Math 58B – Introduction to Biostatistics

Spring 2018

Jo Hardin

iClicker Questions

to go with **Introductory Statistics with Randomization and Simulation**, Diez, Barr, & Centinkaya-Rundel

1. If 16 infants with **no genuine preference** choose 16 toys, what is the most likely number of “helping” toys that will be chosen?
2. 4
3. 7
4. 8
5. 9
6. 10

2. What percent of the time will the simulation produce exactly 8 heads?

1. 0-15%
2. 16-30%
3. 31-49%
4. 50%
5. 51-100%

3. What if we flipped a coin 160 times? What percent of the time will the simulation flip exactly 80 heads?

1. 0-15%
2. 16-30%
3. 31-49%
4. 50%
5. 51-100%

4. Is our actual result of 14 (under the null model)

1. Very surprising
2. Somewhat surprising
3. Not very surprising

5. If gender and promotion are independent variables, how many of each gender (out of 24 each) would you expect to have been promoted?

1. 17 or 18 of each
2. 12 of each
3. 6 or 7 of each
4. 21 men and 14 women
5. 3 men and 10 women

6. Is the actual result of a difference in 29% (under the null model):

1. Very surprising
2. Somewhat surprising
3. Not very surprising

Professor Hardin’s office hours are:

(a) Mon & Wed mornings

(b) Mon & Wed afternoons

(c) Mon & Wed 2:30-5

(d) Mon morning and Wed afternoon

(e) Mon afternoon and Wed morning

7. Hypothesis: the number of hours that grade-school children spend doing homework predicts their future success on standardized tests.

(a) null, one sided

(b) null, two sided

(c) alternative, one sided

(d) alternative, two sided

8. Hypothesis: king cheetahs on average run the same speed as standard spotted cheetahs.

(a) null, one sided

(b) null, two sided

(c) alternative, one sided

(d) alternative, two sided

9. Hypothesis: the mean length of African elephant tusks has changed over the last 100 years.

(a) null, one sided

(b) null, two sided

(c) alternative, one sided

(d) alternative, two sided

10. Hypothesis: the risk of facial clefts is equal for babies born to mothers who take folic acid supplements compared with those from mothers who do not.

(a) null, one sided

(b) null, two sided

(c) alternative, one sided

(d) alternative, two sided

11. Hypothesis: caffeine intake during pregnancy affects mean birth weight.

(a) null, one sided

(b) null, two sided

(c) alternative, one sided

(d) alternative, two sided

1. What are the observational units for your individual study?
2. Color of the candy
3. Piece of candy
4. Cup of candy
5. Hershey’s company
6. Proportion that are orange
7. What are the observational units for the class compilation (dotplot)?
8. Color of the candy
9. Piece of candy
10. Cup of candy
11. Hershey’s company
12. Proportion that are orange
13. How does the sampling distribution for the sample proportion change as n changes (for a fixed p)?
14. The spread changes
15. The symmetry changes
16. The center changes
17. The shape changes
18. How does the sampling distribution change as p changes (for a fixed n)?
19. The spread changes
20. The symmetry changes
21. The center changes
22. The shape changes

16. The Central Limit Theorem says that the distribution of will be approximately normal with what center:

(a)

(b) p

(c) 0.5

(d) 1

(e)

17. The Central Limit Theorem applies as long as (technical conditions):

(a) the trials are independent

(b) n is fixed

(c) the probability of success is constant for each trial

(d) each trial is a success or failure

(e) np ≥ 10 AND n(1-p) ≥10

(see page 124 in your text)

18. The standardized score (z-score) counts:

(a) the number of standard deviations from the mean

(b) the number of standard deviations above the mean

(c) the number of standard deviations below the mean

(d) the distance from the mean

(e) the distance from the standard deviation

19. If the normal distribution is a good model, we would expect our z-scores to be:

(a) within ± 2 of the mean

(b) within ± 3 of the mean

(c) within ± 2

(d) within ± 3

20. Assume n = 100 and p= 0.8

(

What is the largest **reasonable** distance between and p ?

That is, we would expect and p to be no more than \_\_\_\_\_ apart.

(a) 0.04

(b) 0.08

(c) 0.12

(d) 0.16

(e) 0.24

21. Assume n = 100 and p= 0.8 (

Which statement is true?

(a) 95% of are between (0.76, 0.84)

(b) 95% of are between (0.72, 0.88)

(c) 95% of are between (0.68, 0.92)

(d) 95% of p are between (0.76, 0.84)

(e) 95% of p are between (0.72, 0.88)

22. If you want a 90% confidence interval for p, your z\* multiplier should be

(a) less than 1

(b) less than 2 (but greater than 1)

(c) equal to 2

(d) greater than 2 (but less than 3)

(e) greater than 3

23. What is the difference between Z\* and a Z score?

1. Z score comes from the data, Z\* and is a pre-defined unit of measurement.
2. Z\* comes from the data, and Z score is a pre-defined unit of measurement
3. Z score assumes the null hypothesis is true and Z\* doesn’t.
4. Z\* assumes the null hypothesis is true, and Z score doesn’t

24. Let’s say we are making confidence intervals (not doing a hypothesis test), what is your best **guess** for SE()?

(a)

(b)

(c)

(d)

(e)

25. The following is a correct interpretation of the CI:

95% confident that this interval includes the sample proportion who believe that the global poverty rate has doubled.

1. TRUE
2. FALSE

26. The following is a correct interpretation of the CI:

If researchers were to select a new sample of 1005 adult Americans, then we’re 95% confident that between 56% and 62% of those people would answer “doubled” to this question.

1. TRUE
2. FALSE
3. Let’s say that the null hypothesis (e.g., p=0.47) is TRUE. My level of significance is 0.03 (reject if p-value < 0.03).

How often will I reject the null hypothesis?

(a) 1 % of the time

(b) 3% of the time

(c) 5 % of the time

(d) 95% of the time

(e) 97% of the time

What does “of the time” mean???

1. Let’s say that the null hypothesis (e.g., p=0.47) is TRUE. My level of significance is 0.03.

How often will p be in a 97% confidence interval?

(a) 1 % of the time

(b) 3% of the time

(c) 5 % of the time

(d) 95% of the time

(e) 97% of the time

What does “of the time” mean???29. As we’ve seen with the applet, about 5% of all 95% intervals fail to capture the actual value of the population parameter. Do you think the alien just got a “red” interval?

1. Yes
2. No
3. Would it be reasonable for the alien to conclude, with 95% confidence, that between 16.5% and 33.5% of US senators in the year 2019 self-identify as female?
4. Yes
5. No

31. Suppose that you take many samples (i.e., thousands) from a population and graph the distribution of the resulting sample statistics. If the distribution of sample **statistics is centered around the value of the population parameter** then is the sampling distribution of the statistic unbiased?

(a) Yes, definitely

(b) Maybe or maybe not

(c) No, definitely not

32. Suppose that you take many samples (i.e., thousands) from a population and graph the distribution of the resulting sample statistics. If the distribution of sample statistics **appears to be normally distributed** then is the sampling distribution of the statistic unbiased?

(a) Yes, definitely

(b) Maybe or maybe not

(c) No, definitely not

33. Suppose that you take many random samples (i.e., thousands) from a population and graph the distribution of the resulting sample statistics. If **most of the sample statistics are close to the value of the population parameter**, then is the sampling distribution of the statistic unbiased?

(a) Yes, definitely

(b) Maybe or maybe not

(c) No, definitely not

34. Suppose that you take many random samples (i.e., thousands) from a population and graph the distribution of the resulting sample statistics**. If the sampling method is biased, then will increasing the sample size reduce the bias?**

(a) Yes, definitely

(b) Maybe or maybe not

(c) No, definitely not

35. Suppose your sample is 10 times larger. The SE of your statistic:

1. increases
2. stays the same
3. decrease

36. Suppose your population is 10 times larger. The SE of your statistic:

1. increases
2. stays the same
3. decreases

37. Suppose your sample is 10 times larger. The variability of your data:

1. increases
2. stays the same
3. decreases

38. Which of the following are advantages of studies with a **larger sample size** (more than one may be right).

(a) Better represent the population (reduce sampling bias)

(b) More accurately estimate the parameter

(c) Decrease sampling variability of the statistics

(d) Make simulation results more accurate for the theoretical results given the sampling method at hand.

(e) other?

39. In conducting a simulation analysis, why might we take a **larger number of samples**? (more than one answer may be correct).

(a) Better represent the population (reduce sampling bias)

(b) More accurately estimate the parameter

(c) Decrease sampling variability of the statistic

(d) Make simulation results more accurate for the theoretical results given the sampling method at hand.

(e) other?

1. Let’s say that the null hypothesis (e.g., p=0.47) is TRUE. My level of significance is 0.03 (reject if p-value < 0.03).

How often will I fail to reject the null hypothesis?

(a) 1 % of the time

(b) 3% of the time

(c) 5 % of the time

(d) 95% of the time

(e) 97% of the time

What does “of the time” mean???

1. Let’s say that the null hypothesis (e.g., p=0.47) is TRUE. My level of significance is 0.03.

How often will p be in a 97% confidence interval?

(a) 1 % of the time

(b) 3% of the time

(c) 5 % of the time

(d) 95% of the time

(e) 97% of the time

What does “of the time” mean???42. How many hits out of 20 at bats would make you believe him?

1. 5
2. 6
3. 7
4. 8
5. 9

43.Type I error is

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

44. Type II error is

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

45.Power is the probability that:

1. We give him a raise when he deserves it.
2. We don’t give him a raise when he deserves it.
3. We give him a raise when he doesn’t deserve it.
4. We don’t give him a raise when he doesn’t deserve it.

46. The player is more worried about

1. A type I error
2. A type II error

47. The manager is more worried about

1. A type I error
2. A type II error
3. Increasing your sample size
4. Increases your power
5. Decreases your power
6. Making your significance level more stringent (α smaller)
7. Increases your power
8. Decreases your power
9. A more extreme alternative
10. Increases your power
11. Decreases your power
12. Let’s say that the null hypothesis (e.g., p=0.47) is TRUE. My level of significance is 0.03 (reject if p-value < 0.03).

How often will I reject the null hypothesis?

(a) less than 3% of the time

(b) 3% of the time

(c) more than 3% of the time (less than 97%)

(d) 97% of the time

(e) more than 97% of the time

What does “of the time” mean???

1. Let’s say that the null hypothesis (e.g., p=0.47) is FALSE. My level of significance is 0.03 (reject if p-value < 0.03).

How often will I reject the null hypothesis?

(a) less than 3% of the time

(b) 3% of the time

(c) more than 3% of the time (less than 97%)

(d) 97% of the time

(e) more than 97% of the time

What does “of the time” mean???

Math 58:

53. Given the 5 question quiz set-up, how many different ways can you get 1 success?

1. 1
2. 2
3. 3
4. 4
5. 5

Math 58:

1. Given the 5 question quiz set-up, how many different ways can you get 2 successes?
2. 5
3. 7
4. 10
5. 15
6. 20

Math 58:

1. What is the probability that

X ≥ 3? Find: P(X ≥ 3)



1. 0.016
2. greater than 0.016
3. less than 0.016
4. 0.984
5. greater than 0.984
6. What is the probability that

X ≥ 3? Find: P(X ≥ 3)



1. 0.104
2. greater than 0.104
3. less than 0.104
4. 0.896
5. greater than 0.896

Math 58

1. Consider the urban legend about the trip and the tire blow out.

Which tire?

1. Front left
2. Front right
3. Back left
4. Back right

Math 58

1. Shuffle a deck of 52 cards. Turn over the top card. Replace the card. Repeat the process 5 times.

\*\* observe the card you chose

Is this a binomial experiment?

1. Yes
2. No, more than 2 outcomes
3. No, trials are not independent
4. No, p changes for each trial
5. No, n is not fixed

Math 58

1. Shuffle a deck of 52 cards. Turn over the top card. Replace the card. Repeat the process until you have 5 red cards.

\*\* observe the color of the card

Is this a binomial experiment?

1. Yes
2. No, more than 2 outcomes
3. No, trials are not independent
4. No, p changes for each trial
5. No, n is not fixed

Math 58

1. Shuffle a deck of 52 cards. Turn over the top card. Do not replace the card. Repeat the process 5 times.

\*\* observe the color of the card

Is this a binomial experiment?

1. Yes
2. No, more than 2 outcomes
3. No, trials are not independent
4. No, p changes for each trial
5. No, n is not fixed

Math 58

1. Shuffle a deck of 52 cards. Turn over the top card. Replace the card. Repeat the process 5 times.

\*\* observe the color of the card

Is this a binomial experiment?

1. Yes
2. No, more than 2 outcomes
3. No, trials are not independent
4. No, p changes for each trial
5. No, n is not fixed

Math 58

1. Person A flips a coin 2 times, and person B flips a coin 2 times. What is the probability that the proportion of heads for A is greater than the proportion of heads for person B?

Use the BINOMIAL DISTRIBUTION!!

(a) 0.1

(b) 0.375

(c) 0.5

(d) 0.625

(e) .9

Math 58B:

63. Relative Risk is

(a) the difference of two proportions

(b) the ratio of two proportions

(c) the log of the ratio of two proportions

(d) the log of the difference of two proportions

Math 58B:

64. One reason we should be careful interpreting relative risks is if:

(a) we don’t know the difference in proportions

(b) we don’t know the SE of the relative risk

(c) we might be dividing by zero

(d) we don’t know the baseline risk

Math 58B:

65. When we randomly select individuals based on the explanatory variable, we cannot accurately measure

(a) the proportion of people in the population in each explanatory category

(b) the proportion of people in the population in each response group

(c) anything about the population

(d) confounding variables

Math 58B:

66. The odds ratio is invariant to which variable is explanatory and which is response means:

(a) we always put the bigger odds in the numerator

(b) we must collect data so that we can estimate the response in the population

(c) which variable is called the explanatory changes the value of the OR

(d) which variable is called the explanatory does not change the value of the OR

Math 58B:

67. In order to find a CI for the true RR, our steps are:

1. ln(RR)

2. add ± z\* sqrt( 1/A - 1/(A+C) + 1/B - 1/(B+D) )

3. find exp of the endpoints

(a) because the sampling distribution of RR is normal

(b) because RR is typically greater than 1

(c) because the ln transformation makes the sampling distribution almost normal

(d) because RR is invariant to the choice of explanatory or response variable

Math 58B:

68. In finding a CI for p1/p2, why is it okay to exponentiate the end points of the interval for ln(p1/p2)?

(a) Because if ln(p1/p2) is in the natural log-interval, p1/p2 will be in the exponentiated interval.

(b) Because taking the natural log of the RR makes the distribution approximately normal.

(c) Because the natural log compresses values that are bigger than 1 and spreads values that are smaller than 1.

Math 58B:

69. Usually, the CI for p1/p2 is considered to be “significant” if

(a) p1/p2 is not in the interval

(b) 1/2 is not in the interval

(c) 0 is not in the interval

(d) 1 is not in the interval

Math 58B:

70. In order to find a CI for the true OR, our steps are:

1. ln(OR)

2. add ± z\* sqrt( 1/A + 1/B + 1/C + 1/D )

3. find exp of the endpoints

(a) because the sampling distribution of OR is normal

(b) because OR is typically greater than 1

(c) because the ln transformation makes the sampling distribution almost normal

(d) because OR is invariant to the choice of explanatory or response variable

71. When doing a hypothesis test for

H0: p1 - p2 = 0

AND assuming H0 is true, the best guess for is:

(a)

(c)

(d) 0.5

(e) 0

72. A 95% confidence interval for (p<40K−p≥40K), where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02).

At the 5% significance level, the data provide convincing evidence of a real difference in the proportion who are not affected personally between Americans who make less than $40,000 annually and Americans who make $40,000 or more annually.

1. TRUE
2. FALSE

73. A 95% confidence interval for (p<40K−p≥40K), where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02).

We are 95% confident that 16% more to 2% fewer Americans who make less than $40,000 per year are not at all personally affected by the government shutdown compared to those who make $40,000 or more per year.

1. TRUE
2. FALSE

74. A 95% confidence interval for (p<40K−p≥40K), where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02).

A 90% confidence interval for (p<40K−p≥40K) would be wider than the (-0.16, 0.02) interval.

1. TRUE
2. FALSE

75. A 95% confidence interval for (p<40K−p≥40K), where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02).

A 95% confidence interval for (p≥40K−p<40K) is (-0.02, 0.16).

1. TRUE
2. FALSE

Based on the first handwriting study, can we conclude that cursive causes higher scores (on average)?

1. Yes
2. No
3. It depends

Based on the second handwriting study, can we conclude that cursive causes higher scores (on average)?

1. Yes
2. No
3. It depends

76. A possible confounding variable for the handwriting study is:

(a) grade of the student (age)

(b) region of country where the SAT was taken

(c) academic ability of the student

(d) gender identity of the student

(e) number of siblings of the student.

77. The main reason we randomly assign the explanatory variable is:

(a) To get the smallest p-value possible

(b) To balance the expected causal mechanism across the two groups

(c) To balance every possible variable except the causal mechanism across the two groups

(d) So that our sample is representative of the population

(e) So that the sampling process is unbiased

1. The main reason we take random samples from the population is:

(a) To get the smallest p-value possible

(b) To balance the expected causal mechanism across the two groups

(c) To balance every possible variable except the expected causal mechanism across the two groups

(d) So that our sample is representative of the population

(e) So that the sampling process is unbiased

1. The “random” part in clinical trials typically comes from:

(a) random samples

(b) random allocation of treatment

80. The “random” part in polling typically comes from:

(a) random samples

(b) random allocation of treatment

81. You want to collect data to investigate whether teenagers in the United States have read fewer Harry Potter books (from the original series of seven books) than teenagers in the United Kingdom. Would you make use of random sampling, random assignment, both, or neither?

1. Random sampling
2. Random assignment
3. Both
4. Neither
5. An instructor wants to investigate whether using a red pen to grade assignments leads to lower scores on exams than using a blue pen to grade assignments. Would you advise the professor to make use of random sampling, random assignment, both, or neither?
6. Random sampling
7. Random assignment
8. Both
9. Neither
10. A student decides to investigate whether NFL football games played in indoor stadiums tend to have more points scored than games played outdoors. The student examines points scored in every NFL game of the 2019 season. Has the student used random sampling, random assignment, both, or neither?
11. Random sampling
12. Random assignment
13. Both
14. Neither

84. Are there effects of second-hand smoke on the health of children?

(a) definitely obs study

(b) definitely experiment

(c) unhappily obs study

(d) unhappily experiment

85. Do people tend to spend more money in stores located next to food outlets with pleasing smells?

(a) definitely obs study

(b) definitely experiment

(c) unhappily obs study

(d) unhappily experiment

86. Does cell phone use increase the rate of automobile accidents?

(a) definitely obs study

(b) definitely experiment

(c) unhappily obs study

(d) unhappily experiment

87. Do people consume different amounts of ice cream depending on the size of bowl used?

(a) definitely obs study

(b) definitely experiment

(c) unhappily obs study

(d) unhappily experiment

1. Which is more effective: diet A or diet B?

(a) definitely obs study

(b) definitely experiment

(c) unhappily obs study

(d) unhappily experiment

89. If the null hypothesis is true, the observed counts should equal the expected counts.

(a) True

(b) False

90. To reject the null hypothesis we want to see

(a) a small X2 value

(b) a big X2 value

91. A chi-square test is

(a) one-sided alt hypothesis, and we only consider the upper end of the sampling distribution

(b) one-sided alt hypothesis, and we consider both ends of the sampling distribution

(c) two-sided alt hypothesis, and we only consider the upper end of the sampling distribution

(d) two-sided alt hypothesis, and we consider both ends of the sampling distribution

92. For the lighting study, which variable is the explanatory variable?

(a) sleeping light

(b) eye sightedness

(c) child

(d) parent

93. If we sample randomly from a population, the conclusions we can make are about:

(a) causation

(b) population characteristics

1. Based on the night light / myopia example, the correct conclusion is:

(a) the p-value is small, so sleeping in a lit room makes it more likely that you are near-sighted.

(b) the p-value is small, so sleeping in a dark room makes it more likely that you are near-sighted.

(c) the p-value is small, so a higher proportion of children who sleep in light rooms are near-sighted than who sleep in dark rooms.

(d) pnear if lit room = 41/75 = 0.547 and pnear if dark = 18/172 = 0.105, therefore sleeping with the light on is bad for you.

95. A possible confounding variable for the night light study is:

(a) low birth weight

(b) race (70% of the children were white)

(c) region of the country where the clinic was located

96. The standard deviation of weights (mean = 167 lbs) is approximately­­

(a) 1

(b) 5

(c) 10

(d) 35

(e) 100

97. The standard deviation of average weights (mean = 167 lbs) in a sample of size 10 is approximately

(a) 1

(b) 5

(c) 10

(d) 35

(e) 100

98. The standard deviation of average weights (mean = 167 lbs) in a sample of size 50 is approximately

(a) 1

(b) 5

(c) 10

(d) 35

(e) 100

99. The standard deviation of average weights (mean = 167 lbs) in a sample of size 1000 is approximately

(a) 1

(b) 5

(c) 10

(d) 35

(e) 100

Q: what is the most confusing part of understanding the difference between the variability of the weights and the variability of the average of the weights?

100. The sampling distribution of the mean will be

(a) centered below the data distribution

(b) centered at the same place as the data distribution

(c) centered above the data distribution

(d) unrelated to the center of the data distribution

101. The sampling distribution of the mean will be

(a) less variable than the data distribution

(b) the same variability as the data distribution

(c) more variable than the data distribution

(d) unrelated to the variability of the data distribution

102. Why did we switch from talking about total weight to talking about average weight?

(a) So that it is easier to infer from the sample to the population.

(b) Because the Coast Guard certifies vessels according to average weight.

(c) Because the average is less variable than the sum.

(d) Because the average has a normal distribution and the sum doesn’t.

103. When the population is skewed right, the sampling distribution for the sample mean will be

(a) always skewed right

(b) skewed right if n is big enough

(c) always normal

(d) normal if n is big enough

104. What does the CLT say?

Dance of the p-values:

<https://www.youtube.com/watch?v=ez4DgdurRPg>

105. What type of variable is “healthy body temp”?

1. explanatory
2. response

106. We use s instead of because

(a) we know s and we don’t know

(b) s is a better estimate of the st dev

(c) s is less variable than

(d) we want our test statistic to vary as much as possible

(e) we like the letter t better than the letter z

107. The variability associated with is

(a) less than the variability of X

(b) more than the variability of X

(c) the same as the variability of X

(d) unrelated to the variability of X

(e) some other function of X

108. In part (f) you are asked to “determine how many standard errors the sample mean (98.249) falls from the hypothesized value of 98.6”

Which formula should you use?

(a)

(b)

(c)

(d)

109. When we use s instead of in the CI for µ, but still keep z\* (instead of using a t\* multiplier), the resulting CI has coverage

(a) **LESS** than the stated confidence level

(b) **MORE** than the stated confidence level

(c) **OF** the stated confidence level

110. What is the correct interpretation of the 95% CI for μ which is given as (98.121, 98.376) ?

(a) 95% of intervals will be (98.121, 98.376).

(b) 95% of individual temperatures will be between (98.121, 98.376).

(c) There is a 0.95 probability that the true temperature is between (98.121, 98.376).

(d) There is a 0.95 probability that the true average temperature is between (98.121, 98.376).

(e) In repeated samples, 95% of the intervals will contain μ.

Also good interpretation for a CI:

(f) We are 95% confident that the interval (98.121, 98.376) captures the true average temperature, μ.

111. Let’s say you \*truly\* believe that the true average body temp is between (98.121, 98.376). (Your CI is green.)

You record a temp of 98.6 F. Do you think you are sick?

1. Yes, it is outside the range above.
2. No, I still believe μ is 98.6.
3. No, 98.6 isn’t too far above the upper bound.
4. No, the interval isn’t for individuals.

112. The variability associated with is

(a) less than the variability of X

(b) more than the variability of X

(c) the same as the variability of X

(d) unrelated to the variability of X

(e) some other function of X

113. The variability associated with predicting a new value, ,

(a) is less than the variability of

(b) is more than the variability of

(c) is the same as variability of

(d) is less than the variability of X

(e) is more than the variability of X

114. Prediction intervals are

(a) smaller than confidence intervals

(b) about the same width as confidence intervals

(c) larger than confidence intervals

(d) unrelated to confidence intervals

115. Where should a prediction interval for a new value, , be centered?

(a)

(b) μ

(c) 98.6

(d) X1 (the first person in the dataset)

(e) Xn (the last person in the dataset)

116. What is the correct interpretation of the 95% PI for which is given as (96.79, 99.70)?

(a) 95% of intervals will be (96.79, 99.70).

(b) 95% of individual temperatures will be between (96.79, 99.70).

(c) There is a 0.95 probability that the true temperature is between (96.79, 99.70).

(d) There is a 0.95 probability that the true average temperature is between (96.79, 99.70).

(e) In repeated samples, 95% of the intervals will contain μ.

117. Prediction intervals have

(a) the same technical conditions as CIs

(b) stricter technical conditions than CIs

(c) more lenient technical conditions than CIs

(d) technical conditions which are unrelated to CIs

103a. When the population is skewed right, the **sampling distribution for the sample mean** will be

(a) always skewed right

(b) skewed right if n is big enough

(c) always normal

(d) normal if n is big enough

103b. When the population is skewed right, the **distribution for the data** will be

(a) always skewed right

(b) skewed right if n is big enough

(c) always normal

(d) normal if n is big enough

1. Calculating the averages of the observations given in the Investigation (for each of the two conferences) would determine:

(d) because the observations from the investigation \*are\* the entire population!

1. For the observations at hand, which do you expect to be bigger, the mean or the median?
2. The mean
3. The median
4. It isn’t possible to guess
5. In the Investigation, samples are taken from the population. Which is more variable?
6. The eastern population
7. The eastern sample(s)
8. They are equivalent / can’t determine

Variability in the SD goes up when n is small. But the SD itself is not systematically bigger (or smaller).

1. From each of the repeated samples, the sample mean is calculated. Which is wider (more variable)?
2. Distribution of the population(s)
3. Distribution of the sample mean(s)
4. They are equivalent
5. From each of the repeated samples (separately per conference), the sample mean is calculated. Should the distributions of the sample means be centered at the same place?
6. Yes, always.
7. Yes, because H0 is true.
8. No, never.
9. No, because H0 is true.
10. No, because H0 is false.
11. From each of the repeated samples, the sample mean is calculated for each group, and the difference is taken. Which is wider (more variable)?
12. Distribution of the sample mean(s)
13. Distribution of the differences in sample means
14. They are equivalent

124. Are the two samples (lefties and righties) independent?

(a) yes

(b) no

(c) we can’t tell

125. For the handedness example, which has a **lower** p-value?

1. Scenario 2
2. Scenario 3

126. For the handedness example, which has a **lower** p-value?

1. Scenario 3
2. Scenario 4

127. How does each affect the power?

(i) increasing the sample sizes of both groups

(a) increases the power

(b) doesn’t change the power

(c) decreases the power

(ii) increasing the variability within the groups

(a) increases the power

(b) doesn’t change the power

(c) decreases the power

(iii) increasing the difference in actual (population) group means

(a) increases the power

(b) doesn’t change the power

(c) decreases the power

128. The NBA simulation is meant to mirror:

(a) random sampling

(b) random allocation

Yes! If p-value small, we rule out the possibility that the difference in sample means is due to the random samples we took.

129. The sleep deprivation study is meant to mirror:

1. random sampling
2. random allocation

Yes! If p-value is small, we rule out the possibility that we happened to allocate “treatment A” to the people who were going to get low scores.

130. We typically compare means instead of medians because

(a) we don’t know the SE of the difference of medians

(b) means are inherently more interesting than medians

(c) the randomization applet (or R code) doesn’t work with medians

(d) the Central Limit Theorem doesn’t apply for medians.

131. We use the t-distribution (instead of the z-distribution) because:

(a) the CLT makes the test statistic normal

(b) the CLT makes the numerator of the test statistic normal

(c) the variability in the denominator makes the test statistic more variable

(d) the variability in the denominator makes the test statistic less variable

132. SE(is

(a)

(b)

(c)

(d)

(e)

133. If we use the SE and the z-curve (instead of t-curve) to find the p-value (assuming x-bar values are reasonably different):

(a) the p-value will be too small

(b) the p-value will be too big

(c) the p-value will be just right

(d) the p-value is unrelated to the curve

(e) we should use the SD instead134. If we use the SE and the z-curve (instead of t-curve) to find a 95% CI:

1. The capture rate will be at 95% over the long run.
2. The capture rate will be higher than 95% over the long run.
3. The capture rate will be lower than 95% over the long run.

Review

1. The sample standard deviation measures

(a) The distance from the sample mean corresponding to 68% of the data.

(b) The average distance of a point to the sample mean

(c) The square of the distance from a point to the sample mean

(d) A usual distance of a point to the sample mean

Review

2. The line in the interior of the box in a boxplot is

(a) The mode

(b) The mean

(c) The median

(d) The IQR

Review

3. Given the situation where Ha is TRUE. Consider 100 CIs (for true difference in means), the power of the test can be approximated by:

(a) The proportion that contain the true difference in means.

(b) The proportion that do not contain the true difference in means.

(c) The proportion that contain zero.

(d) The proportion that do not contain zero.

Review: Inv 4.10

4. Given the shopping data, can we tell if the technical conditions are met for applying t-prediction intervals

Variable N Mean StDev Min Median Max

Luckys 28 2.45 1.75 0.49 1.99 7.0

Scolaris 28 2.57 1.77 0.50 2.15 6.8

diffs 28 -0.118 0.359 -1.00 0.00 0.8

(a) No

(b) We can’t tell because we need to know if the distribution of the data is normal.

(c) Yes, the mean is very close to the median.

(d) Yes, the sample size is large.

(e) Yes, the standard deviation isn’t very big.

Review: Inv 4.10

5. The variance of the individual observations

(a) is smaller than the variance of the sample mean

(b) is about the same as the variance of the sample mean

(c) is larger than the variance of the sample mean

(d) is not related to the variance of the sample mean

Review: Inv 4.10

6. Prediction intervals are

(a) smaller than confidence intervals

(b) about the same width as confidence intervals

(c) larger than confidence intervals

(d) unrelated to confidence intervals

Review: Inv 4.10

7. Prediction intervals have

(a) the same technical conditions as CIs

(b) stricter technical conditions than CIs

(c) more lenient technical conditions than CIs

(d) technical conditions which are unrelated to CIs

135. Suppose that we record the midterm exam score and the final exam score for every student in a class. What would the value of the correlation coefficient be if every student in the class scored **ten points higher on the final than on the midterm:**

(a) r = -1

(b) -1 < r < 0

(c) r = 0

(d) 0 < r < 1

(e) r = 1

136. Suppose that we record the midterm exam score and the final exam score for every student in a class. What would the value of the correlation coefficient be if every student in the class scored **five points lower on the final than on the midterm:**

(a) r = -1

(b) -1 < r < 0

(c) r = 0

(d) 0 < r < 1

(e) r = 1

137. Suppose that we record the midterm exam score and the final exam score for every student in a class. What would the value of the correlation coefficient be if every student in the class scored **twice as many points on the final than on the midterm:**

(a) r = -1

(b) -1 < r < 0

(c) r = 0

(d) 0 < r < 1

(e) r = 1

138. Suppose you guessed every value correctly (guess the correlation applet), what would be the value of the correlation coefficient between your guesses and the actual correlations?

(a) r = -1

(b) -1 < r < 0

(c) r = 0

(d) 0 < r < 1

(e) r = 1

139. Suppose each of your guesses was too high by 0.2 from the actual value of the correlation coefficient, what would be the value of the correlation coefficient between your guesses and the actual correlations?

(a) r = -1

(b) -1 < r < 0

(c) r = 0

(d) 0 < r < 1

(e) r = 1

140. A correlation coefficient equal to 1 indicates that you are a good guesser.

(a) TRUE

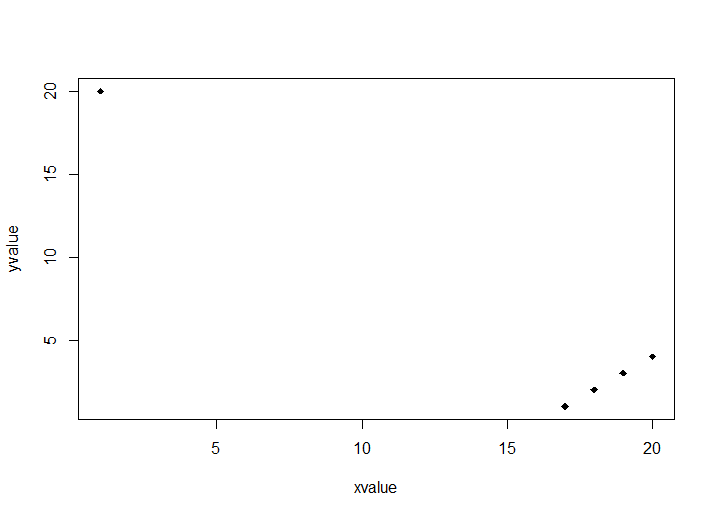
(b) FALSE

141. Perfect Correlation … if not for a single outlier

n = 101 observations:

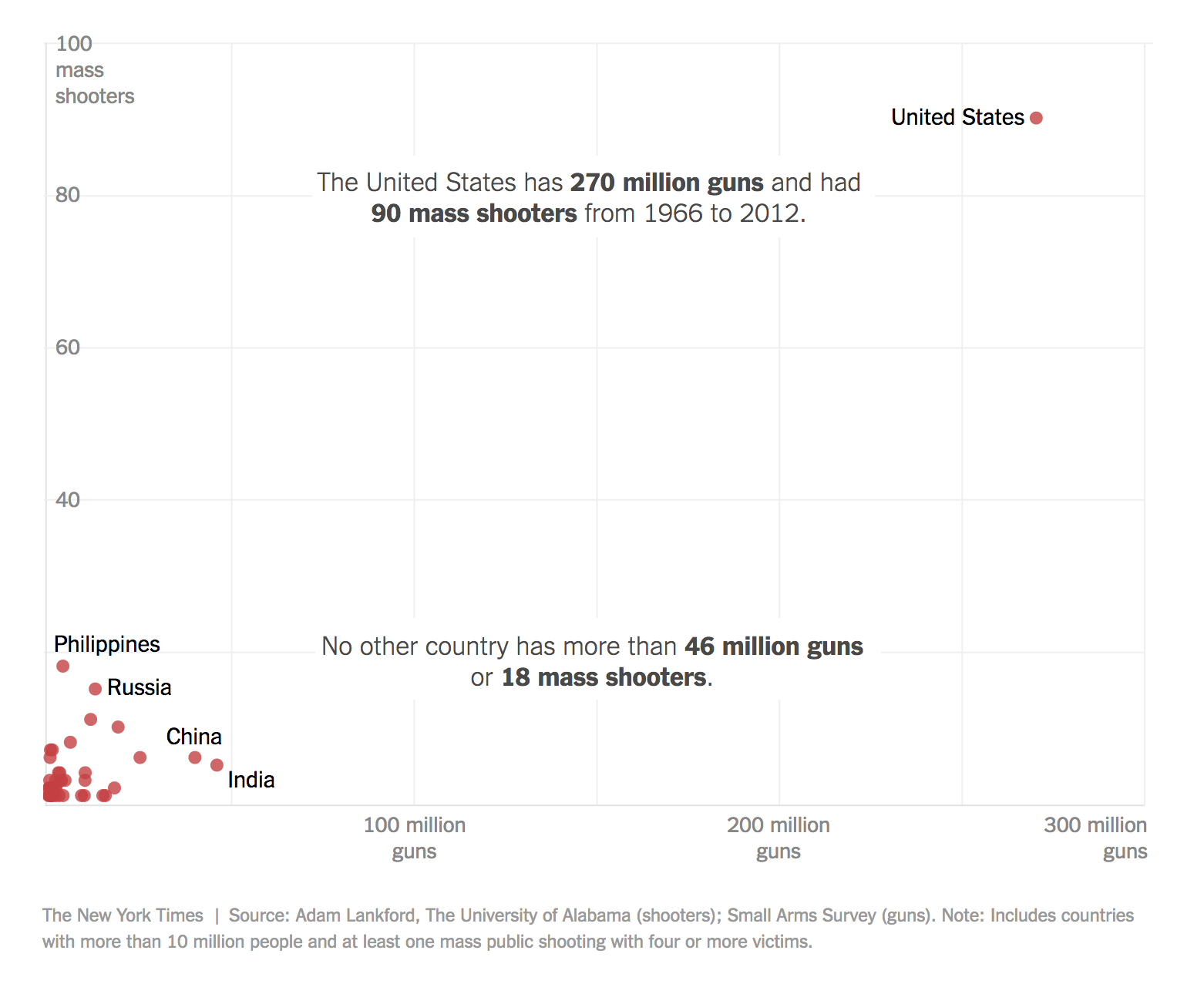
1 observation in top left, 25 observations in each in of the points near the bottom right.

The value of the correlation, r, is:



1. -1 < r < -.9
2. -.9 < r < -.5
3. -.5 < r < .5
4. .5 < r < .9
5. .9 < r < 1

141 follow-up …. Do you think that the above graph is unrealistic? Consider\* this image from the NY Times.



\* although there are other reasons / evidence to connect guns in circulation with mass shooting … much of the research does point to a causal connection.

142. The sum of residuals from the sample mean (no X):

∑ (Yi - )

(a) is positive

(b) is negative

(c) is zero

(d) is different for every dataset

143. A good measure of how well the prediction (of the sample mean) fits the data is:

(a) ∑ (Yi - )

(b) ∑ (Yi - )2

(c) ∑ |(Yi - )|

(d) median (Yi - )

(e) median |(Yi - )|

144. What math is used to find the value of m that minimizes:

∑ (Yi - m)2

(a) combinatorics

(b) derivative

(c) integral

(d) linear algebra

145. When writing the regression equation, why is there a hat ( ^) on the response variable?

(a) because the prediction is an estimate

(b) because the prediction is an average

(c) because the prediction may be due to extrapolation

(d) (a) & (b)

(e) all of the above

146. If there is no relationship in the population (true correlation ρ = 0), then r=0.

(a) TRUE

(b) FALSE

147. If there is no relationship in the population (true slope β1 = 0), then b1=0.

(a) TRUE

(b) FALSE

If we set a parameter equal to XXXX, should we expect the statistic to be XXXX?

No. Because statistics vary from sample to sample.

148. A smaller variability around the regression line (can be thought of as: σ or MSE or variability of the ei):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

149. A smaller variability in the explanatory variable (SD(X) = sx):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

150. A smaller sample size (n):

(a) increases the variability of b1.

(b) decreases the variability of b1.

(c) doesn’t necessarily change the variability of b1.

151. The regression technical assumptions include:

(a) The Y variable is normally distributed at each X

(b) The X variable is normally distributed

(c) The residuals are normally distributed

(d) The slope coefficient is normally distributed

(e) The intercept coefficient is normally distributed

152. The technical conditions do **not** include:

(a) normally distributed residuals

(b) normally distributed response

(c) normally distributed explanatory variable

(d) constant variance

(e) independence of observations

153. What happens if the technical conditions are not met?

1. The line does not minimize the sum of squared residuals.
2. R^2 does not measure the proportion of variability explained by the line.
3. The distribution of b1 is wrong (therefore incorrect p-values and CI).
4. The computer (R) will produce an error when running the linear model.

154. An econometrician is interested in evaluating the relation of demand for building materials to mortgage rates in LA and SF.

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

holding constant the effect of city, each additional increase of 1% in the mortgage rate would lead to an estimated average \_\_\_\_\_\_\_\_ in the mean demand.

(a) predicted $500 more per capita

(b) predicted $500 less per capita

(c) predicted $5 more per capita

(d) predicted $5 less per capita

155. Referring to

Y = 10 + 5X1 + 8X2

where

X1 = mortgage rate in %

X2 = 1 if SF, 0 if LA

Y = demand in $100 per capita

the effect of living in LA rather than SF is a \_\_\_\_\_\_\_\_ demand by an estimated \_\_\_\_\_\_\_\_ holding the effect of mortgage rate constant.

(a) larger; $800 per capita

(b) smaller; $800 per capita

(c) larger, $8 per capita

(d) smaller, $8 per capita

156. Consider the housing model,

= 12.2 + 0.000468⋅sqft − 0.0603⋅# bedrooms

the coefficient on bedrooms (− 0.0603) can be interpreted as the change in predicted ln(price)…

1. for a one unit increase in bedrooms
2. for a home that adds a bedroom to the existing structure
3. for a one unit increase in bedrooms when comparing homes that have identical square feet
4. for a one unit increase in square feet
5. for a one unit increase in square feet when comparing homes that have identical number of bedrooms

157. To test if there is convincing evidence that the slope of the regression line between ln(price) and square feet (only, no bedroom here) is different from zero,

what are the appropriate hypotheses?

(a) H0: b0 = 0

Ha: b0 ≠ 0

(b) H0: b1 = 0

Ha: b1 ≠ 0

(c) H0: β0 = 0

Ha: β0 ≠ 0

(d) H0: β1 = 0

Ha: β1 ≠ 0

158. With # bedrooms in the model, the test is now:

= 12.2 + 0.000468⋅sqft − 0.0603⋅# bedrooms

H0: β1 = 0

Ha: β1 ≠ 0

In words, the test asks if there is convincing evidence that:

1. the slope of the regression line between ln(price) and square feet is different from zero
2. the slope of the regression line between ln(price) and square feet is different from zero when # bedrooms is included in the model
3. adding square footage to your house causes the value to increase
4. the slope of the regression line between ln(price) and bedrooms is different from zero
5. the slope of the regression line between ln(price) and bedrooms is different from zero when square feet is included in the model

159. We created a 95% confidence interval for the **mean GPA** given 10 absences to be (3.20, 3.42). What is the correct interpretation?

(a) There is a 95% chance that the mean GPA of students with 10 absences is between 3.20 and 3.42.

(b) 95% of GPA averages (for students with 10 absences) are between 3.20 and 3.42.

(c) 95% of GPAs (for students with 10 absences) are between 3.20 and 3.42.

(d) We are 95% confident that the true mean GPA (for students with 10 absences) is between 3.20 and 3.42.

(e) 95% of our intervals will have a mean GPA between 3.20 and 3.42.

160. We created a 95% prediction interval for an **individual GPA** given 10 absences to be (3, 3.62). What is the correct interpretation?

(a) There is a 95% chance that the mean GPA of students with 10 absences is between 3 and 3.62.

(b) 95% of GPA averages (for students with 10 absences) are between 3 and 3.62.

(c) 95% of GPAs (for students with 10 absences) are between 3 and 3.62.

(d) We are 95% confident that the true mean GPA (for students with 10 absences) is between 3 and 3.62.

(e) 95% of our intervals will have a mean GPA between 3 and 3.62.

161. Prediction intervals and confidence intervals have the same technical conditions:

(a) TRUE

(b) FALSE

(c) sort of

162. It is often a good idea to transform the variable(s)…

(a) … to find the highest r^2 value.

(b) … when the X variable is not normally distributed.

(c) … to make the model easier to interpret.

(d) … so that the technical conditions are met.