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:

$$\begin{array}{ll} \text{W(BBBW)} \to +1 & \text{BWBW(B)} \to 0 \\ \text{W(BBWB)} \to +1 & \text{BBWW(B)} \to 0 \\ \text{W(BWBB)} \to +1 & \text{BBBWW()} \to -1 \\ \text{W(WBBB)} \to +1 & \text{BBWBW()} \to -1 \\ \text{BWW(BB)} \to +1 & \text{BWBBW()} \to -1 \end{array}$$

so the expected value is $(1+1+1+1+1-1-1-1-1)/10 = 1/5$. 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 0.90000 89000
[2,] 2 0.99000 97000
[3,] 3 0.99900 96900
[4,] 4 0.99990 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