University of California, Berkeley, Statistics 21: Probability and Statistics for Business

Michael Lugo, Spring 2012

Final exam

Wednesday, May 9, 3:10 pm - 6:00 pm

Name: Key (5/3) Student ID:

This exam consists of eleven pages: this cover page and ten pages containing a total of thirty-three questions, some of which are linked to each other. Please pick up separately three pages of tables. You may use a calculator, and notes on three sides of a standard 8.5-by-11-inch (or A4) sheet of paper which you have written by hand, yourself.

You must show your work on all questions except multiple-choice questions. (The multiple-choice questions are questions 1 through 11.) You may use a calculator to do arithmetic but you should indicate what you are asking the calculator to do; you should not use built-in calculator functions to compute standard deviations or correlation coefficients.

Write your name at the top of each page.

The total number of possible points is 85. The number of points for each question are indicated in brackets.

DO NOT WRITE BELOW THIS LINE

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1-6.</td>
<td></td>
<td>21-24.</td>
</tr>
<tr>
<td>12-15.</td>
<td></td>
<td>29.</td>
</tr>
<tr>
<td>16.</td>
<td></td>
<td>30.</td>
</tr>
<tr>
<td>17-20.</td>
<td></td>
<td>31-33.</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>/85</td>
</tr>
</tbody>
</table>
The first part of this test consists of some true-or-false, multiple-choice, and fill-in-the-blank questions. You do not need to show your work and your work will not be graded if you do show it.

1. [2] The correlation between the scores on the midterm and final exams in some course is 0.66. One student scores at the 90th percentile on the midterm. The best prediction of their percentile on the final is (circle one)
   50  60  70  80  90

2. [2] I flip a fair coin 100 times. The probability of getting 50 heads is closest to
   0  1%  4%  8%  20%

3. [2] We want to know whether there is a difference between the rates at which men and women pass Stat 21. Which of the following tests is appropriate? CIRCLE ALL THAT APPLY.
   \( \chi^2 \) test for independence  \( t \)-test
   \( \chi^2 \) test for goodness of fit  two-sample \( z \)-test for averages
   one-sample \( z \)-test for averages  two-sample \( z \)-test for proportions
   one-sample \( z \)-test for proportions

4. [1] The average height of the fathers in some population is 68 inches, with standard deviation 3 inches. The average height of the sons is 69 inches, with standard deviation 4 inches. For each pair of father and son we subtract the father's height from the son's height. The average of these differences is (fill in the blank) \( \underline{1} \) inches.

5. [2] (Continues question 4.) The standard deviation of the differences computed in question 4 is
   \( 1 \) inch  between 1 and 5 inches  \( 5 \) inches  between 5 and 7 inches  \( 7 \) inches
   (positive correlation)

6. [2] You carry out a \( z \)-test and a \( t \)-test on the same data. The \( P \)-value you get from the \( t \)-test is larger than the same as smaller than the \( P \)-value from the \( z \)-test.
Name: __________________________

We have data on the distribution of grades (percentage of A, B, C, D, and F) obtained by students in each major taking Stat 21. True or false [1 point each]:
7. We could use a $\chi^2$ test to determine if different majors have different distributions of grades.

TRUE

8. We could use a $\chi^2$ test to determine which majors tend to get higher grades.

TRUE

Questions 9 through 11 refer to the following situation.
A sabermetrician \(^1\) is interested in how the number of home runs a Major League Baseball player hits is related to the player’s salary. (If you’re not familiar with baseball: hitting home runs is good, so one would expect players who hit more home runs to be more highly paid.)
9. [2] First, he computes the correlation between the number of home runs and the salary for all players. Call this $r_1$. Then he computes it only for first basemen, who are generally among those players who hit more home runs. Call this $r_2$. Which is most likely to be true?

\[ r_1 < r_2 \quad r_1 = r_2 \quad r_1 > r_2 \]

10. [2] Next, for each of the thirty Major League Baseball teams, he computes the average number of home runs hit by players on that team, as well as the average salary of the players on the team. He computes the correlation between the 30 pairs of team averages. Call this $r_3$. Which is most likely to be true?

\[ r_1 < r_3 \quad r_1 = r_3 \quad r_1 > r_3 \]

11. [2] All Major League Baseball teams have the same number of players, namely 25. The sabermetrician computes, for each team, the total number of home runs hit by players on that team and the total salary of all the players on the team. He then finds the correlation between the 30 pairs of team totals. Call this $r_4$. Which of the following statements must be true? (There is exactly one correct answer.)

\[ r_1 < r_4 \quad r_1 = r_4 \quad r_1 > r_4 \]

\[ r_2 < r_4 \quad r_2 = r_4 \quad r_2 > r_4 \]

\[ r_3 < r_4 \quad r_3 = r_4 \quad r_3 > r_4 \]

\[ \text{mixture of constant doesn't change } r. \]

---

\(^1\)This is a whimsical name for people who study baseball statistics, after SABR, the Society for American Baseball Research.
Name: 

The second part of this test consists of questions on which you must show your work.
For questions 12-15, recall that when three fair dice are rolled, the sum of the numbers obtained has probability 25/216 of being 9, and has probability 27/216 of being 10.
12. [3] I roll three dice 1,000 times. What is the probability I roll 9 at least 125 times?

$$\text{EV} = 1,000 \times \frac{25}{216} = 115.7$$

$$\text{SE} = \sqrt{1,000 \times \frac{25}{216} \times \left(1-\frac{25}{216}\right)} = 10.1$$

$$\frac{124.5 - 105.7}{121} = 0.87 \Rightarrow \text{Prob} \left(1 - \frac{0.6047}{2}\right) \approx 20\%.$$  

13. [2] I roll three dice 1,000 times. The difference between the number of 10s rolled and the number of 9s rolled is like the sum of 1,000 draws from a certain box. What box?

$$\begin{bmatrix}
\frac{27}{216} & 4 \\
\frac{25}{216} & -1 \\
\frac{164}{216} & 0
\end{bmatrix}$$

$$216 \times (27+25) = 216 \times 52 = 11,088$$

14. [4] What is the probability that, when rolling three dice 1,000 times, I roll more 10s than 9s?

$$\text{box has EV} = \frac{2}{216}, \text{SE} = \sqrt{\frac{2}{216} - \left(\frac{2}{216}\right)^2} = 0.4906$$

$$\Rightarrow 1,000 \text{ draws have EV } 9.3, \text{ SE } 4906/\sqrt{1000} = 15.5$$

$$z = \frac{0 - 9.3}{15.5} = -0.6 \Rightarrow \text{Prob} \left(1 + \frac{0.4515}{2}\right) \approx 73\%.$$  

15. [2] How many times must I roll three dice for the probability of rolling more 10s than 9s to be 99 percent?

$$\text{need } z = 2.35 \uparrow \text{ from table}$$

$$z = \frac{0.0093}{0.4906\sqrt{n}} \Rightarrow n = \left(\frac{0.4906}{0.0093 \times 2.35}\right)^2 \approx 15,400$$
16. [8] Find the equation of the regression line, in the form \( y = mx + b \), for the data:

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>13</td>
</tr>
</tbody>
</table>

\[
\text{mean } x = \frac{0 + 1 + 3 + 4 + 4 + 6}{6} = 3
\]

\[
\text{SD } x = \sqrt{\frac{3^2 + 2^2 + 0^2 + 1^2 + 1^2 + 3^2}{6}} = 2
\]

\[
\text{mean } y = \frac{0 + 7 + 1 + 2 + 9 + 13}{6} = 7
\]

\[
\text{SD } y = \sqrt{\frac{7^2 + 0^2 + 6^2 + 5^2 + 2^2 + 6^2}{6}} = 5
\]

\begin{tabular}{|c|c|c|c|c|}
\hline
\( x \) & \( y \) & \( \text{Std- } x \) & \( \text{Std- } y \) & \( \text{Product} \) \\
\hline
0 & 0 & -1.5 & -1.4 & 2.1 \\
1 & 7 & -1.0 & 0 & 0 \\
3 & 1 & 0 & -1.2 & 0 \\
4 & 12 & 0.5 & 1.0 & 0.5 \\
4 & 9 & 0.5 & 0.9 & 0.9 \\
6 & 13 & 1.5 & 1.2 & 1.8 \\
\hline
\end{tabular}

\[ r = \frac{4.6}{16} = 0.762 \]

\[ y - \overline{y} = 0.767 \left( \frac{5}{2} \right) (x - 3) \]

\[ y - \overline{y} = 1.92(x - 3) \]

\[ y = 1.92x + 0.24 \]
17. [2] Consider the data set from question 16. You are asked to predict the value of $y$ for a new data point with $x = 2$. What is your prediction?

$$y = (1.92)(2) + 1.24 = 3.84 + 1.24 = 5.08$$

18. [2] By how far is the prediction in question 17 likely to be off?

$$\sqrt{1 - (0.767)^2} \times 5 = 3.21$$

Questions 19 through 24 refer to the following situation.

A recent survey estimated the U.S. per capita consumption of sugar-sweetened beverages (SSBs) among adults 20 to 44 years of age to be 289 calories per day, with a standard deviation of 529 calories. 5713 people in this age range were surveyed: 2742 men and 2971 women. Each one was asked to report all food and beverages that they had consumed the previous day. For the calculations that follow, assume simple random sampling.

19. [3] Find a 90 percent confidence interval for the mean number of calories of SSBs consumed per day.

$$289 \pm (1.64) \frac{529}{\sqrt{5713}} = 289 \pm 11$$

$$= [278, 300]$$

20. [2] Is the number of calories of SSBs consumed by adults in this age range close to normally distributed?

YES

NO

Explain briefly.

The SD is much larger than the mean, but calories consumed cannot be negative!
21. [2] 68 percent of adults between 20 and 44 years old in the sample consumed SSBs at all on the day in question. What was the average consumption of SSBs among those people with nonzero SSB consumption?

\[
\frac{2.39}{0.68} = 4.25
\]

22. [3] Give a 95 percent confidence interval for the proportion of adults aged 20 to 44 who consume SSBs on any given day.

\[
0.68 \pm 1.96 \sqrt{0.68 \times 0.32 \div 5.713} = 0.68 \pm 0.012 = [67\%, 69\%].
\]

23. [1] 64 percent of the women aged 20-44 in the sample, and 74 percent of the men aged 20-44 in the sample, consumed SSBs on the day in question. What hypothesis test could you use to determine whether this difference is statistically significant? (There may be more than one possible answer.)

Two-sample z-test for proportions


\[
SE(\text{diff}) = \sqrt{0.74 \times 0.26 \div 27.92} = 0.0084
\]

\[
SE(\text{male}) = \sqrt{0.64 \times 0.36 \div 27.71} = 0.0088
\]

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{0.0084^2 + 0.0088^2} = 0.0122.
\]

\[
z = \frac{0.74 - 0.64}{0.0122} \approx 9.2
\]

\[
\text{very small, difference significant.}
\]
25. [2] I roll four fair, six-sided dice. What is the probability that at least one of the dice comes up 6?

\[
1 - \left( \frac{5}{6} \right)^4 = \frac{671}{1296} \approx 0.52.
\]

26. [2] I pick four cards from a standard deck of cards. What is the probability of getting at least one king?

\[
1 - \frac{ \binom{48}{4} }{ \binom{52}{4} } = \frac{ \binom{4}{1} \times \frac{47}{51} \times \frac{46}{50} \times \frac{45}{49} }{1} \approx 0.28
\]

27. [2] I flip six biased coins, each of which has probability 0.3 of coming up heads. What is the probability of getting exactly two heads?

\[
\binom{6}{2} (0.3)^2 (0.7)^4 = 15 (0.3)^2 (0.7)^4 = 0.324135
\]

28. [2] A box contains five tickets, two marked with a star and the other three blank. Two draws are made at random without replacement from this box. What is the chance that exactly one of the two tickets with a star is drawn?

\[
\frac{2}{5} \times \frac{3}{4} + \frac{3}{5} \times \frac{2}{4} = \frac{6 + 6}{20} = 0.6
\]
29. [8] A study aimed to determine the relationship between class attendance and grades in certain "gateway courses" (courses which are prerequisites for other key courses). 719 students in gateway courses were randomly selected. Each student was asked what proportion of the time they attended the gateway course. Their grades were recorded in two categories: ABC (passing grades: A, B, or C) and DFW (failing grades: D, F, or withdrawal). The results are recorded in the following table.

<table>
<thead>
<tr>
<th>class attendance</th>
<th>ABC frequency</th>
<th>DFW frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than 50%</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>51% to 74%</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>75% to 94%</td>
<td>135</td>
<td>54</td>
</tr>
<tr>
<td>95% or more</td>
<td>350</td>
<td>92</td>
</tr>
</tbody>
</table>

Test the hypothesis that passing the course is independent of attendance. Name the test to be carried out, do it, and give a P-value and a conclusion.

\[ \chi^2 \text{ test for independence} \]

\[
\begin{array}{cccc}
\text{marginals:} & 11 & 9 & 20 \\
& 43 & 25 & 68 \\
& 135 & 54 & 189 \\
& 350 & 92 & 442 \\
\hline
539 & 180 & 719 \\
\end{array}
\]

\[
\chi^2 = \frac{(11-15)^2}{15} + \frac{(9-5)^2}{5} + \ldots + \frac{(92-111)^2}{111} \\
= 15.02.
\]

\[ (4-1)(2-1) = 3 \text{ df} \Rightarrow P < 1\%. \]

We conclude that passing the course is not independent of attendance.
30. [4] The city of London was bombed by Nazi Germany during the London Blitz in 1940. After the war, one square of twelve kilometers on a side was divided into 576 smaller squares, each half a kilometer on a side, and the number of areas hit exactly $k$ times was counted. The observed number of areas hit $k$ times are given by the following table:

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of areas hit $k$ times</td>
<td>229</td>
<td>211</td>
<td>93</td>
<td>35</td>
<td>8</td>
</tr>
</tbody>
</table>

A certain theoretical model predicts that if the bombings took place uniformly, then the chance that any given square is hit $k$ times is given by $e^{-\lambda} \lambda^k / k!$, where $\lambda = 0.9323$ is the average number of hits per small square. In particular we have the numerical values

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>chance of $k$ hits</td>
<td>0.3936</td>
<td>0.3670</td>
<td>0.1711</td>
<td>0.0532</td>
<td>0.0151</td>
</tr>
</tbody>
</table>

Do the observed values fit the theoretical model? Perform an appropriate statistical test.

\[
\begin{align*}
\chi^2 &= \frac{(229 - 226.7)^2}{226.7} + \ldots + \frac{(8 - 8.7)^2}{8.7} \\
&= 1.03 \\
5-1 &= 4 \text{ df} \rightarrow 90\% < P < 95\%.
\end{align*}
\]

We conclude that the model is fit well.
Since previous studies have reported that elite athletes are often deficient in their nutritional intake, a group of researchers decided to evaluate elite Canadian athletes. 114 male Canadian athletes were surveyed. The average caloric intake was 3077.0 calories per day, with a standard deviation of 987.0. The recommended amount is 3421.7 calories per day. Is there evidence that elite Canadian athletes are deficient in their caloric intake?

31. [2] What are the null and alternative hypotheses?

null: average intake in population = 3421.7
alt: ≤ 3421.7

32. [4] Carry out an appropriate statistical test, give the P-value, and state your conclusion.

\[ z = \frac{344.7 - 3077.0}{987.0} = 3.72 \]
\[ P = 1 - 0.9998 \approx 0.0002 \]

we conclude they are deficient

33. [3] What is a 95 percent confidence interval for the average deficiency in caloric intake?

\[ 344.7 \pm 1.96 \times 987.0 \]
\[ = 345 \pm 181 \]
\[ = [164, 526] \]