence interval for the magnitude of the anchoring effect, based on the CLT.

Solution:

```r
lo <- subset(anchoring, rand.num == "10")$africa.percent
hi <- subset(anchoring, rand.num == "65")$africa.percent
lo <- na.omit(lo)
hi <- na.omit(hi)
SE.hi <- sd(hi)/sqrt(length(hi))
SE <- sqrt(var(hi) / length(hi) + var(lo) / length(lo))
ME <- qnorm(0.975) * SE
LCL <- (mean(hi) - mean(lo)) - ME
UCL <- (mean(hi) - mean(lo)) + ME
c(LCL, UCL)
## [1]  0.4461888 12.7182195
```

(c) Do your results differ from those of the past semester, discussed in the lecture slides?

25. This question is based on a recent paper examining how "organic" labeling changes people's perceptions of different food products. Researchers recruited volunteers at a local mall in Ithaca, New York and gave each two samples of yogurt to taste. Although both yogurts were in fact identical, the volunteers were *told* that one of them was organic while the other was not. After tasting both, each volunteer was asked to estimate how many calories each of the samples of yogurt contained. (Since, unknown to the volunteer, both samples contained exactly the same kind of yogurt, each in fact contained the same number of calories.) To prevent confounding from anchoring or other behavioral effects, the order in which a given volunteer tasted the two yogurts, i.e. "organic" first or "organic" second, was chosen at random. The results of this experiment are stored in an R dataframe called `yogurt`. Here are the first few rows:

```
> head(yogurt)
  regular organic
1      60       40
2       5        0
3     200      100
4      60       40
5     100      100
6      90       90
```

Each row in this dataframe corresponds to a single individual's guess of the number of calories contained in each of the two yogurts. For example, the values 60 and 40 in row 1 mean that volunteer number one guessed that the regular yogurt sample contained 60 calories and the organic sample contained 40. Summary statistics for the two columns are as follows:

	`regular`	`organic`
Sample Mean	113	90
Sample Var	3600	2916
Sample SD	60	54
Sample Corr.	0.8	
Sample Size	115	

(a) Give the units of each of the summary statistics from above.

Solution: calories, calories^2, calories, unitless.

(b) Sara thinks that this experiment should be analyzed as independent samples data. Assume that she is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: The difference of means (regular minus organic) is 23 calories. Sara calculates her standard error assuming independent samples:

$$\sqrt{\sigma_X^2/n + \sigma_Y^2/m} = \sqrt{3600/115 + 2916/115} = \sqrt{6516/115} \approx 7.5$$

so her confidence interval is approximately 23 ± 15, in other words $(8, 38)$.

(c) Kevin thinks that this experiment should be analyzed as matched pairs data. Assume that he is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: Kevin takes into account the sample correlation between columns when calculating his standard error. He does this by using the sample statistics from the table to calculate the sample variance of the *differences*: regular minus organic. In particular, he calculates:

$$s_D^2 = 3600 + 2916 - 2 \cdot 0.8 \cdot 60 \cdot 54 = 1332$$

which gives a standard error of

$$\sqrt{s_D^2/n} = \sqrt{1332/115} \approx 3.4$$

This is the only difference between his procedure and Sara's. Hence, Kevin's confidence interval is approximately 23 ± 6.8, in other words $(16.2, 29.8)$.

(d) How do the confidence intervals constructed by Sara and Kevin differ? Explain the source of the discrepancy. Which of them has constructed the appropriate

confidence interval for this example?

Solution: Kevin is right and Sara is wrong. This is matched pairs data because each row corresponds to a *single individual*. Unsurprisingly, we find a high sample correlation between the two columns: individuals who overestimate caloric content for one yogurt sample tend to do so for the other, as do individuals who underestimate. The only difference between Kevin and Sara's confidence intervals comes from how they calculated their standard errors. Both intervals are correctly centered, but Sara's is *too wide* because she calculated the standard error assuming independence between the two samples. When the sample correlation is positive this results in an *overestimate* of the standard error.

(e) Using what you know about experiments, observational studies, and confidence intervals, what conclusions can we draw from this study?

Solution: It appears that merely labeling a product "organic" causes consumers to assume that this product contains fewer calories. Because this is a randomized experiment (randomly assigning labels to identical samples of yogurt and randomizing the order in which subjects tasted), we don't have to worry about confounding. It is less clear, however, whether this result would generalize to foods other than yogurt. Further, people from Ithaca New York who visit the mall and volunteer for a taste test may not be representative of US consumers as a whole. Ideally we would repeat this experiment using different subject pools and different foods to see how robust the result is.

No Extension Questions for Lecture #18

Lecture #19 – Hypothesis Testing II

Lecture #20 – Hypothesis Testing III

Lecture #21 – Hypothesis Testing IV

Lecture #22 – Regression II

26. Let Y and X be RVs. Find the constants β_0 and β_1 that solve

$$\min_{\beta_0,\beta_1} E[(Y - \beta_0 - \beta_1 X)^2]$$

Hint: For the purposes of this question you may assume that expectation and differentiation can be interchanged, i.e. that $\frac{\partial}{\partial Z}E[f(Z)] = E[\frac{\partial}{\partial Z}f(Z)]$.

Solution: Differentiating with respect to β_0 gives the first order condition

$$-2E[Y - \beta_0 - \beta_1 X] = 0$$

Re-arranging using the linearity of expectation, $\beta_0 = E[Y] - \beta_1 E[X]$ Substituting this expression for β_0 back into the objective function,

$$E[(Y - \mu_Y - \beta_1\mu_X - \beta_1 X)^2] = E\left[\{(Y - \mu_Y) - \beta_1(X - \mu_X)\}^2\right]$$

using the shorthand $E[Y] = \mu_Y$ and $E[X] = \mu_X$. Now, differentiating with respect to β_1 gives the first order condition

$$-2E\left[\{(Y - \mu_Y) - \beta_1(X - \mu_X)\}(X - \mu_X)\right] = 0$$

Finally, rearranging and solving for β_1 using the linearity of expectation, we have

$$\begin{aligned} E[(Y - \mu_Y)(X - \mu_X)] &= E[\beta_1(X - \mu_X)^2] \\ Cov(X, Y) &= \beta_1 Var(X) \\ \beta_1 &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

27. This example is based on 12-1 from WW4, but has been adapted somewhat for you to carry out in R. Suppose that the population regression line is $Y = 2.4 + 0.3X$, i.e. that the population regression parameters are $\beta_0 = 2.4$ and $\beta_1 = 0.3$. Normally we don't know these parameters but rather use data to estimate them. In this question, however, we will pretend that we know these parameters and carry out a Monte Carlo simulation to understand sampling variability in the context of regression.

    (a) Write an R function called `simulate_y` that takes as its input a vector `x` of X-values and returns the corresponding Y values from the above equation *plus a standard normal error term* ε.

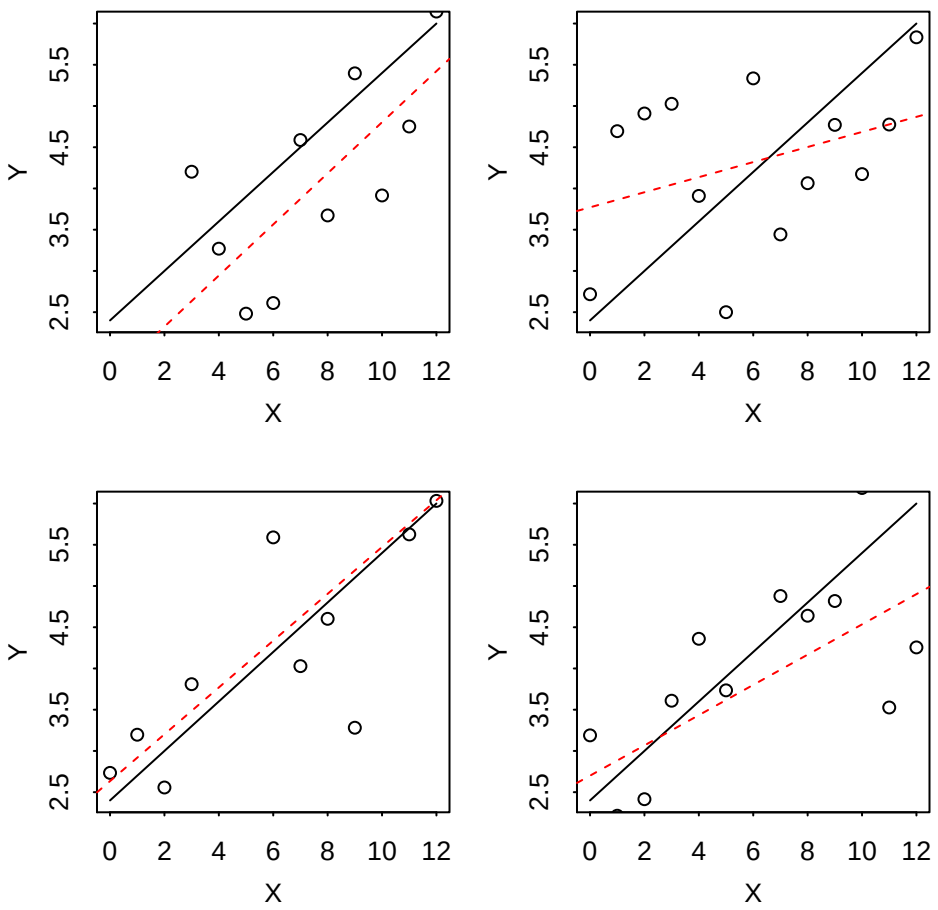
    Solution:

```r
simulate_y <- function(x){
  n <- length(x)
  epsilon <- rnorm(n)
  y <- 2.4 + 0.3 * x + epsilon
  return(y)
}
```

(b) Define `x_test <- 0:12`, a vector containing all the integers from 0 to 12. Test your function from part (a) by inputting `x_test` and assigning the result to `y_sim`. Make a plot of the function $Y = 2.40 + 0.30X$ along with the points `x_test` and `y_sim` and the *estimated* regression line $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ obtained by running a regression in which `x_test` is used to predict `y_sim`. Re-run your code a few times with different random seeds to see how the estimated regression line varies depending on the realizations of the error term. This illustrates *sampling variability* in the estimated regression slope and intercept.

Solution: To make it easier for me to repeat the code several times, I will encapsulate it into a function, repeat it using a loop, and arrange the resulting plots on a grid. I also change some other graphics parameters just to make things fit more easily into the answer key: you don't have to worry about this as long as you understand the code used to make the plot.

```r
set.seed(8372)
x_test <- 0:12
sim_plot <- function() {
  y_sim <- simulate_y(x_test)
  curve(2.4 + 0.3 * x, 0, 12, xlab = 'X', ylab = 'Y')
  points(x_test, y_sim)
  reg <- lm(y_sim ~ x_test)
  abline(coef(reg), col = 'red', lty = 2) # red, dashed line
}
# Make the plot four times with different random draws
par(mfrow = c(2,2), mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
  for(i in 1:4) sim_plot()
```




```
par(mfrow = c(1,1))
```

(c) Write a function called `get_slope` that carries out the following steps:

(i) Use `x_test` and `simulate_y` to generate a vector `y_sim` of simulated y-values from the regression model from above.

(ii) Run a regression using `x_test` to predict `y_sim` and store the result in an R object called `reg`.

(iii) Return the estimated regression slope coefficient from `reg`.

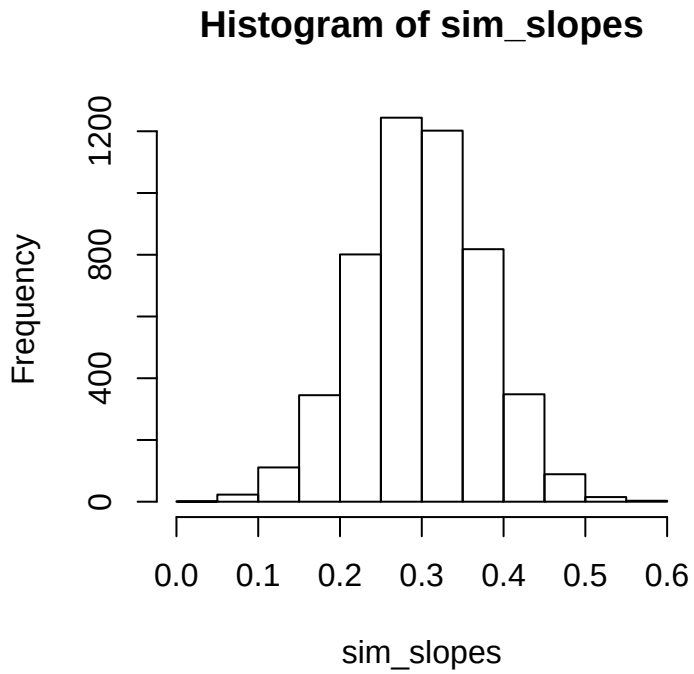
Solution:

```r
get_slope <- function() {
  y_sim <- simulate_y(x_test)
  reg <- lm(y_sim ~ x_test)
  slope <- coef(reg)[2]
  return(slope)
}
```

(d) Use your function `get_slope` and the R function `replicate` to approximate the sampling distribution of the sample regression estimator of β_1 using 5000 simulation draws. Construct a histogram of your results and calculate the approximate bias and standard error of the regression slope estimator. Discuss your findings.

Solution:

```r
sim_slopes <- replicate(5000, get_slope())
mean(sim_slopes)
## [1] 0.2993407
sd(sim_slopes)
## [1] 0.07582773
hist(sim_slopes)
```



We see that the sampling distribution of the estimated regression slope coefficient is centered at the population slope coefficient $\beta = 0.3$ and the sampling distribution is approximately normal. The standard error is around 0.075.

Lecture #23 – Regression III

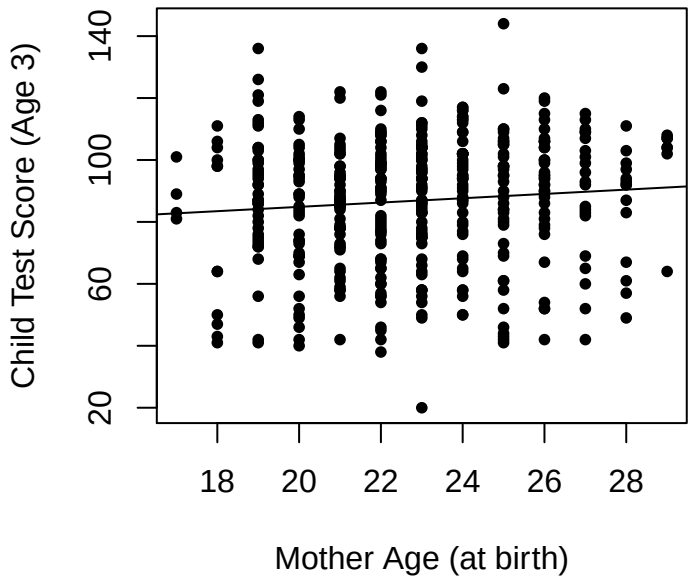
28. This question is based on the dataset on child test scores and mother characteristics we studied during our final lecture of the semester. The columns contained in this dataset are as follows:

Variable Name	Description
`kid.score`	Child's Test Score at Age 3
`mom.age`	Age of Mother at Birth of Child
`mom.hs`	Mother Completed High School? (1 = Yes)
`mom.iq`	Mother's IQ Score

(a) Run a regression of `kid.score` on `mom.age`. Plot both the data and the fitted regression line, making sure to label the axes. Interpret the results.

Solution:

```
source('http://ditraglia.com/econ103/display.R')
data_url <- "http://www.ditraglia.com/econ103/child_test_data.csv"
child <- read.csv(data_url)
reg1 <- lm(kid.score ~ mom.age, child)
display(reg1)
## lm(formula = kid.score ~ mom.age, data = child)
##             coef.est coef.se
## (Intercept) 70.96     8.31
## mom.age      0.70     0.36
## ---
## n = 434, k = 2
## residual sd = 20.35, R-Squared = 0.01
plot(child$mom.age, child$kid.score, pch = 20,
     xlab = 'Mother Age (at birth)', ylab = 'Child Test Score (Age 3)')
abline(coef(reg1))
```

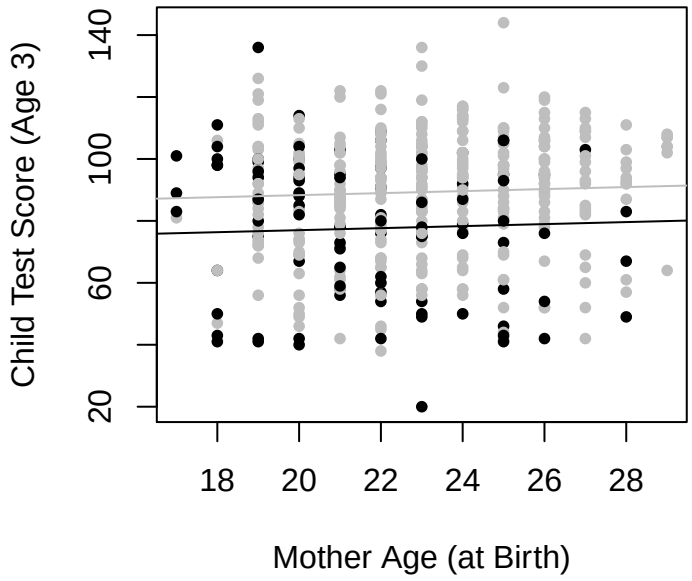


Our model suggests that the children of mothers who were older when they gave birth tend to score higher. In particular, comparing two children whose mothers' age at birth differed by one year, we would predict that the child of the older mother will score, on average, 0.7 points higher. The standard error associated with the estimate, however, is fairly large. An approximate 95% CI would just barely include zero. Nevertheless, this result is suggestive that the children of older mothers do better on the test. A naive reading of these results would be that women should wait to have children until they are as old as possible. The regression results, however, most emphatically do *not* establish this. There are many possible confounders here: for example, teenage pregnancy is correlated with economic disadvantage and lower levels of education.

(b) Augment your model from part (a) by allowing a different intercept for children whose mother completed high school. Plot the data along with the regression lines for each group (those whose mother completed high school and those whose mother did not). Interpret your results and compare them to those you got in part (a).

Solution:

```
reg2 <- lm(kid.score ~ mom.hs + mom.age, child)
display(reg2)
## lm(formula = kid.score ~ mom.hs + mom.age, data = child)
##             coef.est coef.se
## (Intercept) 70.48     8.11
## mom.hs      11.31     2.38
## mom.age      0.33     0.36
## ---
## n = 434, k = 3
## residual sd = 19.86, R-Squared = 0.06
b_both <- coef(reg2)[3]
a_HS <- coef(reg2)[1] + coef(reg2)[2]
a_no_HS <- coef(reg2)[1]
colors <- ifelse(child$mom.hs == 1, 'gray', 'black')
plot(child$mom.age, child$kid.score, pch = 20, col = colors,
     xlab = 'Mother Age (at Birth)', ylab = 'Child Test Score (Age 3)')
abline(a = a_HS, b = b_both, col = 'gray')
abline(a = a_no_HS, b = b_both, col = 'black')
```



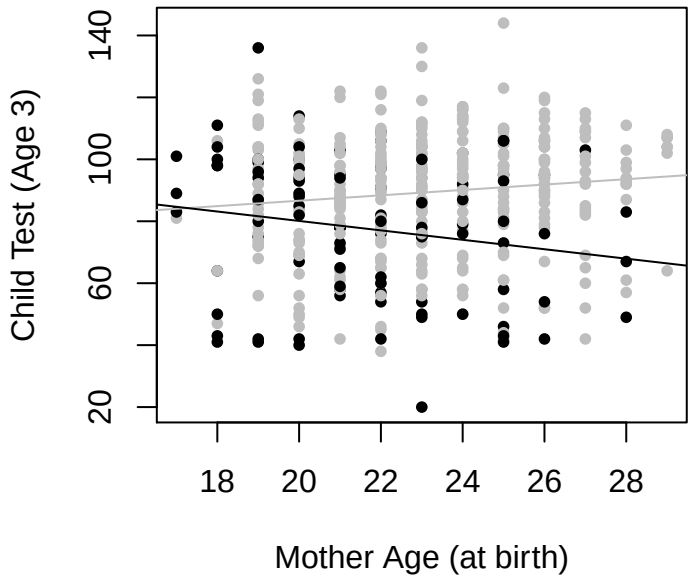
By adding a dummy variable that equals one if a child's mother completed high school, we have controlled for one of the possible confounders from above:

mother's level of education. We have done this by allowing the regression line to have a different intercept depending on mother's education. Comparing two children whose mothers are of the same age but only one whom attended high school, we predict that the child of the better educated mother will score, on average, 11 points higher. The standard error associated with this estimate is quite small, yielding a 95% CI that is is nowhere near zero. We have strong evidence of a large effect from mother's education level. In contrast, once we've controlled from mother's education, the estimated effect of `mom.age` falls substantially while the associated standard error stays the same. This results in an approximate 95% CI that includes many negative values. After controlling for mother's education, there is much less evidence to suggest that older mothers have higher-scoring children. In terms of predictive accuracy, the second model is slightly better but neither is particularly effective: we are only predicting test scores to an accuracy of about 20 points.

(c) Now allow different slopes as well as intercepts for each group (those whose mother completed high school and those whose mother did not). Plot the data and the regression lines for each group and interpret your results.

Solution:

```
reg3 <- lm(kid.score ~ mom.hs + mom.age + mom.hs:mom.age, child)
display(reg3)
## lm(formula = kid.score ~ mom.hs + mom.age + mom.hs:mom.age, data = child)
##                coef.est coef.se
## (Intercept)    110.54    16.45
## mom.hs         -41.29    18.99
## mom.age         -1.52     0.75
## mom.hs:mom.age   2.39     0.86
## ---
## n = 434, k = 4
## residual sd = 19.70, R-Squared = 0.07
a_no_HS <- coef(reg3)[1]
a_HS <- coef(reg3)[1] + coef(reg3)[2]
b_no_HS <- coef(reg3)[3]
b_HS <- coef(reg3)[3] + coef(reg3)[4]
plot(child$mom.age, child$kid.score, pch = 20, col = colors,
     xlab = 'Mother Age (at birth)', ylab = 'Child Test (Age 3)')
abline(a = a_HS, b = b_HS, col = 'gray')
abline(a = a_no_HS, b = b_no_HS, col = 'black')
```



This is very interesting! When we allow for different slopes as well as intercepts,

by adding an *interaction* between `mom.hs` and `mom.hs`, namely `mom.hs:mom.age`, we find very different results depending on mother's education. (There is strong evidence that we should allow for different slopes, since the approximate 95% CI for the interaction does not include zero.) For children whose mothers attended high school, there is a *positive* relationship between mother's age at birth and child's test score. For children whose mothers did not attend high school, the relationship is *negative*. For children whose mothers were 18 then they gave birth, there is essentially *no* impact from mother's education level. As age of mother at birth increases, the impact of mother's education widens.