Spring, 2010

Applied Statistics Comprehensive Exam
Statistical Methods I & II

Calculators are permitted for this exam. Statistical tables are provided.
Answers to all questions require complete explanations to receive full credit.

1. (20 points) For each of the following, identify what <u>type of analysis</u> you would use or was used. In addition, <u>write down the null and alternative hypothesis for each.</u>

   Ex: We are trying to predict a person's glucose level based on his/her age.
   Answer 1: Correlation, H_0: $\rho=0$ vs. H_a: $\rho\neq0$, <u>or</u>
   Answer 2: Regression, H_0: $\beta_1=0$ vs. H_a: $\beta_1\neq0$

   a. A teacher will give the class a test on student attitudes towards statistics on the first day of class and record their name and score. Once the class is complete, he will give them the same exam and again record their name and score. Has taking the class improved the student's attitudes towards statistics?
   b. The average time (in seconds) until a topical anesthesia takes effect is compared for four treatments, along with a control group using the existing anesthesia.
   c. Middle school students were asked which award they'd rather win – an Academy Award, an Olympic Gold, or a Nobel Prize. Do equal proportions of students prefer each award?
   d. An evaluation of psychotherapeutic effectiveness is examined to determine whether three different types of therapy have different impacts on improvement in patient self-esteem. Six therapists are randomly assigned to each of the three therapy groups and they employ their therapy type on all their patients. The self-esteem improvement is then measured on each patient and examined to determine whether the three treatments result in difference changes in self-esteem.

2. (20 points) A random sample of adults was asked whether they believe that astrology is scientific or not. The table below presents a two-way table of their answer to that question versus three levels of higher education. Do these data present evidence that belief about astrology is associated with level of higher education at the 5% level of significance? Make sure to verify the assumptions for this test.

	Associate's	**Bachelor's**	**Master's**	TOTAL
Not Scientific	**169**	**256**	**114**	539
Scientific	**65**	**65**	**18**	148
TOTAL	234	321	132	687

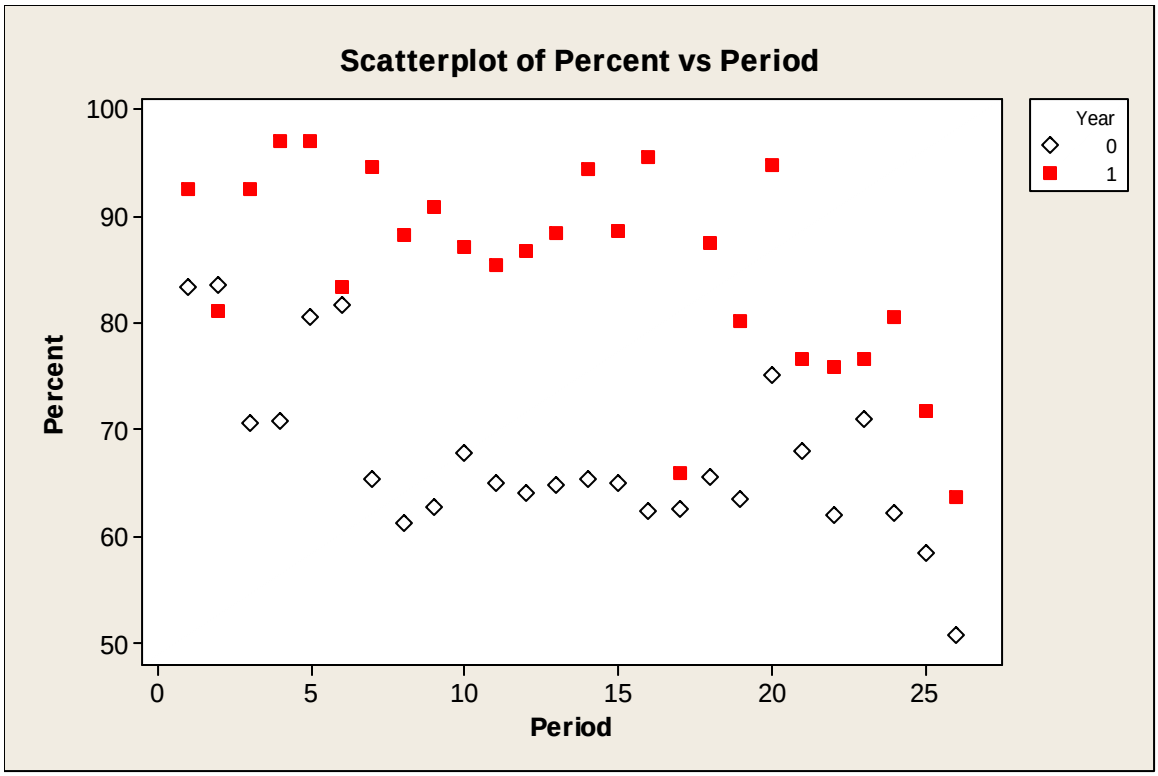
3. (25 points) A study was done to compare the number of minutes per day "lean" vs. "obese" people spend standing or walking. The following table presents the summary data. Let us (perhaps naively) assume that the distribution of time spent standing or walking is normally distributed for each group.

Group	n	Sample Mean	Sample St.Dev.
Lean	10	525.8	107.1
Obese	10	373.3	67.5

   a. Do these data present evidence, at the 5% level of significance, that there is an association between body type (lean vs. obese) and time spent standing or walking? Make sure to state all assumptions necessary for this test.
   b. If the true mean difference was 1 hour, how large a sample size (per group) would be needed in order to have 80% power to detect such a difference? For this part, you may assume that there is a common standard deviation, σ=90, and that the sample size for each group will be the same.

4. (35 points) How many jurors should be summoned in order to achieve the number of jurors that may be needed for a trial? Jury duty for a court in Ohio is two weeks long, so a new group of potential jurors must be brought together twenty-six times a year. Random sampling methods are used to obtain a sample of registered voters in the county every two weeks, and these individuals are sent a summons to appear for jury duty. Not all of the voters who receive a summons actually appear for jury duty. The percent who report for duty in each of the twenty six time periods in 1998 was recorded. New efforts were made to increase participation rates in 2000. The percent who report for duty in each of the twenty six time periods in 2000 were also gathered. The variable *year* takes on the value 0 for 1998 and 1 for 2000. The following pages present some plots and results from the multiple regression analysis. The multiple regression model used is $percent = \hat{\beta}_0 + \hat{\beta}_1 period + \hat{\beta}_2 year$.

   a. Interpret the value for the slope for the *year* variable.
   b. What is the estimated value for σ (the standard deviation of the error terms)?
   c. What is the value of R^2? Interpret this value.
   d. State the assumptions for this regression and evaluate, as best as you can from the given output, whether they are met.
   e. Explain what an interaction term between *year* and *period* would mean, in the context of the study. Do you think that the interaction term would be significant? Explain why or why not.



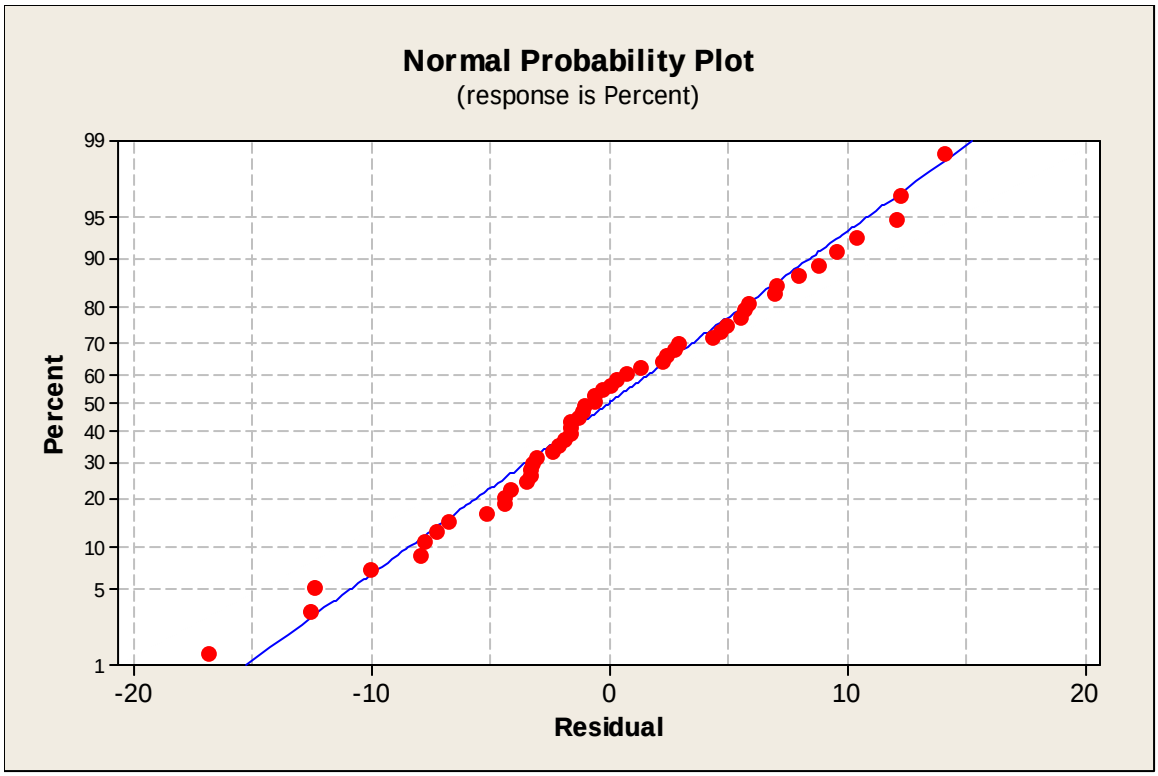
Regression Analysis: Percent versus Period, Year

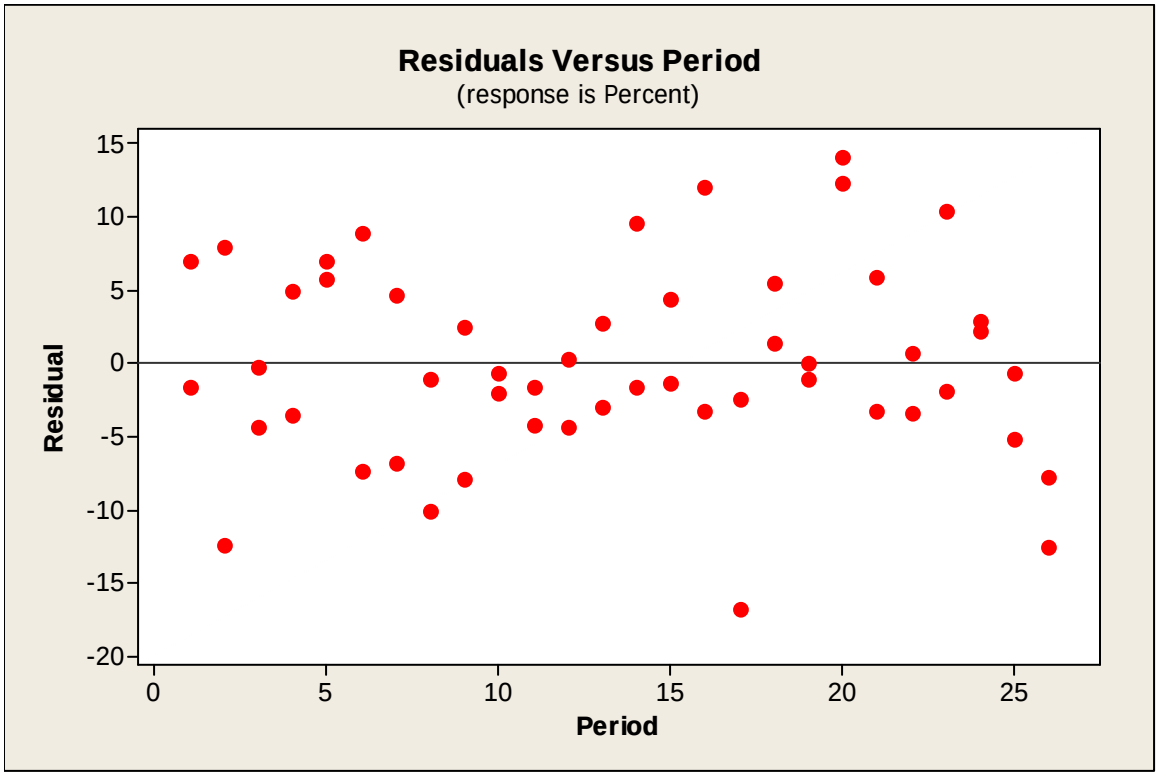
The regression equation is
Percent = 77.1 - 0.717 Period + 17.8 Year

Predictor	Coef	SE Coef	T	P
Constant	77.082	2.130	36.19	0.000
Period	-0.7168	0.1241	-5.78	0.000
Year	17.833	1.861	9.58	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5637.4	2818.7	62.62	0.000
Residual Error	49	2205.6	45.0		
Total	51	7842.9			


Normal Probability Plot
(response is Percent)
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-20
-10
0
10
20
Residual


Residuals Versus Period
(response is Percent)
Residual
15
10
5
0
-5
-10
-15
-20
0
5
10
15
20
25
Period