# Review Questions

Econ 103

Spring 2019

## About This Document

These questions are the "bread and butter" of Econ 103: they cover the basic knowledge that you will need to acquire this semester to pass the course. There are between 10 and 15 questions for each lecture. After a given lecture, and before the next one, you should solve all of the associated review questions. To give you an incentive to keep up with the course material, all quiz questions for the course will be randomly selected from this list. For example Quiz #1, which covers lectures 1–2, will consist of one question drawn at random from questions 1–10 and another drawn at random from questions 12–24 below. We will not circulate solutions to review questions. Compiling your own solutions is an important part of studying for the course. We will be happy to discuss any of the review questions with you in office hours or on Piazza, and you are most welcome to discuss them with your fellow classmates. Be warned, however, that merely memorizing answers written by a classmate is a risky strategy. It may get you through the quiz, but will leave you woefully unprepared for the exams. There is no curve in this course: to pass the exams you will have to learn the material covered in these questions. Rote memorization will not suffice.

## Lecture #1 – Introduction

1. Define the following terms and give a simple example: *population*, *sample*, *sample size*.
2. Explain the distinction between a *parameter* and a *statistic*.
3. Briefly compare and contrast *sampling* and *non-sampling* error.
4. Define a *simple random sample*. Does it help us to address sampling error, non-sampling error, both, or neither?
5. A drive-time radio show frequently holds call-in polls during the evening rush hour. Do you expect that results based on such a poll will be biased? Why?

**Solution:** They will likely be biased. People who are listening to the radio during rush hour are disproportionately likely to be commuters driving home from work. People who are employed and drive to work are not representative of the population at large.

6. Dylan polled a random sample of 100 college students. In total 20 of them said that they approved of President Trump. Calculate the margin of error for this poll.

**Solution:** $2\sqrt{P(1-P)/n} = 2\sqrt{0.2 \times 0.8/100} = 0.08$

7. Define the term *confounder* and give an example.

8. What is a randomized, double-blind experiment? In what sense is it a "gold standard?"

9. Indicate whether each of the following involves experimental or observational data.

   (a) A biologist examines fish in a river to determine the proportion that show signs of disease due to pollutants poured into the river upstream.

   **Solution:** Observational

   (b) In a pilot phase of a fund-raising campaign, a university randomly contacts half of a group of alumni by phone and the other half by a personal letter to determine which method results in higher contributions.

   **Solution:** Experimental

   (c) To analyze possible problems from the by-products of gas combustion, people with with respiratory problems are matched by age and sex to people without respiratory problems and then asked whether or not they cook on a gas stove.

   **Solution:** Observational

   (d) An industrial pump manufacturer monitors warranty claims and surveys customers to assess the failure rate of its pumps.

   **Solution:** Observational

10. Based on information from an observational dataset, Amy finds that students who attend an SAT prep class score, on average, 100 points better on the exam than students who do not. In this example, what would be required for a variable to *confound* the relationship between SAT prep classes and exam performance? What are some possible confounders?

**Solution:**

# Lecture #2 – Summary Statistics I

11. For each variable indicate whether it is nominal, ordinal, or numeric.
    (a) Grade of meat: prime, choice, good.

    **Solution:** ordinal

    (b) Type of house: split-level, ranch, colonial, other.

    **Solution:** nominal

    (c) Income

    **Solution:** numeric

12. Explain the difference between a histogram and a barchart.
13. Define *oversmoothing* and *undersmoothing*.
14. What is an *outlier*?
15. Write down the formula for the sample mean. What does it measure? Compare and contrast it with the sample median.
16. Two hundred students took Dr. Evil's final exam. The third quartile of exam scores was 85. Approximately how many students scored *no higher* than 85 on the exam?
17. Define *range* and *interquartile range*. What do they measure and how do they differ?
18. What is a boxplot? What information does it depict?

19. Write down the formula for variance and standard deviation. What do these measure? How do they differ?

20. Suppose that $x_i$ is measured in inches. What are the units of the following quantities?

    (a) Sample mean of $x$

    **Solution:** inches

    (b) Range of $x$

    **Solution:** inches

    (c) Interquartile Range of $x$

    **Solution:** inches

    (d) Variance of $x$

    **Solution:** square inches

    (e) Standard deviation of $x$

    **Solution:** inches

21. Evaluate the following sums:

    (a) $\sum_{n=1}^{3} n^2$

    **Solution:** $\sum_{n=1}^{3} n^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$

    (b) $\sum_{n=1}^{3} 2^n$

    **Solution:** $\sum_{n=1}^{3} 2^n = 2^1 + 2^2 + 2^3 = 2 + 4 + 8 = 14$

    (c) $\sum_{n=1}^{3} x^n$

**Solution:** $\sum_{n=1}^{3} x^n = x + x^2 + x^3$

22. Evaluate the following sums:

(a) $\sum_{k=0}^{2}(2k+1)$

**Solution:** $\sum_{k=0}^{2}(2k+1) = (2\times 0+1)+(2\times 1+1)+(2\times 2+1) = 9$

(b) $\sum_{k=0}^{3}(2k+1)$

**Solution:** $\sum_{k=0}^{3}(2k+1) = \left[\sum_{k=0}^{2}(2k+1)\right] + (2\times 3+1) = 9+7 = 16$

(c) $\sum_{k=0}^{4}(2k+1)$

**Solution:** $\sum_{k=0}^{4}(2k+1) = \left[\sum_{k=0}^{3}(2k+1)\right] + (2\times 4+1) = 16+9 = 25$

23. Evaluate the following sums:

(a) $\sum_{i=1}^{3}(i^2+i)$

**Solution:** $\sum_{i=1}^{3}(i^2+i) = (1^2+1)+(2^2+2)+(3^2+3) = 20$

(b) $\sum_{n=-2}^{2}(n^2-4)$

**Solution:** $\sum_{n=-2}^{2}(n^2-4) = \left[(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2\right]-(4\times 5) = -10$

(c) $\sum_{n=100}^{102} n$

**Solution:** $\sum_{n=100}^{102} n = 100 + 101 + 102 = 303$

(d) $\sum_{n=0}^{2}(n + 100)$

**Solution:** $\sum_{n=0}^{2}(n + 100) = (0 + 1 + 2) + 3 \times 100 = 303$

24. Express each of the following using $\Sigma$ notation:

(a) $z_1 + z_2 + \cdots + z_{23}$

**Solution:** $\sum_{i=1}^{23} z_i$

(b) $x_1y_1 + x_2y_2 + \cdots + x_8y_8$

**Solution:** $\sum_{i=1}^{8} x_iy_i$

(c) $(x_1 - y_1) + (x_2 - y_2) + \cdots + (x_m - y_m)$

**Solution:** $\sum_{i=1}^{m}(x_i - y_i)$

(d) $x_1^3f_1 + x_2^3f_2 + \cdots + x_9^3f_9$

**Solution:** $\sum_{i=1}^{9} x_i^3f_i$

# Lecture #3 – Summary Statistics II

25. Show that $\sum_{i=m}^{n}(a_i + b_i) = \sum_{i=m}^{n} a_i + \sum_{i=m}^{n} b_i$. Explain your reasoning.

26. Show that if $c$ is a constant then $\sum_{i=m}^{n} cx_i = c\sum_{i=m}^{n} x_i$. Explain your reasoning.

27. Show that if $c$ is a constant then $\sum_{i=1}^{n} c = cn$. Explain your reasoning.

28. Mark each of the following statements as True or False. You do not need to show your work if this question appears on a quiz, although you should make sure you understand the reasoning behind each of your answers.

    (a) $\sum_{i=1}^{n}(x_i/n) = \left(\sum_{i=1}^{n} x_i\right)/n$

    (b) $\sum_{k=1}^{n} x_k z_k = z_k \sum_{k=1}^{n} x_k$

    (c) $\sum_{k=1}^{m} x_k y_k = \left(\sum_{k=1}^{m} x_k\right)\left(\sum_{k=1}^{m} y_k\right)$

    (d) $\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{m} y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j$

    (e) $\left(\sum_{i=1}^{n} x_i\right) / \left(\sum_{i=1}^{n} z_i\right) = \sum_{i=1}^{n}(x_i/z_i)$

29. Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Justify all of the steps you use.

30. Re-write the formula for skewness in terms of the z-scores $z_i = (x_i - \bar{x})/s$. Use this to explain the original formula: why does it involve a cubic and why does it divide by $s^3$?

**Solution:**

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3$$

31. How do we interpret the sign of skewness, and what is the "rule of thumb" that relates skewness, the mean, and median?

32. What is the distinction between $\mu, \sigma^2, \sigma$ and $\bar{x}, s^2, s$? Which corresponds to which?

33. What is the empirical rule?

34. Define *centering*, *standardizing*, and *z-score*.

35. What is the sample mean $\bar{z}$ of the z-scores $z_1, \ldots, z_n$? Prove your answer.

36. What is the sample variance $s_z^2$ of the z-scores $z_1, \ldots, z_n$? Prove your answer.

37. Suppose that $-c < (a - x)/b < c$ where $b > 0$. Find a lower bound $L$ and an upper bound $U$ such that $L < x < U$.

**Solution:** Rearranging,

$$-bc - a < -x < bc - a$$

and multiplying through by $-1$,

$$a - bc < x < a + bc$$

38. Compare and contrast *covariance* and *correlation*. Provide the formula for each, explain the units, the interpretation, etc.

39. Suppose that $x_i$ is measured in centimeters and $y_i$ is measured in feet. What are the units of the following quantities?

    (a) Covariance between $x$ and $y$

    **Solution:** centimeters $\times$ feet

    (b) Correlation between $x$ and $y$

    **Solution:** unitless

    (c) Skewness of $x$

    **Solution:** unitless

    (d) $(x_i - \bar{x})/s_x$

    **Solution:** unitless

# Lecture #4 – Regression I

40. In a regression using height (measured in inches) to predict handspan (measured in centimeters) we obtained $a = 5$ and $b = 0.2$.
    (a) What are the units of $a$?
    (b) What are the units of $b$?
    (c) What handspan would we predict for someone who is 6 feet tall?

41. Plot the following dataset and calculate the corresponding regression slope and intercept *without* using the regression formulas.

| $x$ | $y$ |
|---|---|
| 0 | 2 |
| 1 | 1 |
| 1 | 2 |

42. Write down the optimization problem that linear regression solves.

43. Prove that the regression line goes through the means of the data.

44. By substituting $a = \bar{y} - b\bar{x}$ into the linear regression objective function, derive the formula for $b$.

45. Consider the regression $\widehat{y} = a + bx$.
    (a) Express $b$ in terms of the sample covariance between $x$ and $y$.
    (b) Express the sample correlation between $x$ and $y$ in terms of $b$.

46. What value of $a$ minimizes $\sum_{i=1}^{n}(y_i - a)^2$? Prove your answer.

47. Suppose that $s_{xy} = 30$, $s_x = 10$, $s_y = 6$, $\bar{y} = 12$, and $\bar{x} = 4$. Calculate $a$ and $b$ in the regression $\widehat{y} = a + bx$.

**Solution:**

$$b = s_{xy}/s_x^2 = 30/10^2 = 30/100 = 0.3$$
$$a = \bar{y} - b\bar{x} = 12 - 0.3 \times 4 = 12 - 1.2 = 10.8$$

48. Suppose that $s_{xy} = 30$, $s_x = 10$, $s_y = 6$, $\bar{y} = 12$, and $\bar{x} = 4$. Calculate $c$ and $d$ in the regression $\widehat{x} = c + dy$. Note: we are using $y$ to predict $x$ in this regression!

**Solution:**

$$b = s_{xy}/s_y^2 = 30/6^2 = 30/36 = 5/6 \approx 0.83$$
$$a = \bar{y} - b\bar{x} = 12 - 5/6 \times 4 = 12 - 10/3 = 26/3 \approx 8.7$$

49. A large number of students took two midterm exams. The standard deviation of scores on midterm #1 was 16 points, while the standard deviation of scores midterm #2 was 17 points. The covariance of the scores on the two exams was 124 points squared. Linus scored 60 points on midterm #1 while Lucy scored 80 points. How much higher would we predict that Lucy's score on the midterm #2 will be?

50. Suppose that the correlation between scores on midterm #1 and midterm #2 in Econ 103 is approximately 0.5. If the regression slope when using scores on midterm #1 to predict those on midterm #2 is approximately 1.5, which exam had the larger *spread* in scores? How much larger?

# Lecture #5 – Basic Probability I

51. What is the definition of probability that we will adopt in Econ 103?

52. Define the following terms:
    (a) *random experiment*
    (b) *basic outcomes*
    (c) *sample space*
    (d) *event*

53. Define the following terms and give an example of each:
    (a) *mutually exclusive events*
    (b) *collectively exhaustive events*

54. Suppose that $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 3\}$, $B = \{3, 4, 6\}$, and $C = \{1, 5\}$.
    (a) What is $A^c$?

(b) What is $A \cup B$?

(c) What is $A \cap B$?

(d) What is $A \cap C$?

(e) Are $A, B, C$ mutually exclusive? Are they collectively exhaustive?

55. A family has three children. Let $A$ be the event that they have less than two girls and $B$ be the event that they have exactly two girls.

(a) List all of the basic outcomes in $A$.

(b) List all of the basic outcomes in $B$.

(c) List all of the basic outcomes in $A \cap B$

(d) List all of the basic outcomes in $A \cup B$.

(e) If male and female births are equally likely, what is the probability of $A$?

56. Let $B = A^c$. Are $A$ and $B$ mutually exclusive? Are they collectively exhaustive? Why?

57. State each of the three axioms of probability, aka the *Kolmogorov Axioms*.

58. Suppose we carry out a random experiment that consists of flipping a fair coin twice.

(a) List all the basic outcomes in the sample space.

**Solution:** $S = \{HH, HT, TT, TH\}$

(b) Let $A$ be the event that you get at least one head. List all the basic outcomes in $A$.

**Solution:** $A = \{HH, HT, TH\}$

(c) List all the basic outcomes in $A^c$.

**Solution:** $A^c = \{TT\}$

(d) What is the probability of $A$? What is the probability of $A^c$?

**Solution:** $P(A) = 3/4 = 0.75$ and $P(A^c) = 1/4$

59. Calculate the following:

(a) 5!

**Solution:** 120

(b) $\dfrac{100!}{98!}$

**Solution:** 9900

(c) $\binom{5}{3}$

**Solution:** 10

60. (a) How many different ways can we choose a President and Secretary from a group of 4 people if the two offices must be held by different people?

    (b) How many different committees with two members can we form a group of 4 people, assuming that the order in which we choose people for the committee doesn't matter.

61. Suppose that I flip a fair coin 5 times.

    (a) How many basic outcomes contain exactly two heads?

    (b) How many basic outcomes contain exactly three tails?

    (c) How many basic outcomes contain exactly one heads?

    (d) How many basic outcomes contain exactly four tails?

62. Explain why $\binom{n}{r} = \binom{n}{n-r}$.

63. Suppose that I choose two distinct numbers at random from the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. What is the probability that both are odd?

**Solution:** This solution assumes that order doesn't matter. You could also assume that it does matter and get the same answer. There are $\binom{9}{2} = 36$ equally likely ways to choose 2 items from a set of 9. Of these, there are $\binom{5}{2} = 10$ ways to choose 2 of the 5 odd numbers. Hence the probability is $10/36 = 5/18$.

# Lecture #6 – Basic Probability II

64. State and prove the *complement rule*.

65. State the *multiplication rule*, and compare it to the definition of conditional probability.

66. Mark each statement as TRUE or FALSE. If FALSE, give a one sentence explanation.

    (a) If $A \subseteq B$ then $P(A) \geq P(B)$.

    **Solution:** FALSE: this is the logical consequence rule with the inequality sign going in the *wrong direction*.

    (b) For any events $A$ and $B$, $P(A \cap B) = P(A)P(B)$.

    **Solution:** FALSE: this only holds if $A$ and $B$ are independent.

    (c) For any events $A$ and $B$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

    **Solution:** TRUE: this is the addition rule.

67. Suppose that $P(B) = 0.4$, $P(A|B) = 0.1$ and $P(A|B^c) = 0.9$.

    (a) Calculate $P(A)$.

    **Solution:** By the law of total probability,

    $$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = 0.1 \times 0.4 + 0.9 \times 0.6 = 0.58$$

    (b) Calculate $P(B|A)$.

    **Solution:** By Bayes' rule,

    $$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.1 \times 0.4}{0.58} = 2/29 \approx 0.07$$

68. Define statistical independence. How is it related to conditional probability, and what does it mean intuitively?

69. State and prove the law of total probability for $k = 2$.

70. Find the probability of getting *at least* one six if you roll a fair, six-sided die three times.

**Solution:** Using the complement rule:

$$P(\text{At Least One Six}) = 1 - P(\text{No Sixes})$$

And by independence:

$$P(\text{No Sixes}) = 5/6 \times 5/6 \times 5/6 = 125/216$$

Hence,

$$P(\text{At Least One Six}) = 1 - 125/216 = 91/216 \approx 0.42$$

71. Suppose a couple decides to have three children. Assume that the sex of each child is independent, and the probability of a girl is 0.48, the approximate figure in the US.

    (a) How many basic outcomes are there for this experiment? Are they equally likely?

    **Solution:** There are two possible outcomes for each birth, so by the multiplication rule for counting, the total number of possibilities is $2 \times 2 \times 2 = 8$. They are not equally likely because each child is more likely to be a boy than a girl. The outcome BBB is most likely, followed by outcomes with two boys, and then outcomes with one boy. The outcome GGG is least likely.

    (b) What is the probability that the couple has *at least one* girl?

    **Solution:** Use the Complement Rule and independence to calculate the probability of no girls, i.e. all boys:

    $$0.52 \times 0.52 \times 0.52 \approx 0.14$$

    Hence, the probability of at least one girl is approximately $1 - 0.14 = 0.86$

72. Let $A$ and $B$ be two arbitrary events. Use the addition rule and axioms of probability to establish the following results.

    (a) Show that $P(A \cup B) \leq P(A) + P(B)$. (This is called *Boole's Inequality.*)

    **Solution:** By the Addition Rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. The result follows since $P(A \cap B) \geq 0$ by the first axiom of probability.

    (b) Show that $P(A \cap B) \geq P(A) + P(B) - 1$. (This is called *Bonferroni's Inequality*)

**Solution:** Rearranging the Addition Rule, $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. The result follows since $P(A \cup B)$ is at most one by the first axiom of probability.

73. Let $A$ and $B$ be two mutually exclusive events such that $P(A) > 0$ and $P(B) > 0$. Are $A$ and $B$ independent? Explain why or why not.

**Solution:** They are not independent: knowing that one has occurred means that the other *cannot have occurred*. You can also show this mathematically. Since $A$ and $B$ are mutually exclusive, $P(A \cap B) = 0$. But independence requires that $P(A \cap B) = P(A)P(B)$. Since neither $P(A)$ nor $P(B)$ is zero, it follows that the events cannot be independent.

74. Molly the meteorologist determines that the probability of rain on Saturday is 50%, and the probability of rain on Sunday is also 50%. Adam the anchorman sees Molly's forecast and summarizes it as follows: "According to Molly we're in for a wet weekend. There's a 100% chance of rain this weekend: 50% on Saturday and 50% on Sunday." Is Adam correct? Why or why not?

**Solution:** Adam is incorrect. Let $A$ be the event that it rains on Saturday, $B$ be the event that it rains on Sunday, and $C$ be the event that it rains on the weekend. By the addition rule $P(C) = P(A) + P(B) - P(A \cap B)$, so Adam is only correct if $P(A \cap B) = 0$, in other words he is only correct if it is *impossible* for it to rain on both Saturday and Sunday. There is no way to know that this is the case solely from Molly's information about the probabilities of $A$ and $B$.

75. Suppose I throw two fair, six-sided dice once. Define the following events:

$$\begin{aligned} E &= \text{The first die shows 5} \\ F &= \text{The sum of the two dice equals 7} \\ G &= \text{The sum of the two dice equals 10} \end{aligned}$$

(a) Calculate $P(F)$.

**Solution:** Of the 36 basic outcomes of the experiment, the pairs (1,6), (6,1), (2,5), (5,2), (3,4), and (4,3) sum to 7. Hence the probability is 1/6.

(b) Calculate $P(G)$.

**Solution:** Of the 36 basic outcomes of this experiment, the pairs (5,5), (4,6), and (6,4) sum to 10. Hence the probability is $3/36 = 1/12$.

(c) Calculate $P(F|E)$.

**Solution:** By the definition of conditional probability,

$$P(F|E) = \frac{P(F \cap E)}{P(E)}$$

We know that $P(E) = 1/6$. The only way that $F \cap E$ can occur is if we roll (5,7). Hence $P(F \cap E) = 1/36$. Thus, $P(F|E) = (1/36)/(1/6) = 6/36 = 1/6$.

(d) Calculate $P(G|E)$.

**Solution:** Again, by the definition of conditional probability,

$$P(G|E) = \frac{P(G \cap E)}{P(E)}$$

As before, $P(E) = 1/6$. The only way for $G \cap E$ to occur is if we roll (5,5). Hence $P(G|E) = (1/36)/(1/6) = 6/36 = 1/6$.

# Lecture #7 – Basic Probability III / Discrete RVs I

76. What is the base rate fallacy? Give an example.

77. Derive Bayes' Rule from the definition of conditional probability.

78. What are two names for the *unconditional* probability in the numerator of Bayes' rule?

79. When is it true that $P(A|B) = P(B|A)$? Explain.

80. Of women who undergo regular mammograms, two percent have breast cancer. If a woman has breast cancer, there is a 90% chance that her mammogram will come back positive. If she does *not* have breast cancer there is a 10% chance that her mammogram will come back positive. Given that a woman's mammogram has come back positive, what is the probability that she has breast cancer?

**Solution:** Let $B$ be the event that a given woman has breast cancer and $M$ be the event that her mammogram comes back positive. By Bayes' Rule,

$$P(B|M) = \frac{P(M|B)P(B)}{P(M)}$$

By the law of total probability,

$$\begin{aligned} P(M) &= P(M|B)P(B) + P(M|B^c)P(B^c) \\ &= 0.9 \times 0.02 + 0.1 \times 0.98 = 0.018 + 0.098 = 0.116 \end{aligned}$$

Hence,

$$P(B|M) = \frac{0.9 \times 0.02}{0.116} = \frac{0.018}{0.116} \approx 0.16$$

81. The Triangle is a neighborhood that once housed a chemical plant but has become a residential area. Two percent of the children in the city live in the Triangle, and fourteen percent of these children test positive for excessive presence of toxic metals in the tissue. For children in the city who do not live in the Triangle, the rate of positive tests is only one percent. If we randomly select a child who lives in the city and she tests positive, what is the probability that she lives in the Triangle?

**Solution:** By the law of total probability

$$\begin{aligned} P(M) &= P(M|T)P(T) + P(M|T^c)P(T^c) \\ &= 0.14 \times 0.02 + 0.01 \times 0.98 \\ &= 0.0028 + 0.0098 \\ &= 0.0126 \end{aligned}$$

and by Bayes' Rule:

$$\begin{aligned} P(T|M) &= \frac{P(M|T)P(T)}{P(M)} \\ &= \frac{0.0028}{0.0126} = 2/9 \approx 0.22 \end{aligned}$$

82. Three percent of *Tropicana* brand oranges are already rotten when they arrive at the supermarket. In contrast, six percent of *Sunkist* brand oranges arrive rotten. A local

supermarket buys forty percent of its oranges from *Tropicana* and the rest from *Sunkist*. Suppose we randomly choose an orange from the supermarket and see that it is rotten. What is the probability that it is a *Tropicana*?

**Solution:** By the law of total probability:

$$\begin{aligned} P(R) &= P(R|T)P(T) + P(R|T^c)P(T^c) \\ &= 0.03 \times 0.4 + 0.06 \times 0.6 \\ &= 0.012 + 0.036 \\ &= 0.048 \end{aligned}$$

and by Bayes' Rule:

$$\begin{aligned} P(T|R) &= \frac{P(R|T)P(T)}{P(R)} \\ &= \frac{0.012}{0.048} = 1/4 = 0.25 \end{aligned}$$

83. Define the terms *random variable*, *realization*, and *support set*.

84. What is the probability that a RV takes on a value outside of its support set?

85. What is the difference between a *discrete* and *continuous* RV?

86. What is a *probability mass function*? What two key properties does it satisfy?

# Lecture #8 – Discrete RVs II

87. Define the term *cumulative distribution function* (CDF). How is the CDF of a discrete RV $X$ related to its pmf?

88. Let $X$ be a RV with support set $\{-1, 1\}$ and $p(-1) = 1/3$. Write down the CDF of $X$.

**Solution:** $F(x_0) = \begin{cases} 0, \, x_0 < -1 \\ 1/3, \, -1 \leq x_0 < 1 \\ 1, \, x_0 \geq 1 \end{cases}$

89. Write out the support set, pmf, and CDF of a Bernoulli($p$) RV.

90. Define the term *parameter* as it relates to a random variable. Are parameters constant or random?

91. Let $X$ be a RV with support set $\{0, 1, 2\}$, $p(1) = 0.3$, and $p(2) = 0.5$. Calculate $E[X]$.

> **Solution:** $E(X) = 0 \times 0.2 + 1 \times 0.3 + 2 \times 0.5 = 1.3$

*L*et $X$ be a discrete RV. Define the expected value $E[X]$ of $X$. Is $E[X]$ constant or random? Why?

92. Suppose $X$ is a RV with support $\{-1, 0, 1\}$ where $p(-1) = q$ and $p(1) = p$. What relationship must hold between $p$ and $q$ to ensure that $E[X] = 0$?

> **Solution:** By the complement rule $p(0) = 1 - p - q$. Hence,
>
> $$E[X] = -1 \cdot q + 0 \cdot (1 - p - q) + p \cdot 1 = p - q$$
>
> so that $E[X] = 0$ if and only if $p = q$.

93. Let $X$ be a discrete RV and $a, b$ be constants. Prove that $E[a + bX] = a + bE[X]$.

94. Suppose that $E[X] = 8$ and $Y = 3 + X/2$. Calculate $E[Y]$.

> **Solution:** $E(Y) = 3 + E(X)/2 = 7$

95. Suppose that $X$ is a discrete RV and $g$ is a function. Explain how to calculate $E[g(X)]$. Is this the same thing as $g\left(E[X]\right)$?

96. Let $X$ be a RV with support set $\{-1, 1\}$ and $p(-1) = 1/3$. Calculate $E[X^2]$.

> **Solution:** $E[X^2] = (-1)^2 \times 1/3 + (1)^2 \times 2/3 = 1$

97. Let $X$ be a RV with support set $\{2, 4\}$, $p(2) = 1/2$ and $p(4) = 1/2$. Mark each of the following claims as TRUE or FALSE, either by appealing to a result from class, or by directly calculating both sides of the equality.

(a) $E[X + 10] = E[X] + 10$

**Solution:** TRUE by the linearity of expectation.

(b) $E[X/10] = E[X]/10$

**Solution:** TRUE by the linearity of expectation.

(c) $E[10/X] = 10/E[X]$

**Solution:** FALSE. By direct calculation:

$$E[10/X] = 1/2 \times 10/2 + 1/2 \times 10/4 = 15/4 = 3.75$$
$$10/E[X] = 10/(1/2 \times 2 + 1/2 \times 4) = 10/3$$

(d) $E[X^2] = (E[X])^2$

**Solution:** FALSE. By direct calculation:

$$E[X^2] = 1/2 \times 2^2 + 1/2 \times 4^2 = 10$$
$$(E[X])^2 = (1/2 \times 2 + 1/2 \times 4)^2 = 9$$

(e) $E[5X + 2]/10 = (5E[X] + 2)/10$

**Solution:** TRUE by the linearity of expectation.

# Lecture #9 – Discrete RVs III

98. Define the *variance* and *standard deviation* of a RV $X$. Are these constant or random?

99. Explain how to use our formula for $E[g(X)]$ to calculate the variance of a discrete RV.

100. Write down the shortcut formula for variance, and use it to calculate $Var(X)$ where $X \sim$ Bernoulli$(p)$.

101. Let $X$ be a random variable and $a, b$ be constants. Prove that $Var(a+bX) = b^2 Var(X)$.

102. Define the Binomial$(n, p)$ RV in terms of independent Bernoulli trials, and write down its support set and probability mass function.

103. Substitute $n = 1$ into the pmf of a Binomial$(n, p)$ RV and show that you obtain the pmf of a Bernoulli$(p)$ RV.

**Solution:** The pmf for a Binomial$(n, p)$ RV is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

with support $\{0, 1, 2 \ldots, n\}$. Setting $n = 1$ gives,

$$p(x) = p(x) = \binom{1}{x} p^x (1-p)^{1-x}$$

with support $\{0, 1\}$. Plugging in each realization in the support, and recalling that $0! = 1$, we have

$$p(0) = \frac{1!}{0!(1-0)!} p^0 (1-p)^{1-0} = 1 - p$$

and

$$p(1) = \frac{1!}{1!(1-1)!} p^1 (1-p)^0 = p$$

which is exactly how we defined the Bernoulli Random Variable.

104. A multiple choice quiz has 12 questions, each of which has 5 choices. To pass you need to get at least 8 of them correct. Nina forgot to study, so she simply guesses at random.

(a) Let the random variable $X$ denote the number of questions that Nina gets correct on the quiz. What kind of random variable is $X$? Specify all parameter values.

(b) Calculate the probability that Nina passes the quiz.

# Lecture #10 – Discrete RVs IV

105. What is the difference between a joint pmf and a marginal pmf? Can you calculate a marginal pmf from a joint? How? Can you calculate a joint pmf from a marginal pmf?

106. Suppose that $X$ is a random variable with support $\{1, 2\}$ and $Y$ is a random variable with support $\{0, 1\}$ where $X$ and $Y$ have the following joint pmf:

$$p_{XY}(1, 0) = 0.20, \qquad p_{XY}(1, 1) = 0.30$$
$$p_{XY}(2, 0) = 0.25, \qquad p_{XY}(2, 1) = 0.25$$

(a) Express the joint pmf in a table with $X$ in the *rows*, as we did in class.

**Solution:**

| | | $X$ | |
|---|---|---|---|
| | | 1 | 2 |
| $Y$ | 0 | 0.20 | 0.25 |
| | 1 | 0.30 | 0.25 |

(b) Using the table, calculate the marginal pmfs of $X$ and $Y$.

**Solution:**

$$\begin{aligned} p_X(1) &= p_{XY}(1,0) + p_{XY}(1,1) = 0.20 + 0.30 = 0.50 \\ p_X(2) &= p_{XY}(2,0) + p_{XY}(2,1) = 0.25 + 0.25 = 0.50 \\ p_Y(0) &= p_{XY}(1,0) + p_{XY}(2,0) = 0.20 + 0.25 = 0.45 \\ p_Y(1) &= p_{XY}(1,1) + p_{XY}(2,1) = 0.30 + 0.25 = 0.55 \end{aligned}$$

107. The question relies upon the following joint pmf:

| | | $X$ | |
|---|---|---|---|
| | | 0 | 1 |
| $Y$ | 1 | 0.1 | 0.2 |
| | 2 | 0.3 | 0.4 |

(a) Calculate the conditional pmf of $Y$ given that $X = 0$.

(b) Calculate the conditional pmf of $X$ given that $Y = 2$.

108. This question relies on the following joint pmf:

| | | $Y$ | | |
|---|---|---|---|---|
| | | -1 | 0 | 1 |
| $X$ | 0 | 1/9 | 1/9 | 0 |
| | 1 | 2/9 | 1/9 | 1/9 |
| | 2 | 0 | 1/9 | 2/9 |

(a) Calculate $p_Y(0)$.

**Solution:** $p_X(0) = p_{XY}(0,0) + p_{XY}(1,0) + p_{XY}(2,0) = 1/3$

(b) Calculate $p_{X|Y}(2|0)$.

**Solution:** $p_{X|Y}(0|0) = (1/9)/(1/9 + 1/9 + 1/9) = 1/3$

(c) Calculate $E[XY]$.

**Solution:** $E[XY] = (1\times-1\times2/9)+(1\times1\times1/9)+(2\times1\times2/9) = (-2+1+4)/9 = 1/3$

(d) Calculate $Cov(X, Y)$

(e) Are $X$ and $Y$ independent? Why or why not?

**Solution:** They are not independent: for example, if we know $Y = -1$ then $X$ cannot take on the value 2.

109. Prove the shortcut formula for variance: $Var(X) = E[X^2] - (E[X])^2$.

110. Prove that $Cov(X, Y) = E[XY] - E[X]E[Y]$. Hint: the steps are similar to our derivation of the shortcut formula for variance from class.

**Solution:** By the Linearity of Expectation,

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= E[XY] - \mu_x E[Y] - \mu_Y E[X] + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

111. Let $X$ and $Y$ RVs with $E[X] = 2$ and $E[Y] = 1$. Calculate $E[X - Y]$.

**Solution:** By the linearity of expectation: $E[X - Y] = E[X] - E[Y] = 1$.

112. Suppose $E[X] = 2$ and $Var(X) = 5$. Calculate $E[X^2]$.

**Solution:** By the shortcut formula: $Var(X) = E[X^2] - (E[X])^2$, so $E[X^2] = 9$.

113. Let $X$ and $Y$ be RVs with $Var(X) = 2$ , $Var(Y) = 1$, and $Cov(X, Y) = 0$. Calculate $Var(X - Y)$.

**Solution:** $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y) = 3$

114. Let $X$ and $Y$ be two RVs with $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$, and $Cov(X, Y) = \sigma_{XY}$. If $a, b, c$ are constants, what is $Var(cX + bY + a)$?

115. Suppose that $X$ and $Y$ are two RVs with correlation $\rho = 0.3$, and standard deviations $\sigma_X = 4$ and $\sigma_Y = 5$.

(a) Calculate $Cov(X, Y)$.

(b) Let $Z = (X + Y)/2$. Calculate $Var(Z)$.

116. What does it mean for a sequence of random variables $X_1, X_2, \ldots, X_n$ to be "independent and identically distributed (iid)?"

117. Mark each statment as TRUE or FALSE. If FALSE, explain.

(a) The expected value of a sum $E[X_1 + X_2 + \cdots + X_n]$ is *not* in general equal to the sum of the expected values $E[X_1] + E[X_2] + \cdots + E[X_n]$. But when $X_1, X_2, \ldots, X_n$ are independent then the two are equal.

(b) The variance of a sum $Var(X_1 + X_2 + \cdots + X_n)$ is always equal to the sum of the variances $Var(X_1) + Var(X_2) + \cdots + Var(X_n)$.

118. Suppose that $X \sim \text{Binomial}(n, p)$.

(a) Explain how $X$ can be defined in terms of Bernoulli($p$) RVs.

(b) Using the preceding part, derive $E[X]$.

(c) Using the preceding part, derive $Var(X)$.

119. Suppose that $X \sim \text{Binomial}(9, 1/3)$ and $Y \sim \text{Binomial}(4, 1/2)$. Calculate $E[(Y - X)/2]$.

# Lecture #11 – Continuous RVs I

120. If $X$ is a continuous RV and $a, b$ are constants, how do we calculate $P(a \leq X \leq b)$?

121. What are the two properties of a probability density function?

122. True or False: since $f(x)$ is a probability, $0 \leq f(x) \leq 1$. If false, correct the statement.

123. How is the PDF of a continuous RV related to its CDF?

124. Let $X$ be a continuous RV with CDF $F$. Express $P(-2 \leq X \leq 4)$ in terms of $F$.

125. Let $X$ be a continuous RV with CDF $F$. Express $P(X \geq x_0)$ in terms of $F$.

126. Suppose that $X$ is a continuous RV with support set $[-1, 1]$.
    (a) Is 2 a possible realization of this RV?
    (b) What is $P(X = 0.5)$?
    (c) True or False: $P(X \leq 0.3) = P(X < 0.3)$. Explain.

127. Let $X$ be a Uniform$(0, 1)$ RV. Calculate the CDF of $X$.

128. Let $X$ be a Uniform$(0, 1)$ RV. Calculate $Var(X)$.

129. Let $X$ be a Uniform$(a, b)$ RV. Calculate $E[X]$

130. Let $X$ be a continuous RV with support $[0, 1]$ and $f(x) = Cx^2(1 - x)$. Find $C$.

**Solution:** 12

131. Let $X$ be a continuous RV with support $[0, 1]$ and $f(x) = 3x^2$. Find the CDF of $X$.

132. Let $X$ be a continuous RV with support $[0, 1]$ and $f(x) = 3x^2$. Calculate $Var(X)$.

133. Let $X$ be a continuous RV with support $[0, 1]$ and $f(x) = 3x^2$. Calculate the probability that $X$ takes a value in the interval $[0.2, 0.8]$.

134. Let $X$ be a RV with support set $[-2, 2]$ and the following CDF:

$$F(x_0) = \begin{cases} 0, & x_0 < -2 \\ x_0/4, & -2 \leq x_0 \leq 2 \\ 1, & x_0 > 2 \end{cases}$$

    (a) Calculate the PDF of $X$.
    (b) Is $X$ an example of one of the "named" RVs from the lecture slides? If so, which one? Be sure to specify the values of any and all parameters of the distribution.

# Lecture #12 – Continuous RVs II

135. Suppose that $X$ is a $N(\mu, \sigma^2)$ RV.
    (a) What is $E[X]$?
    (b) What is $Var(X)$?
    (c) What is the support set of $X$?
    (d) What is the median of $X$?

136. Suppose that $X \sim N(\mu, \sigma^2)$. Approximately what are the values of the following probabilities?
    (a) $P(\mu - \sigma \leq X \leq \mu + \sigma)$
    (b) $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$
    (c) $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$

137. Write R code to accomplish the following tasks:
    (a) Calculate the height of the standard normal PDF at $x = 0$.
    (b) Make 10 random draws from a standard normal distribution.
    (c) Calculate the 20th percentile of a standard normal distribution.
    (d) Calculate $P(X \leq 0.5)$ if $X \sim N(0, 1)$.

138. Write R code to plot the PDF and CDF of a standard normal RV between -5 and 5.

139. Approximately what result would you get if you entered `pnorm(1)` at the R console? Hint: use symmetry and the "empirical rule."

> **Solution:** $\approx 0.84$

140. Suppose that $X \sim N(0, 1)$. What is $E[X^2]$?

141. Define the quantile function $Q(p)$ of a continuous RV $X$. How is it related to the CDF $F(x_0)$ of $X$?

142. Let $X \sim N(\mu = -2, \sigma^2 = 25)$. Without using R, find the approximate value of

$$P(-12 \leq X \leq 8)$$

**Solution:**

$$P(-12 \leq X \leq *) = P(-2 \leq Z \leq 2) \approx 0.95$$

where $Z \sim N(0,1)$.

143. Write a line of R code that calculates the probability that $P(-0.2 \leq Z \leq 0.4)$ if $Z$ is a standard normal random variable.

144. Suppose that $Y \sim N(\mu = 2, \sigma^2 = 4)$.
   (a) Write R code to calculate $P(-1 \leq Y \leq 6)$.
   (b) Write R code to calculate $P(Y \geq 6)$.

145. Suppose that $Z \sim N(0,1)$. Write the line of R code you would use to find $c > 0$ such that $P(-c \leq Z \leq c) = 0.8$.

**Solution:** `qnorm(0.9)` for $c$ or `qnorm(0.1)` for $-c$

146. Suppose that $X_1 \sim N(0,1)$ independently of $X_2 \sim N(\mu = 2, \sigma^2 = 9)$.
   (a) What kind of RV is $\frac{1}{3}(X_2 - 2)$? Specify the values of any and all of its parameters.
   (b) What kind of random variable is $X_1 + X_2$? Specify any and all of its parameters.

147. Suppose that $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and define $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$. What kind of RV is $\bar{X}_n$? Specify the values of any and all of its parameters.

148. Suppose that $X_1, \ldots, X_n \sim$ iid $N(\mu_X, \sigma_X^2)$ independently of $Y_1, \ldots, Y_m \sim$ iid $N(\mu_Y, \sigma_Y^2)$ and define $\bar{X}_n = (X_1 + \cdots + X_n)/n$ and $\bar{Y}_m = (Y_1 + \cdots + Y_m)/m$. What kind of RV is $\bar{X}_n - \bar{Y}_m$? Specify the values of any and all of its parameters.

# Lecture #13 – Sampling and Estimation I

149. We gave a verbal definition of the term *random sample* in lecture #1. What is the mathematical definition?

150. Why do draws made without replacement *fail* to constitute a random sample? Under what circumstances does it become for all practical purposes irrelevant whether one draws with or without replacement?

151. Suppose that we have a vector `x` that we will treat as a *population* for the purpose of carrying out a simulation exercise. Write R code to generate a histogram of 1000 sample means, each of which is constructed from a random sample of size 10 drawn from `x`.

152. Let $X_1, \ldots, X_n \sim$ iid with mean $\mu$ and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculate $E[\bar{X}_n]$.

153. Let $X_1, \ldots, X_n \sim$ iid with mean $\mu$ and variance $\sigma^2$, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculate $Var(\bar{X}_n)$

154. Define the term *standard error*. What is the standard error of the sample mean under random sampling?

155. True or False: under random sampling, the population size $N$ affects the sampling distribution of $\bar{X}_n$. If false, explain.

156. Define the term *statistic*.

157. Explain the difference between an *estimator* and an *estimate*, using the sample mean as an example.

158. Suppose that 20% of registered Democrats plan to vote for Bernie Sanders in the 2020 Democratic Primary. You poll a random sample of 3 Democrats and calculate the proportion $\widehat{p}$ who support Sanders. What is the sampling distribution of $\widehat{p}$? (Write out the support set and pmf.)

# Lecture #14 – Sampling and Estimation II

159. Define the *bias* of an estimator. What does it mean for an estimator to be *unbiased*?

160. Why do we divide by $n-1$ in our definition of the sample variance?

161. Define the concept of *efficiency*. What does it mean to say that one estimator is *more efficient* than another?

162. Define *mean-squared error*. Why is it a useful concept?

163. Explain the difference between the *finite sample* and *asymptotic* properties of an estimator.

164. What does it mean to say that an estimator $\widehat{\theta}_n$ is *consistent* for $\theta_0$?

165. Show that $\bar{X}_n$ is consistent for the population mean under random sampling.

For the following five questions, let $X_1, X_2, X_3, \ldots, X_n \sim$ iid with mean $\mu$ and variance $\sigma^2$ and define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

166. Is $\bar{X}_n$ is an unbiased estimator of $\mu$? Why or why not?

167. Is $(0.1X_1 + 0.9X_2)$ is an unbiased estimator of $\mu$? Why or why not?

**Solution:** Yes: $E[0.1X_1 + 0.9X_2] = 0.1E[X_1] + 0.9E[X_2] = 0.1\mu + 0.9\mu = \mu$

168. Is $(0.1X_1 + 0.9X_2)$ is a more efficient estimator of $\mu$ than $(0.5X_1 + 0.5X_2)$? Explain.

**Solution:** No: both estimators are unbiased, so it makes sense to talk about "efficiency," but $\text{Var}(0.1X_1 + 0.9X_2) = 0.01\sigma^2 + 0.81\sigma^2 = 0.82\sigma^2$ which is much larger than $\text{Var}(0.5X_1 + 0.5X_2) = 0.5\sigma^2$.

169. Suppose $\mu$ is *known* and we want to estimate $\sigma^2$. Is $\widetilde{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$ an unbiased estimator of $\sigma^2$? Justify your answer.

**Solution:** It is a *biased estimator*: $E[\widetilde{\sigma}] = \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \mu)^2] = \frac{1}{n-1}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{n}{n-1}\sigma^2 \neq 0$

170. Calculate the bias and variance of the estimator $\widehat{\mu} = \dfrac{n\bar{X}_n}{1+n}$. Is $\widehat{\mu}$ consistent for $\mu$?

**Solution:** Yes: $\text{Bias}(\widehat{\mu}) = \left[\left(\frac{n}{n+1}\right)E[\bar{X}_n] - \mu\right] = \left[\left(\frac{n}{n+1}\right)\mu - \mu\right]$ and $\text{Var}(\widehat{\mu}) = \left(\frac{n}{n+1}\right)^2 \text{Var}(\bar{X}_n) = \left(\frac{n}{n+1}\right)^2 \frac{\sigma^2}{n}$. Both of these converge to zero as $n \to \infty$.

# Lecture #15 – Confidence Intervals I

171. Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and define $Y = \sqrt{n}(\bar{X}_n - \mu)/\sigma$. What kind of random variable is $Y$? Specify any and all parameters of its distribution.

172. Give the formula for an approximate 95% confidence interval for the mean $\mu$ of a normal population with known variance $\sigma^2$, based on a random sample of size $n$.

173. Give the formal definition of a confidence interval and the "rough intuition" that we can use to interpret it.

174. Define the term *confidence level* in the context of constructing a confidence interval. What symbol do we use to represent the confidence level of an interval?

175. Define the following terms related to confidence intervals:

    (a) Margin of error

    (b) Lower confidence limit (LCL)

    (c) Upper confidence limit (UCL)

    (d) Width

176. Suppose that Alice and Bob each draw independent random samples of size $n = 100$ from a normal population with unknown mean $\mu$ and known variance $\sigma^2 = 9$. Both of them construct 95% confidence intervals for $\mu$.

    (a) Will the widths of Alice and Bob's confidence intervals be the same? Explain.

    (b) Will the Alice and Bob's confidence intervals be identical? Explain.

177. Suppose that we draw a random sample of $n = 25$ observations from a normal population with known variance $\sigma^2 = 4$ and unknown mean $\mu$.

    (a) What is the margin of error for an approximate 95% CI for $\mu$?

    (b) Say we observe $\bar{x} = 2.5$. Construct an approximate 95% CI for $\mu$.

178. Give the formula for a $(1 - \alpha) \times 100\%$ confidence interval for the mean $\mu$ of a normal population with known variance $\sigma^2$, based on a random sample of size $n$.

179. Explain how $\alpha$, $\sigma$, and $n$ affect the width of a confidence interval for the mean of a normal population with known variance $\sigma^2$.

180. Suppose that I draw a random sample of size 64 from a normal population with known variance 16 and unknown mean $\mu$. My sample mean equals $-1.8$. Construct an approximate 68% confidence interval for $\mu$.

# Lecture #16 – Confidence Intervals II

181. In what sense can we say that the values near the center of a symmetric confidence interval are "more plausible" than those near the LCL and UCL?

182. Suppose we observe $X_1, \ldots, X_2 \sim N(\mu, \sigma^2)$. If we want to construct a confidence interval for $\mu$, does it make a difference if $\sigma^2$ is unknown and has to be estimated? Explain.

183. Let $X_1, \ldots, X_9 \sim N(\mu, \sigma^2)$ and define $\bar{X} = (\sum_{i=1}^9 X_i)/9$ and $S^2 = \left[\sum_{i=1}^9 (X_i - \bar{X})^2\right]/8$. If $S$ is the positive square root of $S^2$, then what kind of random variable is $3(\bar{X} - \mu)/S$? Be sure to specify the values of any and all parameters of its distribution.

**Solution:** $t(8)$.

184. What is the support set of a Student-t random variable?

185. Under what circumstances is the Student-t random variable practically identical to the standard normal random variable?

186. Suppose that $X$ is a Student-t random variable with degrees of freedom equal to 10. Write a line of R code to find $c$ such that $P(-c \leq X \leq c) = 0.68$.

187. What is the median of a Student-t random variable with 17 degrees of freedom?

**Solution:** 0

188. Write out the formula for a $(1 - \alpha) \times 100\%$ CI for the mean $\mu$ of a normal population with unknown variance $\sigma^2$, based on a random sample of size $n$.

189. Alice and Bob each observe the same random sample $X_1, \ldots, X_n$ from a normal population with mean $\mu$ and variance $\sigma^2$. Each of them constructs a 95% confidence interval for $\mu$. Alice knows the true value of $\sigma^2$ while Bob does not. Each researcher uses the appropriate confidence interval based on the information that she has available.

(a) Will Alice and Bob's intervals be centered in the same place?

(b) Whose interval would we expect to be *wider*?

190. Suppose that $X_1, \ldots, X_n$ are iid draws from some unknown population. If $n$ is large, what is the approximate sampling distribution of $\frac{\bar{X}_n - \mu}{S/\sqrt{n}}$?

191. TRUE or FALSE: the Central Limit Theorem says that large populations are approximately normally distributed. If FALSE, correct the statement.

192. Suppose that $X_1, \ldots, X_n \sim$ iid Bernoulli$(p)$ and let $\widehat{p}$ be the sample proportion of ones. Show that:

(a) $E(\widehat{p}) = p$

(b) $Var(\widehat{p}) = p(1-p)/n$

193. Camilo wants to know the proportion $p$ of US voters who favor legalizing marijuana, so he carries out a poll based on a random sample. Of the 100 individuals in his sample, 60 favor legalizing marijuana. Based on this information, construct an approximate 95% confidence interval for $p$.

**Solution:** $\widehat{p} = 60/100 = 0.6$ and $ME = 2\sqrt{\frac{0.6 \times 0.4}{100}} \approx 0.1$ so the CI is $0.6 \pm 0.1$ or equivalently $(0.5, 0.7)$.

194. Suppose $X_1, \ldots, X_n \sim$ iid Bernoulli(1/2), and define $\widehat{p} = (\sum_{i=1}^n X_i)/n$. If $n = 100$, approximately what is the probability that $0.45 \leq \widehat{p} \leq 0.55$?

**Solution:** Since $n$ is large, $\widehat{p}$ is approximately normal by the Central Limit Theorem. Its mean is $p = 1/2$ and its standard error is $\sqrt{p(1-p)/n} = \sqrt{0.5^2/100} = 0.5/10 = 0.05$, and the probability that a normal RV is within $\pm$ standard errors of its mean is about 0.68.

# Lecture #17 – Confidence Intervals III

195. Let $X_1, \ldots, X_6 \sim$ iid N$(\mu_x = 200, \sigma_x^2 = 54)$ and $Y_1, \ldots, Y_{10} \sim$ iid N$(\mu_y = 150, \sigma_y^2 = 160)$ where the $X$ and $Y$ observations are independent. Approximately what is the probability that $\bar{X} - \bar{Y} > 55$?

**Solution:** $\bar{X} - \bar{Y} \sim$ N(mean $= 50$, SD $= 5$) so this is just the probability that a normal is at least one standard deviation above its mean, which is approximately 16%.

196. Let $X_1, \ldots, X_n$ be a random sample from a population with mean $\mu_X$ and variance $\sigma_X^2$, and $Y_1, \ldots, Y_m$ be a random sample from a *different* population with mean $\mu_Y$ and variance $\sigma_Y^2$. Suppose that the $X$ and $Y$ observations are independent of one another.

(a) Derive the standard error of $\bar{X}_n - \bar{Y}_m$.

(b) Suppose that we do not know $\sigma_X^2$ or $\sigma_Y^2$. How can we estimate $SE(\bar{X}_n - \bar{Y}_m)$?

(c) Let $\widehat{SE}$ be your proposed estimator from part (b). If $n$ and $m$ are both large, what is the approximate sampling distribution of $[(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)]/\widehat{SE}$?

197. Suppose that we observe hourly wages for a random sample of 20 college graduates: the sample mean is \$31 with a standard deviation of \$15. In contrast, the sample mean wage is \$17 with a standard deviation of \$10 for a sample of 30 non-college graduates. Construct an approximate 95% confidence interval for the difference in population mean wages ($\mu_X - \mu_Y$) between college graduates ($X$) and non-college graduates ($Y$).

198. Xanthippe wants to know which university has a higher proportion of philosophy majors: Penn or Princeton. She polls a random sample of 100 Penn students – 10 are philosophy majors. In contrast, 7 of the 50 Princeton students she polls are philosophy majors. Construct an approximate 95% confidence interval for the difference in proportions of philosophy majors: Penn minus Princeton.

199. TRUE or FALSE: Regardless of whether our dataset consists of independent samples or matched pairs, the confidence interval turns out to be exactly the same. If FALSE, correct the statement.

200. Suppose we observe two datasets: $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ with sample standard deviations $s_x = 3$ and $s_y = 4$ and sample correlation $r_{xy} = 0.5$. Calculate the sample variance $s_d^2$ of $d_1, \ldots, d_n$ where $d_i = x_i - y_i$.

**Solution:** $s_d^2 = s_x^2 + s_y^2 - 2 s_x s_y r_{xy} = 9 + 16 - 2 \times 3 \times 4 \times 0.5 = 25 - 12 = 13$

201. Let $D_i = X_i - Y_i$. Show that $\bar{D}_n = \bar{X}_n - \bar{Y}_n$, where $\bar{D}_n, \bar{X}_n, \bar{Y}_n$ denote the sample means of $D, X$, and $Y$.

202. Let $D_i = X_i - Y_i$. Show that $S_D^2 = S_X^2 + S_Y^2 - 2 S_X S_Y r_{XY}$ where $S_D, S_X, S_Y$ denote the sample standard deviations of $D, X$, and $Y$, and $r_{XY}$ is the sample correlation between $X$ and $Y$.

203. For each example, indicate whether it involves *matched pairs* or *independent samples*.

(a) To compare the performance of the two brands, Alice installs Firestone tires on half of the cars in the *Consumer Reports* test garage, and Michelin tires on the rest.

(b) To determine the effect of listening to music on his workers' productivity, Bob installs a radio in the office. During the month of March, he keeps the radio turned on all day. During the month of April, he keeps it turned off.

(c) To test the effectiveness of a new marketing campaign, Charlotte takes out new advertisements in half of the cities where her firm has a retail presence. She leaves the old advertisements in place in the remaining cities.

(d) Dan compares the wages of male and female high school teachers in Philadelphia.

(e) To determine the effect of college attendance on wages, Elise studies a sample of identical twins in which one twin attended college and the other didn't.

204. What are the two equivalent ways to construct a matched pairs CI?

205. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of matched pairs. Suppose that you erroneously construct an independent samples confidence interval for $\mu_X - \mu_Y$ using these observations. If the observations $(X_i, Y_i)$ within a given pair are *negatively correlated*, will your interval be too wide or too narrow? Explain.

# Lecture #18 – Hypothesis Testing I

206. Define the term *type I error*.

207. Define the term *type II error*.

208. What was our null hypothesis in the Pepsi Challenge experiment from class?

209. What was our alternative hypothesis in the Pepsi Challenge experiment from class?

210. In the "Pepsi Challenge" experiment from class there were four cups of Coke and four of Pepsi. In this question, consider a modified version of the experiment with *three* cups of each kind of soda. Everything else is unchanged. Calculate the probability that our test statistic, the number of cokes correctly identified, will equal two *under the null hypothesis*.

**Solution:**

$$\frac{\binom{3}{2} \times \binom{3}{1}}{\binom{6}{3}} = \frac{3 \times 3}{20} = 9/20 = 0.45$$

# Lecture #19 – Hypothesis Testing II

Note: if asked "what is the distribution of ..." you must specify the values of any and all parameters for full credit.

211. Let $X_1, \ldots, X_9 \sim$ iid $\mathrm{N}(\mu = 5, \sigma^2 = 81)$ and define $\bar{X} = (X_1 + \cdots + X_9)/9$.

(a) What is the distribution of $\bar{X}$?

**Solution:** $\bar{X} \sim N(\mu, \sigma^2/n) = N(5, 9)$

(b) What is the distribution of $(\bar{X} - 5)/3$?

**Solution:** $N(0, 1)$

212. Let $Z_1, \ldots, Z_{16} \sim$ iid $\mathrm{N}(\mu = -2, \sigma^2 = 4)$, $Y = Z_1/16 + \cdots + Z_{16}/16$, and $X = 2Y + 4$.

(a) What is the distribution of $X$?

**Solution:** Since $X \sim N(\mu, \sigma^2/n) = N(-2, 1/4)$,

$$X = 2Y + 4 = 2(Y + 2) = \frac{Y - (-2)}{\sqrt{1/4}} \sim N(0, 1)$$

(b) What is $P(|X| > 1)$?

**Solution:**

$$\begin{aligned} P(|X| > 1) &= 1 - P(|X| < 1) = 1 - P(-1 \leq X \leq 1) \\ &= 1 - [\texttt{pnorm(1) - pnorm(-1)}] \approx 0.32 \end{aligned}$$

213. Suppose that I observe a random sample of $n$ observations from a normal population with unknown mean and known variance and decide to test $H_0 \colon \mu = \mu_0$ vs. $H_0 \colon \mu \neq \mu_0$ where $\mu_0$ is some hypothesized value of $\mu$.

(a) If I set $\alpha = 0.05$, what is the critical value for my test?

**Solution:** 2

(b) Write a line of R code to calculate the critical value for my test if I set $\alpha = 0.3$.

**Solution:** `qnorm(1 - 0.3/2) = qnorm(1 - 0.15) = qnorm(0.85)`

214. Suppose that $X_1, \ldots, X_n \sim$ iid $\mathrm{N}(\mu, \sigma^2)$ where $\sigma$ is known, and you test $H_0 \colon \mu = \mu_0$ vs. $H_1 \colon \mu \neq \mu_0$ where $\mu_0$ is some hypothesized value of $\mu$.

(a) Say choose $\alpha = 0.05$ and reject. Would you still have rejected if you had instead chosen $\alpha = 0.1$? Explain.

**Solution:** Yes: the critical value for a test with $\alpha = 0.1$ is smaller than that for a test with $\alpha = 0.05$.

(b) Say you choose $\alpha = 0.01$ and fail to reject. Would you have failed to reject if you had instead chosen $\alpha = 0.05$? Explain.

**Solution:** There is not enough information to determine: the critical value for the $\alpha = 0.05$ test would indeed be smaller, but we don't know the value of the test statistic so we do not know if it exceeds this threshold.

For the following three questions (Alice and Bob), you may assume any tests and confidence intervals are based on a random sample from a normal population with known variance.

215. Alice constructs a 95% CI for $\mu$: $[-3, -1.5]$. Bob tests $H_0 \colon \mu = -1$ vs. $H_1 \colon \mu \neq -1$ with $\alpha = 0.05$ using the same dataset as Alice. Will he reject $H_0$? Explain.

**Solution:** Since $-1 \notin [-3, -1.5]$ he will reject.

216. Alice constructs a 90% CI for $\mu$: $[5.1, 6.7]$. Bob tests $H_0 \colon \mu = 6$ vs. $H_1 \colon \mu \neq 6$ with $\alpha = 0.1$ using the same dataset as Alice. Will he reject $H_0$? Explain.

**Solution:** Since $6 \in [5.1, 6.7]$ he will not reject.

217. Alice constructs a 95% CI for $\mu$: $[-0.5, 0.3]$. Bob tests $H_0 \colon \mu = 0$ vs. $H_1 \colon \mu \neq 0$ with $\alpha = 0.01$ using the same dataset as Alice. Will he reject $H_0$? Explain.

**Solution:** We cannot carry out Bob's test directly using Alice's interval since the two have a different value of $\alpha$. Instead we'll introduce a third character: Cheryl. Suppose that Cherly constructed a 99% confidence interval using the same data as Alice. To determine the result of Bob's test, we could simply check whether zero is contained in Cheryl's confidence interval. Unfortunately the question doesn't provide Cheryl's confidence interval. From our discussion of confidence intervals, however, we know that Alice's interval must be a *subset* of Cheryl's interval. Since 0 is in Alice's interval, this implies that it will also be in Cheryl's interval. Hence, Bob will fail to reject $H_0$.

218. Suppose that $X_1, \ldots, X_{25} \sim$ iid $\mathrm{N}(\mu, \sigma^2 = 4)$ and we want to test $H_0\colon \mu = 1$ vs. $H_1\colon \mu \neq 1$ with $\alpha = 0.05$.

(a) For what range of values for $\bar{X}$ would we *fail to reject* $H_0$?

**Solution:** Our test statistic is $T_n = 5(\bar{X} - 1)/2$ and we will reject if $|T_n| > 2$. Hence, we will *fail to reject* when $|T_n| \leq 2$, i.e. when

$$\begin{aligned} -2 \leq 5(\bar{X} - 1)/2 &\leq 2 \\ -4/5 \leq \bar{X} - 1 &\leq 4/5 \\ -4/5 + 1 \leq \bar{X} &\leq 4/5 + 1 \\ 0.2 \leq \bar{X} &\leq 1.8 \end{aligned}$$

(b) For what range of values for $\bar{X}$ would we *reject* $H_0$?

**Solution:** $\bar{X} > 1.8$ or $\bar{X} < 0.2$

219. Let $X_1, \ldots, X_{25} \sim$ iid $\mathrm{N}(\mu, \sigma^2 = 100)$. We want to test $H_0\colon \mu = -1$ against $H_1\colon \mu \neq -1$ using the fact that $\sigma^2$ is known.

(a) Suppose $\bar{x} = -0.6$. Calculate the value of our test statistic.

**Solution:**

$$\left| \frac{\bar{x} - (-1)}{\sqrt{100/25}} \right| = \frac{|\bar{x} + 1|}{2} = 0.5 \times |\bar{x} + 1| = 0.2$$

(b) Continuing from the preceding part, suppose that we set $\alpha = 0.1$. Without consulting R, determine whether we should reject the null hypothesis. Explain.

**Solution:** We don't have `qnorm`(0.95) memorized, but the critical value is definitely larger than 1 since since 68% of the probability for a standard normal is between -1 and 1. Since the test statistic is less than 1, we fail to reject.

220. Let that $X_1, \ldots, X_n \sim$ iid $\mathrm{N}(\mu, \sigma^2)$ where $\sigma$ is known and suppose that you want to test $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu \neq \mu_0$ where $\mu_0$ is some hypothesized value of $\mu$.

(a) If $\alpha = 0.1$, what line of R code would you use to calculate the critical value?

**Solution:** `qnorm(0.95)`

(b) Suppose that your test statistic is 1.2. Write out the line of R code that you would use to calculate the p-value for the test.

**Solution:** `2 * (1 - pnorm(1.2))`

221. Define the term *p-value*, and explain how you can use one to carry out a hypothesis test with a significance level of $\alpha$.

222. For each part, indicate whether we would *reject* or *fail to reject* the null hypothesis.

(a) Alice chose a significance level of 0.05 and the p-value for her test was 0.95.

**Solution:** Fail to reject.

(b) Bob chose a significance level of 0.1 and the p-value for his test was 0.01.

**Solution:** Reject.

223. Mark each statement as True or False, and if False correct it:

(a) "The smaller the p-value, the weaker the evidence against $H_0$."

**Solution:** False: the smaller the p-value, the *stronger* the evidence against $H_0$.

(b) "The larger the p-value, the larger the size of the effect we have discovered."

**Solution:** False: a p-value only tells us the strength of evidence against $H_0$, it tells us nothing about the size of an effect.

224. Lindsay observes a random sample of size 400 from a normal population with variance 4 and uses this information to test $H_0 : \mu = 50$ vs. $H_1 : \mu \neq 50$. Her sample mean is 50.3.

(a) Without using R, what is the p-value for Lindsay's test? If $\alpha = 0.01$ would she reject the null hypothesis?

**Solution:** Lindsay's standard error is $\sqrt{4/400} = 0.1$ so her test statistic is $|(50.3 - 50)/0.1| = 3$. The probability that a standard normal RV takes a value within $[-3, 3]$ is 0.997 so her p-value is 0.003. Hence, she will reject the null hypothesis.

(b) Explain why Lindsay's p-value is so small even though 50.3 is very close to 50.

**Solution:** The test statistic depends on two things: how close $\bar{x}$ is to 50, and the size of the standard error. In Lindsay's example, 50.3 is very close to 50 but the standard error is extremely small: 0.1. This makes the test statistic large, resulting in a small p-value.

# Lecture #20 – Hypothesis Testing III

225. According to the CDC, 32% of US births in 2017 were by Caesarian delivery. Prajwal decides to carry out a study to determine whether Philadelphia hospitals have a *higher* rate of Caesarian delivery that the US as a whole. State his null and alternative hypotheses in words.

**Solution:** The null hypothesis is that the rate of Caesarian deliveries in Philadelphia equals 0.32. The alternative hypothesis could either be: (i) the rate *does not* equal 0.32 or (ii) the rate is *greater than* 0.32, depending on whether the students interpret this as a one-sided or two-sided problem. The wording is intended to suggest one-sided, but either is fine as an answer.

226. Suppose that you wish to test $H_0\colon \mu = 4$ with a significance level $\alpha = 0.05$. Your sample mean equals 4.8 with a standard error of 0.3.

(a) If the alternative hypothesis is $H_1\colon \mu < 4$, would you reject $H_0$?

**Solution:** No: for this alternative we should only reject $H_0$ when the sample mean is sufficiently far *below* 4.

(b) If the alternative hypothesis is $H_1\colon \mu > 4$, would you reject $H_0$?

**Solution:** Yes. The test statistic is $8/3 \approx 2.6$, and the critical value for this test is less than 2, the critical value for a two-sided test.

227. Suppose that you wish to test $H_0\colon \mu = -1$ with a significance level of $\alpha = 0.05$. Your sample mean equals $-2.1$ with a standard error of 0.6. To help you answer this question, note that `qnorm(0.95)` $\approx 1.64$.

(a) If the alternative hypothesis is $H_1\colon \mu \neq -1$ would you reject $H_0$?

**Solution:** The test statistic is $|-2.1-(-1)|/0.6 \approx 1.83$, but the critical value for the two-sided test is approximately 2. Hence we would fail to reject.

(b) If the alternative hypothesis is $H_1\colon \mu < -1$ would you reject $H_0$?

**Solution:** The test statistic is $[-2.1-(-1)]/0.6 \approx -1.83$. By the symmetry of the normal distribution `qnorm(0.05)` $\approx -1.64$ is the critical value for the one-sided test. Hence we should reject $H_0$.

228. Suppose I test $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu > \mu_0$ with $\alpha = 0.2$. Write a line of R code to calculate the critical value for my test.

**Solution:** `qnorm(0.8)`

229. Suppose I test $H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu < \mu_0$ with $\alpha = 0.15$. Write a line of R code to calculate the critical value for my test.

**Solution:** `qnorm(0.15)`

230. Suppose I test $H_0\colon \ \colon \mu = 4$ vs. $H_1\colon \mu > 4$. My sample mean equals 4.8 with a standard error of 0.3. Write a line of R code to calculate the p-value of my test.

**Solution:** The test statistic is $(4.8-4)/0.3 \approx 2.67$ Hence the one-sided p-value is `1 - pnorm(2.67)`.

231. Suppose I test $H_0\colon \ \colon \mu = -1$ vs. $H_1\colon \mu < -1$. My sample mean equals $-2.1$ with a standard error of 0.6. Write a line or R code to calculate the p-value of my test.

**Solution:** The test statistic is $[-2.1-(-1)]/0.6 \approx -1.83$. Hence the one-sided p-value is `pnorm(-1.83)`.

232. When and why would we want to use a one-sided rather than a two-sided test?

233. Danae observes a random sample of 25 observations from a normal population with unknown mean $\mu$ and known variance $\sigma^2 = 9$. She wants to test $H_0\colon \mu = 0$ with $\alpha = 0.05$. Danae knows that `qnorm(0.95)` $\approx 1.64$, so she decides to use the following procedure. First she will look at the sample mean $\bar{X}$. If $\bar{X}$ is positive, then she will use the rule "reject $H_0$ if $5\bar{X}/3 > 1.64$." If instead $\bar{X}$ is negative, she will use the rule "reject $H_0$ if $5\bar{X}/3 < -1.64$." What is the problem with Danae's procedure? Explain.

**Solution:** Danae's procedure is equivalent to carrying out a *two-sided test* with critical value 1.64. But since `qnorm(0.95)` $\approx 1.64$ it follows from the symmetry of the normal distribution that `pnorm`$(1.64) -$ `pnorm`$(-1.64) \approx 0.9$. Hence, the significance value of her test is actually 0.1 rather than 0.05. This shows that you can't decide your alternative hypothesis after looking at the data; doing so is equivalent to using the wrong critical value.

234. Don wants to know whether Philadelphia cab drivers are less likely to accept fares from African American males compared to white males. State his null and alternative hypotheses in words.

**Solution:** His null hypothesis is that there is no difference in the rate at which cab drivers accept fares from African American versus while males. His alternative hypothesis could either be that there *is a difference* (two-sided), or that they are *less likely* to accept fares from African American males (one-sided).

235. Suppose that $X_1, \ldots, X_5 \sim$ iid N$(1, 4)$ independently of $Y_1, \ldots, Y_{20} \sim$ iid N$(-1, 24)$. Write a line of R code to calculate $P(\bar{X} - \bar{Y} > 0)$.

**Solution:** We have $\bar{X} \sim N(1, 4/5)$ independently of $\bar{Y}(-1, 6/5)$. Thus, it follows that $\bar{X} - \bar{Y} \sim N(2, 2)$ and accordingly

$$P(\bar{X} - \bar{Y} > 0) = P\left(\frac{\bar{X} - \bar{Y} - 2}{\sqrt{2}} > \frac{-2}{\sqrt{2}}\right) = P(Z > -\sqrt{2})$$

where $Z \sim N(0, 1)$. Therefore the desired probability is `1 - pnorm(-sqrt(2))`.

236. Suppose we observe two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from populations with unknown means $\mu_X$ and $\mu_Y$. We wish to test the null hypothesis that $\mu_X = \mu_Y$ against the two-sided alternative at the 5% significance level. Suppose $\bar{x} = 4.1$

and $\bar{y} = 2.7$ and we reject the null. What is the smallest possible value that the standard error of $\bar{X} - \bar{Y}$ could have been in this example?

**Solution:** Since we rejected at the 5% level, the test statistic must have been at least 2. Since the absolute difference of means is 1.4, this means that the standard error cannot have been larger than 0.7.

# Lecture #21 – Hypothesis Testing IV

237. Alejandro wants to test whether a penny is equally likely to come up heads or tails when it is *spun* on its side against the two-sided alternative. He spins a penny 25 times, and it comes up tails 20 times.

(a) State Alejandro's null and alternative hypotheses.

**Solution:** Let $p = P(\text{Heads})$. Then the null is $H_0 \colon p = 1/2$ and the alternative is $H_1 \colon p \neq 1/2$.

(b) Calculate the test statistic. Would Alejandro reject at the 1% significance level?

**Solution:**

$$\text{Test Statistic} = \frac{0.8 - 0.5}{\sqrt{0.5 \times (1 - 0.5)/25}} = 3$$

Since $P(-3 \leq Z \leq 3) = 0.997$ if $Z \sim N(0, 1)$, the critical value for a two-sided test with $\alpha = 0.01$ is *less than* 3. Therefore Alejandro will reject $H_0$.

238. Liz and Prof. DiTraglia are shooting free throws at the Palestra. Prof. DiTraglia takes 10 shots and makes 5 of them; Liz takes 20 shots and makes 15 of them. Test the null hypothesis that Prof. DiTraglia is just as good at making free throws as Liz against the two-sided alternative at the 5% significance level.

**Solution:** Let $p$ be Prof. DiTraglia's probability of success and $q$ be Liz's probability of success. We will test $H_0 \colon p = q$ against $H_1 \colon p \neq q$. Because $\alpha = 0.05$, our critical value is 2. First we calculate the pooled standard error estimate. The pooled proportion estimate is

$$\widehat{\pi} = \frac{5 + 15}{10 + 20} = \frac{20}{30} = 2/3$$

and hence

$$\widehat{SE}_{\text{pooled}} = \sqrt{\frac{2}{3} \times \frac{1}{3} \times \left(\frac{1}{10} + \frac{1}{20}\right)} = \sqrt{\frac{2}{9} \times \frac{3}{20}} = 1/\sqrt{30}$$

Finally we calculate the test statistic:

$$\text{Test Statistic} = \left|\frac{0.5 - 0.75}{1/\sqrt{30}}\right| = \frac{\sqrt{30}}{4} \approx 1.4$$

Hence we would fail to reject $H_0$.

239. Compare and contrast *statistical significance* and *practical importance*.

240. Suppose you test 500 null hypotheses with significance level equal to 0.05. Unbeknownst to you, all of these null hypotheses are in fact *true*. On average, how many will you reject?

**Solution:** $500 \times 0.05 = 25$

241. Write an R function called `prop_test` to calculate the p-value for a two-sided test of the null hypothesis $H_0 \colon p = p_0$ based on a sample proportion $\widehat{p}$. Your test should take three input arguments: the estimated sample proportion `phat`, the sample size `n`, and the hypothesized value `p0`. It should return the two-sided p-value for the test.

**Solution:**

```
prop_test <- function(phat, n, p0) {
  SE <- sqrt(p0 * (1 - p0) / n)
  test_stat <- abs(phat - p0) / SE
  p_value <- 2 * (1 - pnorm(test_stat))
  return(p_value)
}
```

242. Suppose I have an R dataframe called `econ103` with two columns: `grade` is a numeric vector containing each student's course grade, and `class` is a character vector indicating a student's class standing (Freshman, Sophomore, Junior, or Senior). Write R code that uses `econ103` to construct two vectors: `x` should contain the non-missing grades for Sophomores and `y` should contain the non-missing grades for Juniors in the class.

**Solution:**

```
grades <- na.omit(econ103$grades)
x <- subset(grades, class == 'Sophomore')
y <- subset(grades, class == 'Junior')
```

243. The Fibonacci sequence has, so far as I know, nothing to do with statistics but provides a nice example of using for loops in R. The sequence is defined as follows: $F_1 = 1, F_2 = 1$, and $F_i = F_{i-1} + F_{i-2}$ for $i \geq 3$. In other words: $1, 1, 2, 3, 5, 8, 13, 21, 34, 55 \ldots$ and so on. Write R code to calculate the first 20 terms of the Fibonacci sequence $(F_1, F_2, \ldots, F_{20})$ and store them in a vector called `fib`.

**Hint #1:** create an "empty" vector and fill it with values (see slide 21).

**Hint #2:** fill in the first two values of the sequence *before starting your loop.*

**Hint #3:** check your code by running it in R.

**Solution:**

```
fib <- rep(NA, 20)
fib[1] <- 1
fib[2] <- 1
for(i in 3:20) {
  fib[i] <- fib[i - 1] + fib[i - 2]
}
fib
```

# Lecture #22 – Regression II

The following five questions refer to the population regression model: $Y = \beta_0 + \beta_1 X + \varepsilon$.

244. Write $\beta_1$ in terms of the appropriate features of the distributions of $X$ and $Y$.

**Solution:** $\beta_1 = \dfrac{Cov(X, Y)}{Var(X)}$

245. Write $\beta_0$ in terms of $\beta_1$ and the appropriate features of the distributions of $X$ and $Y$.

**Solution:** $\beta_0 = E[Y] - \beta_1 E[X]$

246. Use the expression for $\beta_0$ from above to prove that $E[\varepsilon] = 0$.

**Solution:**

$$\begin{aligned} E[\varepsilon] &= E[Y - \beta_0 - \beta_1 X] = E[Y] - \beta_0 - \beta_1 E[X] \\ &= E[Y] - (E[Y] - \beta_1 E[X]) - \beta_1 E[X] = 0 \\ &= 0 \end{aligned}$$

247. Use the expression for $\beta_1$ from above to prove that $Var(\varepsilon) = Var(Y) - Cov(X,Y)^2/Var(X)$.

**Solution:**

$$\begin{aligned} Var(\varepsilon) &= Var\left(Y - \beta_0 - \beta_1 X\right) = Var(Y - \beta_1 X) \\ &= Var(Y) + \beta_1^2 Var(X) - 2\beta_1 Cov(X,Y) \\ &= Var(Y) + \frac{Cov(X,Y)^2}{Var(X)^2} Var(X) - 2\frac{Cov(X,Y)}{Var(X)} Cov(X,Y) \\ &= Var(Y) - \frac{Cov(X,Y)^2}{Var(X)} \end{aligned}$$

248. Use $E[\varepsilon] = 0$ and the expressions for $\beta_0$ and $\beta_1$ from above to prove that $Cov(X, \varepsilon) = 0$.

**Solution:** By the shortcut formula, and the fact that $E[\varepsilon] = 0$,

$$Cov(X, \varepsilon) = E[X\varepsilon] - E[X]E[\varepsilon] = E[X\varepsilon]$$

Now, substituting $\varepsilon = Y - \beta_0 - \beta_1 X$ and $\beta_0 = E[Y] - \beta_1 E[X]$,

$$\begin{aligned} E[X\varepsilon] &= E[X(Y - \beta_0 - \beta_1 X)] = E[XY] - \beta_0 E[X] - \beta_1 E[X^2] \\ &= E[XY] - (E[Y] - \beta_1 E[X])\, E[X] - \beta_1 E[X^2] \\ &= (E[XY] - E[X]E[Y]) - \beta_1 \left(E[X^2] - E[X]^2\right) \\ &= Cov(X, Y) - \beta_1 Var(X) \end{aligned}$$

Therefore, substituting $\beta_1 = Cov(X, Y)/Var(X)$ we have

$$\begin{aligned} Cov(X, \varepsilon) &= Cov(X, Y) - \beta_1 Var(X) = Cov(X, Y) - \frac{Cov(X, Y)}{Var(X)} Var(X) \\ &= Cov(X, Y) - Cov(X, Y) = 0 \end{aligned}$$

249. Suppose we run a linear regression of the form $Y_i = \beta_0 + \varepsilon_i$ where $Y_i$ is student i's grade on midterm #2. In terms of the sample data, what will be our estimate of $\beta_0$?

**Solution:** It will be the sample mean grade on midterm #2.

250. Suppose we run a linear regression of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $Y_i$ is student i's grade on midterm #2 and $X_i$ is a dummy variable that takes the value 1 if student i is an Economics major. In terms of our sample data, what will be our estimates of the regression parameters $\beta_0$ and $\beta_1$?

**Solution:** Our estimate of $\beta_0$ will be the sample mean grade on midterm #2 for *non-Econ majors*, while our estimate of $\beta_1$ will be the difference of sample means on midterm #2: Econ majors minus non-Econ majors.

251. Suppose we run a linear regression of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$ where $Y_i$ is a student's grade on midterm #2 and $X_i$ is her grade on midterm #1. What is the interpretation of $\beta_0$ and $\beta_1$ in this example?

**Solution:** $\beta_0$ is the grade we would predict on midterm #2 for someone who scored zero on midterm #1: this is a meaningless quantity. If Alice scored $x$ points on midterm #1 and Bob scored $x + 1$ points, then we would predict that Bob will score $\beta_1$ points better on midterm #2 then Alice.

252. Define the following terms:

(a) Residual standard deviation.

(b) R-squared

253. Define the following terms:

(a) Fitted value

(b) Residual

The following 4 questions rely on a dataset called `mother_child_weight.csv` available from `http://ditraglia.com/econ103/`. The dataset contains two columns: `mother` gives a mother's weight during pregnancy (kg) while `child` give's her child's birthweight (kg).

254. Write R code to accomplish the following tasks:

(i) Download the mother-child dataset and store it in a dataframe called `weight`.

(ii) Run a linear regression using mother's pregnancy weight to predict her child's birthweight and store the results in an object called `reg`.

(iii) Output the regression results.

**Solution:** This solution assumes that you have already loaded the `display` command from my website. If you haven't you'll need to do so first.

```
data_url <- 'http://ditraglia.com/econ103/mother_child_weight.csv'
weight <- read.csv(data_url)
reg <- lm(child ~ mother, data = weight)
display(reg)
```

The next 3 questions rely on the regression results from the preceding question:

```
            coef.est coef.se
(Intercept) 1.50     0.63
mother       0.03     0.01
---
n = 25, k = 2
residual sd = 0.40, R-Squared = 0.27
```

255. Answer each part:

(a) Interpret the estimated intercept from the preceding set of regression results.

> **Solution:** The estimated intercept 1.5 says that we would predict a birthweight of 1.5kg for a child whose mother weighed 0kg during pregnancy: this is a totally meaningless quantity!

(b) Consider two mothers whose weight during pregnancy differs by one kg. How much more would would we predict that the child of the heavier mother will weigh at birth?

> **Solution:** 0.03 kg heavier, i.e. 30 grams heavier

256. Answer each part:

(a) What is the sample correlation between a mother's weight during pregnancy and her child's birthweight?

> **Solution:** $\sqrt{0.27} \approx 0.52$

(b) Approximately how accurately does a mother's weight during pregnancy predict her child's birthweight?

> **Solution:** To an accuracy of about 0.4 kg, i.e. 400 grams.

257. Answer each part:

(a) Construct an approximate 95% CI for the population regression slope $\beta_1$ based on the preceding set of regression results.

> **Solution:** $0.03 \pm 0.02$

(b) Suppose you wanted to test $H_0 \colon \beta_1 = 0$ against the two-sided alternative at the 1% significance level. Would you reject the null hypothesis?

> **Solution:** Yes we would reject $H_0$: the observed test statistic is 3 and the critical value for a two-sided test with $\alpha = 0.01$ is between 2 and 3.

258. Consider a linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$. Each $i$ is a professor who teaches introductory statistics: $Y_i$ is Professor $i$'s average student rating, $X_{1i}$ is the average grade in Professor $i$'s course, and $X_{2i}$ is a dummy variable that takes the value 1 if Professor $i$ is female.

(a) Suppose you learn that $\beta_1$ is positive. Explain what this means.

> **Solution:** If we consider two statistics professors of the same sex (i.e. both male or both female), we would predict that the professor who awards higher grades will have higher student ratings.

(b) Suppose you learn that $\beta_2$ is negative. Explain what this means.

> **Solution:** If we consider two statistics professors who give the same average grade to students, we would predict that the female professor will receive lower average student ratings.

# Lecture #23 – Regression III

All of these questions refer to a dataframe called `houses`. The columns are as follows: `price` is the sale price of a house in US dollars, `brick` is a dummy variable that equals one if the house is made of brick, and `sqft` is the size of the house in square feet.

259. For each part, provide the R code needed to run the appropriate regression.

(a) Run a regression predicting sale price from house size allowing a different intercept for brick and non-brick houses, but constraining the slope for both kinds of house to be the same.

> **Solution:**
> ```
> lm(price ~ brick + sqft, houses)
> ```

(b) Run a regression predicting sale price from house size allowing a different slope *and* intercept for brick and non-brick houses.

> **Solution:**
> ```
> lm(price ~ brick + sqft + brick:sqft, houses)
> ```

260. I ran the regression $\texttt{price}_i = \beta_0 + \beta_1 \times \texttt{brick}_i + \beta_2 \times \texttt{sqft}_i + \varepsilon_i$. My estimates were:

$$\widehat{\beta}_0 = -9000 \quad \widehat{\beta}_1 = 24000 \quad \widehat{\beta}_2 = 80$$

(a) Consider two brick houses that differ in size by 500 square feet. What difference sale price would we predict between these two houses?

> **Solution:** We would predict that the larger house will sell for $80 \times 500 = \$4000$ more than the smaller one.

(b) Consider two houses: both are 2000 square feet, but one is brick while the other is not. What difference in sale prices would we predict between these two houses?

> **Solution:** We would predict that the brick house will sell for \$24,000 more than the non-brick house.

261. I ran the regression $\texttt{price}_i = \beta_0 + \beta_1 \times \texttt{brick}_i + \beta_2 \times \texttt{sqft}_i + \varepsilon_i$. My estimates and standard errors were:

$$\begin{array}{lll} \widehat{\beta}_0 = -9000 & \widehat{\beta}_1 = 24000 & \widehat{\beta}_2 = 80 \\ SE(\widehat{\beta}_0) = 17000 & SE(\widehat{\beta}_1) = 4000 & SE(\widehat{\beta}_2) = 10 \end{array}$$

(a) Construct an approximate 95% CI for the price premium for a brick house.

> **Solution:** There is a substantial price premium for brick houses: the confidence interval is $24000 \pm 8000$.

(b) Is there convincing evidence that larger houses command higher prices? Explain.

> **Solution:** Yes. The test statistic for $H_0\colon \beta_2 = 0$ is $80/10 = 8$. Regardless of whether we are carrying out a one-sided or two-sided test, the p-value would be far below 0.003. Another way of saying the same thing is that a 99.7% confidence interval for the per-square-foot premium is $80 \pm 30$. This interval is very far from zero.

262. I ran the regression $\texttt{price}_i = \beta_0 + \beta_1 \times \texttt{brick}_i + \beta_2 \times \texttt{sqft}_i + \beta_3 \times \texttt{brick}_i \times \texttt{sqft}_i + \varepsilon_i$. My estimates were:

$$\widehat{\beta}_0 = 5000 \quad \widehat{\beta}_1 = -30000 \quad \widehat{\beta}_2 = 60 \quad \widehat{\beta}_3 = 30$$

(a) Suppose we compare two brick houses that differed in size by 500 square feet. Based on these results, what difference in prices would we predict between these two houses?

**Solution:** The estimated slope for brick houses is $\widehat{\beta}_2 + \widehat{\beta}_3 = 90$. Hence, we would predict that the larger house will sell for $90 \times 500 = \$4500$ more than the smaller one.

(b) Suppose we compare two houses: both are 2000 square feet, but one is brick while the other is not. What difference in sale prices would we predict between these two houses?

**Solution:** Our prediction for the brick house is

$$(\widehat{\beta}_0 + \widehat{\beta}_1) + (\widehat{\beta}_2 + \widehat{\beta}_3) \times \texttt{sqft} = -25000 + 90 \times 2000 = 155000$$

while our prediction for the non-brick house is

$$\widehat{\beta}_0 + \widehat{\beta}_2 \times \texttt{sqft} = 5000 + 60 \times 2000 = 125000$$

Hence, we predict that the brick house will sell for \$30,000 more.

263. I ran the regression $\texttt{price}_i = \beta_0 + \beta_1 \times \texttt{brick}_i + \beta_2 \times \texttt{sqft}_i + \beta_3 \times \texttt{brick}_i \times \texttt{sqft}_i + \varepsilon_i$. My estimates and standard errors were:

$$\begin{array}{cccc} \widehat{\beta}_0 = 5000 & \widehat{\beta}_1 = -30000 & \widehat{\beta}_2 = 60 & \widehat{\beta}_3 = 30 \\ SE(\widehat{\beta}_0) = 20000 & SE(\widehat{\beta}_1) = 40000 & SE(\widehat{\beta}_2) = 10 & SE(\widehat{\beta}_3) = 40 \end{array}$$

Is there compelling evidence that brick houses command a higher per-square-foot price premium? Discuss briefly.

**Solution:** Our estimate for the difference of slopes (brick minus non-brick) is 30, but the approximate 95\$ confidence interval is $30 \pm 80$ which comfortably includes zero and indeed some fairly large negative values. Even a 68% interval ($30 \pm 40$) would comfortably include zero. So while our estimates are suggestive of a difference, the evidence is not particularly compelling.