

STATISTICS 210 – Fall 2009

Midterm Exam: Wednesday November 4, 2009

- The exam is closed book, closed notes. You may use a calculator.
- Tables are provided separately.
- Problem values are written to the left of each problem. The total number of points is equal to 120.
- Two important reminders:
 - show your work so that you can receive partial credit;
 - budget your time so that you don't miss problems you know how to do.
- Please do not write your solutions on the exam paper.
- Good luck!

1. Everyone has had the experience of telling a story to someone and wondering if you have already told it to them. This is related to the idea that destination memory (remembering items we have told to others) is worse than source memory (remembering item we have heard from others). A psychologist carries out a randomized experiment to compare the two kinds of memory. Seventy individuals participate in the experiment. Thirty-five are randomized to a “destination” condition in which they learn facts by telling them to pictures of famous people. The other thirty-five are assigned to a “source” condition in which they learn facts by hearing them from pictures of famous people. Later the subjects are tested on the facts they have learned to see if they can identify who told them (or who they told) the fact. The response for each subject is the percentage correct. Summaries from the data are provided in the table below.

	Destination condition	Source condition
n	35	35
mean	61.3	74.7
s.d.	14.3	13.9

- (a) The quantity 61.3 given in the table is an estimate. Carefully identify the parameter that it estimates.
- (b) Give a 95% confidence interval for the difference in mean percentage correct for the two conditions.
- (c) Explain what the confidence interval tells us in terms that a a non-statistical audience could understand.
- (d) A critic complains that age is strongly associated with memory and that this experimental design does not control for this important factor. The psychologist says that she doesn't need to worry about age. What feature of the experimental design guarantees that the conclusion is still valid even though the study did not not account for age? Explain.
- (e) Even though the study is valid as carried out it might still be a good idea to control for age by creating pairs of individuals of the same age and randomizing one member of each pair to each condition. Explain the benefits of this idea.

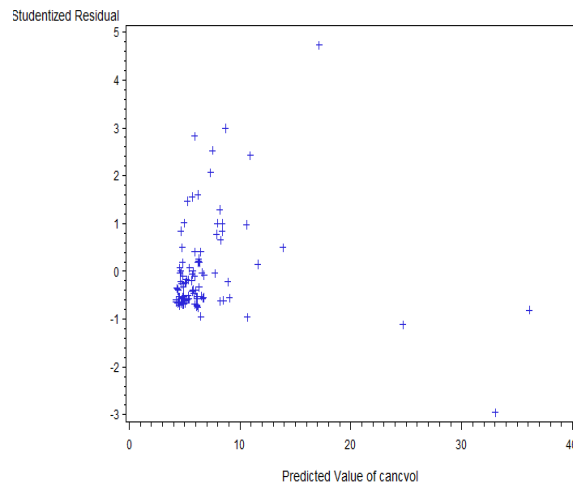
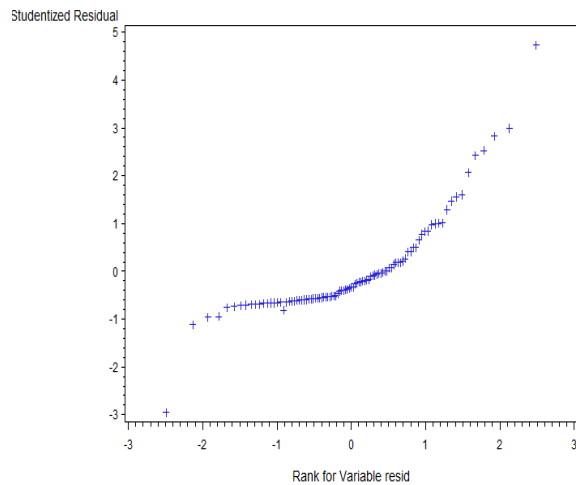
2. The PSA (prostate-specific antigen) test is used as a screening test for prostate cancer. (You may remember we examined some data related to the PSA test on one of the homeworks.) The justification for using the PSA as a screening test comes from a series of studies in which the size of detected cancers were regressed on PSA test scores for a sample of men. This question considers data from one such study. There are 97 men; the response variable is the volume of the cancer detected and the predictor is the PSA test score. Summary statistics for the two variables are provided in the table below and so is an incomplete regression table. Plots of the studentized residuals are also provided.

Summary statistics

variable	mean	s.d.
psa	23.73	40.78
cancer	7.00	7.88
correlation = .624		

Regression results

variable	param.est.	std.err.
intercept		0.728
psa		0.016



- Calculate the regression line for predicting cancer volume from PSA test score.
- Interpret the slope parameter that you have estimated. What does it tell us about the relationship of the variables?
- Do the assumptions of the simple linear regression model appear to be satisfied? Explain.

As an alternative to the above regression a transformed model is considered next. The logarithm of cancer volume is regressed on the logarithm of PSA count. Summary statistics and regression output is provided for this regression below.

Summary statistics

Variable	mean	s.d.
logpsa	2.479	1.154
logcanc	1.350	1.179
correlation = .734		

Regression results

Variable	Param.Est.	Std.Err.
intercept	-0.509	0.195
logpsa	0.750	0.071

- Give a 95% confidence interval for the slope.
- What does this estimated regression equation tell us about the relationship of PSA test score and cancer volume on the original (untransformed) scale? Explain.

3. Exposure to lead when young can produce long-lasting effects on mental and physical performance. One study considered 99 children in El Paso who lived close to a lead smelter and thus were exposed to lead in the air. The 99 children were split into three groups based on their exposure during a two year period: group 1 consists of the children with low (< 40 mg) exposure in both 1972 and 1973; group 2 consists of the children with high exposure in 1972 and low exposure in 1973; group 3 consists of the children with high exposure in both 1972 and 1973. The children were given a battery of tests – we focus here on a finger tapping task which measures the number taps done by each child in a fixed period of time. Summaries of the data are provided in the table below.

Group	Lead exposure		n	mean	s.d.
	(1972, 1973)				
1	$< 40, < 40$		64	49.8	11.2
2	$\geq 40, < 40$		16	47.7	11.2
3	$\geq 40, \geq 40$		19	40.7	12.3

- (a) An incomplete analysis of variance table is provided below. Complete the table and carry out a test of the hypothesis that all population means are the same. Be sure to carefully state the null and alternative hypotheses, obtain the P -value, and state your conclusion.

Source	d.f.	SS
Groups		
Error		
Total	98	13720.76

- (b) The contrast defined by $c = (2, -1, -1)$ compares the group with low lead exposure in both years to the other two groups. Test whether the data support the pattern encoded in this set of contrast weights. Carefully state the null and alternative hypotheses and interpret your result.
- (c) Identify a set of contrast weights that compares children with low exposure in 1973 to those with high exposure in 1973. Would this contrast be independent of the contrast in (b)?
- (d) Does this study by itself prove a causal link between lead exposure and deteriorating fine motor skills? Explain.
- (e) One possible concern with this analysis is that more than one child from each of a number families are included. Explain:
- why this is a potential problem for the analysis?
 - how you could tell if it is a problem?
 - what can be done to fix the analysis if there is a problem?