Econ 57
Gary Smith
Spring 2022

Final Examination (150 minutes)

No calculators allowed. Just set up your answers, for example, P = 49/52. 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. A 2021 study of 56 cities compared the average daily temperature in each city during the years 1981-2010 with the average daily temperature in that city during the years 1971-2000, giving 56 matched-pair observations. What big problem do you see with this methodology?

2. Identify the most appropriate null hypothesis and statistical test for each of these studies (for example,

   H_0: the average difference is zero

   test: matched-pair t-test:

   a. Time series data were used to predict interest rates based on the rate of inflation and the government deficit.

      H_0:

      test:

   b. The daily returns on a professionally managed stock portfolio were compared to the daily returns on the S&P 500 index during the year 2000-2020.

      H_0:

      test:

   c. "As January goes, so goes the year" is an old stock market adage. Data were collected for 1970-2020 on the S&P 500 return in January of each year and the S&P 500 return for February through December of that year.

      H_0:

      test:

   d. In golf, each hole is assigned a *par* score (typically between 3 and 5) that a good golfer should be able to achieve. A study of several professional golf tournaments recorded how often a player's score on a hole was below-par, par, or above-par, and how often this player's score on the next hole was below-par, par, or above-par.

      H_0:

      test:

   e. To test a mentalist's claim to be able to influence dice rolls, two standard 6-sided dice were rolled 200 times and the number of doubles was recorded.

      H_0:

      test:

3. Overall, 40% of U.S.adults are firstborn or only children. A survey of 100 Pomona College students found that 48 were first born. The researchers did a chi-square test:

	Observed		Expected		
	First or Only	Other	First or Only	Other	Total
Pomona students	48	52	44	56	100
U.S. population	40	60	44	56	100
Total	88	112	88	112	200

$$\chi_1^2 = \frac{(48-44)^2}{44} + \frac{(52-56)^2}{56} + \frac{(40-44)^2}{44} + \frac{(60-56)^2}{56} = 1.2987$$

$$p = 0.2544$$

a. What error did the researchers make?

b. How would you do a test, using only the data available here?

4. All of the women in the country OneOfEach who want to have children want to have at least one boy and one girl and will stop having children after they have one of each. For example, a woman who has a boy, another boy, and then a girl will stop with three children. Assuming boy and girl babies are equally likely and independent, how many children, on average, do the women with children have?

5. Explain:

This is also known as the Will Rogers effect, after the US comedian who reportedly quipped: "When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states."

6. Identify two ways in which this report of the results of the estimation of a multiple regression equation could be improved:

Variable	T-Statistic	P-Value
Variable1	2.47	0.02
Variable2	1.70	0.09
Variable3	6.31	<0.001

7. Bulgaria has a weekly national lottery in which participants choose 6 different numbers from the 42 numbers 1 through 42 and win the grand prize if their 6 numbers match (not necessarily in order) the 6 numbers picked on live television. On September 10, 2009, the winning numbers (4, 15, 23, 24, 35, 42) were the same numbers picked the week before (though in different order). No one had the winning numbers on September 6, but a record 18 people had the winning numbers on September 10. If the game is fair,
   a. What is the probability that the 6 numbers picked in a lottery will be the same 6 numbers picked in the previous lottery (not necessarily in the same order)?

   b. If the lottery is held 1,000 times (50 times a year for 20 years), what is the probability that there will be at least one occasion where the numbers repeat?

8. What is the most important thing wrong with these multiple regression results that examined the relationship between year in college and mental depression?

Model

Dependent variable:
Depression = survey, measured on a scale of 0 to 10
Explanatory variables:
Frosh = 1 if first year, 0 otherwise
Sophomore = 1 if second year, 0 otherwise
Junior = 1 if third year, 0 otherwise
Senior = 1 if fourth year or longer, 0 otherwise

Results

Constant	4.720***
	(1.664)
Frosh	0.669
	(0.818)
Sophomore	0.802
	(0.850)
Junior	0.923
	(0.838)
Senior	0.819
	(0.814)
observations	113
R-Squared	0.531

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

9. Americans drive about 3 trillion miles a year and average about 77 reported injuries and 1.09 fatalities per 100 million miles driven. In comparison, a Google fleet of 55 cars had 11 crashes (2 injuries and no fatalities) while driving 1.3 million miles in autonomous mode between 2009 to 2015.

    If a test of autonomous vehicles has no fatalities, how many miles would have to be driven fatality free to reject at the 5% level the null hypothesis that the probability of a fatality is 1 per 100 million miles?

10. *The Millionaire Next Door* by Thomas J. Stanley and William D. Danko, promises to reveal the "surprising secrets of America's wealthy." To uncover these secrets, the authors sent an 8-page survey, a return envelope, and a dollar bill to 3,000 Americans living in wealthy neighborhoods. A total of 1,115 surveys were returned, including 385 from people with a reported net worth of more than a million dollars. The authors then identified seven characteristics that these 385 wealthy people had in common. As a statistician, what is the biggest problem that you see with this study?

11. One hundred randomly selected students were asked this question:

    You have a coach ticket for a nonstop flight from Los Angeles to New York. Because the flight was overbooked, randomly selected passengers will be allowed to sit in open first-class seats. You are the first person selected. Would you rather sit next to: (a) the U.S. president; (b) the president's wife; or (c) Michael Jordan? Is the difference between the female and male responses statistically persuasive?

    Here are the results:

	Joe Biden	Jill Biden	Michael Jordan
females	12	25	10
males	13	0	40

    a. What is the most appropriate null hypothesis?

    b. Set up the appropriate test of this null hypothesis.

12. How would you, as a statistician, explain the fact that, of the 228 golfers who have won at least one of the four major men's golf championships, 144 (63%) only won once?

13. A multiple regression model was estimated to see if a company's Environmental, Social, & Governance (ESG) score affects its revenue (TR). The possible confounding explanatory variables considered were the year, risk (measured by the company's debt-to-asset ratio), and innovation (measured by corporate investment in new products). The reported results were:

Dependent variable: TR

	Coefficient	Std. err.	|t|
constant	-.9793100	64.55512	1.52
ESG	-.0031310	.0076487	0.41
year	-.0001902	.0321261	0.01
risk	-.1219829	.4110409	0.30
innovation	1.086324	1.868541	0.58

R-squared = –.0185

a. Identify 2 errors in the reported results.

b. Explain the error in the author's conclusion: "there were no variables which significantly affected total returns, which means this theory must be rejected."

14. Traffic fines for speeding in California are supposed to be determined by a formula based on how many miles over the speed limit the driver was driving, but there is flexibility in various extra fees. A researcher wants to see if total speeding fines are higher for people under the age of 30 than for older drivers.
    a. Specify a multiple regression equation that might be used to investigate this research question.

    b. Which coefficients in your model would show whether there is discrimination against younger drivers?

    c. Identify a situation in which a difference-in-means test would show discrimination against younger drivers, but a multiple regression model would not.

    d. Identify a situation in which a difference-in-means test would show no discrimination against younger drivers, but a multiple regression model would.

15. What are the problems with these statements?
    a. "I have data on 11 possible explanatory variables for my multiple regression model. I will use the three variables with the highest t-values."

    b. The high R-squared shows that these explanatory variables have a lot of oomph.

    c. "Since the p-value for the coefficient of income is 0.0000000427, we are very confident that there is less than a 5% probability that the null hypothesis is true."

    d. "There is a multicollinearity problem: The explanatory variables are correlated with each other but the multiple regression model assumes that the explanatory variables are independent."

16. A high school student is going to apply to 10 top-tier colleges.
    a. If each college has a 0.2 probability of accepting this student, and college acceptances are independent, what is the probability that this student will be accepted by at least one of these 10 colleges?

    b. Will the probability of acceptance by at least one college be higher or lower if acceptances are not independent; i.e., if this student is accepted (or rejected) by one school this increases the probability of being accepted (or rejected) by other schools?

17. A computer program simulated a player's first 20 moves in Monopoly and recorded whether or not the player landed on Free Parking at least once in those 20 moves. The program was run 1 billion times and in 387,444,281 of those 1 billion simulations, the player landed on Free Parking at least once. Give a 95% confidence interval for the probability of landing on Free Parking at least once in the first 20 moves.

18. As a statistician, what are the two most important issues you would raise with this estimation of a model intended to see whether investors who buy stocks with high PEG ratios receive below-average returns:

The model is

$$R = \alpha + \beta PEG + \varepsilon$$

using these data for the 30 stocks in the Dow-Jones Industrial Average on December 31 of each year, during the period December 31, 2004, through December 31, 2021.

P = per share stock price on December 31 of that year
E = per share earnings during the preceding year
G = growth rate of earnings over the next 5 years
PEG = (P/E)/G
R = stock rate of return over next 7 months; January 1 through July 31

19. Give three ways in which this presentation of the results of the estimation of the model described in the preceding exercise could be improved:

Least squares estimates of our model show a statistically significant relationship between PEG and R:

	Coefficient	*Std. Err.*
_constant	*0.0198368*	*0.0238002*
PEG	*0.0050861***	*0.0017132*

**: p<0.05 **: p<0.01*

20. How would improve this display of the breakdown of the sports played by the varsity athletes who answered a questionnaire?

Figure 1: Breakdown of Each Sport