AP Statistics
3rd Six weeks (2010-2011)

MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
November 8	9	10	11	12
Power Models HW: 4.15 Skip "D"	**Quiz 4.1** Project Assigned	Interpreting Correlation and Regression HW: 4.33,4.37,4.38, 4.42,4.44,4.76	Relations in Categorical Data HW: 4.51,4.53-4.58 (Worksheet)	Computer Lab Review
15	16	17	18	19
Review HW: 4.67, 4.81,	Intro to Random Variables HW: 7.3, 7.4, 7.5	**Chapter 4 Test** HW: AP Problem Set	Continuous Random Variables HW: 7.6-7.8, 7.13-7.17odd	Means and Variances of Random Variables HW: 7.22, 7.28, 7.34, 7.36, 7.37, 7.41,7.46, 7.60, 7.61
22	23	24	25	26
Calculating Expected Value/Law of Large Numbers Worksheet 7.24,7.25,7.32,7.33	**Quiz** Project Due HW: Review Sheet	Thanksgiving Holiday	Thanksgiving Holiday	Thanksgiving Holiday
29	30	December 1	2	3
Review	Binomial Distributions HW: Worksheet	**Chapter 7 Test**	Binomial Formula HW: 8.7,8.8,8.13 8.16-8.18	Normal Approx. to a Binomial Distribution HW: WS #1
6	7	8	9	10
Normal Approx. to a Binomial Distribution HW: WS #2	**Quiz** HW: 8.1-8.6	Intro to Geometric Distribution HW: Worksheet	Geometric and Binomial Mixed HW: Worksheet	Final Exam Review
13	14	15	16	17
Final Exam Review	**Final Exams**	**Final Exams**	**Final Exams**	**Final Exams**

****All assignments subject to last minute changes!**

Example: Cell Phone Usage (found on page 205 of the textbook).

The following table shows the growth in cell home usage per year.

Year	Subscribers	Log Subscribers
1990	5283	
1993	16,009	
1994	24,134	
1995	33,786	
1996	44,043	
1997	55,312	
1998	69,209	
1999	86,047	

1. Construct a scatterplot on the grid provided. Calculate the ratios of the subscriber values in order to determine if the data is exponential – show the results below. Is the data exponential? Why?

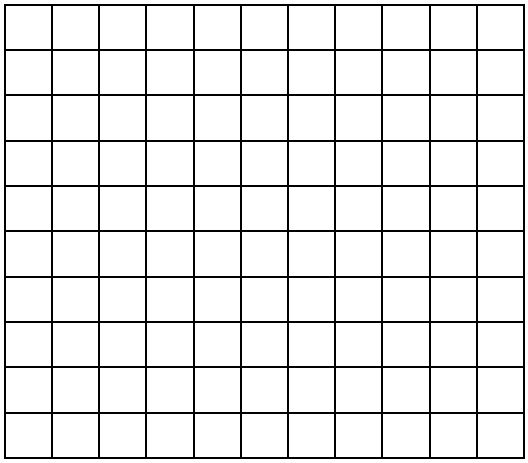
2. Perform an appropriate transformation on the data in order to linearize the data. Show the transformed values on the table above. Then perform least-squares regression on the transformed data. Write the LSRL for the transformed data.

   What is the correlation coefficient?

   Interpret what the coefficient of determination is telling you in the context of this problem.

3. Convert the LSRL to exponential form ($y = ab^x$).

4. Calculate the residuals for the transformed data and make a residual plot. Explain the growth patterns over time which is reflected in this plot. Be sure to identify which years have growth slower or faster than the model predicts.



5. Use your model to predict cell phone usage in the year 1992. Would you feel comfortable using your model to predict cell phone usage for the year 2000? Why or why not?

Power Regression Model Example

We want to evaluate the lead in the soil when distance from the high point in the ground is measured. Soil is measured in parts per million and distance is measured in meters.

Distance	Soil Content	Log Distance	Log Soil Content	Residuals
.3	62.75			
1	37.51			
5	29.70			
10	20.71			
15	17.65			
20	15.41			
25	14.15			
30	13.50			
40	12.11			
50	11.40			
75	10.85			
100	10.85			

1. Plot the data on your calculator. What do you see?

2. Take the log(L1) and log (L2)

3. Use the transformed data and plot the new graph. What do you see?

4. Write the LSRL line to fit the new data. Be sure to find r and r squared.

5. Interpret the slope and r squared.

6. Make a residual plot on you calculator. What does this tell you about the data?

7. Convert the transformed LSRL line back into a power model so we can make predictions.

8. Make a prediction for soil content when the distance from the highway is 23 meters.

Graphical Methods for Describing Univariate Data

You are a dietician who has been asked to make a presentation to a high school class. You would like to discuss the nutritional value of fast food, using graphical analysis as we have done previously. You are to:

1. Pick a fast food restaurant (McDonalds, Domino's Pizza, Taco Bell, etc.) of your choosing. Use data obtained either from company brochures or from data obtained from company web sites. Look at a nutritional value* across all food products – entrees, salads, and sides. If there is more than one size for that product, use the middle size. Prepare a histogram and a modified boxplot of the nutritional value of the data set. You may also include a Normal Probability Plot, but you must refer to it in your analysis. Include number summaries such as the mean, median, standard deviation, IQR, five-number summary. Use the CUSS method of describing your distributions (Center, Unusual Characteristics, Shape and Spread). Identify outliers from your modified boxplot and explain them.

   Do the same for a different restaurant using the same nutritional value as for the first restaurant. In addition, compare the two restaurants on this nutritional value. Answer the question "Which is healthier?" Your answer should be supported with graphical and number summaries. Compare your findings to established nutritional guidelines. You should look them up and cite them.

2. Pick two other restaurants (they must differ from the ones you analyzed above) and repeat the process from #1 with a different nutritional value than the one you chose previously.

Your report must be proof-read and typed. Use a cover sheet and avoid "tacky" embellishments such as clip art and "coloring" with markers. Graphs and number summaries must be done using Minitab. A 30-day trial version for PC may be downloaded at:

http://www.minitab.com/en-US/support/downloads/default.aspx

Copies of company brochures and/or data printouts from company web pages are to be attached to the BACK of the report. Include an introduction for each restaurant with historical and other information. Include a bibliography of sources.

A professional job is expected. Samples are provided of past student work. It must be stapled, bound, or placed in a folder before you come to class. It must ready to hand in the minute you walk in the door – staplers will not be available the day your project is due.

See the back for a list of project "No-No's".

*nutritional values include things such as fat grams, sodium content, grams of carbohydrates, etc.

Project No-No's

1. **Don't be late.** Each 24-hour time period after the due date is a letter-grade deduction. Also, don't miss class the day it is due and still attend your other classes. That is still a penalty.

2. Don't fill space with giant graphs. Your graphs should be adjusted to fit within the bounds of the paper.

3. Don't leave too much blank space in your project – no more than an inch or two per page.

4. Don't rely on spell-check. Proof-read your work.

5. Don't leave out required elements.

6. Don't wait until the last minute to complete your project. The product you turn in will show it.

7. You must have four different restaurants, each with a minimum of two graphs.

8. Check to see that you have included the correct titles and labels on your graphs.

9. Do not make reference to the "range" as a legitimate numerical summary. It isn't.

10. Check your printer ink cartridge well in advance. Last minute printer problems aren't an adequate excuse for late work. There is always Kinko's.

11. The project must be attached together either with binding, staples, binder clips, etc. **No paper clips!** Loose pages are a deduction, even if they are placed in a folder.

12. Wrinkled pages are unprofessional and will force a deduction. Covers which are second-hand, worn, etc will also be penalized.

13. Leave the brochures/internet printouts attached to the **end** of your report. Do not intersperse it throughout the project. No-one cares about the raw data, unless there is a question about your results/conclusions.

14. Do not include graphs which you don't or can't interpret in your analysis.

Relations in Categorical Data

In this section, you will study some basic techniques for comparing distributions of categorical variables. These techniques involve the analysis of two-way tables of counts – we use the counts or percents of individuals that fall into various categories to analyze categorical data.

Activity #1 – On an index card, answer the following questions:

1. What is your gender – male or female?
2. If you could choose one of the following accomplishments for your life, which would you choose:
   - to win an Olympic gold medal
   - to become President of the United States
   - to win an Academy Award

Activity #2 – Consider the two **categorical** variables we just collected, gender and accomplishment. Which of these variables would you consider as the explanatory variable and which as the response variable?

- Explanatory variable –

- Response variable –

Count how many students fall into each of the six possible pairs of responses to these questions. Record these counts in the appropriate calls of the table below.

	Male	Female
Gold Medal		
President		
Academy Award		

This table is called a **two-way table** since it classifies each person according to two variables. In particular, it is a 3 x 2 table; the first number represents the number of categories of the row variable (accomplishment), and the second number represents the number of categories of the column variable (gender). The explanatory variable should be in columns and the response variable in rows.

Activity #3 –

In a national survey of adult Americans in 1998, people were asked to indicate their age and to classify their interest in politics as very much, somewhat, or not much. While age is typically a quantitative variable, it was categorized into three groups for this analysis: 18 – 35, 36 – 55, and 56 – 94 (the oldest subject in the survey). The results are summarized in the following table of counts; notice that row and column totals are also provided.

	18 – 35	36 – 55	56 – 94	total
not much	146	146	89	381
somewhat	192	260	154	606
very much	47	125	106	278
total	385	531	349	1265

(a) What proportion of the survey respondents were between ages 18 and 35?

(b) What proportion of the survey respondents were between 36 and 55 years of age?

(c) What proportion of the survey respondents were over age 55?

You have calculated the **marginal distribution** of the age variable. When analyzing two-way tables, one typically starts by considering the marginal distribution of each of the variables by itself before moving on to explore possible relationships between the two variables.

To study possible relationships between two categorical variables, one examines **conditional distributions**; i.e., distributions of one variable for given categories of the other variable.

(d) Restrict your attention (for the moment) to just the respondents under 35 years of age. What proportion of these young respondents classify themselves as having not much interest in politics?

(e) What proportion of the young respondents classify themselves as somewhat interested in politics?

(f) What proportion of the young respondents classify themselves as having very much interest in politics?

(g) Record the conditional distribution that you have just calculated in the "18 – 35" column of the table below:

	18 – 35	36 – 55	56 – 94
not much		.275	.255
somewhat		.490	.441
very much		.235	.304
total	1.000	1.000	1.000

(h) Based on the calculations you have performed, does there seem to be any relationship between age and political interest? in other words, does the distribution of political interest seem to differ among the three age groups? If so, describe key features of the differences.

(i) Draw a bar chart that compare the percent of people in each age group that do not have much interest in politics.

In dealing with conditional proportions, it is very important to keep straight which category is the one being conditioned on. For example, the proportion of American males who are US Senators is very small, yet the proportion of US Senators who are American males is very large.

Refer to the original table of counts to answer the following:

(j) What proportion of respondents aged 36 – 55 classified themselves as not much interested in politics?

(k) What proportion of those with not much interest in politics are of age 36 – 55?

(l) What proportion of the people surveyed identified themselves as being both between the ages of 36 – 55 and having not much political interest?

HW – 4.51, 4.53-4.58

AP Statistics
(Chapter 4 Review) **Name _____**

Part I - Multiple Choice (Questions 1-10) - Circle the answer of your choice.

1. The relationship between population (y) and year (x) was determined to be exponential. The least-squares regression equation of the appropriately transformed data was $\hat{y} = .05 + .004x$. What would be the predicted population in the year 1990?

(a) 8.46
(b) 288,403,150
(c) 3.21
(d) 102,329,299
(e) There is insufficient information to make a prediction.

2. Suppose that the scatterplot of (log x, log y) shows a strong positive correlation close to 1. Which of the following are true?

I. The variables x and y also have a correlation close to 1.
II. A scatterplot of (x, y) shows a strong nonlinear pattern.
III. The residual plot of the variables x and y shows a random pattern.

(a) I only
(b) II only
(c) III only
(d) I and II
(e) I, II, and III

3. What is the purpose of residual plots?

(a) To determine causation.
(b) To assess the type of relationship that exists between x and y.
(c) To check the appropriateness and fit of the regression equation for the data.
(d) To measure the variability in the residuals.
(e) To provide predictions for the response variable.

4. Fourth grade children were asked what emotion they associated with the color red. The responses for emotion and gender of the children are summarized in the following two-way table.

	Anger	**Pain**	**Happiness**	**Love**
Male	35	27	12	38
Female	27	17	19	39

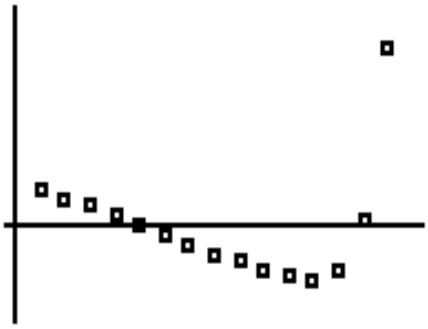
What proportion of the males associate the color red with love?

(a) 0.5234
(b) 0.3598
(c) 0.3393
(d) 0.1822
(e) 0.1775

5. A strong negative association between Average State SAT scores and Percentage of students taking the SAT reflects which underlying relationship?

(a) causation
(b) correlation
(c) common response
(d) extrapolation
(e) confounding

6. The following residual plot was generated after fitting a LSRL to a set of data. The most likely conclusion would be:



(a) The LSRL is an appropriate model since the residuals are randomly scattered.
(b) There is a pattern in the residuals which indicates an exponential model would be more appropriate.
(c) There is a pattern in the residuals which indicates a power model would be more appropriate.
(d) There is a pattern in the residuals which indicates a nonlinear model would be more appropriate, but the type cannot be determined from the residual plot.
(e) The residuals indicate there cannot be a relationship between the variables, so finding a model would be inappropriate.

7. Two variables are confounded when:

(a) The effect of one variable on the response variable is dependent upon the effect of the other variable.
(b) The effect of one variable on the response variable cannot be separated from the other variable.
(c) The effect of one variable on the response variable changes the impact of the other variable on the response variable.
(d) Both variables are classified as lurking or extraneous variables.
(e) They interact in their effects on the response variable.

8. Which of the following are true statements?

I. High correlation does not necessarily imply causation.
II. A lurking variable is a name given to variables that cannot be identified or explained.
III. Successful prediction requires a cause and effect relationship.

(a) I only
(b) II only
(c) III only
(d) I and III only
(e) I and II only

9. If the model for the relationship between the score on AP Statistics Test #4 (y) and the number of hours spent preparing for the test (x) was $\log y = 0.1 + 1.9 \log x$, determine the residual if a student studied 9 hours and earned an 85.

(a) 6.53
(b) 3.14
(c) 15.23
(d) 0
(e) –4.86

10. A study was conducted to determine the effectiveness of varying amounts of vitamin C in reducing the number of common colds. A survey of 450 people provided the following information:

	Daily amount of Vitamin C taken		
	None	**500 mg**	**1000 mg**
No colds	57	26	17
At least one cold	223	84	43

What conclusion can be made?

(a) The data proves that vitamin C reduces the number of common colds.
(b) The data proves that vitamin C has no effect on the number of common colds.
(c) There appears to be a strong association between consumption of vitamin C and the occurrence of common colds.
(d) There appears to be little association between consumption of vitamin C and the occurrence of common colds.
(e) Since common colds are caused by viruses, there is no reason to conclude that vitamin C could have any effect.

Part II – Free Response (Questions 11-13) – Show your work and explain your results clearly.

11. The table below describes the data comparing the relationship between age groups and localities of residence.

		Localities of Residence			
		Urban	Suburban	Rural	Totals
Age Groups	Under 25	110	150	65	
	25-50	240	220	75	
	Over 50	53	112	58	
Totals					

(a) Compute the marginal frequencies. **Place the answers in the table.**

(b) What percent of the urban dwellers are over 50? _____

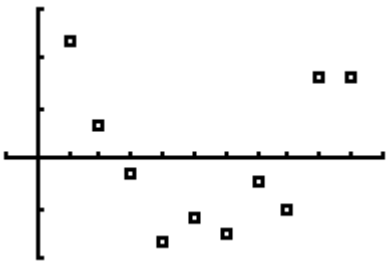
(c) What percent of the 0ver-50 residents live in rural areas? _____

(d) Compute the percentage of:

		Urban	Suburban	Rural
i.	Dwellers Category	_____	_____	_____
ii.	,given they are Under 25	_____	_____	_____
iii.	,given they are 25-50	_____	_____	_____
iv.	,given they are Over 50	_____	_____	_____

(e) Based on your analyses, do you believe that these data indicate there is a relationship between locality of residence and the ages of the residences? Explain your answer.

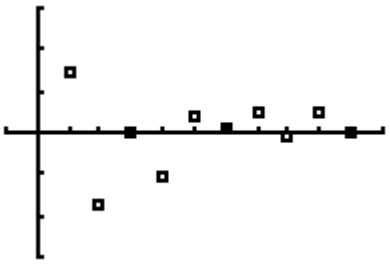
12. A linear model for a set of data was $y = -20.6 + 9.72x$ and produced the following residual plot. [Xscl = 1, Yscl = 5]



(a) Predict the value of y when x = 7.5.

(b) Determine the residual if x = 5.

A nonlinear model for the same set of data was $\log y = -.14 + 2.07 \log x$ and produced the following residual plot. [Xscl = 1, Yscl = 0.1]



(c) Predict the value of y when x = 7.5.

(d) Determine the residual if x = 5.

(e) Which model is better and why?

13. The following data represent the Woodward Academy Upper School enrollment over the past 35 years.

Year	1965	1970	1975	1980	1985	1990	1995	2000
Enrollment	650	690	740	790	840	900	960	1025

(a) Sketch a scatterplot of the data.

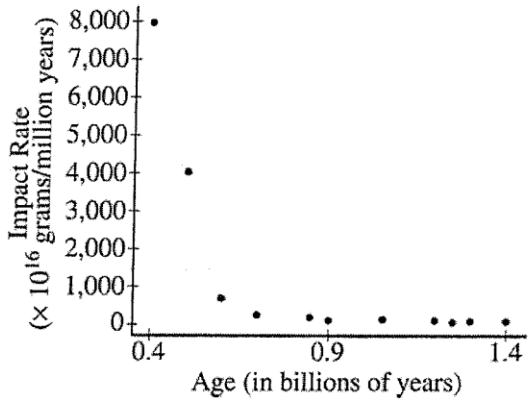
(b) Using an appropriate transformation, sketch a scatterplot of the transformed data.

(b) Determine an appropriate model for the data. Justify your answer.

(c) Use your model to predict the enrollment in 2010.

Directions: Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

1. The Earth's Moon has many impact craters that were created when the inner solar system was subjected to heavy bombardment of small celestial bodies. Scientists studied 11 impact craters on the Moon to determine whether there was any relationship between the age of the craters (based on radioactive dating of lunar rocks) and the impact rate (as deduced from the density of the craters). The data are displayed in the scatterplot below.



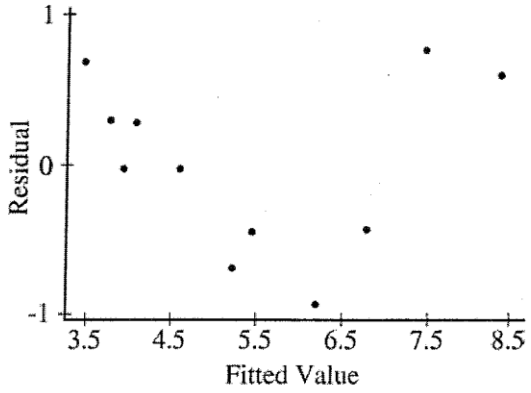
(a) Describe the nature of the relationship between impact rate and age.

Prior to fitting a linear regression model, the researchers transformed both impact rate and age by using logarithms. The following computer output and residual plot were produced.

Regression Equation: ln(rate) = 4.82 – 3.92 ln(age)

Predictor	Coef	SE Coef	T	P
Constant	4.8247	0.1931	24.98	0.000
ln(age)	–3.9232	0.4514	–8.69	0.000

S = 0.5977 R-Sq = 89.4% R-Sq (adj) = 88.2%



(b) Interpret the value of r^2.

(c) Comment on the appropriateness of this linear regression for modeling the relationship between the transformed variables.