University of California, Berkeley, Statistics 20, Lecture 1
Michael Lugo, Fall 2010

Final Exam

December 13, 2010, 8:10 am - 11:00 am

Name:  

Signature: 

Student ID: 

Section (circle one): 

101 (Joyce Chen, TR 9) 

102 (Moorea Brega, TR 10) 

103 (Moorea Brega, TR 11) 

104 (Joyce Chen, TR 11) 

This exam consists of fifteen pages: this cover page, ten pages each containing one question, one page containing five true-false questions, and three pages of tables. You may use a calculator, and notes on three sides of a standard 8.5-by-11-inch sheet of paper which you have written by hand, yourself. You must show all work other than basic arithmetic.

The total number of points available is 200.

DO NOT WRITE BELOW THIS LINE

<table>
<thead>
<tr>
<th>Question: 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points available:</td>
<td>25</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Your score:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. [25 points: \(5 + 6 + 5 + 5 + 4\)] In the land of Encinal, squirrels take two tests upon graduating high school: test C, which tests their ability to climb trees, and test H, which tests their ability to hide acorns. Scores on these tests are normally distributed. The average score on test C is 10, with a standard deviation of 3; the average score on test H is 12, with a standard deviation of 5. The scores on tests C and H are uncorrelated.

a. Draw the scatter diagram for scores on tests C and H.

b. On your diagram, draw the SD line, the regression line for predicting C-scores from H-scores, and the regression line for predicting H-scores from C-scores.

c. Oak College only accepts squirrels whose total score on the two tests is at least 25. What percentage of squirrels are they willing to accept?

d. Among squirrels whose scores are high enough to get into Oak College, the coefficient of correlation between C-score and H-score is closest to (circle one)

\[-0.9 \quad -0.5 \quad 0 \quad +0.5 \quad +0.9\]

Explain your answer.

e. A researcher working at Oak College finds the coefficient of correlation of a simple random sample of squirrels at the college. The researcher publishes a paper commenting on this coefficient of correlation. Should they do this? If so, why? If not, why not?

(d) See the dotted line on the diagram. "All admitted students are above the lines."

(e) No; because we're restricted to just those students who could get in, this population conclusion won't hold up in the population of all squirrels.
2. \([20: 5 + 5 + 5 + 5]\) Four people are chosen at random. What is the probability that:
   (a) No two of them have their birthday in the same month?
   (b) All four have their birthday in the same month?
   (c) One of the first three people chosen has their birthday in the same month as the fourth person?
   (d) The first and second people chosen have their birthdays in the same month, given that there is some pair of people having their birthdays in the same month?

\[(a) \quad \frac{12 \times 11 \times 10 \times 9}{12 \times 12 \times 12 \times 12} = \frac{55}{96} = 0.5729\]

\[(b) \quad \frac{1}{12^3}\]

\[(c) \quad = 1 - P(\text{none of first three in same month})\]
   
   \[= 1 - \left(\frac{11}{12}\right)^3 = \frac{397}{1728} = 0.2297\]

\[(d) \quad \frac{P(\text{first and second in same month})}{P(\text{some pair in same month})}\]

\[= \frac{\frac{1}{12}}{1 - \frac{55}{96}} = \frac{8/96}{41/96} = \frac{8}{41} = 0.1951\]
3. [15: 7 + 3 + 5]
(a) A coin is flipped \( N \) times. 52% of the time it comes up heads, and 48% of the time it comes up tails. How large must \( N \) be for this observation to be statistically significant evidence that the coin is not fair?
(b) Repeat (a) if the coin comes up heads 49% of the time.
(c) Comment on the difference between the answers in (a) and (b).

\[
(a) \ z = \frac{0.52N - 0.50N}{\sqrt{N \times \frac{1}{2} \times \frac{1}{2}}} = \frac{0.02N}{\frac{1}{2} \sqrt{N}} = \frac{1}{25} \sqrt{N}.
\]

To have \( z = 2 \), need \( \sqrt{N} = 50 \). \( N = 2500 \).

\[
(b) \ z = \frac{-0.01N}{\frac{1}{2} \sqrt{N}} = \frac{-1}{50} \sqrt{N}.
\]

To get \( z = -2 \), need \( \sqrt{N} = 100 \), \( N = 10,000 \).

(c) A difference of half the size requires \( 2^2 = 4 \) times as many same observations.
4. [20: 10 + 10]
One hundred words were chosen at random from all words in Shakespeare’s *Hamlet* having at least two letters, and the first two letters of each were examined. The results were:

- 30 begin with two consonants (for example, “shuffled”)
- 48 begin with a consonant followed by a vowel (for example, “perchance”)
- 21 begin with a vowel followed by a consonant (for example, “against”)
- 1 begins with two vowels (“outrageous”)

Among words in ordinary English text:

- 24% begin with two consonants
- 49% begin with a consonant followed by a vowel
- 24% begin with a vowel followed by a consonant
- 3% begin with two vowels

(a) Test the hypothesis that in *Hamlet*, whether the first letter of a given word is a vowel is independent of whether the second letter in that word is a vowel.

(b) Using the data available to you here, test the hypothesis that Shakespeare wrote *Hamlet* by picking words at random from ordinary English text.

(You should assume that all samples are simple random samples.)

\[
\begin{array}{c|cc|cc}
& C & V & \text{Observed} & \text{Expected} \\
\hline
\text{1st letter} & \text{2nd letter} & & & \\
C & 30 & 48 & 78 & 39.78 & 38.22 \\
V & 21 & 1 & 22 & 11.22 & 10.78 \\
\hline
 & 51 & 49 & 100 & \\
\end{array}
\]

\[
\chi^2 = \frac{(30-39.78)^2}{39.78} + \frac{(48-38.22)^2}{38.22} + \frac{(21-11.22)^2}{11.22} + \frac{(1-10.78)^2}{10.78}
\]

\[
= 22.30, \quad 1 \text{ df} \rightarrow P \text{ very small, reject hypothesis}
\]
5. [15] A simple random sample of 200 households renting apartments or houses in the 94704 zip code (which includes the areas immediately south and west of the UC Berkeley campus) was taken. The average rent reported was $1,150, with a standard deviation of $483.

A simple random sample of 100 households renting apartments or houses in the 94709 zip code (which includes areas immediately north of campus) was also taken. The average rent reported was $1,331, with a standard deviation of $610.

Test the hypothesis that the average rent is higher in the 94709 zip code than in the 94704 zip code.

\[
Z = \frac{1331 - 1150}{\sqrt{\frac{483^2}{200} + \frac{610^2}{100}}} = \frac{181}{196} = 0.924.
\]

\[
P = 1 - 0.43145 = 28\%.
\]

We accept (or fail to reject) the null hypothesis that average rents are the same.
6. [15: 10 + 5] I make 5 measurements of the temperature in my refrigerator; they are 35, 36, 34, 37, and 33 degrees Fahrenheit.

(a) Test the hypothesis that the actual temperature in my refrigerator is 32 degrees Fahrenheit.

(b) To convert a temperature from Fahrenheit to Celsius, we have the formula \( {}^\circ C = \frac{5}{9}({}^\circ F - 32) \). (For example, 50 degrees Fahrenheit is \((5/9)(50-32) = (5/9)(18) = 10\) degrees Celsius.) What is the mean and SD of the temperature measurements in degrees Celsius?

\[
\text{avg} = \frac{35 + 36 + 34 + 37 + 33}{5} = 35
\]

\[
SD = \sqrt{\frac{0^2 + 1^2 + 1^2 + 2^2 + 2^2}{5}} = \sqrt{2} = 1.41
\]

\[
SD^t = \sqrt{\frac{5}{4}} \quad SD = \sqrt{2} = 1.58
\]

\[
t = \frac{35 - 32}{1.58} = 1.90
\]

From table (4 df): \( 5% < P < 10\% \) (one-tailed) we accept the null hypothesis that the temperature is 32°.

(b) \( \frac{5}{9}(35 - 32) = 1.67 \)

\[ SD \frac{5}{9}(1.41) = 0.78 \]
7. [20: 8 + 4 + 8] Before the US presidential election, a polling firm surveys 1,000 people in each of the 50 states and asks who they will vote for. Assume simple random sampling, that everyone either votes for the Democrat or the Republican, and that everyone actually votes for the candidate they say they would vote for.

For each state, they report a 96% confidence interval for the percentage of people in that state who will vote for the Democrat.

(a) In one state, the survey shows 530 people voting for the Democrat and 470 voting for the Republican. What confidence interval does the polling firm report?

(b) What is the expected number of states for which the confidence interval reported does not include the true percentage of people in that state who will vote for the Democrat?

(c) What is the probability that in every state, the confidence interval reported includes the true percentage of people in that state who will vote for the Democrat?

\[
\text{(a) } SE \sqrt{\frac{0.53 \times 0.47}{1000}} = 0.0158.
\]

96% CI \[ \frac{0.53 \pm 0.0158 \times 2.05}{2.05} \]
= \[ 0.53 \pm 0.03241 \]
= \[ (49.8\%, 56.2\%) \]

(b) \( 50 \times (1 - 0.96) = 2 \)

(c) \( 0.96^{50} \approx 0.13 \).
8. [15] Consider the following three sums of draws:
(1) twenty-five draws from a box containing nine 0s and one 1
(2) fifteen draws from a box containing one 0 and nine 1s
(3) five draws from a box containing one 0 and one 1
(4) ten draws from a box containing one 0, one 1, and one 5
Match these with the three probability histograms (with axes unlabeled) below by circling the number of the corresponding description under each histogram. Explain your answers. One number will be unused.

Histogram (ii) is not smooth—it must be (4), since the result of this process depends heavily on the number of 5s drawn.

(iii) shows some right skew, which we expect from a process like (1)’s, with average 2.5 and SE 1.5; we should see some squashing up against zero.

(i) is symmetric, as expected from (3). In particular, we have the same numbe of 0s and 5s, 1 and 4, 2 and 3.

(ii) is unused.
9. \([20: 10 + 10]\) (a) I flip a coin ten times. It comes up heads twice. Should I conclude that it is biased to turn up tails, or could this just be chance variation?

(b) I flip a coin one hundred times. It comes up heads twenty times. Should I conclude that it is biased to turn up tails, or could this just be chance variation?

(a) \(\text{EV (\# of heads)} = \frac{10}{2} = 5\)

\(\text{SE (\# of heads)} = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = 1.58\)

\(Z = \frac{2 - 5}{1.58} = 1.90\)

\(\frac{2.5 - 5}{1.58} = -1.58\).

\(\rightarrow P \approx 6\% \). This could be chance variation.

Or: \(P(\leq 2 \text{ heads}) = \frac{C(10)^{10} + \binom{10}{0} + \binom{10}{2}}{2^{10}} = \frac{1 + 10 + 45}{1024} \approx 5.5\%\).

(b) \(\text{EV (\# of heads)} = \frac{100}{2} = 50\)

\(\text{SE (\# of heads)} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5\)

\(Z = \frac{20 - 50}{5} = -6\) \(121\) very large

\(\rightarrow P \text{ very small. } \rightarrow \text{ We conclude that bias is likely.}\)
10. [20: 10 + 10] The distribution of midterm and final exam scores for a class both follow the normal curve, and the scatter diagram is football-shaped.
   The average score on the midterm is 70, with SD 10.
   The 25th percentile of scores on the final is 50, and the 90th percentile is 78.
   The correlation coefficient is 0.5.
   (a) Find the regression line for predicting final exam scores from midterm exam scores.
   (b) A student scores 80 on the midterm. Give two numbers such that the probability that the student’s final exam score falls between them is approximately 95%.

(a) We need the SD, for final exam scores.
   50 is 0.7 SD below average
   78 is 1.3 SD above average
   So SD is \( \frac{78 - 50}{1.3 + 0.7} = 14 \).
   mean \ 50 + (10 \cdot 0.7)(14) = 60.

Regression line: 
\[
y - 60 = 0.5 \cdot \frac{14}{10} \cdot (x - 50)
\]
\[
y = 0.7x + 11.
\]

(b) Predicted score is \( 0.7 \cdot 80 + 11 = 67 \).

\[\text{RMSE} = \sqrt{1 - r^2} \cdot SD(y) \]
\[= \sqrt{1 - 0.5^2} \cdot (14) \approx 12\]

Prediction interval: \( 67 \pm 2 \times 12 = 67 \pm 24 = [43, 91] \).
11. [15 points: 3 each] Answer the following true-false questions. No explanation is necessary. Give your answers by circling the appropriate choice for each question. You’ll receive 3 points for the right answer, 0 for the wrong answer, and 1 if you circle “I don’t know”.

(a) To determine the distance from B to C, I measure the distance from A to B and from A to C, where A, B, and C lie on a straight line. The standard error of my measurement of the A-B distance is 3; the standard error of my measurement of the A-C distance is 4. True or false: The standard error of the B-C distance is 5.

TRUE / FALSE / I DON’T KNOW

(b) True or false: Given a histogram for a data set, we can compute the average of the data set exactly.

TRUE / FALSE / I DON’T KNOW

(c) I take a simple random sample of 100 Berkeley students and ask them how many books they have read in the last year, and compute a 95% confidence interval for the average number of books read by all Berkeley students in the last year. I then repeat this experiment with a simple random sample of size 400. True or false: The size of the second confidence interval will be about half that of the first.

TRUE / FALSE / I DON’T KNOW

(d) A surveyor asks a large number of people two yes/no questions, and wishes two test the hypothesis that their answers to these two questions are independent. She computes \( \chi^2 = 5 \). True or false: She rejects the hypothesis of independence at the 95% confidence level.

TRUE / FALSE / I DON’T KNOW

(e) The average American male is 69 inches tall, with a standard deviation of 3 inches. True or false: Among men who are between 72 and 78 inches tall, the average height is 75 inches.

TRUE / FALSE / I DON’T KNOW