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

Michael Lugo, Spring 2011

Final exam

May 9, 2011, 7:10 - 10:00 pm

Name: Key (5/1/12)

Student ID:

This exam consists of sixteen pages: this cover page; twelve pages containg problems; 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.

On questions 1 through 8, you need not show any work, and any work you do show will not be graded. On questions 9 through 16, you must show all work other than basic arithmetic.

Write your name at the top of each page.

DO NOT WRITE BELOW THIS LINE

Question:	1	2	3	4	5	6	7	8
Points available:	5	5	5	5	5	5	5	5
Your score:								

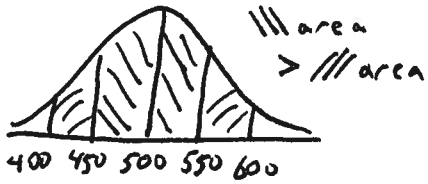
Question:	9	10	11	12	13	14	15	16
Points available:	20	20	20	20	20	20	20	20
Your score:								

Name: _____

For questions 1 through 8, you only need to circle the answer. Your work will not be graded. Questions 1 through 8 are worth 5 points each.

1. The scores on a certain test follow the normal curve, with average 500 and unknown SD. The chance that a random person has a score between 400 and 600 is _____ the chance that a random person has a score between 450 and 550. Circle the choice that best fills in the blank:

(a) less than twice
(b) exactly twice
(c) more than twice
(d) cannot be determined from the information given.



2. There are 100,000 people in Berkeley, and 900,000 in San Jose. A simple random sample consisting of one percent of all people in Berkeley is taken, and these people are asked if they recycle regularly. A simple random sample consisting of one percent of all people in San Jose is taken, and these people are asked if they recycle regularly. The proportions of people in the two cities that recycle regularly are approximately the same.

A 95-percent confidence interval for the proportion of people in Berkeley who recycle is found from these results. The same is done in San Jose.

Divide the width of the confidence interval for the proportion of recyclers in San Jose by the width of the confidence interval for the proportion of recyclers in Berkeley. The answer is closest to:

0.1 (0.3) 0.5 1 2 3 5 9

$\sqrt{\frac{100{,}000}{900{,}000}} = \frac{1}{3}$

3. Six people live in a house. Four of them are male. Three were born in California. One is female and not born in California.

I pick a person at random from this house. The events "this person is male" and "this person was born in California" are:

(a) mutually exclusive and independent
(b) mutually exclusive, but not independent
(c) independent, but not mutually exclusive
(d) neither mutually exclusive nor independent.

	male	female	
born CA	2	1	(3)
not born CA	2	(1)	3
	(4)	2	(6)

circled numbers are given

$P(\text{male and born in CA}) = \frac{2}{6} = \frac{1}{3}$

$P(\text{male}) \times P(\text{born in CA}) = \frac{4}{6} \times \frac{3}{6} = \frac{1}{3}$

Name: _____

Questions 4 and 5 refer to the following hypothetical situation.

Do jelly beans cause acne? The leader of a university research lab thinks they might, and does a study to show this; when the results from the study come back, jelly beans appear to not cause acne. After further brainstorming, the leader hypothesizes that perhaps only certain colors of jelly bean cause acne. Twenty graduate students are each assigned a color of jelly bean and asked to determine if it causes acne. Nineteen of the graduate students come back and say that their color of jelly bean does not appear to cause acne ($P > 0.05$); one student, who tested green jelly beans, says that green jelly beans do appear to cause acne ($P < 0.05$). The leader of the lab contacts the media, receives much acclaim, and shares a prize with the grad student who was assigned green jelly beans, who goes on to graduate and get a good job; the other nineteen grad students drown their sorrows in pizza and many of them, ironically, develop bad cases of acne.

4. Assuming that there is absolutely no connection between jelly beans of any color and acne, the probability that *at least* one of the twenty students would come back with a statistically significant result is closest to:

0 0.2 0.4 (0.6) 0.8 1 $1-(1-0.05)^{20}=0.6415$

5. This is an example of which of the following errors in reasoning?

(a) Simpson's paradox
(b) regression fallacy
(c) data snooping
(d) ecological fallacy

5.5. **[0 points]** I did not conceive of this story myself; I took it from a source that is well-known in certain circles. What is that source? (If you don't know, remember, this question is worth ZERO POINTS.) XKCD

6. The coefficient of correlation between the height of a man and the height of his son is +0.5. The coefficient of correlation between the height of a man and the height of his grandson is therefore (**circle one**):

(a) -1
(b) between -1 and -0.5
(c) -0.5
(d) between -0.5 and 0
(e) 0
(f) between 0 and +0.5
(g) +0.5
(h) between +0.5 and +1
(i) +1

correlation should be weaker but still positive.

Name: _____

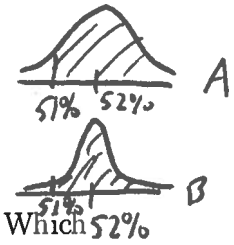
7. Hospital A has 310 live births during the month of June. Hospital B has 814. Which of the following is true?

(a) Hospital A is more likely than hospital B to have 51% or more male births.

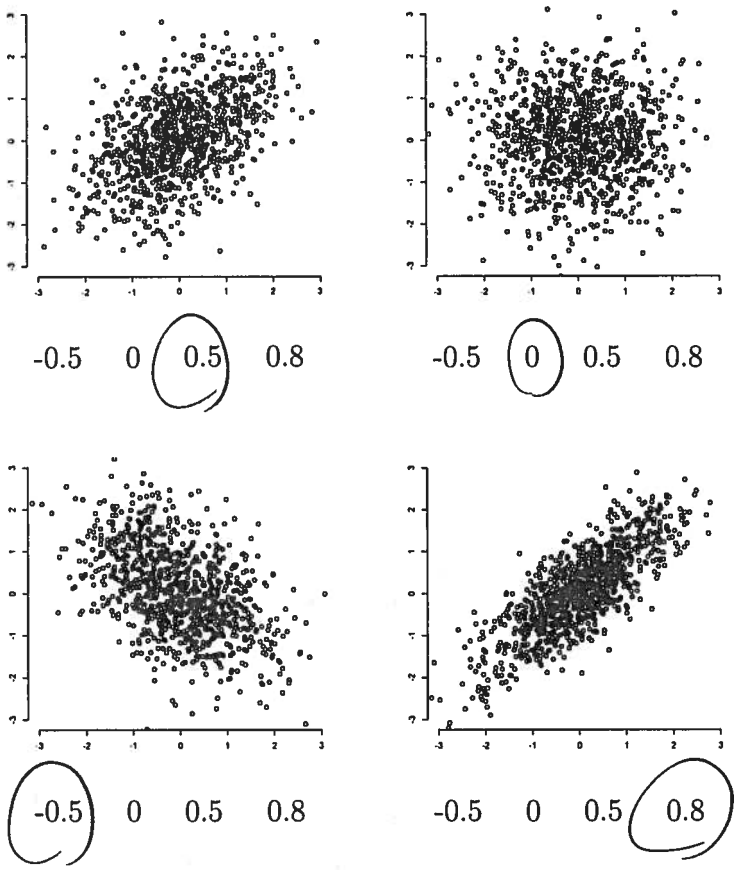
(b) Hospital B is more likely than hospital A to have 51% or more male births.

(c) Both hospitals are equally likely to have 51% or more male births.

The chance that a live-born infant is male is about 52%.



8. The four scatterplots below have correlation coefficients $-0.5, 0, 0.5$, and 0.8. Which is which? Under each scatterplot are four numbers; circle the coefficient of correlation for that scatterplot.



Name: _____

For questions 9 through 16, please show all work. Questions 9 through 16 are worth 20 points each; the value of each part of each question is given.

9. In a certain population, the heights of husbands and wives are as follows:
average height of husband: 69 inches. standard deviation: 3 inches.
average height of wife: 63 inches. standard deviation: 2.5 inches.
correlation between heights of husband and wife: 0.3.
The scatter diagram is football-shaped.

(a) [10] Give the equation of the regression line for predicting the wife's height from the husband's height.

y = wife's height x = husband's height

$$y - 63 = 0.3 \cdot \frac{2.5}{3} \cdot (x - 69)$$

$$y - 63 = 0.25(x - 69)$$

$$y = 0.25x + 45.75.$$

(b) [10] A man is 66 inches tall. What is the chance that his wife is between 63 and 65 inches tall?

$x = 66 \rightarrow y = 62.25.$

RMS error of prediction $\sqrt{1-r^2}\, SD(y)$

$= \sqrt{1-0.3^2}\,(2.5) = 2.38.$

area under normal curve between

$$\frac{65 - 62.25}{2.38}, \quad \frac{63 - 62.25}{2.38}$$

$$1.16, \quad 0.32 \quad \approx \frac{1}{2}(.7499 - .2358)$$

$$= \boxed{26\%}$$

Name: _____

10. A fair, six-sided die (with numbers 1, 2, 3, 4, 5, 6 on the sides) is rolled four times. Give your answers as a single fraction or decimal. Fractional answers need not be in lowest terms. So, for example, 1/3, 2/6, and .333 are all acceptable, but $(2/5) \times (5/6)$ is not.

(a) [5] What is the probability that the four numbers rolled are all different?

$$\frac{6 \times 5 \times 4 \times 3}{6 \times 6 \times 6 \times 6} = \frac{5}{18} = 0.2778.$$

(b) [5] What is the probability that the second number rolled is even, given that the first number rolled is odd?

$\frac{1}{2}$. (first number doesn't matter!)

(c) [5] What is the probability that the four numbers rolled add up to exactly 5?

Four ways: 1112, 1121, 1211, 2111.

$$\frac{4}{6^4} = \frac{1}{324} = 0.0031$$

(d) [5] What is the probability of rolling two 3s and two 5s?

$\binom{4}{2}$ = six ways: 3355, 3535, 5335, 5533, 5353, 3553.

$$\frac{6}{6^4} = \frac{1}{216} = 0.0046.$$

Name: _____

11. A box contains the numbers 1, 2, 4 and 9.

(a) [6] Find the average and SD of this box.

$$\text{average} = \frac{1+2+4+9}{4} = \frac{16}{4} = \boxed{4}$$

$$\text{SD} = \sqrt{\frac{(4-1)^2+(4-2)^2+(4-4)^2+(4-9)^2}{4}} = \sqrt{\frac{3^2+2^2+0^2+5^2}{4}} = \boxed{3.08.}$$

$$= \sqrt{9.5}$$

(b) [7] Find the approximate chance that the sum of 38 draws from this box will be in the range from 133 to 171.

~~sum~~ of 38 draws: EV $4 \times 38 = 152$

SE $\sqrt{9.5}\sqrt{38} = \frac{38}{\sqrt{4}} = 19.$

133 to 171, in SU: $\frac{171-152}{19} = 1$, $\frac{133-152}{19} = -1$

$\rightarrow \boxed{68\%}$

(c) [7] Find the approximate chance that in ~~48 draws~~ from this box, the number 9 is drawn at least 16 times.

EV $48 \times \frac{1}{4} = 12$

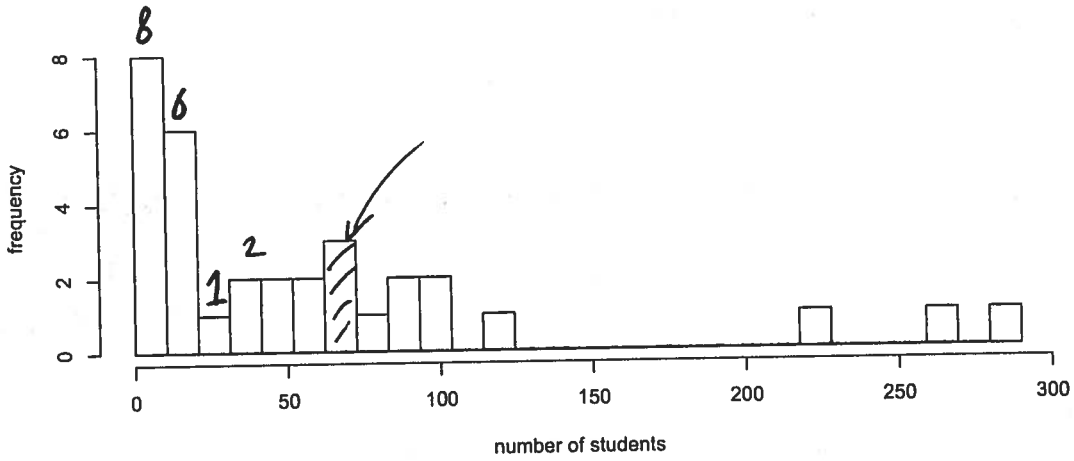
SE $\sqrt{48 \times \frac{1}{4} \times \frac{3}{4}} = \sqrt{9} = 3.$

$\frac{15.5-12}{3} = 1.17.$ area to right of 1.17

$\approx \frac{1-.7498}{2} = \boxed{13\%}$

Name: _____

12. Below is a histogram of the enrollments in all classes offered by the Department of Statistics this semester.



There are 33 courses represented here, with a total of 1974 students. Statistics 21 lecture 1, which is our class, has 69 people.

(a) [4] What is the median (50th percentile) of the class sizes? If you cannot say exactly, give as small a range as possible into which it falls.

17th smallest is between 40 and 50.

(b) [6] The average class size is 60, and the standard deviation is 72. I claim that about 68 percent of all classes in the university have enrollment in the range 60 ± 72. There are two reasons why this claim does not follow from the facts given. What are they? (Your answer should be two sentences long, one for each reason.)

The distribution of class sizes in the Stat dept. is not normal. Also, Statistics classes may not be representative of all classes in the university.

Name: _____

(c) [3] I pick twenty students at random from all the students taking statistics classes this semester. (Assume that no student takes more than one statistics class.) I ask each of them how many people are in their statistics class and average the twenty numbers. I get a number that is (**circle one, no explanation necessary**):

(i) less than 60

(ii) very close to 60 (class size paradox)

(iii) more than 60

For parts (d) and (e), consider the following scenario.

I would like to take a simple random sample of size 1 of students in statistics classes. I propose the following procedure for doing this:

- first, pick a statistics class at random, so that all classes are equally likely to be picked.
- second, pick a student from that class at random, so that all students in the class are equally likely to be picked.

(d) [3] What is the probability I pick you? (Again, assume you are taking only one statistics class.)

$\frac{1}{33} \times \frac{1}{69} = \frac{1}{2277}$

(e) [4] Is the sample obtained in this way a simple random sample of size 1? Why or why not?

Not a simple random sample. Students in smaller classes are ~~less~~ more likely to be picked.

Name: _____

13. For the data set given below, find the regression equation for predicting y from x. **Give your answer in the form $y = mx + b$.**

x	y
2	1
5	4
9	6
9	12
11	7
18	12

$$\text{avg}(x) = \frac{2+5+9+9+11+18}{6} = 9$$

$$SD(x) = \sqrt{\frac{(2-9)^2+(5-9)^2+(9-9)^2+(9-9)^2+(11-9)^2+(18-9)^2}{6}}$$

$$= 5.$$

$$\text{avg}(y) = \frac{1+4+6+12+7+12}{6} = 7$$

$$SD(y) = \sqrt{\frac{(1-7)^2+(4-7)^2+(6-7)^2+(12-7)^2+(7-7)^2+(12-7)^2}{6}}$$

$$= 4.$$

x	y	std-x	std-y	product
2	1	-1.4	-1.5	+2.1
5	4	-0.8	-0.75	+0.6
9	6	0	-0.25	0
9	12	0	+1.25	0
11	7	+0.4	0	0
18	12	+1.8	+1.25	+2.25
				4.95

$$\rightarrow r = \frac{4.95}{6} = 0.825$$

$$y - 7 = 0.825(x - 9)$$

$$y - 7 = \frac{4}{5}\,0.825(x - 9)$$

$$y - 7 = 0.66(x - 9)$$

$$y - 7 = 0.66x - 5.94$$

$$y = 0.66x + 1.06.$$

Name: _____

14. A simple random sample of 100 Berkeley students is taken. The average of their GPAs is 2.90, and the SD is 0.40.

Another simple random sample of 64 Stanford students is taken. The average of their GPAs is 2.97, and the SD is 0.24.

(a) [5] What is a 95 percent confidence interval for the average GPA of all Berkeley students?

(b) [10] Test the hypothesis that the average GPA of Stanford students is higher than the average GPA of Berkeley students. Give a P-value and clearly state which hypothesis test you are using and your conclusion.

(c) [5] I pick a random Berkeley student. What is the chance that their GPA lies in the 95 percent confidence interval in (a)?

(a) $2.90 \pm 2\frac{0.40}{\sqrt{100}} = 2.90 \pm 0.08$

$= [2.82, 2.98]$.

(b) two-sample z-test. H_0: Stanford mean $\leq$ Berkeley mean.

H_A: Stanford mean $>$ Berkeley mean

$SE = \sqrt{\left(\frac{0.24}{\sqrt{64}}\right)^2 + \left(\frac{0.40}{\sqrt{100}}\right)^2}$

$= \sqrt{0.03^2 + 0.04^2} = 0.05$.

$z = \frac{2.97 - 2.90}{0.05} = 1.4$. $\frac{1 - 0.8384}{2} \approx 8\%$

Conclusion: the difference in GPA is not statistically significant.

(c) $\frac{2.82 - 2.90}{0.40} = -0.2$, $\frac{2.98 - 2.90}{0.40} = +0.2$.

area between -0.2, $0.2 \approx 16\%$.

Name: _____

15. In 1965, a newspaper carried a story about a high school student who reported getting 9,207 heads and 8,743 tails in 17,950 tosses of a coin.

(a) [6] Test the hypothesis that the coin is fair. Give a P-value and clearly state which hypothesis test you are using and your conclusion.

one-sample z-test, test $H_0: p = 0.5$ against $H_A: p \neq 0.5$.

$$z = \frac{9{,}207 - \frac{17{,}950}{2}}{\sqrt{17950 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{232}{67.0} = 3.46.$$

$\rightarrow P \approx 1 - 0.999417 = 0.00053$ (two-tailed).

Conclude coin is unfair.

A statistician contacted the student and asked exactly how the experiment had been performed. To save time, the student had tossed groups of five coins at a time, and a younger brother recorded the results; these are shown in the following table.

number of heads	0	1	2	3	4	5
frequency	100	524	1080	1126	655	105

(b) [6] What is the probability that tossing five coins will give n heads, for $n = 0, 1, 2, 3, 4, 5$? Your answer may be either six numbers (one for each possible number of heads) or of a general formula in terms of n. If you choose to give six numbers, give your answers as a single fraction or decimal. Fractional answers need not be in lowest terms. So, for example, 1/3, 2/6, and .333 are all acceptable, but $(2/5) \times (5/6)$ is not.

$\binom{5}{n}/2^n$. $\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32}$.

(c) [8] Using the probabilities from part (b) and the table given above, test the hypothesis that the five coins are fair.

freq.	100	524	1080	1126	655	105
exp. freq.	112.4	562.2	1124.4	1124.4	562.2	112.4

$3598 \times \frac{1}{32}$

$$\chi^2 = \frac{(100-112.4)^2}{112.4} + \frac{(524-562.2)^2}{562.2} + \ldots + \frac{(105-112.4)^2}{112.4}$$

$= 21.52.$

5 df: $P < 1\%$.

coins are not fair.

Name: _____

16. A study was done to investigate whether people could briefly postpone their deaths until after the occurrence of a significant occasion. The senior woman of the household plays a central ceremonial role in the Chinese Harvest Moon Festival. The researchers compared the mortality patterns of old Jewish women and old Chinese women who died of natural causes for the weeks immediately preceding and following the festival, using records from California for the years 1960-1984. The results are given in the table below.

time of death	Chinese	Jewish
second week before festival	55	141
first week before festival	33	145
first week after festival	70	139
second week after festival	49	161

Test the hypothesis that time of death (relative to the festival) is independent of background (Chinese vs. Jewish). Give a P-value and clearly state which hypothesis test you are using and your conclusion. χ^2 test for independence.

marginals:

55	141	196
33	145	178
70	139	209
49	161	210
207	586	793

expected: $\frac{196 \times 207}{793} = 51.16$ 144.84

46.46 131.54

54.56 154.44

54.82 155.18

$\chi^2 = \frac{(55-51.16)^2}{51.16} + \ldots + \frac{(161-155.18)^2}{155.18}$ (8 terms)

$= 12.41.$

$(4-1)(2-1) = 3$ df. $p < 1\%$.

Conclude that time of death depends on background.