Study Guide 2 key

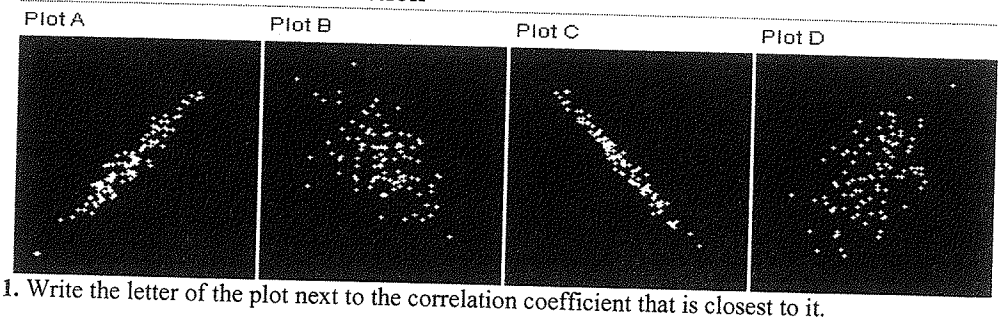
STUDY GUIDE FOR EXAM 2

For each of the topics below look over the relevant problems in your lecture notes and homework.

Chapter 6: Correlation

- Match scatter plots with correlation coefficients. Practice "guessing correlation" game
- Scatter plot has 5 summary statistics: average and SD of the x-values, average and SD of the y-values and the correlation coefficient. Be able to estimate these 5 statistics as well as the SD and the regression line from looking at a scatter plot.
- Compute the correlation coefficient.
- Correlation coefficient is not affected by: adding a constant to all values of one variable, multiplying all values of one variable by a positive constant, interchanging all values of x and y, or changing units (i.e. from inches to centimeters).
- Ecological Correlations are based on averages and tend to be higher than those based on individuals.
- Correlation is NOT causation

Sample Question on Correlation

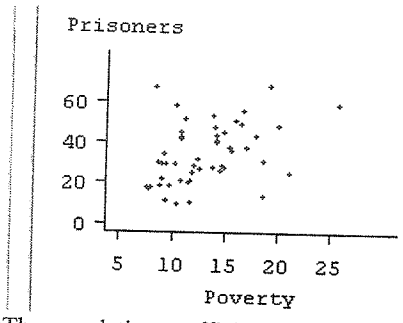


1. Write the letter of the plot next to the correlation coefficient that is closest to it.

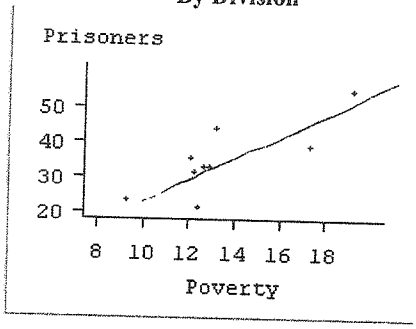
r = 0.52 D r = 0.96 A r = -0.54 B r = -0.99 C

2. The following scatter plots show the relation between poverty level (percentage of people living below the poverty line) and number of prisoners (per 100,000 people) by state and by geographical region. The graph on the left has 50 points, one for each **individual** state's poverty and prisoner level. The graph on the right has the same information condensed into 9 points, one for each of the 9 geographical regions. (In other words, the 50 states were divided into 9 geographical regions. The average poverty and prisoner level was computed for each region.)

By State



By Division



pts. hug line more closely since scatter within region eliminated

a) The correlation coefficient for the graph on the left is 0.4. The correlation for the graph on the right is
i) less than 0.4 ii) equal to 0.4 iii) greater than 0.4

b) The scatter plots above are an illustration of
i) The Regression Effect ii) Simpson's Paradox iii) Ecological Correlation iv) Negative Correlation

KEY

3. Compute the correlation coefficient between X and Y by filling in the table below. The average of X = 4 and the SD of X = 2 and the average of Y= 4 and the SD of Y = 2.

X	Y	X in Standard Units	Y in Standard Units	Products
1	3	$\frac{(1-4)}{2} = \frac{-3}{2}$	$\frac{(3-4)}{2} = -1/2$	$-\frac{3}{2} \cdot -\frac{1}{2} = \frac{3}{4}$
3	1	$-1/2$	$-3/2$	$-1/2 \cdot -3/2 = \frac{3}{4}$
4	5	0	$1/2$	0
5	4	$1/2$	0	0
7	7	$3/2$	$3/2$	$3/2 \cdot 3/2 = 9/4$

column sums to 0 column sums to 0

a) The correlation coefficient = 0.75

$r = \frac{\frac{3}{4} + \frac{3}{4} + 0 + 0 + \frac{9}{4}}{5} = .75$

b) If all the y values are increased by 3 the correlation coefficient would
i)stay the same ii)increase ii)decrease

c) If all the y values are doubled the correlation coefficient would
i)stay the same ii)double ii)decrease

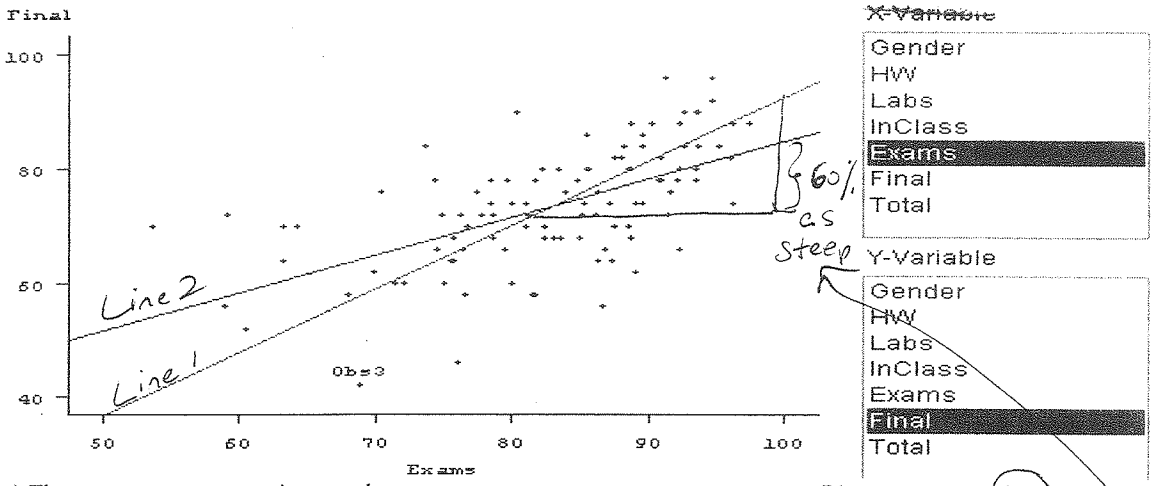
d) If BOTH the x and y values are all multiplied by -1 then the correlation coefficient would
i) stay the same ii) change sign

e) If all the X and Y values were switched the correlation coefficient would
i)stay the same ii)increase ii)decrease

f) If the top and bottom x values (the 1 and 7) were switched, the average and the SD of x would stay the same. Would the correlation coefficient stay the same? YES NO

KEY

Question 4 pertains to the scatter plot below which shows the exam and final scores for 107 students in a previous Stat 100 class.



a) The average exam score is around 74 85 (85) 90

b) The average final score is around (74) 85 90

c) Which is the regression line? Line 1 (Line 2) less steep

d) If a student was exactly average on both the exam and the final which line would he fall on?
Only the SD Line Only the Regression Line (Both) Neither

e) If a student was exactly 1 SD above average on both the exam and the final which line would he fall on?
Only the SD Line Only the Regression Line Both Neither

f) The correlation between exam and final score is closest to 0 -0.2 (0.6) 0.27

g) The prediction error for Obs 3 is closest to 0 -5 -15 15 (-20)
→ Actual − Pred = 42 − 62

h) If a new scatter plot was drawn with 10 pts. added to everyone's final score then the correlation between exam and final scores would_____ i) increase ii) decrease iii) stay the same

i) If a new scatter plot was drawn with 10 % added to everyone's final score then the correlation between exam and final scores would_____ i) stay the same ii) decrease iii) increase

j) The following 3 numbers: **0, 10.27 and 8.25** are, in scrambled order, the SD of the final, the average of the prediction errors and the SD of the errors (when predicting final from exam.) Write the number next to its description.
i) average of the errors 0 ii) SD of the final 10.27 iii) SD of the errors 8.25
ave of errors ALWAYS 0

k) If Obs 3 was removed, the correlation coefficient would i) stay the same ii) decrease iii) increase
since Obs 3 has large error

SD of y is ALWAYS ≥ SD errors
10.27 > 8.25

KEY

Chapter 7: Regression

- Know which is the regression line and which is the SD line. (The regression line is always less steep.)
- Computing the regression estimate given the 5 summary statistics by converting independent variable to Standard Units, multiplying by r and then converting back to the units of the dependent variable
- Computing the regression estimate for percentiles
- The regression effect and the regression fallacy.

Sample Questions on Regression

1. A large group of high school students took the ACT twice. The averages and the SDs of both the first and the second sets of ACT scores were the same: **Ave = 20 SD= 5 and the correlation was r = 0.8** The scatter plot was foot-ball shaped.

a) Predict the 2nd ACT score of someone who got a 15 on the first ACT. (Use the 3 step procedure. Show work.)

1st ACT value	1st ACT z	× r	2nd ACT z	2nd ACT value
15	$\frac{15-20}{5} = -1$	× .8	$\rightarrow$	$-.8(5)+20 = 16$

(5 = SD, 20 = ave)

b) A student got a 15 on his first ACT and an 18 on his second ACT. What is the prediction error (when predicting 2nd ACT from 1st ACT)?_____

error = actual − predicted = 18 − 16 = (2)

c) A group of students all got 30on their first ACT. Estimate what their average 2nd ACT score would be.
(You may use either the 3 step procedure or the regression equation. Show work)

value	z	× r	z	value
30	$\frac{30-20}{5} = 2$	× .8	$= 1.6(5)+20 =$	(28)

d) Of course, they won't all get exactly the score you predicted them to get in your answer to (c). But about 2/3 of those who scored 30 on their first ACT will get between 25 and 31 on their second ACT.
(Show work)

2/3 will get will 1 SD errors $= \sqrt{1-r^2}\,SD_y$

$= \sqrt{1-.8^2}(5) = 3$

28 +/- 3

Key

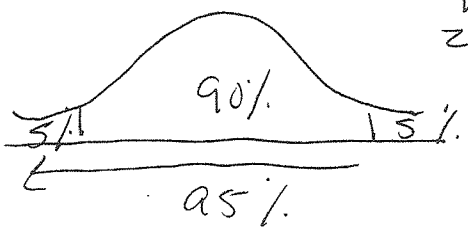
2. For a past Stat 100 class the correlation between Exam 1 and Exam 2 was **r = 0.6** and the scatter plot was football-shaped.

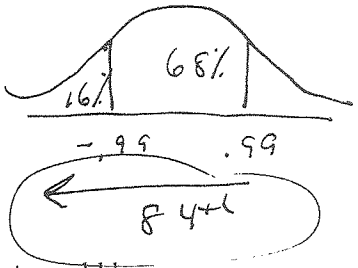
a) If a student scored in the 95th percentile on Exam 1, estimate his Exam 2 percentile. (Draw a picture. Show work You may round z scores to fit the closest line on the table and you may round percents on the table to the nearest whole number.)

percentile → z × r → z → percentile

look up z for middle area = 90%.

$z = 1.65 \times .6 = .99$





b) If there was NO correlation between Exam 1 and Exam 2 scores then what would be your estimate for the Exam 2 percentile of someone who was in the 95th percentile on Exam 1?

50 percentile

b/c it's the ave

c) If there was a perfect correlation (r=1) between Exam 1 and Exam 2 scores then what would be your best estimate for the Exam 2 percentile of someone who in the 95th percentile on Exam 1?

95 percentile

d) If a student scored below the 50th percentile on Exam 1 and r = 0.6, then the best estimate for his Exam 2 percentile is

i) Above 50th
ii) 50th
iii) Below his Exam 1 percentile
iv) Above his Exam 1 percentile but below 50th

regression estimate is ALWAYS closer to the 50th percentile

e) Students who did the best on Exam 1 did somewhat worse on Exam 2 and students who did the worst on Exam1 did somewhat better on Exam 2. This is an example of

i) Ecological correlations
ii) The Regression effect
iii) Negative correlation
iv) Simpson's paradox

Key

Study Guide for Exam 2

Chapter 8: Prediction Errors

- **Prediction Error = Actual Y value- Predicted Y value**
- **Average of the Prediction Errors is ALWAYS 0**
- **SD of the Prediction Errors gives the size of the typical error and is calculated by SD of errors= $\sqrt{1-r^2}$ SD of y**
- **If the scatter plot is roughly elliptical about 2/3 of the predictions will be within 1 SD_{errors} and 95% will be within 2 SD_{errors}**

Chapter 9: Equation of the Regression Line

- Know how to calculate the slope and the intercept of the regression line. p.207 #3
- Slope of the regression line for predicting *y* from *x* is r (SD of *y*)/ (SD of *x*).
- Intercept may be solved for, since you know the point of averages is always on the line.
- Make predictions using the equation for the regression equation. P.207 #1,2

Sample questions on Prediction Errors and the Equation of the Regression Line

1. In a large population study, the ages of husbands and wives were recorded. Here are the 5 summary statistics:
Average Age of Husbands = 45, SD= 15
Average Age of Wives = 40, SD = 10 **r = 0.9** The scatter plot was foot-ball shaped.

a) The **slope** of the regression equation for **predicting a husband's** age from his wife's age is

i) 0.9 ii)2/3 iii) 3/2 iv) .6 v) 1.35 (circled)

slope = $r\frac{SD_y}{SD_x}$ = $.9\left(\frac{15}{10}\right)$

b) The **y-intercept** of the regression equation for **predicting a husband's** age from his wife's age is

i) -13 ii) 21 iii) -9 (circled) iv) 4 v) 0

H = 1.35W + b → 45 = 1.35(40) + b → b = 45 − 1.35(40) = −9

c) In the study, husbands of 60 year old women have an average age of 72 years. Can you conclude that wives of 72 year old men have an average age of 60? YES NO (circled)

d) The regression equation for **predicting a wife's** age from her husband's age is
Wife's Age = (0.6) (Husband's Age) + 13 years

.6(35) + 13 = 34

One wife in the study has a husband who is 35 years old. How old would you estimate her to be? 34

e) There's about a 2/3 chance that your estimate in (d), give or take _____years is right.

i) $\sqrt{1-0.9^2} * 10$ (circled) ii) $\sqrt{1-0.9^2} * 15$ iii) 10 iv) 15

f) Suppose you don't know the husband's age, what is your best guess for the age of a wife you've never seen?

40 b/c it's the ave

g) There's about a 2/3 chance that your guess in (f), give or take _____years is right.

i) $\sqrt{1-0.9^2} * 10$ ii) $\sqrt{1-0.9^2} * 15$ iii) 10 (circled) iv) 15

b/c it's the SD

Chapter 10: Chance

- Chance of an outcome = # of that outcome/total # of possibilities
- Chance of something = 100% - Chance of the opposite
- Multiplication Rule- The chance that 2 events will both happen = (Chance that the first will happen) x (Chance that the second will happen, given that the first has happened)
- If a situation involves 2 or more conditions (e.g. diagnostic tests) a table is helpful

Sample Questions:

1. What is the chance of getting no heads on 4 tosses of a fair coin? $1/2 \cdot 1/2 \cdot 1/2 \cdot 1/2 = (1/2)^4$

2. What is the chance of getting at least 1 head? $1 - P(\text{No heads}) = 1 - (1/2)^4$

3. A die is rolled 3 times. What is the chance of getting no 4's? $(5/6)^3$

4. What is the chance of getting at least one 4? $1 - P(\text{No 4's}) = 1 - (5/6)^3$

- **Independence/Dependence-** 2 events are **independent** if the chance of one event does not change if the other event happens. They're **dependent** if the chance of one changes depending on whether the other happened or not.

Sample Questions (Independence):

5. What is the chance that the first 10 flips of a fair coin are all heads? $(1/2)^{10}$

6. What is the chance that the 11th flip is a head? 1/2 What is the chance that the 11th flip is a tail? 1/2

7. A box contains 5 tickets--2 reds, 2 blues and 1 white. Two draws will be made **with replacement** from the box. [R R B B W]

a. Suppose the first draw is a red ticket. What is the chance of getting a blue ticket on the second draw? 2/5

b. Suppose the first draw is a blue ticket. What is the chance of getting a blue ticket on the second draw? 2/5

c. Are the draws independent? Yes

Sample Questions (Dependence):

8. Do #7a-c again, this time drawing **without** replacement.

7c) Ch(B) = 2/4

b) Ch(B) = 1/4

c) No, chance changes depending on what the 1st draw was.

Chap 10 cont.

9. Shuffle a deck of 52 cards. What is the chance that the first 2 cards are hearts?

$13/52 \cdot 12/51$

10. A drawer contains 4 red socks and 6 white socks. If 2 socks are chosen at random without replacement, what is the chance that both will be white?

$6/10 \cdot 5/9$

11. Two draws are made at random from a box containing the numbers 1, 2, and 3. What is the chance of getting a 1 at least once?

a) with replacement $1 - Ch(\text{None}) = 1 - \left(\frac{2}{3}\right)^2$

b) without replacement $1 - 2/3 \cdot 1/2$

12. Randomly choose one card from a deck of 52.

a. What is the chance that the card is a heart? $13/52 = 1/4$

b. What is the chance that the card is a heart given that it is an ace? $1/4$

c. Is getting a heart independent of getting an ace? Yes, chances are the same.

13. If you draw 2 cards what is the chance that both are hearts?

a. with replacement $13/52 \cdot 13/52$

b. without replacement $13/52 \cdot \frac{12}{51}$

Problems with 2 or more conditions—It's useful to draw a table.

14. There are 10 cans of Coke and 10 cans of Pepsi in a cooler.
2 of the Cokes are diet and 5 of the Pepsis are diet.

	Coke	Pepsi	
Diet	2	5	7
Regular	8	5	13
Total	10	10	20

a. You draw 1 can at random from the cooler. What is the chance that it is diet?

$7/20$

b. You draw 2 cans at random without replacement, what is the chance that both are diet?

$7/20 \cdot 6/19$

c. You draw 1 can, what is the chance that it is diet given that it is Coke? $2/10$

d. Is ~~drawing a diet independent of drawing a Coke~~? What's the chance that it's Coke given it's Diet? $2/7$

Chap 10 cont.

15. Suppose 10% of men aged 70 in routine screening have prostate cancer. Further suppose that a diagnostic test is 95% accurate in correctly giving a positive result for those who have cancer and also 95% accurate in correctly giving a negative result to those who do *not* have cancer.

	Cancer	No Cancer	
Test +	.95(100) = 95	45	140
Test −	5	.95(900) = 855	860
	100	900	1000

a) What's the chance that a 70 year old man who gets a positive result truly has cancer?

$$\frac{95}{140} \approx 68\%$$

b) Suppose 50% of 90 year old men have cancer. What's the chance that a 90 year old man who gets a positive result truly has cancer?

	Cancer	No Cancer	
Test +	.95(500) = 475	25	500
Test −	25	.95(500) = 475	500
	500	500	1000

$$\frac{475}{500} = 95\%$$

Key

Chapter 11: More on Chance, the Addition Rule

- Be able to figure the chance by counting all possible way a chance process can turn out. Then figure the chance of an event by (# of that event)/(# of total possibilities).
- Addition Rule- used to figure the chance that **either** of 2 events occurs. If the 2 events are mutually exclusive, simply **add** the chances. If the 2 events are not mutually exclusive, add the chances and then subtract the chance that both occur.
- To figure the chance that at least one of several events happen, it is often easier to calculate the opposite and subtract from 100%.

Sample Questions:

1. What's the chance of rolling 2 dice and getting doubles? (Doubles means having both dice show the same number of spots.) 6/36
2. What's the chance of drawing 1 card from a ~~fair~~ deck and getting either a face card or a 7?

$$\frac{12}{52} + \frac{4}{52} = \frac{16}{52}$$

3. What's the chance of rolling 2 dice and getting either doubles ~~of~~ or a sum of 8?

$$\frac{6}{36} + 5/36 - 1/36 = \frac{10}{36}$$

(Both)

2,6 6,2 3,5 5,3 4,4

4. What's the chance of rolling 2 dice and getting either doubles or a sum of 9?

1,1 2,2 3,3 4,4 5,5 6,6

$$\frac{6}{36} + \frac{4}{36} = \frac{10}{36}$$

3,6 4,5 6,3 5,4