CMM262/BGGN237 Midterm Exam 2018

Part 3

Email questions and final answers to Prof. Scott Rifkin: sarifkin@ucsd.edu

**Use the following to answer questions 1-3.**

Suppose that a 95% confidence interval for the slope of a regression line based on a sample of size *n* = 100 and the percentiles of the slopes for 1,000 bootstrap samples goes from 2.50 to 2.80. For each change described (with all else staying the same), indicate which of the three confidence intervals would be the most likely result.

1) Decrease the sample size to *n* = 60.

a) 2.53 to 2.77 (narrower)

b) 2.50 to 2.80 (the same)

c) 2.46 to 2.84 (wider)

2) Increase the number of bootstrap samples to 5,000

a) 2.53 to 2.77 (narrower)

b) 2.50 to 2.80 (the same)

c) 2.46 to 2.84 (wider)

3) Increase the confidence level to 99%.

a) 2.53 to 2.77 (narrower)

b) 2.50 to 2.80 (the same)

c) 2.46 to 2.84 (wider)

4) A scientist tests 200 drugs on a cancer cell line to test whether they stop cell proliferation. Unbeknownst to the researcher, 80 of them really work, while 120 are ineffective. The scientist sets a type I error rate of 0.05. This leads to a power of 0.8 for the drugs that really work. Out of the 200 drugs tested, how many type I (false alarm) and type II (missed opportunity) errors are expected in the scientist’s results?

5) Which one of the following statements *best* describes the purpose of scrambling (or shuffling) to generate a null distribution regardless of the study design when there are two variables?

a) Scrambling breaks any relationship between the two variables.

b) Scrambling gives each group a 50% chance of being selected.

c) Scrambling simulates no cause-effect relationship.

d) None of the above

6) Suppose that a well-established and validated model predicts that 60% of the dung beetles of a particular species A should have long horns and 40% should have short horns.



You discover a new species B of dung beetles and want to test whether the model is valid for this species too. You have no reason to suspect either that there will be more or that there will be fewer long-horned beetles for species B but you want to test the conjecture that model for species A is not appropriate for species B. You count the number of long- and short-horned dung beetles in a random sample of the new species B and find that there are 25 long- and 5 short-horned dung beetles in your sample.

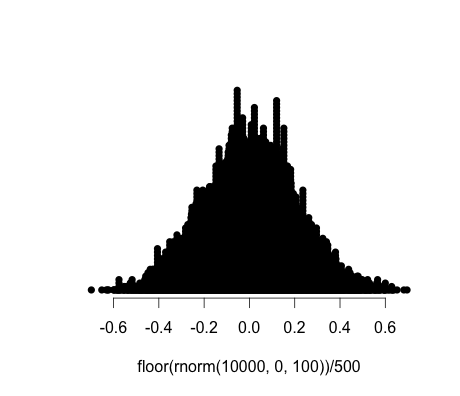
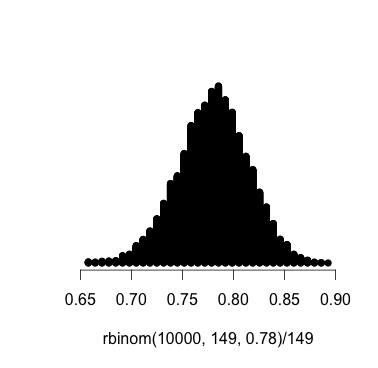
You decide to do a hypothesis test by simulation and get the following null distribution for the number of long-horned beetles based on the model.

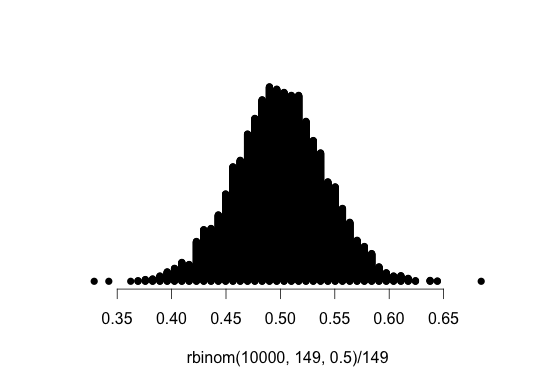
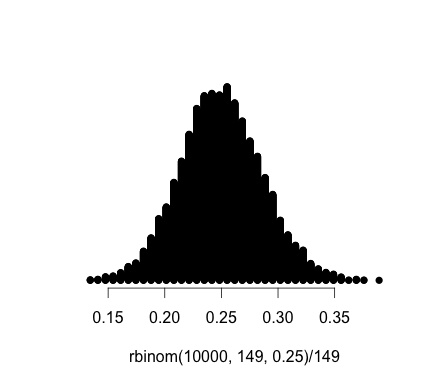
|  |  |
| --- | --- |
| Number of long-horned beetles | Frequency out of 10,000 repetitions |
| ≤6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  ≥29 | 0  0  1  6  25  49  145  271  481  741  1140  1438  1454  1424  1139  779  471  279  104  42  9  2  0  0 |

a) What are the null and alternative hypotheses framed in terms of your statistic?

b) What is the p-value for your sample statistic based on your null distribution? Show how you determined it.

7) A multiple choice question on a midterm exam has 4 possible answers. 78% of the 126 students get the question correct. You want to test whether students choose the correct answer significantly more than chance. Circle the appropriate null distribution





8) Suppose that a student is working on a statistics project using data on height collected from a random sample of 100 students from her college, a small, selective school in Colorado. She ﬁnds a 95% conﬁdence interval for mean height to be (65.5,71.8) inches. Discuss how each of the statements below would indicate an improper interpretation of this interval.

(a) I am 95% sure that all students will have heights between 65.5 and 71.8 inches.

(b) I am 95% sure that the mean height for this sample of students will be between 65.5 and 71.8 inches.

(c) I am 95% sure that the conﬁdence interval for the mean height of all students at this college is 65.5 to 71.8 inches.

(d) I am sure that 95% of all students at this college will have heights between 65.5 and 71.8 inches.

(e) I am 95% sure that the mean height for all US college students is between 65.5 and 71.8 inches.

(f) Of random samples of this size taken from students at this college, 95% will have mean heights between 65.5 and 71.8 inches.

9) If the width of my 95% confidence interval is 6 and I increase my sample size from 300 to 1200 and my sample statistic remains the same, what is the new width of my 95% confidence interval? [Note – if you do not know how to figure out this answer, feel free to experiment and look for patterns with the bootstrapping applet on http://www.lock5stat.com/StatKey/ ]

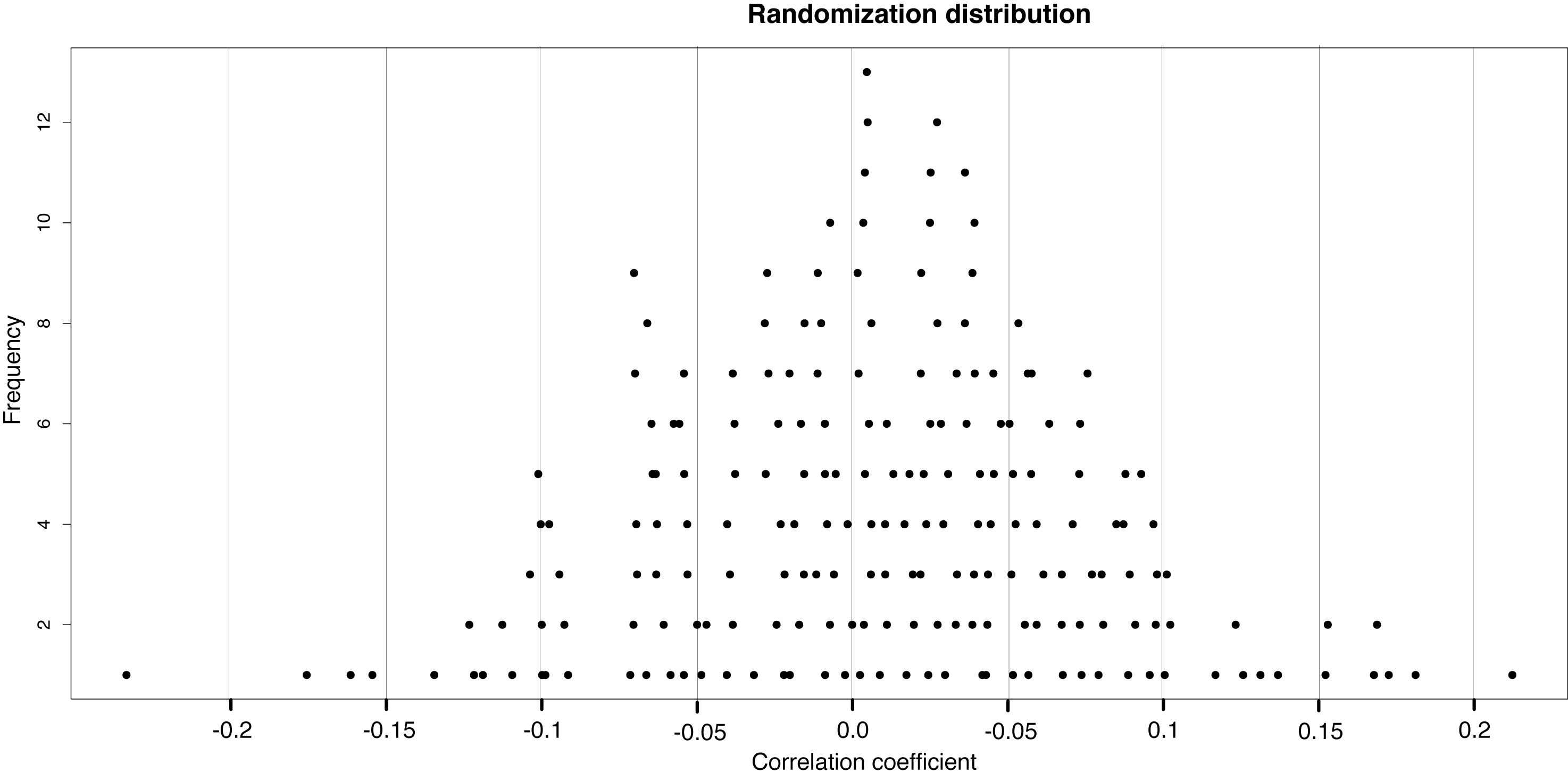
10) Suppose you are testing the hypotheses H0: p= 0.6 versus HA: p ≠ 0.6. Your sample proportion is 0.44 and your p-value is 0.12. Now suppose the study were different in each of the following ways. Would your new p-value be larger or smaller than 0.12?

* 1. You used a smaller sample size but your sample proportion was still 0.44. Your new p-value would be:
  2. You used the same sample size, but your sample proportion was 0.45. Your new p-value would be:
  3. You hypothesized that p should be greater than 0.6 and so instead of a two-tailed test you decided to use a one-sided test with HA: p > 0.6. Your sample proportion was still 0.44. Your new p-value would be:

11) Some neurobiologists measured the concentration of the neurotransmitter dopamine in the brains of a random sample of 100 rats. The mean concentration was 890 ng/gm and the standard deviation of the sample was 200 ng/gm. They estimate the width of a 95% confidence interval for the mean concentration in rats as 800 ng/gm. Is this too wide, too narrow, or about right? Explain.

12) You use a randomization technique to construct the null distribution below for testing a correlation. You only took 200 randomized samples. The correlation coefficient from your actual data is r = -0.15. Your alternative hypothesis is: *ρ* ≠ 0

What is the p-value based on this randomization distribution? Explain how you determined it.



13) For each item below that you report, write whether it is a data value, a parameter, or a statistic:

a) You sequence a bunch of mouse gametes at a polymorphic nucleotide and report the fraction of gametes with a C at that nucleotide position.

b) You report the mean height of your biological parents.

c) You measure and report your pulse rate while you are answering this question.

d) You report your estimated value for the 97th percentile of diastolic blood pressure among dialysis patients.