

## Midterm



IOMS Department

# Statistics and Data Analysis

## Professor William Greene

Phone: 212.998.0876

Office: KMC 7-78

Home page: [www.stern.nyu.edu/~wgreene](http://www.stern.nyu.edu/~wgreene)

Email: [wgreene@stern.nyu.edu](mailto:wgreene@stern.nyu.edu)

Course web page: [www.stern.nyu.edu/~wgreene/Statistics/Outline.htm](http://www.stern.nyu.edu/~wgreene/Statistics/Outline.htm)

## Midterm Preparation

The midterm for Statistics and Data Analysis, for Block 5 (and all the other blocks), will be on Thursday, November 1, 2007, in Room 3-70, from 9:00 – 10:20 AM. The following are some notes that (I hope) will help you prepare for the test.

1. What will be covered on the exam? Overall, the course consists of the material that I have covered in class. The Powerpoint slide sets numbered 1 to 15 on the course home page are the material that will be included. The first numbered slide in each set contains a short outline of the material that was to be covered that day. The last numbered slide in each set contains a very brief summary of where we were that day. These outlines tell you “what was covered” that day. Implications:

- a. The various handouts, pamphlets, other notes, etc. that are posted on the course home page are intended to provide background material and extensions of the class material for those interested. Nothing in these notes that does not appear in my class notes will be covered on the midterm.
  - b. The textbook is suggested for background and extensions. But, again, nothing that appears in the text that was not covered in class will appear on the midterm.
  - c. The midterm will be based on what was covered in class (excluding the short research presentation that I will give on October 16). If you were comfortable with all of that material, you are ready.
2. What resources can I use during the test, and what will I need?
- a. You may refer to your notes, your textbook and the other materials that I distributed.
  - b. The arithmetic on the test will consist of basic addition, multiplication, and a few simple partial differential equations. (OK, no PDEs.) You should bring a calculator. For computing probabilities, I will provide the information you need. You should not need a computer.
3. About the questions below: Listed below are a few sample questions that are at the level of difficulty that I would hope you can handle on the midterm. Please understand, these are not the actual midterm, and it is certainly possible that the midterm questions will look different and even require different computations. But, these suggest the range of material, the level of difficulty and the style of questions that you can expect.

II. Sample Questions (Provided by Jeffrey Simonoff)

## Midterm

1. The story “Tough Times in Japan,” which appeared in the February 9, 2003 issue of Parade magazine, contained various statistics purporting to illustrate tough economic times in that country. The article included the following sentence: “A recent survey by Japan’s Ministry of Health, Labor and Welfare indicates that as many as 60% of all households now have annual incomes below the national average of \$52,339.” Explain why it would not be surprising to find that more than half of all households have annual incomes below the national average, even if a country was not having tough economic times.

Since income data are typically skewed to the right, so that the mean is greater than the median (the 50<sup>th</sup> percentile), it might not be so surprising that the 60<sup>th</sup> percentile is also below the mean. However, in fact, as suggested by the income example we examined in class, a distribution with these statistics is, in fact, very highly skewed.

2. In December 1993, the U.S. Census Bureau issued a report on college enrollments in the United States. The report stated that 56% of all college students in 1992 were women. Further, 17% of all college students were aged 35 or older. Overall, 11% of all college students were women aged 35 or older.

(a) What is the probability that a randomly selected student is 35 or older, given that they are a woman?

(b) Are the events “being a woman” and “being aged 35 years or older” independent?

(c) What is the probability that a randomly selected college student is a man under the age of 35? Denoting by “W” being a woman and “A” being 35+, the data give  $P(W) = .56$ ,  $P(A) = .17$ , and  $P(A,W) = .11$ .

Part (a) asks for  $P(A|W)$  which we compute as  $P(A,W)/P(W) = .11/.56 = 0.196$ .

Part (b) asks whether  $P(A,W) = P(A)P(W)$ , which it clearly does not; 0.11 is not equal to 0.56 times 0.17 (which is 0.0952). Alternatively, we would need  $P(A|W) = P(A)$  which, again is not true. Part (c) asks for  $P(\sim W, \sim A)$ .

Part (c) can be solved with Bayes theorem, but it is also straightforward just to deduce the result from the information given. The bold underlined values below are given directly. We know that the marginal probabilities have to be the respective row or column sums, and we know that all four probabilities have to sum to 1.00. So, the missing values are easy to see.

	Woman	Man	
35+	<b><u>0.11</u></b>	0.06	<b><u>0.17</u></b>
<35	0.45	0.38	0.83
	<b><u>0.56</u></b>	0.44	1.00

The desired probability is 0.38.

## Midterm

3. Consider two investments that you are looking at for your portfolio. The one-month return for each is normally distributed, with a mean  $\mu = .01$  (that is, a 1% return). Investment A's return has standard deviation  $\sigma_A = .005$ , while investment B's return has standard deviation  $\sigma_B = .015$ .

(a) What is the probability that investment A's return next month will be positive?

$$P(A > 0) = P[(A-\mu)/\sigma_A > (0-\mu)/\sigma_A] = P[z > (0-.01)/.005] = P[z > -2] = P[z < +2] = 0.9772$$

(b) Is investment B's probability of positive return higher, lower, or the same?

$$P(B > 0) = P[(B-\mu)/\sigma_B > (0-\mu)/\sigma_B] = P[z > (0-.01)/.015] = P[z > -.667]. \text{ The answer is } 0.7475.$$

But, if you just draw a picture of the two distributions, A with the smaller variance and B with the larger, with the same mean, you could see the outcome immediately.

(c) Which investment would a "rational" investor prefer?

You would think that the same expected return with lower variance (risk) would be preferred. Thus, A should be the preferred investment.

4. Say the SAT verbal exam is graded along a strict curve, such that the scores in the population of students have mean 500 and standard deviation 100.

(a) If a random sample of 20 high school seniors is taken, what is the probability that the mean of their test scores exceeds 530?

The mean will have a mean of 500 and a standard deviation of  $100/\sqrt{20} = 22.36$ . Assuming we can use the normal distribution, then, the probability is

$$P[(\text{mean} - 500)/22.36 > (530 - 500)/22.36] = P[z > 30/22.36] = P[z > 1.34] = 0.0901.$$

(b) What assumptions are you making here?

Since I used the normal distribution, I assumed either that the scores themselves were normally distributed, or that whatever conditions I would need to have the central limit would apply. 20 observations is a little low for this assumption, but people do do this all the time.

5. Alcohol-related problems on college campuses are widespread, even though most college students are under the national drinking age of 21. A study funded by the Harvard School of Public Health, released on November 4, 1995, examined some issues relating to college alcohol abuse on campuses in the northeastern United States. The study stated that 50% of college men are binge drinkers, where a binge-drinking man is defined as one who consumed five drinks one after the other at least once in the previous two weeks. Consider 100 randomly selected northeastern U.S. college men.

(a) What is the exact probability that at least 35 of these men are binge drinkers? A formula is an adequate answer here.

The exact probability would be obtained from the binomial distribution with  $\pi = 0.5$ . That is

$$P(k) = \binom{100}{k} 0.5^k (1-0.5)^{100-k}, k = 0, 1, \dots, 100$$

So, the exact answer is

$$P(k \geq 35) = \sum_{k=35}^{100} \binom{100}{k} 0.5^k (1-0.5)^{100-k}$$

This is going to be extremely difficult, even with a copy of Pascal's triangle available.

(b) Give a numerical approximation to the value in part (a).

We can use the normal approximation. The mean is  $100(0.5) = 50$ . The variance is  $100(.5)(1-.5) = 25$ , so the standard deviation is 5. Using the continuity correction, the approximate answer is then  $P[x \geq 35] \approx P[z > (34.5 - 50)/5] = P[z > -3.1] = 0.9991$ .

## Midterm

### III. More Sample Questions

1. (Assume) that study of pet ownership has provided the following information about pets and television ownership (Note,  $X = 0$  if no televisions, 1 if at least one,  $X =$  number of pets, 0, 1 or 2.)

	P=Number of cats and/or dogs		
T=TVs	0	1	2
0	.05	.15	.20
1	.15	.20	.25

a. In this population, what is the expected number of pets?

We deduce the distribution of  $P$ , the number of pets by computing the column sums. Thus,  $P(P=0)=0.2$ ,  $P(P=1) = 0.35$ ,  $P(P=2)=0.45$ . Then,  $E[P] = 0(.2) + 1(.35) + 2(.45) = 1.25$

b. Are pet ownership and television ownership positively correlated, negatively correlated, or not correlated at all? Justify your answer. (Hint: The sign of the covariance is all you need to answer the question.)

To compute the covariance, we are going to need the expected value of  $T$ . We compute the marginal probabilities,  $P(T=0) = .05+.15+.20 = .40$  and  $P(T=1) = .60$ . So,  $E[T] = 0(.40) + 1(.60) = 0.60$ . Now, the covariance is

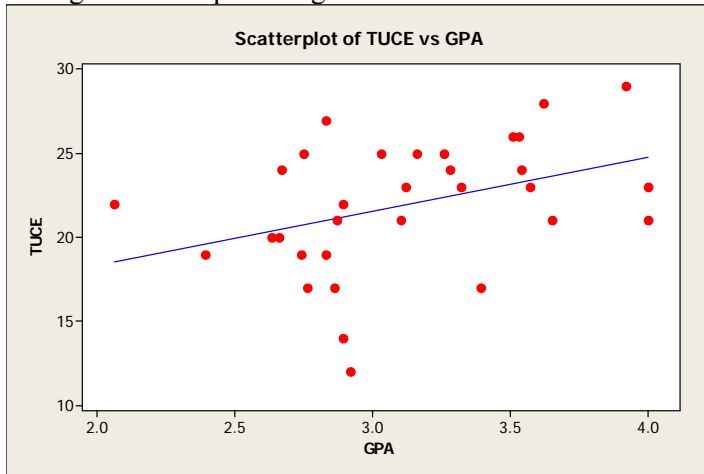
$.05(0)(0) + .15(0)(1) + .20(0)(2) + .15(1)(0) + .20(1)(1) + .25(1)(2) = 0.70$ . The covariance is therefore,  $0.70 - 1.25(0.60) = -0.05$ . This is negative so the correlation is negative also.

c. What is the expected number of pets for people with no televisions?

The distribution of number of pets for people with  $T=0$  is  $P[P=1|T=0] = .05/.4 = 0.125$ ,  $P[P=1|T=0]=.15/.4 = 0.375$ , and  $P[P=2|T=0]=.2/.4 = 0.50$ . The mean is  $E[P|T=0] = 0(.125) + 1(.375) + 2(.5) = 1.375$ .

## Midterm

2. The results below show a scatter plot and the regression of a test score on student grade point average for a sample of high school students.



### Regression Analysis: TUCE versus GPA

The regression equation is

$$\text{TUCE} = 11.9 + 3.24 \text{ GPA}$$

Predictor	Coef	SE Coef	T	P
Constant	11.853	4.434	2.67	0.012
GPA	3.235	1.407	2.30	0.029

S = 3.65699 R-Sq = 15.0% R-Sq(adj) = 12.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	70.67	70.67	5.28	0.029
Residual Error	30	401.21	13.37		
Total	31	471.87			

Unusual Observations

Obs	GPA	TUCE	Fit	SE Fit	Residual	St Resid
4	2.92	12.000	21.300	0.704	-9.300	-2.59R
21	2.06	22.000	18.517	1.622	3.483	1.06 X
23	2.89	14.000	21.203	0.721	-7.203	-2.01R

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large influence.

(a) How many observations were used? Explain.

You could count the dots in the figure, though two lie on top of one another, so this is risky. In the analysis of variance table, tot "Total DF" is  $n-1$ . So,  $n = 31+1 = 32$ .

(b) What proportion of the variation in TUCE is explained by the covariation with GPA?

That is  $R^2 = 15.0\%$  or  $0.15$ .

(c) What general conclusion does the regression suggest about the relationship between TUCE and GPA?

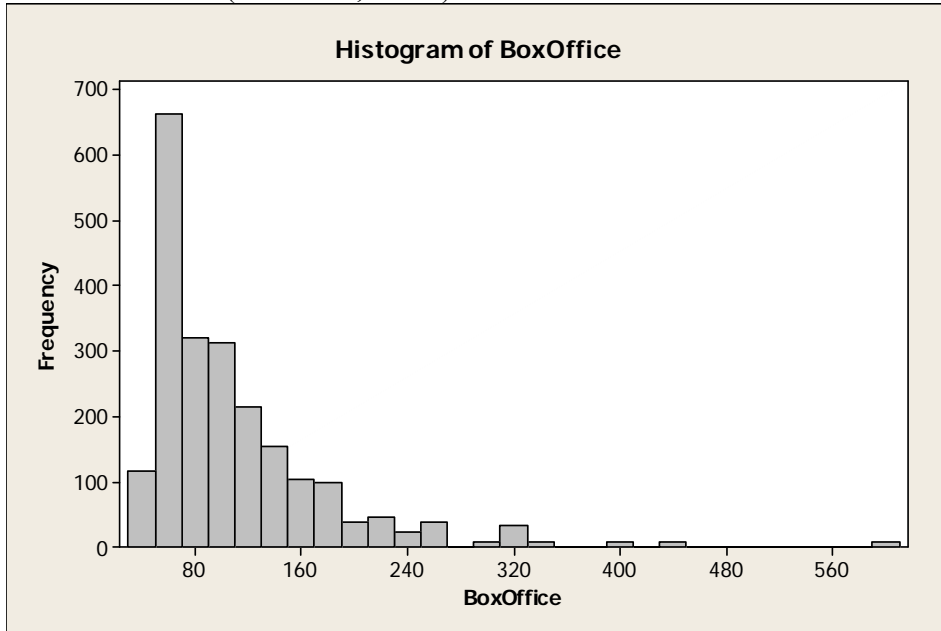
The positive slope of the line suggests that higher test scores are associated with higher GPAs. The fairly low  $R^2$  suggests that the relationship is not especially strong. But, the t statistic for the slope estimate of 2.30 is greater than 2, which does imply that the relationship is "significant."

(d) There are two points with exceedingly low values of TUCE. Would you describe these as outliers or highly influential observations? Explain.

By the definitions discussed in class, these would be classified as outliers. Influential data points have unusual values of the "x" variable, in this case GPA, not the "y" variable, TUCE.

## Midterm

3. The figure below is a histogram of the US Box office receipts for the 1,000 most successful movies of all time (some time, that is).



Box office receipts are measured in millions of dollars. The average for the sample is \$109 million.

a. Which of the following three values is the best estimate of the sample standard deviation? Explain your answer.

- \$5 million
- \$71 million
- \$300 million.

The empirical rule that we have used in many situations tells us that the mean plus and minus one standard deviation should include about two thirds of the data. The \$5M and \$300M do not come close to this outcome, but \$109 plus and minus \$71M, or about \$30M to \$180M does, indeed, include about 2/3 of the data, so this is the best answer.

b. Based on the figure and the description in the sample, suggest a reasonable estimate of the sample median of the observations.

Given the rightward skewness in the data, we would expect the mean to be greater than the median. The mean is \$109M, so we would guess that the median is less than \$109M. Looking at the histogram, we need for half of the observations to be less than the median. A good guess would be, say, \$90M or \$95M – for certain something more than \$80M, and as we know, less than \$109M.

4. *The* central principle of classical statistics (what we are studying in this class), is that the characteristics of a random sample resemble the characteristics of the population from which the sample is drawn. And, the resemblance becomes closer as the sample has more observations. Explain this principle to an educated person who has never had a statistics course.

This question will be on the exam. As part of your exam submission, I will ask you to turn in an answer to this question – one carefully worded paragraph with no more than 100 words.