

Math 110
Final Exam Sample Solutions
April 30, 2010

Do all your work in the blue book. Be sure to show your calculations and provide explanations as indicated. You may use your formula sheet (half of one-side of one $8\frac{1}{2} \times 11$ piece of paper), your calculator, and the z , t , and χ^2 tables that accompany the test.

1. (30 pts.) Explain the following statistical concepts:

(a) Residual plot.

The residual plot for a test-retest data set by replacing the second or y -coordinate of each point in the plot by y -value minus the residual line value for the x -value of the point.

(b) Bootstrap method.

The bootstrap method is used in estimating population averages from sample averages. In computing the SE for the sample, the SD of the sample is used in the place of the SD of the population in the formula for the SE.

(c) Probability histogram.

A probability histogram for a chance process repeated N times is the histogram whose blocks have width 1 and have area (=height in this case) computed by the binomial formula.

(d) Regression fallacy.

The regression fallacy is to attribute regression to the mean in a test-retest situation to causation rather than the statistics of test-retest.

2. (30 pts.) Short Answer:

(a) When drawing without replacement from a box, why is it necessary to use a correction factor when calculating the standard error (SE)?

Drawing without replacement lessens the variation or spread of the numbers in the box. The correction factor accounts for this.

(b) (Hypothetical.) Two Holy Cross students, Gannon and Oglethorpe, have to design a study of student cell-phone use for class. Gannon insists on using a simple random sample but Oglethorpe counters that a simple random sample is too simple and a more sophisticated technique is required. Why would Oglethorpe say a simple random sample is too simple when surveying Holy Cross students?

In a simple random sample, different groups of students might not be appropriately represented. A quota sample or cluster sample might be better.

(c) (Hypothetical.) A town near Worcester is the location of a Superfund* toxic waste site. A doctor in the town observes an unusually high number of cases of autism in his town and calculates that for a town that size, there is a 1 in 15 chance of having so many cases. Is this reason to conclude that there is a connection between the Superfund site and the number of cases of autism? Why or why not?

No. A 1 in 15 chance is more than 5%. We would not reject the null hypothesis that this is due to chance.

*From Wikipedia: Superfund is the common name for the United States environmental policy officially known as the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, 42 U.S.C. 9601-9675), enacted by the United States Congress on December 11, 1980 in response to the Love Canal disaster and the environmental contamination at the Valley of the Drums. The Superfund law was created to protect people, families, communities and others from heavily contaminated toxic waste sites that have been abandoned.

- (d) (Hypothetical.) A major university health service agrees to participate in a trial of new medication for acne. It contacts all of the students who are currently on prescription medication for acne and offers them the opportunity to try the new medication under the supervision of a health service doctor. Of the 145 students contacted, 53 agree to take the new medication. The health service uses these 53 students for the treatment group and the remaining 92 students as controls. Is this a well-designed study? If so, why? If not, why not?

No. There is the possibility of bias because patients who volunteer for studies are more aware of health issues and likely to be in overall better health. Further, they would be more likely to adhere to a drug regime. Both factors would bias the results of the study towards saying the treatment was effective.

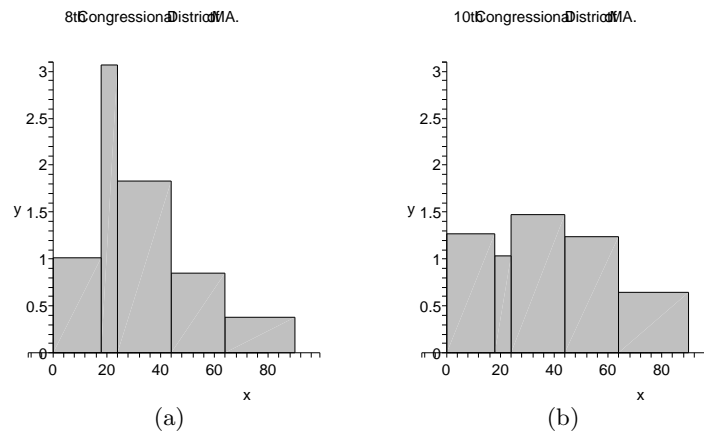


Figure 1: (a) Boston and Cambridge. (b) Cape Cod, Nantucket, and Martha's Vineyard.

3. (20 pts.) The above histograms are for the population by age in the 8th and 10th Congressional Districts of Massachusetts for the 106th Congress. There were 620,372 residents of the 8th District and 663,508 residents of the 10th District. (Data from the US Census 2000.) The 8th District contains most of Boston and all of Cambridge. The 10th District contains Cape Cod, Nantucket, and the Martha's Vineyard.

- (a) Based on the histograms, which district has a higher median age. Explain your answer.

The 10th has a higher median age. The median age in the 10th District appears to be approximately 40, with roughly half the area on either side. The histogram is skewed noticeably to the left of 40 in the 8th district, so its median would be lower.

- (b) Based on the histograms, how do the population distributions in the 8th District and 10th Districts compare?

The 10th has an older population than the 8th because the density of the population over 45 is greater than that of the 8th.

- (c) Is this difference to be expected? Why or why not?

Yes, the 8th contains Boston and Cambridge, which have larger immigrant populations and very large student populations. Many people retire to Cape Cod causing greater percentage of residents 65 and above.

4. (20 pts.) (Hypothetical) The dean at a nearby engineering school wants to analyze the grades in a two course introduction to engineering sequence that must be taken by all first year students. The course grade is based on a raw total of 500 points in each of the courses. In the fall semester the average point total is 385 with an SD of 30 and in the spring semester, the average point total is 350 with an SD of 25. The correlation coefficient for the data is $r = .6$.

- (a) Find the formula for the regression line for this data.

Use the formula:

$$r \frac{SD_y}{SD_x} (x - x_{mu}) + y_{mu} = .6 \frac{25}{30} (x - 385) + 350.$$

- (b) If a student is chosen at random what would you estimate their score to be in the second course?

The average score for the second course, 350.

- (c) If a student scores 400 in the first semester, what would you estimate his or her score to be in the second semester?

Use the regression line formula from (a) with $x = 400$:

$$.6 \frac{25}{30} (400 - 385) + 350 = 357.5.$$

- (d) If a student scores in the 70th percentile first semester, what would you estimate for his or her percentile in the second semester?

The 70th percentile corresponds to a symmetric area of 40 % and a z -score of approximately .53. To obtain a z -score for the second test, multiply by $r = .6$. The result is $.53 \times .6 \approx .32$. This corresponds to a symmetric area of approximately 25% and a percentile of $25 + \frac{1}{2}(100 - 25) \approx 63\%$.

5. (20 pts.) Answer the following questions about rolling a fair die six times:

- (a) What are the chances of rolling a 1 or a 2 on the first three rolls and *no* 1 or 2 on the last three rolls?

$$\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 \times 100\% \approx 1.1\%.$$

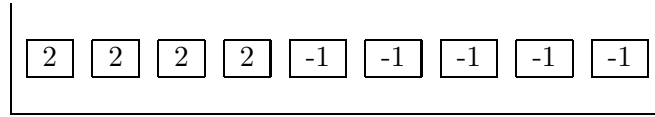
- (b) What are the chances of rolling 1 or a 2 exactly 3 times out of six rolls?

$$\frac{6!}{3!3!} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 \times 100\% \approx 21.9\%.$$

(c) What are the chances of not rolling a 1 or 2 at all?

$$\left(\frac{2}{3}\right)^6 \times 100\% \approx 8.8\%.$$

6. (20 pts.) Sixty draws with replacement are made from the following box:



(a) What are chances that the sum of the draws will be between 15 and 18 inclusive?

Convert both totals, 15 and 18, to standard units, subtract their symmetric areas, and divide by two to get the answer. Start by computing the box average, SD, and SE. The box average is $\frac{1}{3}$. The EV is $60 \times \frac{1}{3} = 20$. The SD of the box is $(2 - (-1))\sqrt{\frac{4}{9} \frac{5}{9}} \approx 1.5$. The SE for sum is $\sqrt{60} \times 1.5 \approx 11.6$. Then $\frac{20-15}{11.6} \approx .43$ and $\frac{18-15}{11.6} \approx .17$. The corresponding areas are 33% and 13%. The answer is $\frac{1}{2}(33 - 13) = 10\%$.

(b) What are chances that number of 2s drawn will be 30 or more? (*Hint*: What is the box model for this question?)

This requires the use of a 0-1 box with four 1's and five 0's. The expected value for the number of 2's is $\frac{4}{9} \times 60 = 26\frac{2}{3}$. The SD is also $\approx .5$ and the SE for count is $\sqrt{60} \times 0.5 \approx 3.87$. The z -score is $\frac{30-26.66}{3.87} \approx .86$. We want the area of the corresponding right tail, $\frac{1}{2}(100 - 61) = 19.5\%$.

7. (20 pts.) (Hypothetical) After a number of complaints, the librarians in Dinand decide to conduct a poll to determine whether study rooms in the library should be kept open 24 hours a day during study period and exam period. They conducted a simple random sample of 78 library users. Of the 78, 43 said study rooms should remain open 24 hours a day and 35 said it should not. Before changing policy the librarians put a plus-minus number on the result to make sure the favorable response is not due to the chance in sampling.

(a) What box model should the librarians use?

A 1-0 box with a ticket for each student. The percentage of 1's and 0's is not known. A total of 78 draws made from the box.

(b) Compute a 95% confidence interval for this poll.

To compute the SE, use the sample SD: $(1 - 0)\sqrt{\frac{43}{78}\sqrt{3578}} \approx .5$. The SE for percentage is $\frac{.5}{\sqrt{78}} \times 100 \approx 5.6\%$. The sample percentage is 55%, so the 95% confidence interval is $55 \pm 11.2\%$.

(c) Is it likely that a majority of all students favor keeping the library open based on the results of this poll? Explain.

Since 50% lies within the 95% confidence interval, in fact within the 68% confidence interval, we cannot conclude that a majority favor keeping the library open.

8. (20 pts.) A large university mathematics department has a policy that students in a calculus class should work on average at least 7 hours per week on calculus outside of class. If the average in an instructor's class falls below this number, the instructor is told that he or she must assign more homework. Since monitoring every student's hours is time consuming and expensive, the department takes sample numbers from each class. If the average falls below 7 hours per week, the department runs a test of significance and if the difference in the sample average is statistically significant, the homework policy is applied. In one instructor's section, which had 32 students, the out-of-class hours worked in a six person sample were

5, 3, 4, 9, 3, 7

- (a) What test of significance should be used to determine if this instructor will need to assign more homework? Why?

A one sample t -test, since the sample size is less than 25.

- (b) What are the null and alternative hypotheses for this test?

The null hypothesis is that the difference between the sample average, 5.17, and 7 is due to the chance in sampling. The alternative hypothesis is that the difference is real.

- (c) Carry out your test of significance.

Compute the SD, SE, and SE+ of the sample. The SD is

$$\sqrt{\frac{1}{6}((5 - 5.17)^2 + (3 - 5.17)^2 + (4 - 5.17)^2 + (9 - 5.17)^2 + (3 - 5.17)^2 + (7 - 5.17)^2)} \approx 2.2.$$

The SE for average is $\frac{2.2}{\sqrt{6}} \approx .90$. The SE+ is $\sqrt{\frac{32-6}{32-1}}.9 \approx .82$. The observed t -statistics is $\frac{7-5.17}{.82} \approx 2.23$. Using t -table for degrees of freedom = 5, we obtained an observe P -value $< 5\%$. We conclude the difference is statistically significant and we reject the null hypothesis.

- (d) Will the department require the instructor to assign more homework?

Yes.

9. (20 pts.) The following question and responses appear in the SDA archive:

Question: Do you ever read a horoscope or your personal asatitrology report?

The valid responses to this question by sex among 18-22 year old men and women are as follows:

	Male	Female	Row Total
Yes	31	64	95
No	27	11	38
Col Total	58	75	133

Statistically, we would like to know whether the difference in percentages of responses for men and women is "real" or due to the chance error in sampling.

- (a) Formulate a null and alternative hypothesis for this question.

The null hypothesis is that the difference in percentages are due to the chance of the sampling. The alternative hypothesis is that the difference is real.

- (b) What significance test should you use to determine the answer?

Use a χ^2 test.

- (c) Carry out your test of significance. What do you get for a test statistic and observed significance level?

The total sample percentages are 71.4% yes and 28.6% no. We use these to compute the expected frequencies. These are: 41.4 for male for yes; 16.6 for male for no; 53.6 for female for yes; and 21.5 for female for no. The χ^2 statistic is:

$$\frac{(31 - 41.4)^2}{41.4} + \frac{(27 - 16.6)^2}{16.6} + \frac{(64 - 53.6)^2}{53.6} + \frac{(11 - 21.5)^2}{21.5} = 16.3$$

The degrees of freedom are $(2-1) \times (2-1) = 1$. Using the χ^2 table, the observed significance level is $< 1\%$.

- (d) Is the difference real or not?

The difference is real.