

Final Examination (150 minutes)

No calculators allowed. Just set up your answers, for example, $P = 49/52$. BE SURE TO EXPLAIN YOUR REASONING. If you want extra time, you can buy time at a price of 1 point a minute; for example, if your test is handed in 10 minutes after the scheduled finish time, 10 points will be subtracted from the test score.

1. In a class of 15 students, 6 students will be chosen randomly every week to do a project. There are 10 weeks in the term. Cory is one of the students. What is the probability that Cory will be chosen more than 5 times?

2. In a letter to the *New England Journal of Medicine*, Dr. Sanders Frank reported that 20 of his male patients with creases in their earlobes had many of the risk factors (such as high cholesterol levels, high blood pressure, and heavy cigarette usage) associated with heart disease. For instance, the average cholesterol level for his patients with earlobe creases was 257 (mg per 100 ml), compared to an average of 215 with a standard deviation of 10 for healthy middle-aged men. If these 20 patients were a random sample from a population with a mean of 215 and a standard deviation of 10, what is the probability their average cholesterol level would be 257 or higher? Explain why these 20 patients may be a biased sample.

3. A *Wall Street Journal* (July 6, 1987) poll asked a 35 economic forecasters to predict the interest rate on 3-month Treasury bills in June 1988. These 35 forecasts had a mean of 6.19 and a variance of 0.47. Assuming these to be a random sample, give a 95 percent confidence interval for the mean prediction of all economic forecasters and then explain why each of these interpretations is or is not correct:
 - a. There is a 0.95 probability that the actual Treasury-bill rate on June 1988 will be in this interval.

 - b. Approximately 95 percent of the predictions of all economic forecasters are in this interval.

 - c. If the *Journal* took another random sample, there is a 0.95 probability that the new confidence interval would include 6.19.

4. The board game, Settlers of Catan, is played with each player taking turns rolling two standard 6-sided dice and adding the numbers on the two dice. To win a 6-person game of Settlers on his next turn, Professor Smith needs at least one of the six players to roll either a 5 or a 10. What are his chances of winning?
5. A presidential candidate figures that she has a 0.7 probability of winning in Wisconsin, a 0.6 probability of winning in Michigan, and a 0.4 probability of winning in Ohio. If the outcomes are independent, what is the probability that she will win exactly 2 of these three states?
6. After 162 regular season baseball games, 8 teams qualify for playoff games leading to the World Series, where the two remaining teams play a best-of-7 series to determine the world champion. A sportswriter noted that the San Francisco Giants team won three world championships (in 2010, 2012, and 2014) even though they did not have the best regular-season record, a result he attributed to survivorship bias:
- The concept of “survivorship bias” holds that we often give victors more credit for their tactics than we should, even if they won by very slim margins.*
- Explain why “survivorship bias” isn’t the most appropriate statistical label, and choose the label that is the most appropriate: law of averages, paradox of luck and skill, regression to the mean, or self-selection bias.

7. A study of merger-and-acquisition (M&A) activity between Japanese companies and foreign companies reported the following estimated equation, using quarterly data for 2001 - 2018:

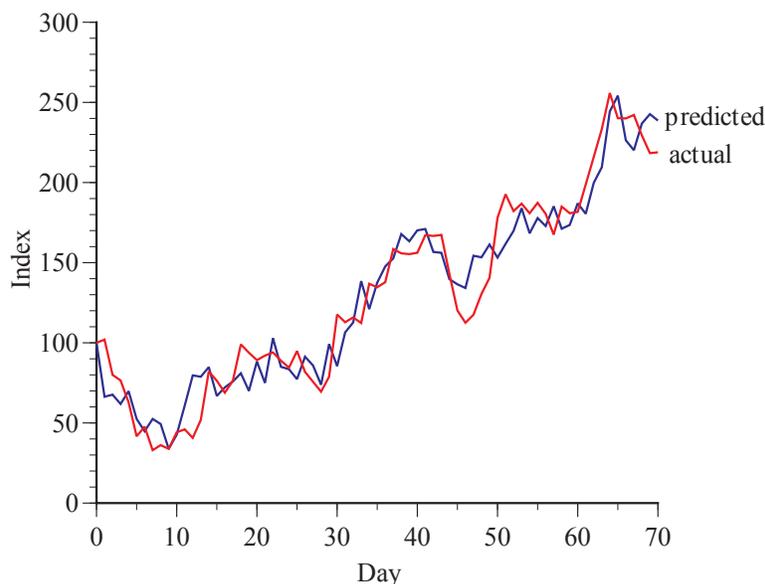
$$M = 5292.2 + 15.27R - 0.0125S + 1.068Y + 48.60P, R^2 = 0.88$$

(999.8)
(24.73)
(0.01)
(0.89)
(43.27)

where M = number of M&A deals with foreign companies, R = interest rate on 10-year bonds, S = Nikkei index of Japanese stock prices stock market, Y = GDP, P = population, and the standard errors are in parentheses. Explain why each of the following interpretations of the results is incorrect.

- a. "The substantial and highly statistically significant intercept (5292.2) shows that this equation is highly successful in predicting M&A activity."
- b. "The population has a negative effect on M&A activity because the declining Japanese population has reduced the growth of GDP, forcing Japanese companies to look elsewhere for growth opportunities."
- c. "88% of the volatility in the explanatory variables is explained by the error term."

8. A data-savvy police department in Minnesota data-mined the Facebook accounts of local residents to see if surges and slumps in the use of certain words might be helpful in predicting criminal activity. They started their investigation by identifying the 100 most popular nouns, 50 most popular adjectives, and 50 most popular adverbs in the English language. Then they collected daily data for 10 weeks on the frequency with which each of these 200 words were used in Facebook status updates and the number of burglaries committed the next day. All the data were scaled to equal 100 at the start of the study, thus a value of 101 means 1 percent more than initially, and 99 means 1 percent less. They found that the two most helpful words for predicting burglaries were *day* and *most*. Why do you doubt that their model will be a useful tool for predicting burglaries?



9. A study of initial public offerings (IPOs) compared the 3-year return Y following the first day of trading with rate of return X on the first day of trading:

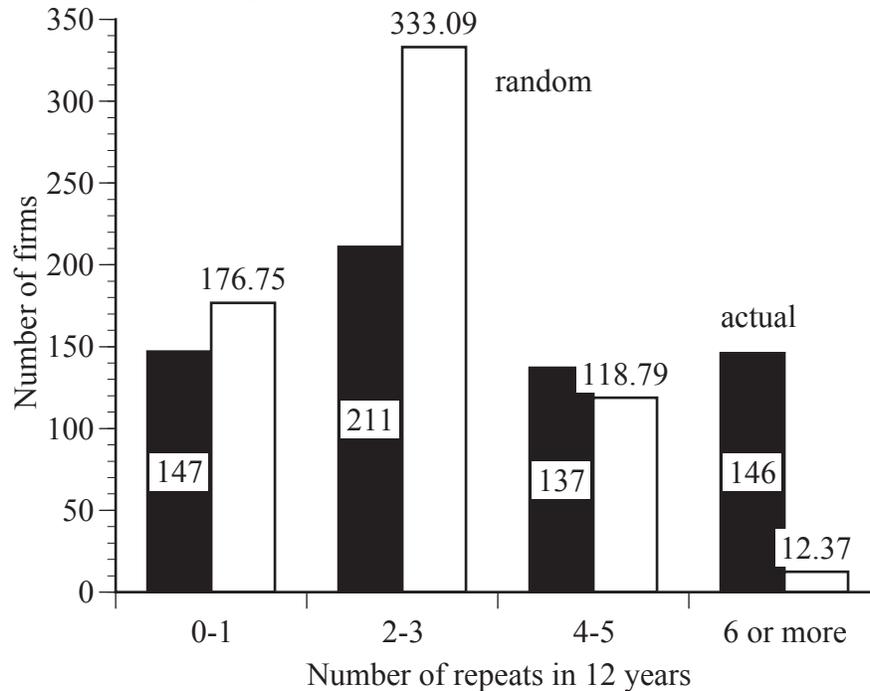
$$Y = 0.087 + 0.630X, \quad R^2 = 0.009$$

(0.122) (0.309)

The standard errors are in parentheses. Explain why you either agree or disagree with these conclusions:

- a. “The coefficient of X had a two-sided p-value of 0.042, indicating it was close to a 5% significant threshold level.”
- b. “The R-squared value was under 1%. This means that we should reject my initial hypothesis of finding a correlation between a stock’s performance on the first day and a stock’s 3-year performance.”
10. Jerry’s time in the 100 freestyle is normally distributed with a mean of 56 seconds and a standard deviation of 2 seconds; Kerry’s time is normally distributed with a mean of 57 seconds and a standard deviation of 2.5 seconds. If their swim times are independent (perhaps they swim in lanes 1 and 8), what is the probability that Kerry will beat Jerry?

11. Sam Glazer collected data for 641 U.S. companies over a 12-year period on their annual research efficiency, an index of how much their revenue was increased by their research spending. In order to test the persistence of research efficiency, he identified the top 128 companies (20%) each year, and compared how frequently companies appeared among the top 20% to how often they could be expected to appear if the research rankings were completely random. For example, we expect 12.37 firms to appear in the top 20% six or more times in 12 years if rankings are random, but 146 did. How would you test if the differences in the bar chart below are statistically significant? You do not need to do the test, but you need to set it up.



12. A treatment group was given a cold vaccine, while the control group received a placebo. Doctors then recorded the fraction of each group that caught a cold and calculated the two-sided p-value to be 0.08. Explain why you either agree or disagree with each of these interpretations of these results:

- “There is an 8 percent probability that this cold vaccine works.”
- “If a randomly selected person takes this vaccine, the chances of getting sick fall by about 8 percent.”
- “These data do not show a statistically significant effect at the 5 percent level; therefore, we are 95 percent certain that this vaccine doesn’t work.”

13. Scientists are trying to develop a test for a genetic disorder that occurs in roughly 1 in 100,000 babies. Suppose that the test is $X\%$ accurate in that it gives a positive result for $X\%$ of the babies who have the disorder and gives a negative result for $X\%$ of the babies who do not have the disorder. How high must X be for there to be a 0.90 probability that a baby who tests positive actually has the disorder?

14. Jack and Jill each roll a standard 6-sided die. Jack pays Jill \$1 if the number on her die is equal to or larger than the number on his die; Jill pays Jack \$ X if the number on Jack's die is larger than the number on her die. What value of X makes this a bet with an expected value for of 0 Jill?

15. A researcher looked at annual U.S. population data (in millions) for the 58-years 1960-2017 and calculated

$$\bar{X} = \frac{\sum_{i=1}^{58} X_i}{58} = 251.27$$

$$s = \sqrt{\frac{\sum_{i=1}^{58} (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{58} (X_i - 251.27)^2}{58-1}} = 43.56$$

$$\text{Standard error} = \frac{s}{\sqrt{n}} = \frac{43.56}{\sqrt{58}} = 5.72$$

What is bizarre about this calculation of the standard error?

16. A study compared the daily returns for a portfolio of stocks that have clever ticker symbols to the performance of the overall stock market for the years 1984-2005. Part of the study involved the estimation of this regression equation,

$$Y = 0.00049 + 0.81X, R^2 = 0.29$$

[3.45] [39.52]

where Y is the clever-ticker daily return, X is the stock market daily return, and the t-values are in brackets. Naomi Baer and Erica Barry updated this study, using data for 2006-2018:

$$Y = 0.00026 + 0.88X, R^2 = 0.68$$

[1.78] [64.56]

How could you test if the differences in the estimated coefficients (0.00049 versus 0.00026 and 0.81 versus 0.88) are statistically significant? Be specific.

17. Identify the most appropriate statistical test for each of these research hypotheses. You do not need to show any formulas, just identify the test, for example, “difference-in-means t test.”

a. A study of injuries suffered by Pomona varsity athletes during the years 1980-1995 obtained these data:

	Females	Males
Basketball	214	344
Soccer	155	236
Swimming	161	34
Tennis	82	82
Track	151	233

b. Data for 116 California communities were used to see how the percent change in home prices between 2005 and 2010 was affected by each county’s percent change in home prices between 2000 and 2005 and each county’s ratio of the median home price in 2005 to median rent in 2005.

c. A double-blind study of the effectiveness of a certain medication found a 20% success rate for those taking the medication and a 10% success rate for those taking a placebo.

d. Survey data were collected from 120 college students on the average time they went to bed each night during the previous semester and their GPA that semester. Bedtimes were divided into 10-12, 12-1, 1-2, and 2-4 with respective average GPAs of 3.68, 3.53, 3.56, and 3.38.

e. Jordan Hawkins and Jack Storrs predicted a college football team’s success based on the recruiting rankings for its first-year, second-year, third-year, fourth-year, and fifth-year players.

18. In 2019, *Business Insider* reported: “if you possess a certain set of characteristics, you may be more likely to become wealthy, according to Sarah Stanley Fallaw, director of research for the Affluent Market Institute.” Ms. Fallaw surveyed more than 600 American millionaires, and identified 6 characteristics she calls “wealth factors” that are most predictive of net worth, regardless of age or income; for example, frugality and confidence in financial management. As a statistician, what problems do you see with this study?

19. A study of profitability of the 30 stocks in the Dow Jones Industrial Average over a 10-year period reported these least-squares estimates,

$$R = -3.25 + 0.56R_{-1} \quad R^2 = 0.587$$

(0.46) (0.14)

where R is the firm’s percent rate of return on assets that year, and R_{-1} is the firm’s percent rate of return on assets the previous year. Is there evidence of regression toward the mean? How do you know whether there is or isn’t?

20. How would you explain this 0.9926 correlation?

