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?

   (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.

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

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.

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

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

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

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

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.

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

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

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

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.

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?

   (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

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

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

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?

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

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

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

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

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

   (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.

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.

   (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.

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*.

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?

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 $p_{XY}(x, y) = p_X(x)p_Y(y)$.

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.

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.
    (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*.

(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`.

(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.

(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.

(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.)

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.

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 $E[Z]$

    (b) Calculate $Var(Z)$.

    (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?

    (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?

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.

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

    (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

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

    (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.

    (e) Back to Garth's problem: *amazingly* it's still possible to construct an unbiased estimator of $\mu_X - \mu_Y$ *without knowning* 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 $E[\alpha\bar{Z}_1 + \beta\bar{Z}_2] = \mu_X - \mu_Y$.

    (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.

Lecture #14 – Sampling and Estimation II

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

19. Problem 7-18 from Wonnacott & Wonnacott.

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

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.

    (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.)

    (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 correponds 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?

(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.

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.

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.
    (b) Is there evidence that male and female voters differ in their beliefs about Roswell and UFOs?
    (c) Is there evidence that Romney voters differ from Obama voters in their beliefs about Roswell and UFOs?
    (d) How should we interpret the results of the preceding two parts?

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.

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

    (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.

(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.

(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.

(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?

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

No Extension Questions for Lecture #18

No Extension Questions for Lecture #19

Lecture #20 – Hypothesis Testing III

26. This question revisits the data from this semester's anchoring experiment. In a previous question you used these data to construct confidence intervals. In this question you will carry out hypothesis tests. You may assume throughout that the sample size is large enough to use the Central Limit Theorem. Details of how to load the data from my website appear in Lecture 19. Be sure to properly account for missing values.

    (a) Suppose we want to test the null hypothesis of equality of population means across the two groups. What is the value of our test statistic?

    (b) Suppose we want to test the equality of population means against the one-sided alternative that the the "Hi" group has a higher mean at the 10% level. What is our critical value, and what is our decision rule? Do we reject the null hypothesis?

    (c) Calculate the p-value for a test of the equality of population means against the one-sided alternative that the "Hi" group has a higher mean.

    (d) Suppose we wanted to test the equality of population means against the two-sided alternative at the 10% level. What is our critical value, and what is our decision rule? Do we reject the null hypothesis?

    (e) Calculate the p-value for a test of the equality of population means against the two-sided alternative.

Lecture #21 – Hypothesis Testing IV

27. 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`
In an earlier extension question you used these data to construct confidence intervals. In this question you'll use them to carry out hypothesis tests. Throughout you may assume that the sample size is large enough for the approximate based on the central limit theorem to be valid.

    (a) Suppose we wanted to test the null hypothesis that 20% of registered voters believe that a UFO crashed at Roswell, New Mexico in 1947 and the US Government covered it up. Calculate our test statistic.

    (b) Suppose that we wanted to test the null hypothesis from the preceding part against the one-sided alternative that *more than* 20% of registered voters believe in the UFO conspiracy. Calculate the p-value for this test.

    (c) Repeat the preceding part for the *two-sided* alternative.

    (d) Calculate the p-value for a test of the null hypothesis that equal proportions of Romney and Obama voters believe in the UFO conspiracy against the two-sided alternative.

28. This question is based on a dataset containing the results of the tae kwon do event in the 2004 Athens Olympics. (In case this event is unfamiliar to you, my dictionary defines tae kwon do as "a modern Korean martial art similar to karate.") The competition is a tournament consisting of a number of bouts. In each bout, a pair of competitors fight each other, points are awarded, and a winner is declared by the judges. In accordance with Olympic regulations, one of the competitors in each bout is *randomly chosen* to wear blue body protectors. The other wears red body protectors. This question investigates whether wearing one color or the other gives an advantage in the competition. The data are stored in an R data frame called `taekwondo`. Each row corresponds to a *single bout* in the competition. The columns are as follows:

`class`	weight class of the bout
`red.id`	competitor id number for the fighter who wore red
`blue.id`	competitor id number for the fighter who wore blue
`round`	round of the tournament (i.e. semifinals, finals, etc.)
`winner`	color worn by the fighter who won the bout
`method`	method of win (i.e. points, knockout, etc.)
`red.points`	number of points awarded to the fighter who wore red
`blue.points`	number of points awarded to the fighter who wore blue

(a) We'll restrict attention to the "last 16" round of the competition. This ensures that each row contains a *unique* pair of fighters. Write R code to extract only those rows of `taekwondo` for which the value in the column `round` is "last 16" and store the result in a data frame called `last16`.

(b) To begin, we'll analyze the *proportion* of bouts won by the blue fighter. Write R code to: (i) count the number of elements in the column `winner` of `last16` and store the result in a variable called `n`, and (ii) count the number of bouts won by the blue fighter and store the result in a variable called `n_blue`.

As it happens there are 32 bouts in `last16`, 8 bouts for each weight class times 4 weight classes, of which 19 were won by the blue fighter. Using this information, test the null hypothesis that the population proportion of bouts won by fighters wearing blue equals 0.5 against the two-sided alternative. Approximately what is the p-value for this test? Interpret your results.

(c) (6 points) For the remainder of the question, we will examine the relative difference in the number of *points* scored by the blue and red fighters in each bout. Write R code accomplish the following: (i) select only those rows of `last16` for which the value in the column `method` is `Points` and store the result in a data frame called `last16.points`, (ii) create a vector called `D` whose entries contain the *difference* in the number of points scored by blue versus red (Blue - Red) in each bout.

(d) I calculated the mean of the column `red.points` in `last16.points` and got 10.1. Similarly, I calculated the mean of the column `blue.points` and got 11.7. If I were to run the command `mean(D)` at the R console what result would I get?

(e) I entered the command `var(D)` at the R console and got 25. Next I entered `var(last16_points$red_points)` and `var(last16_points$blue_points)` and got 17 and 31, respectively. Calculate the sample correlation between the columns `red_points` and `blue_points` of the data frame `last16_points`.

(f) (10 points) To test the null hypothesis that red and blue fighters are awarded, on average, the same number of points against the two-sided alternative, should we use a test for independent samples or matched pairs data? Explain briefly and then carry out the appropriate test at the 5% level based on the CLT. To answer, you

will need the fact that there are 29 rows in the data frame `last16.points`. Be sure to report: (i) the test statistic, (ii) the decision rule, and (iii) the result of the test.

Lecture #22 – Regression II

29. 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 $\frac{\partial}{\partial Z} E[f(Z)] = E[\frac{\partial}{\partial Z} f(Z)]$.

30. 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.

    (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.

    (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`.

    (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.

Lecture #23 – Regression III

31. 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.

(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).

(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.