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:

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

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