

Review Questions to Prepare for the Exam

1) George Constanza, the character on Seinfeld, is a perpetual loser who keeps trying. As long-time viewers of the show know, only 10% of the women George approaches are willing to go out with him. George often will introduce himself, and then preface his invitation by saying “So waddya doin’ Friday night?” Three quarters of the women who don’t want to go out with him tell him they are busy – “*I’m giving my Doberman a bath.*” Only one quarter of the women who are willing to go out with him say they are busy. If George finds a woman who admits to being free on Friday night, what is the probability she will go out with him?

2) A random sample is drawn from a population believed to be approximately normally distributed. The sample values are

6, 18, 3, 21, 16, 9, 13, 14, 17

- a) From this sample, compute the sample mean, sample median, sample standard deviation.
- b) Compute a **98%** confidence interval for μ .
- c) If you wanted to verify the claim that the population is approximately normally distributed, what graphical procedure could you use? Explain what you would be looking for. <Hint: While a histogram or dotplot isn't a bad idea, the right answer is a more sophisticated tool.>

3) A student is rushing off to class when she remembers she needs to bring two floppy disks with her. She has 10 disks on her desk, but unfortunately four of them are defective, and she doesn’t know which four. Being in a hurry she grabs three disks at random, figuring that there will probably be two good ones among the three.

- a) What is the probability that *at least two* of the three disks are good?
- b) She has three hours of work to do on the computers at the lab. During the day, network crashes occur randomly and independently at the rate of a crash every 35 hours. What is the probability she will experience exactly one network crash while working in the computer lab?

4) Suppose the GPA’s of prospective Commerce majors at Virginia are normally distributed with a mean of 3.05 and a standard deviation of .31.

- a) What percentage of prospective Commerce majors have at least a B (that is, 3.0) GPA
- b) If the Commerce department wanted to pick a minimum GPA to use as a cutoff for admission to their program, and wanted a rule that would eliminate the weakest 35% of its applicants, what minimum GPA would it select?

- 5) A couple involved in a long-distance romance speak on the phone once a day, every day. The length of their calls are a random variable with the following distribution (in minutes):

x	f(x)
10	.2
20	.3
30	.3
40	.2

- a) What is the mean and variance of the length of their calls?
- b) The couple subscribes to Sprint's "ten cents a minute" plan. What is the mean and variance of their daily long-distance phone bill (measured in dollars rather than cents) ?
- c) In a month containing 30 days, what is the probability their long-distance phone bill exceeds \$72.00?
- d) Did you use the central limit theorem in answering this question, and if so, how?
- 6) If you bet \$1 on a color in roulette, the chance of winning a dollar is $18/38$, and the chance of losing a dollar is $20/38$.
- a) If you place a series of 5 one dollar bets on a color, and then stop, what is the probability you will walk away from the table a net winner? {Note: You are not to use the normal distribution in solving this part of the problem. }
- b) If you place a series of 25 one dollar bets on a color, and then stop, what is the probability you will walk away from the table a net winner? {Use Continuity Correction. }
- 7) Suppose 23% of all patients with high blood pressure have bad side effects from a certain medicine.
- a. Write down an expression (but do not solve it) for the exact probability that 32 patients from a total of 120 patients will have bad side effects.
- b. Use an approximation to find the probability that more than 32 patients of the 120 will have bad side effects. {Use continuity correction. }
- 8) The number of really pointless stories a particular professor tells in any one class meeting is a random variable, X , with the following distribution.

x	p(X)
0	.2
1	.1
2	.4
3	.1
4	.2

- a) Find the mean and standard deviation of X .
- b) Since each pointless story wastes 5 minutes of class time, the time left for "real" work is given by

$$W = 75 - 5x$$

Find the mean and standard deviation of W .

c) Knowing the professor's reputation, one student decided to cut almost all the time. She only came to class five times the whole semester.

i) What is the probability that every single time she attended class she had to listen to at least one of these pointless stories?

ii) What is the probability that she attended exactly one class where the professor told 3 or more stories?

9) A barber at Lackland Air Force base finds that the time it takes him to give a haircut to a new enlistee is a random variable with a mean of 43 seconds and a standard deviation of 12 seconds.

a. If it is now precisely 10:59 am, and he must finish 80 enlistees before leaving for lunch, what is the probability he will be ready to leave by noon?

Have you relied on the Central Limit Theorem to answer this question? Explain.

10) Students leaving a test were asked to assess it on a scale from 1 (easy) to 3 (difficult). Subsequently their scores were set on a scale from 1 (high) to 3 (low). Let X denote assessment and Y score. The table shows proportions of students in each of the nine possible categories.

		Score (Y)		
		1	2	3
Assessment (X)	1	.14	.12	.07
	2	.12	.11	.11
	3	.07	.11	.15

- a. Find the mean of X and the mean of Y . Find the standard deviation of X .
- b. Find the covariance between X and Y . Interpret your result in a sentence or two.
- c. If a student assesses the degree of difficulty as 1, what is the probability his or her score will be in class 3?

11) The mean rate of return on stock A is 10% and on stock B is 20%. The standard deviation of the rate of return is 8% on stock A and 24% on stock B. The correlation coefficient, ρ , between stocks A and B is .4. What is the standard deviation of the return on the portfolio that has an expected return of 16%?

Please read this background information. It should help you to interpret the regressions that follow.

In the December 17, 1996 issue of *PC Magazine*, there are reviews of 18 new home personal computers. The speed of computers is measured using a Winstone score. This measures how fast a computer can complete a set of standard tasks using Windows programs. The higher the score, the faster the computer. In testing computers, *PC Magazine* measures the speed of several subsystems and assigns them scores. The Graphics score measures how fast images are drawn on the computer monitor. The

graphics score depends on how fast the video card is, and also on how fast the CPU/memory subsystem is. The Disk score measures how fast the hard drive is. The CPU mark measures how fast the CPU/memory subsystem is. The CPU mark depends on how much memory is installed, on the raw speed of the CPU, and on the size of the secondary cache. More cache is better, and should improve the CPU mark. Typically, new computers have 16 megabytes of memory, but all the Pentium 200 systems reviewed had 32. There is also CD-ROM score that measures how fast the CD-ROM on the computer is. In each case, a higher score is better than a lower score.

Below is a regression that explains the Winstone score with the scores of each of the subsystems.

The regression equation is

$$\text{Winstone} = 2.50 + 0.245 \text{ graphics} + 0.0158 \text{ disk} + 0.0423 \text{ cpu mark} - 0.00434 \text{ cdRom}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	2.504	2.479	1.01	0.331
graphics	0.24481	0.04999	4.90	0.000
disk	0.015835	0.003209	4.93	0.000
cpu mark	0.042295	0.007963	5.31	0.000
cdRom	-0.004341	0.002171	-2.00	0.067

s = 1.117 R-sq = 97.8% R-sq(adj) = 97.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	716.06	179.01	143.49	0.000
Error	13	16.22	1.25		
Total	17	732.28			

Question 12

- Does this regression do a good job of explaining the Winstone score?
Explain.
- Carefully interpret the coefficient of the graphics score.
- Compute a 95% Confidence Interval for the coefficient of the graphics score.
- Let's call the population coefficient for the cdRom score β_4 . Use the

information to compute the p-value for the following hypothesis:
 $H_0 : \beta_4 \leq 0$
 $H_A : \beta_4 > 0$

Do you accept or reject the null hypothesis?

Would you want to leave the cdRom score in the regression or not? **Explain.**

13) There were four different video card manufacturers whose products appeared in these computers: ATI, Virge, Matrox, and Diamond. To try to determine which of these manufacturers has the best product, three dummy variables were introduced, to represent the ATI, Virge, and Matrox cards respectively. Then the following regression was run to explain the speed of the graphics subsystem.

MTB > regress c3 4 c5 c7-c9

The regression equation is
 $\text{graphics} = 23.7 + 0.0726 \text{ cpu mark} - 17.9 \text{ ati} - 13.8 \text{ virge} - 4.71 \text{ matrox}$

Predictor	Coef	Stdev	t-ratio	p
Constant	23.652	4.554	5.19	0.000
cpu mark	0.07257	0.01129	6.43	0.000
ati	-17.934	2.243	-7.99	0.000
virge	-13.808	2.225	-6.21	0.000
matrox	-4.709	2.763	-1.70	0.112

$s = 2.685$ $R\text{-sq} = 94.3\%$ $R\text{-sq(adj)} = 92.5\%$

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	1548.79	387.20	53.72	0.000
Error	13	93.69	7.21		
Total	17	1642.48			

- Explain how the dummy variable “ATI” would have been formed. In particular, if the first 5 computers in the sample had respectively, ATI, Matrox, Virge, ATI, and Diamond video products, what would be the first 5 numbers entered as data for the variable “ATI”?
- Who seems to be making the best video card(s)? Explain your answer.
- Carefully interpret the coefficients in front of “Virge” and “cpu mark.”
- If a new computer were introduced with a cpu mark score of 350, and a Matrox video card, what is the predicted graphics score?

14) Now let’s turn our attention to the Cpu mark. The two most common CPUs found in these machines were the Pentium 166 and the Pentium 200 (the 200 should be faster than the 166). For simplicity, I eliminated the few machines using other CPUs from the sample. Some of the difference in the Cpu mark score was due to the difference between the Pentium 166 and 200, some was due to the difference between having 256K of cache and 512K of cache, and some was due to the fact that some machines shipped with 32 meg of memory, while others had only 16 meg. To try to figure out the relative importance of these three factors, I defined a dummy variable for Pentium 200 Cpus, a dummy variable for 512K cache, and a dummy variable for 32 meg of memory, and then ran the following regressions.

REGRESSION A

MTB > Regress 'cpu mark' 2 'cache' 'P200'.

The regression equation is
 cpu mark = 309 + 26.7 cache + 35.7 P200

Predictor	Coef	Stdev	t-ratio	p
Constant	308.870	7.803	39.59	0.000
cache	26.65	11.03	2.42	0.036
P200	35.717	9.556	3.74	0.004

s = 16.73 R-sq = 67.7% R-sq(adj) = 61.2%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	5857.5	2928.7	10.46	0.004
Error	10	2800.5	280.1		
Total	12	8658.0			

REGRESSION B

MTB > Regress 'cpu mark' 2 'cache' '32meg'.

The regression equation is
 cpu mark = 306 + 30.2 cache + 34.1 32meg

Predictor	Coef	Stdev	t-ratio	p
Constant	306.458	9.853	31.10	0.000
cache	30.17	12.33	2.45	0.034
32meg	34.06	11.25	3.03	0.013

s = 18.72 R-sq = 59.5% R-sq(adj) = 51.4%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	5154.8	2577.4	7.36	0.011
Error	10	3503.2	350.3		
Total	12	8658.0			

REGRESSION C

MTB > Regress 'cpu mark' 3 'cache' 'P200' '32meg'.

The regression equation is
 cpu mark = 307 + 27.3 cache + 29.4 P200 + 7.8 32meg

Predictor	Coef	Stdev	t-ratio	p
Constant	307.167	9.217	33.33	0.000
cache	27.33	11.66	2.34	0.044
P200	29.42	18.78	1.57	0.152
32meg	7.83	19.77	0.40	0.701

s = 17.49 R-sq = 68.2% R-sq(adj) = 57.6%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	5905.5	1968.5	6.44	0.013
Error	9	2752.5	305.8		
Total	12	8658.0			

- a) In regression A, does the Pentium 200 CPU cause a statistically significant increase in the cpu mark, using $\alpha = .05$? Explain how you reached your conclusion.
- b) In regression C, does the Pentium 200 CPU cause a statistically significant increase in the cpu mark, using $\alpha = .05$? Explain using $\alpha = .05$? Explain how you reached your conclusion.
- c) In regression B, does having 32 meg of memory instead of 16 cause a statistically significant increase in the cpu mark, using $\alpha = .05$? Explain how you reached your conclusion.
- d) In regression C, does having 32 meg of memory instead of 16 cause a statistically significant increase in the cpu mark, using $\alpha = .05$? Explain how you reached your conclusion.
- e) What most likely accounts for these apparently inconsistent conclusions? Explain your reasoning.

15) The average shopper in an express check-out lane has 6.2 items, with a standard deviation of 2.75 items. If 625 shoppers go through an express checkout lane during a shift, what is the probability the cashier will have checked out more than 2400 items? (Assume these shoppers are a random sample from the population of all shoppers)

16) Examine the regression output given below, and use it to answer the following questions. In the regression, PRICE is the selling prices of houses in a city, measured in thousands of dollars. SQFT is the number of square feet of floor space the house has; ROOMS is the number of rooms it has; AGE is the house's age in years, and BEDROOMS is the number of bedrooms it has. The regression was based on 63 observations. (Note: a bedroom is also a room.)

Regression Analysis

The regression equation is

$$\text{PRICE} = 10.4 + 0.0500 \text{ SQFT} + 6.32 \text{ ROOMS} - 0.432 \text{ AGE} - 11.1 \text{ BEDROOMS}$$

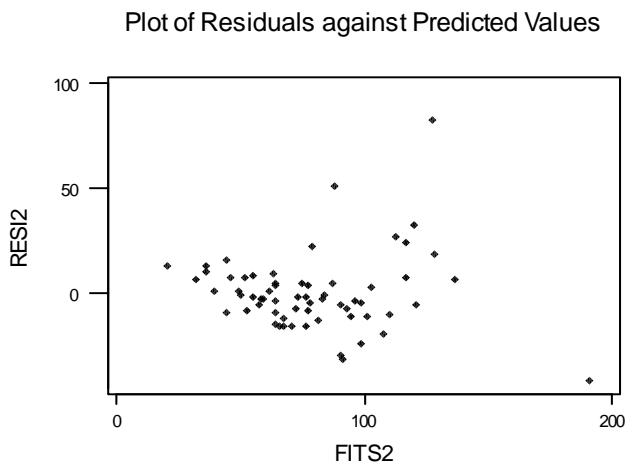
Predictor	Coef	StDev	T	P
Constant	10.37	11.50	0.90	0.371
SQFT	0.050011	0.008104	6.17	0.000
ROOMS	6.322	2.528	2.50	0.015
AGE	-0.4319	0.1097	-3.94	0.000
BEDROOMS	-11.103	5.868	-1.89	0.063

S = 18.96 R-Sq = 72.6% R-Sq(adj) = 70.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	55184	13796	38.37	0.000
Error	58	20853	360		
Total	62	76037			

- Would you say that this regression fits the data well or not? *Explain.*
- Carefully interpret the coefficient of AGE. Phrase your answer so it is appropriate for this *particular* problem.
- If you test the null hypothesis that the coefficient of the variable ROOMS is zero, against the alternative that it is positive, what is the correct p-value? Do you accept or reject the null hypothesis using $\alpha = .01$?
- Your son or daughter has just gone off to college. If this model is correct, and causal, what would be the effect on the value of your home if you converted their bedroom into a den? Explain, with any qualifications you believe are important.
- After fitting this model, you store the residuals and get the following residual plot. Which assumption of the regression model appears to be violated?



- 17) In an observational study to determine the effects of a drug on blood pressure, the diastolic blood pressure of 6 patients taking the drug were compared with the diastolic blood pressure of 9 patients who were not taking the drug. The 9 patients not taking the drug showed a sample mean blood pressure of 77.44 with $s=3.81$, while the 6 taking the drug showed a sample mean blood pressure of 93.00 with $s = 7.16$.
- Compute a 95% confidence interval for the difference in the population means of the two groups.
 - Since it was an observational study, and not a designed experiment, there was a possibility that some factor other than the drug accounted for the difference in blood pressures. Since the treatment group was somewhat heavier than the non-treatment group, weight was one variable that might account for the difference. To see if this might be the case, a multiple regression was run,

using a dummy variable “Drug Tkr” to identify those patients treated with the drug. This is the result of the regression.

The regression equation is
 Bld Pres = 54.0 + 10.6 Drug Tkr + 0.139 weight

Predictor	Coef	Stdev	t-ratio	p
Constant	54.040	8.927	6.05	0.000
Drug Tkr	10.596	2.984	3.55	0.004
weight	0.13950	0.05248	2.66	0.021

s = 4.418 R-sq = 81.2% R-sq(adj) = 78.0%
 Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	1009.06	504.53	25.84	0.000
Error	12	234.27	19.52		
Total	14	1243.33			

- b) If the first 9 people in the data set were those who did not take the drug, and the last 6 were those who did, write down the column of data you would input for the variable “Drug Tkr.”
 - c) Using the results of the multiple regression, compute a 95% confidence interval for the impact of the drug on blood pressure.
 - d) The answers you got in parts (b) and (c) are not the same, nor should they be, because they measure subtly different things. Explain the difference.
 - e) In the regression above, what is the estimated variance of the regression’s error term?
- 18) A frustrated web surfer is annoyed with the long delays in downloading material from the World Wide Web (or World Wide Wait, as it is also known.) He gets the idea that it might be quicker to do his surfing in the morning, while users on the West Coast are still asleep. To test the hypothesis that things should be quicker in the morning, he selects 6 web pages and measures the waiting time to access each one in the morning, and then to access the same pages in the afternoon. This is his data:

Web Site	Wait in morning	Wait in Afternoon
ST “First Contact”	45 seconds	62 seconds
Virgo	12 seconds	13 seconds
Baywatch	30 seconds	35 seconds
Whitehouse.gov	21 seconds	26 seconds
Whitehouse.net	42 seconds	35 seconds
Monty Python	32 seconds	43 seconds

Is this convincing evidence that web surfing is quicker in the morning? Use $\alpha = .10$.

- a) Write down the null and alternative hypotheses, being sure to use the correct notation.

b) Perform an appropriate test of the hypothesis. Do you accept or reject the null hypothesis?

19) a) Suppose you come across a report of an experiment done to determine the value of anti-lock brakes in preventing collision damage to automobiles. According to the report, 800 prospective buyers of a particular brand of car that does not ordinarily have anti-lock brakes were randomly assigned to either a control or treatment group. The buyers in the control group received the ordinary version of the automobile, while the buyers in the treatment group were given anti-lock brakes as a free option in an otherwise identical car. After 5 years, the 400 buyers in the control group had experienced accidents with an average collision damage of \$688 and a standard deviation of \$250, while the 400 buyers in the treatment group experienced accidents with an average collision damage of \$625 with a standard deviation of \$220. Perform a hypothesis test to see whether this is convincing evidence that anti-lock brakes reduce collision damage. Use $\alpha = .05$.

b) In reading the report above, you notice something odd. Of the 400 people in the treatment group, 291 were over 40 years old, while of the 400 people in the control group, only 254 were over 40 years old. The report stated that the assignment was random. However, if the statement in the report is not exactly true, and assignment was only haphazard, the researchers might have given anti-lock brakes to people who lobbied hard for them, or people they felt “ought” or have them, in which case, there is no way to scientifically test the hypothesis in part (a). Perform a 2-sided hypothesis test to see if there is convincing evidence of non-random assignment. Because you have faith in published reports, and wouldn’t have even thought to do the test before you looked at the data, you feel it is only fair to insist on a small chance of type I error, $\alpha = .01$. Compute the p-value associated with this hypothesis test. Do you accept or reject the null hypothesis?

20) A mail order firm advertises that on average customers must wait no more than 3 days before receiving delivery of their orders. A skeptical consumer’s group decided to test this claim, so they order 50 packages to be shipped to 50 different addresses, and measure the number of days before each was delivered. They discover that the sample mean wait was 3.08 days. It is known that $\sigma = .5$ days.

a) State the null and alternative hypotheses -- write them out formally, being careful to define your notation.

b) Using $\alpha = .05$, describe the acceptance and rejection region in this case. Given the observed sample mean, do you accept or reject the null hypothesis?

c) Suppose that in fact, the *true* mean shipping time was 3.1 days. What would be the chance of committing a type II error if a new sample were drawn, and a test performed using the procedure above?

21) A researcher in parapsychology wants to test a subject for ESP. He proposes to use the card-guessing experiment described in class. Being skeptical, the researcher plans to use $\alpha = .005$ and a one tailed greater-than test of the null hypothesis that $p = .20$. The test subject is to be allowed to guess 1000 cards.

- a) What fraction of the subject's responses must be correct before the researcher will be willing to believe he or she is psychic?
- b) Suppose the subject is genuinely psychic and capable of getting $p = .23$ of the cards right. What is the power of the proposed test?

22) Below you will find regression output used to predict the asking price of used Chevrolet Camaros. The variable "askprice" is the asking price of the car in dollars; the variable "age" is the car's age in years; the variable "mileage" is the number of miles the car has on it; the variable "mile sq" is the square of the mileage. The car's condition was identified as either excellent, average, or poor. The variable "ave cond" is a dummy equal to one when the car's condition is average and is equal to zero otherwise; the variable "poorcond" is a dummy equal to one when the car's condition is poor, and equal to zero otherwise.

```
MTB > regress c1 5 c2 c3 c4 c5 c7
```

The regression equation is
 askprice = 20171 - 902 age - 192 mileage - 929 ave cond - 2248 poorcond
 + 1.09 mile sq

Predictor	Coef	Stdev	t-ratio	p
Constant	20170.9	681.1	29.62	0.000
age	-902.0	149.4	-6.04	0.000
mileage	-191.65	30.86	-6.21	0.000
ave cond	-929.0	520.7	-1.78	0.087
poorcond	-2248.0	508.1	-4.42	0.000
mile sq	1.0921	0.1978	5.52	0.000

s = 818.8 R-sq = 96.3% R-sq(adj) = 95.5%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	5	414788608	82957720	123.75	0.000
Error	24	16089027	670376		
Total	29	430877632			

Unusual Observations

Obs.	age	askprice	Fit	Stdev.Fit	Residual	St.Resid
24	3.00	14350	12825	336	1525	2.04R

If our regression model is given by

$$\text{askprice} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{mileage} + \beta_3 \text{ave cond} + \beta_4 \text{poorcond} + \beta_5 \text{mile sq}$$

- a) Interpret the coefficients β_1 and β_4 in this particular example. *Be precise and be sure your answer is in terms of these particular variables!*
- b) Compute a 99% confidence interval for β_1 .
- c) In what way was is the 24th observation "unusual." Is it very unusual, or only moderately so? Explain how you reached your conclusion.
- d) As you can see from the output, a quadratic term in mileage is very significant. I did not recognize the need for a quadratic term at first, and fit the linear model:

$$\text{askprice} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{mileage} + \beta_3 \text{ave cond} + \beta_4 \text{poorcond}$$

What diagnostic tool do you think suggested the need for the quadratic term?
What would the graph (for the inappropriate linear model) look like?

e) Use the output below to test the null hypothesis, using $\alpha = .05$, that

$$H_0: \beta_3 = \beta_4 = 0$$

H_A : at least one coefficient $\neq 0$

```
MTB > regress c1 3 c2 c3 c7
```

The regression equation is

```
askprice = 20404 - 767 age - 247 mileage + 1.30 mile sq
```

30 cases used 1 cases contain missing values

Predictor	Coef	Stdev	t-ratio	p
Constant	20404.4	915.2	22.29	0.000
age	-767.2	198.0	-3.88	0.001
mileage	-246.89	36.91	-6.69	0.000
mile sq	1.3050	0.2416	5.40	0.000

s = 1115 R-sq = 92.5% R-sq(adj) = 91.6%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	398532384	132844128	106.78	0.000
Error	26	32345234	1244048		
Total	29	430877632			

You should also do a few chapter 11 questions. Numbers 23, 29 and 31 on pages 440-441 of Anderson, Sweeny and Williams would be reasonable examples.