Econ 57
Gary Smith
Spring 2013

Final Examination Answers

1. a. Difference in proportions or chi-square
   b. Matched-pair
   c. Chi-square
   d. Regression
   e. Binomial

2. a. People who marry and divorce self-select.
   b. Women may be less likely to marry men who are in poor health and more likely to divorce men whose health suffers the most during marriage.

3. (Bart K. Holland, What are the Chances, Johns Hopkins University Press 2002, p. 33). This is a binomial problem:

$$P[X > 1] = \sum_{x=2}^{20} \binom{20}{x} 0.05^x 0.95^{20-x} = 0.264$$

4. It weakened both. By moving players who would have been above-average in minors and below-average in majors from the minors to the majors, they lowered the average quality of both divisions.

5. The conclusion refers to the probability of having an accident if the driver is within 10 miles of home, but the 70% statistic is the reverse probability: the probability of being close to home if an accident occurs. To go from one to the other, we need to know the fraction of all driving that is within 10 miles of home.

6. p values cannot be larger than 1.

7. There should be separate 0-1 dummy variables for each state, instead of one dummy variable for all 50 states. There is no reason why D going from 1 to 2 (from Alabama to Alaska) has the same effect on household income as does D going from 2 to 3 (from Alaska to Arizona).

8. Using Bayes' Rule,

$$P[\text{high-risk} \mid \text{dies}] = \frac{P[\text{high-risk}]P[\text{die} \mid \text{high-risk}]}{P[\text{high-risk}]P[\text{die} \mid \text{high-risk}] + P[\text{low-risk}]P[\text{die} \mid \text{low-risk}]}$$
$$= \frac{0.20(0.006)}{0.20(0.006) + 0.80(0.003)} = \frac{6}{18} = 0.333$$

Using a contingency table with 100,000 women:

	Die	Doesn't Die	Total
low-risk	240	79,760	80,000
high-risk	120	19,880	20,000
total	360	99,640	100,000

The probability that a woman who died was in the high-risk group is 120/360 = 1/3. This 1/3 figure is larger than the 20% of the total population that is high-risk, but it is still far from certain that a woman who died was in the high-risk group.

9. a. The expected value of the payout is \$1,000,000(0.003) + \$0(0.997) = \$3,000 for the 80% of the women who are low-risk and \$1,000,000(0.006) + \$0(0.994) = \$6,000 for the 20% who are high-risk. The overall expected value is 0.80(\$3,000) + 0.20(\$6,000) = \$3,600, which is less than \$5,000.
   b. For a high-risk woman, the expected value of the payoff is \$6,000, which is larger than \$5,000.
   c. For a low-risk woman, the expected value of the payoff is \$3,000, which is smaller than \$5,000.
   d. If only high-risk women buy policies, the insurance company will lose money because the expected value of the payoff is \$6,000, which is larger than \$5,000.

10. We can't put a probability on the null hypothesis being true unless we do a Bayesian analysis. The p value relates to the probability of observing certain data if the null hypothesis is true, not the probability that the null hypothesis is true if we observe certain data.

11. There are 36 possible letters and digits. The probability of a perfect match with randomly chosen characters is $(1/36)^6 = 4.6 \times 10^{-10}$. The probability that at least one of three tries will be successful is equal to one minus the probability that none will be successful $1 - (1 - (1/36)^6)^3 = 1.4 \times 10^{-9}$, or about 1 in 725,594,150. With one try a day, the expected wait is 725,594,150 days (almost two million years).

12. This is an example of Simpson's paradox, which occurs if a pattern in aggregated data is reversed when the data are disaggregated. Here, the type of user—U.S. or international—is a *confounding* factor in that RPM is related not only to the click format, but also to the type of user. U.S. users have a higher RPM than do international users and U.S. users are more likely to go to 2-click pages, which boosts the overall RPM for 2-clicks. If we take this confounding factor into account—by separating the data into U.S. and international users—we see that the 1-click RPM is higher for both types of users.

13. a. chi-square
    b. regression to the mean

14. The probability of a particular Venus sequence is (0.39)(0.37)(0.12)(0.12). There are (4)(3)(2)(1) = 24 possible Venus sequences, since any of the four sides can appear on the first astragalus, any of the three remaining sides on the second astragalus, two on the third astragalus, and only the last side on the fourth astragalus. Thus, the probability of a Venus is 24(0.39)(0.37)(0.12)(0.12) = 0.0499, almost exactly 5%.

15. a. There is survivor bias in that these data exclude people who raced as Junior and, burned out, did not race as juniors. Also, this does not control for the fact that 22-year-olds will have more races than 19-year olds simply because they are older.
    b. U23 races should be the dependent variable.
    c. Yes, because the 7.91 t value is much larger than 2.
    d. No, the regression results shows that people who raced more as a junior tended to race more as a U23. (We can see from the estimated slope and intercept that Y is not equal to X.

16. This probability calculation assumes that the traits were specified before looking at the companies, not

afterward. Unless the companies are completely identical, the probability that you will find some differences between the two groups is 1.

17. The F test tests the null hypothesis that the samples were drawn from populations with the same mean. The F test is always 1-tailed, and it does consider the possibility that one mean may be either higher or lower than another. Not rejecting H0 does not prove that H0 is true. The differences do seem substantial. The F test takes into account the unequal sample sizes.

18. The difference in means test assumes that these are two independent samples. The matched-pair test assumes that observations in each sample are matched with observations in the other sample. Here, we are looking at the sport-season and off-season GPAs for each of these 26 athletes, so a matched-pair test is appropriate. It is also more persuasive because with independent samples we have to consider the possibility that the students in one group were better students than those in the other group.

19. Perfect multicollinearity since M = 1 - U. Either M or U should be dropped from the equation. It doesn't matter which one is dropped as long as researcher interprets the parameters carefully.

20. When any two things increase over time, there can be a statistical correlation without any causal relationship; for example, beer sales and the number of married people in the United States.