Econ 57
Gary Smith
Spring 2014

Final Examination Answers

1. a. chi-square
   b. multiple regression
   c. matched-pair
   d. difference-in-means
   e. difference-in-proportions or chi-square

2. a. False. The t-value can be high, even though R^2 is low, if X has a statistically significant effect on Y but does not explain much of the variation in Y
   b. False. $R^2 = 1$ means there is a perfect linear relationship between X and Y, but the slope need not be 1.
   c. True.
   d. True.

3. Galileo was persuaded by the evidence to rethink this person's calculations, which mistakenly assume that each of these possible outcomes is equally likely. There are, for example, three ways to roll a 1, 4, and 4 (1-4-4, 4-1-4, and 4-4-1), six ways to roll a 1, 3, and 5 (1-3-5, 1-5-3, 3-1-5, 3-5-1, 5-1-3, and 5-3-1), only one way to roll 3-3-3. The correct number of ways to roll these numbers is

   9: 3 + 6 + 6 + 6 + 3 + 1 = 25
   10: 6 + 6 + 3 + 6 + 3 + 3 = 27
   11: 3 + 6 + 6 + 6 + 3 + 3 = 27
   12: 6 + 6 + 3 + 6 + 3 + 1 = 25

   Galileo's correct calculations are in accord with the gamblers' empirical evidence.

4. This is like the in-class example of pedestrian traffic-accident victims wearing dark clothes. We need to know the percentage of couples who have trouble conceiving but do not adopt and then eventually conceive a child.[Thomas Gilovich, Some systematic biases of everyday judgment, *Skeptical Inquirer*, March/April 1997, pp. 31-35.]

5. This is a binomial problem

$$P[X \geq 9] = \binom{13}{9}.5^9.5^4 + \binom{13}{10}.5^{10}.5^3 + \ldots = 0.1334$$

6. First, it sure sounds like data mining. (Why 5 days? Why 1928?) Second, the first 5 trading days are part of the year's (approximately) 250 trading days, so that the biases the calculations. He should have compared the first 5 days with the other 245 days. Third, the incorrect impression is that it is 50-50 whether the market goes up. In fact, the market usually goes up, 73% of the years 1926 - 2013, 78% of the years 1950-2013.

7. This is a Bayesian problem:

$$P[\text{X if beach}] = \frac{P[\text{X}]P[\text{beach if X}]}{P[\text{X}]P[\text{beach if X}] + P[\text{Y}]P[\text{beach if Y}] + P[\text{Z}]P[\text{beach if Z}]}$$

$$= \frac{(1/3)(.7)}{(1/3)(.7) + (1/3)(.5) + (1/3)(.3)} = \frac{0.7}{1.5} = 0.467$$

Similarly, P[Y if beach] = 0.5/1.5 = 0.333 and P[Z if beach] = 0.3/1.5 = 0.200.

8. [Roger Parloff, "The Gray Art of Not Quite Insider Trading," Fortune. 9/2/2013, Vol. 168 Issue 4, p. 60.]
   a. A 95% confidence interval is
   $$\bar{X} \pm t\frac{s}{\sqrt{n}} = 6.2 \pm 1.972\frac{20}{\sqrt{200}} = 6.2 \pm 2.8$$
   b. Because of the central limit theorem we do not have to assume that percentage increase in sales at individual stores is normally distributed.
   c. The confidence interval would't change if there were 32,000 stores.
   d. Unless the sample is a much larger percentage of the population (so that a finite population correction factor is needed), the accuracy of the estimate depends on the number of stores surveyed, not whether they are 1% of 5% of all stores.

9. The estimated coefficients are *ceteris paribus*.

10. Sales data for single-family houses in a Dallas suburb were used to estimate the following regression equation:

$$\begin{array}{rcrcrcrcrcrcl} Y & = & 60{,}076 & + & 75.1S & + & 36.4G & - & 3{,}295A & + & 4{,}473B & - & 14{,}632T \qquad R^2 = 0.84 \\ & & (14.4) & & (19.7) & & (12.1) & & (1{,}078) & & (1{,}708) & & (3{,}531) \end{array}$$

   where P = sale price, S = square feet of living area; G = garage square feet; A = age of house in years; B = number of baths; and T = 1 if 2-story, 0 if not. The standard errors are in parentheses.
   a. All of the estimated coefficients are statistically significant at the 5% level because all of the t-values (estimate divided by standard error) are much larger than 2.
   b. They all see reasonableAn extra square foot of living area is predicted to increase the sales price by $75.1. An extra square foot of garage is predicted to increase the sales price by $36.4. An extra year of age is predicted to reduce the price by $3,295. An extra bathroom is predicted to increase the sales price by $4,473. Being two-story *ceteris paribus* (holding total square footage constant) is predicted to reduce the price by $14,632.

11. This could happen if low-income women have more children than high-income women. Consider four 1990 women, three earning $20,000 and one earning $120,000. Average income is $45,000. Now suppose the low-income women have a total of 14 daughters (hypothetically!), each earning $30,000 and the high-income woman has one daughter, who earns $180,000. The average income is $40,000. Even though each daughter earns 50% more than her mother, the average income of the daughters is lower than the average income of the mothers.

12. Regression to the mean. The author of this quotation goes on:

    Rewards *do* work, and punishment doesn't work as well. They got the wrong impression because of something called "statistical regression." This means that statistically, an exceptionally good performance is usually followed by a performance that is not as good. A high point, in other words, is statistically likely to be followed by a regression of some sort. Of course! It is "exceptionally" good because it is *exceptional*. It is the exception.

    And in the same way and for the same reason, an exceptionally bad performance is usually followed by something somewhat better.

13. The model doesn't estimate a single value for ε. The t-values are for the coefficients (like β), not an

explanatory variable (like X)

14. [Dan Ariely, *Predictably Irrational*, pp. 1-6.] Using a chi-square test with a 2-by-2 table:

	2 choices	3 choices	Total
Online	68	16	84
Print/Online	32	84	116
Total	100	100	200

The expected values are:

	2 choices	3 choices	Total
Online	42	42	84
Print/Online	58	58	116
Total	100	100	200

The chi-square value is 55.5, and the p-value is less than 0 .000000001.

Alternatively, a difference in proportions test using $\hat{p} = \frac{68+16}{100+100} = 0.42$ gives a Z-value of

$$Z = \frac{68/100 - 16/100}{\sqrt{\frac{\hat{p}(1-\hat{p})}{100} + \frac{\hat{p}(1-\hat{p})}{100}}} = 7.4499$$

This Z-squared is 55.5 (the chi-square value) and the two-sided p-value for the Z is the same as the p-value for the chi-square test.

15. Z = (310 – 266)/16 = 2.75. P = 0.00298.

16. a. The judge's equation forces the male and female equations to have the same coefficients of S and E
    b. $Y = \alpha + \beta_1 S + \gamma E + \delta D + \beta_2 D * S + \beta_3 D * E$

17. The odds that the first four numbers chosen will be your selected numbers, followed by two misses is

$$\frac{6}{46}\frac{5}{45}\frac{4}{44}\frac{3}{43}\frac{40}{42}\frac{39}{41}$$

This is the probability of any single sequence of four hits and two misses. So we just have to count the number of winning sequences. This is "6 choose 4," or 6*5/2. So the answer is

$$\frac{6*5}{2}\frac{6}{46}\frac{5}{45}\frac{4}{44}\frac{3}{43}\frac{40}{42}\frac{39}{41} = 0.00001249$$

or about 1 in 801.

Here is an interesting article about the Massachusetts Cash Winfall lottery which, like many lotteries, rolls over jackpots when there is no winner—which can make the expected value of a ticket in the rollover lottery positive. The Winfall lottery had the unusual feature that if there was no winner in the rollover lottery, the prize money rolled down to smaller prizes (like picking 4 of 6 number). This feature made the positive expected value more attractive to ticket buyers because it increased the chances of winning an enhanced prize. You didn't have to win the rollover jackpot to get an enhanced prize. This also made it more feasible for syndicates to profit from buying lots of tickets. http://www.businessinsider.com/heres-how-an-mit-senior-a-michigan-retiree-and-two-biomedical-researchers-beat-the-lottery-2012-8.

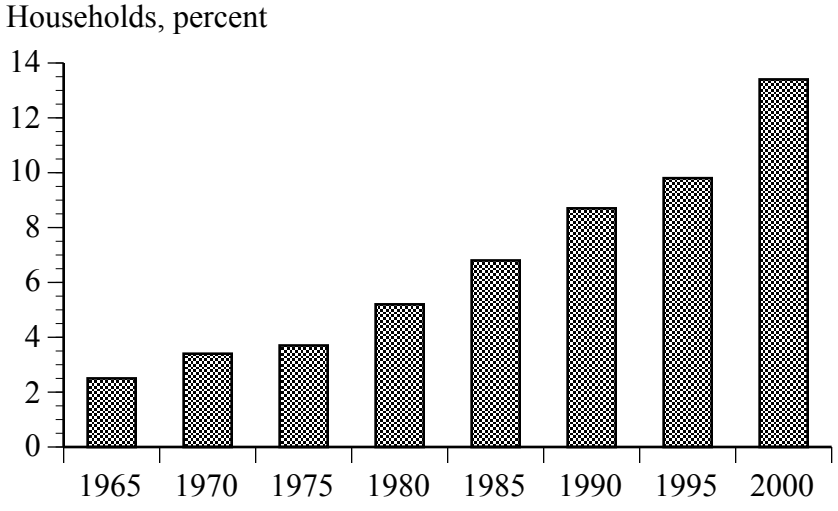
18. The model should generally come before the data. It is misleading to test a model with the data used to

derive the model.

19. The fallacious law of averages.

20. There is a 5-year difference between each of the first four bars, but a *10-year difference* between the fourth and fifth bars (1980 and 1990). If the bars had been spaced properly and a 1985 bar inserted, the increase over time would appear gradual, without an abrupt jump between 1980 and 1990. In addition, $100,000 in 1990 is not the same as $100,000 in 1965. Prices were about four times higher, so that $100,000 in 1990 was roughly equivalent to $25,000 in 1965. We should compare the number of 1965 families earning $25,000 with the number of 1990 families earning $100,000. We should also take into account the increase in the population between 1965 and 1990. It is not surprising that more people have high incomes when there are more people.

    The figure below fixes all these problems by showing the percentage of households that earned more than $100,000 in inflation-adjusted dollars, with 1985 inserted. Data for 1995 and 2000 are included to give more historical context. Viewed with appropriately adjusted data, the 1980s are unremarkable. What does stand out is the end of the 1990s, during the Internet Bubble



Percent of Households Earning More than $100,000 a Year, adjusted for inflation