STATISTICS 20 Practice FINAL EXAM

There are 10 questions, with a total of 45 points. Most explanations require only 1 or 2 sentences. On calculations, show your work, and work through to a numerical answer.

1. [6 points]. About 1 million high school students took the SATs in 1987. The summary statistics are given below (changed slightly to keep the arithmetic simple)

SAT verbal: ave = 430 s.d. = 100

SAT mathematical: ave = 475 s.d. = 100

correlation = 0.6

The histograms follow the normal curve, and the scatter diagram is football-shaped.

(a) Approximately how many students scored between 525 and 575 on the mathematical test?

(b) About how many scored exactly 525 on the mathematical test? [Hint: consider height of normal curve]

(c) Find the equation for the regression line for predicting verbal score from mathematical score.

2. [6 points]. Continuing with the data from question 1:

Of the students who scored exactly 525 on the mathematical test, about what percentage scored

(d) over 490

(e) between 490 and 500

on the verbal test?

(f) Using the results above, estimate the number of students (out of the total 1 million) who got exactly 525 on the math test and 490-500 on the verbal. If the actual data showed that the exact number was 20% larger than your estimate, how would you explain the discrepancy?

3. [3 points]. In a study of a representative group of men, the correlation between height and weight was 0.43. One man in the study was both 1 s.d. below average in weight and 1 s.d. below average in height. His weight will be

(i) larger than

(ii) about the same as
or (iii) smaller than
the average weight of all men of his height in the study.
Choose one option, and explain briefly.

4. [3 points]. An instructor standardizes her midterm and final so that the class average in 50 with a s.d. of 10 on each test. The correlation between the tests is always around 0.5. On one occasion, she took all the students who scored below 30 on the midterm, and gave them special tutoring. They all scored above 50 on the final. Can this be explained by the regression effect?

5. [4 points]. A simple random sample of 200 voters in a Congressional district is taken to discover voting intentions in an upcoming election. It is found that 105 of the sampled voters intend to vote Democrat and the other 95 Republican. Based on this, calculate a 68% confidence interval for the the percentage of all voters in the district intending to vote Republican.

6. [4 points]. A second statistician looks more closely at the results of the sample in question 5, separating the sampled votes into those resident in urban or suburban areas.

	Republican	Democrat
urban	26	87
suburban	69	18

It is known that 60% of the total voters live in urban areas and 40% in suburban areas. By chance, the sample turned up with fewer urban voters than expected (113 instead of 120, which is a typical amount of chance error). The second statistician argues as follows.

> The data shows that urban voters are more likely to vote Democrat, and so our particular sample is biased toward Republicans because it happens to have too few urban voters. So let's adjust the figures: instead of 113 we should have got 120 urban voters, so instead of 26 urban Republicans we should have got $26 \times \frac{120}{113} = 27.6$. Similarly, instead of 69 suburban Republicans

we should have got $69 \times \frac{80}{87} = 63.4$. Thus we should have got $27.6 + 63.4 = 91$ Republicans in our sample, and so a better 68% confidence interval is 45.5% plus or minus the same amount as you calculated in question 5.

(a) What do you think of this argument?

(b) Is there a different way of doing the sampling which would avoid the problem of different urban and suburban voting preferences?

7. [6 points]. Newspapers like to print stories saying how little today's students know. One such story recently reported a test in which students were asked to identify various people (e.g. given the name "Mark Twain" the student was expected to say something like "Nineteenth century author"). The newspaper article was headlined "More students can identify Erica Kane than Andrew Jackson". Suppose (hypothetically) the test was given to a simple random sample of 100 students at Cal State Hayward and to another simple random sample of 100 U.C. Berkeley students, with the following numbers of correct identifications.

	Kane	Jackson
Hayward	83	65
Berkeley	69	86

Based on this data, one might ask

(a) Can more Hayward students identify Erica Kane than Andrew Jackson?

(b) Can a larger proportion of Berkeley students than Hayward students identify Andrew Jackson?

If possible, do tests of significance to answer these questions. If not possible, explain why.

8. [5 points] A 1979 survey of alcohol consumption contained the following data comparing marital status and number of drinks per month.

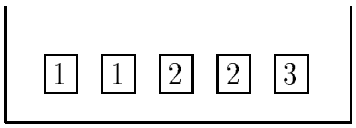
	number of drinks			
	0	$1-60$	61+	total
single	67	213	74	354
married	411	633	129	1173
total	478	846	203	

Assuming this data comes from a simple random sample, can we conclude there is a real association between alcohol consumption and marital status?

9. [3 points] A hospital is testing a new drug for a disease RPS. It is thought that the drug might be harmful to fetuses, so women of child-bearing age make up the control group; the treatment group is older and younger women. In the control group, 20 of the 100 patients (20%) get better; in the treatment group, 70 out of 200 (35%) get better. The hospital concludes that the new drug helps RPS in women. The main reason this conclusion is unreliable is
(i) The control and treatment groups were of different sizes
(ii) The treatment group included both older and younger women
(iii) The treatment and control groups were not similar with respect to age
(iv) Pregnancy might reduce RPS, and only the women in the control group might have been pregnant.

Explain briefly.

10. [5 points]. A computer program simulates drawing 50 times with replacement from the box



The program prints out the results from left to right, in two rows of 25 digits each, like

$$x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x$$

$$x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x\ x$$

Calculate the EV and the SE of the number of times that a digit in the second row is the same as the digit directly above it in the first row.