

---

Please show work! The score you earn on each problem is based on your complete solution, not only on the final answer. You may use your textbook, a calculator, and two sheets of prepared notes.

---

**Problem 1:** (10 points)

Investigators are interested in the effect of marijuana use during pregnancy on birth weight. Low birth weight is associated with many problems that occur later in life. The investigators take a sample of 17 pregnant marijuana users from a volunteers in a local shelter and take a second sample of 26 women, presumed not to be marijuana users, from an area hospital. A 95% confidence interval for the difference in head circumference (associated with body weight) is  $0.90 \pm 0.29$  cm.

- (a) Can it be concluded that using marijuana during pregnancy decreases a baby's head circumference? Explain.
- (b) If cost and ethical considerations were irrelevant and statistical concerns were the only consideration, briefly describe an experiment design that would provide unbiased estimates of the effects of marijuana use during pregnancy.

**Problem 2:** (15 points)

There were 27 players drafted in the first round of the 1991 NBA draft. (Their draft positions ranged from 1 to 27 where 1 is the first person drafted.) Starting salaries ranged from a low of \$180,000 to a high of \$3,333,333. Two players, the 15th and 25th players selected did not sign with the teams that drafted them. For the 25 players who signed contracts, the mean and standard deviation of the starting salaries are \$1,320,000 and \$885,000 respectively. The mean and standard deviation of the draft positions are 13.52 and 7.93 respectively. The correlation coefficient is  $-0.887$ .

- (a) Write down the regression equation for predicting starting salary based on draft position.
- (b) Use the equation to predict what salary the player drafted 15th might have expected.
- (c) Use the equation to predict what salary the player drafted 25th might have expected.
- (d) For each position lower in the draft, by how much does the starting salary decrease?
- (e) What graph would you make to check on the validity of a linear fit to this data?

**Problem 3:** (15 points)

Does living in a rural area decrease total cholesterol levels? Three urban dwellers have total serum cholesterol measurements of 205, 196, and 241. Two rural dwellers have total serum cholesterol measurements of 129 and 175.

- (a) Use the  $t$ -tools to test the null hypothesis that mean cholesterol levels are equal in rural and urban areas versus the alternative that mean cholesterol levels are lower in rural areas than in urban areas. Express the p-value as an area under a  $t$  curve.
- (b) Test the same hypotheses with a permutation test. (Note. You can find the p-value without calculating the test statistic for all cases).
- (c) Test the same hypotheses with a rank-sum test. (Note. You can find the p-value without calculating the test statistic for all cases).
- (d) The data was not randomly sampled from the populations of interest. How does this affect any inferences you may make?
- (e) What confounding variables might affect conclusions from a similar study random samples of much greater size?

**Problem 4:** (10 points)

A biologist wishes to see if brain size in mammals is associated with average litter size. The biologist divides species into a group of 51 for whom the average litter size is less than two and a group of 45 for whom the average litter size is at least two. For each species, a relative brain size is calculated as  $1000 \times \text{Brain Weight} / \text{Body Weight}$ . Summary statistics for the sample data and the log-transformed sample data are shown below.

	n	mean	std. dev.	min	$Q_1$	median	$Q_3$	max
avg. litter < 2								
raw data	51	6.886	5.460	0.42	2.48	5.00	10.48	20.00
log-transformed data	51	1.552	0.952	-0.868	0.908	1.609	2.348	2.996
avg. litter $\geq$ 2								
raw data	45	10.968	9.837	0.94	3.39	7.97	18.61	36.35
log-transformed data	45	1.949	1.016	-0.062	1.221	2.076	2.924	3.593

- Sketch side-by-side boxplots of the raw data and a separate side-by-side boxplot of the log-transformed data.
- Decide if the transformation is appropriate before further analysis. Explain your decision.
- Find a 95% confidence interval for the difference in population means based either the raw data or the transformed data. Summarize your inference in the units of the original problem.

**Problem 5:** (10 points)

A researcher suspects that an antibody (CCK) may differ with gastrointestinal health. In a study with 25 guinea pigs, there are 9 healthy controls, 8 with gall stones, and 8 with ulcers.

Partial output from S-PLUS is below.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
treat		0.4328218			
Residuals		0.5989222			

- Use the information in the table and the problem description to complete an ANOVA table similar to the one on Display 5.19 on page 136 of *The Sleuth*.
- Use the  $F$  tables beginning on page 712 to find a range for the p-value.
- Does it appear that CCK is associated with gastrointestinal health in guinea pigs? Write a brief paragraph that interprets the results of the ANOVA study in the context of the problem.

**Problem 6:** (10 points)

Biologists are interested in determining a relationship between the number of species on an island as a function of the log of area of the island (A), the average elevation of the island (B), and the distance to the nearest island (C). They gather data from 30 islands in the Galapagos Archipelago. The fits from all possible models with only main effects are shown below.

Model	RSS	df
none	381081	29
A	146848	28
B	173547	28
C	381006	28
AB	136309	27
AC	130055	27
BC	173534	27
ABC	126681	26

- Which model is the best according to the  $C_p$  criterion?
- Which model is the best according to BIC?

**Problem 7:** (15 points)

Researchers in Europe assess the effect of the Chernobyl disaster by measuring the amounts of radioactive cesium in plants and soil. They wish to find an equation to predict the cesium amounts in mushrooms based on the soil measurements. Five samples yield the following concentrations (in Bq/kg).

						mean	standard deviation
mushroom	9	20	15	46	190	56.0	76.2
soil	55	415	475	82	1310	467.4	507.8

The regression equation is (mushroom conc.) =  $-6.4115 + 0.1335$  (soil conc.) and the residual sum of squares is 4852.4.

- Sketch a scatter plot of the data with the regression line drawn in.
- Which point do you expect to be most influential? Explain.
- Calculate the leverage of this point.
- Calculate the Cook's distance of the point.
- Comment on the validity of the regression line for predicting mushroom cesium concentration when the soil concentration is in the range from 0 to 500 Bq/kg.

**Problem 8:** (15 points)

Corn yield and rainfall was measured and totaled in six states in the U.S. from 1890 to 1927. The S-PLUS output from a multiple regression to predict corn yield (bu/acre) as a function of rainfall (in/year), year, and many power and interaction terms is shown below.

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	30400.6311	120215.8274	0.2529	0.8021
year	-32.2822	126.4338	-0.2553	0.8003
I(year <sup>2</sup> )	0.0086	0.0332	0.2572	0.7988
rainfall	-4317.1256	10870.4058	-0.3971	0.6942
I(rainfall <sup>2</sup> )	8.1688	17.2981	0.4722	0.6403
I(rainfall <sup>3</sup> )	0.0439	0.0297	1.4783	0.1501
rainfall:year	4.5111	11.4593	0.3937	0.6967
year:I(rainfall <sup>2</sup> )	-0.0052	0.0092	-0.5618	0.5786
I(year <sup>2</sup> ):rainfall	-0.0012	0.0030	-0.3879	0.7009

- If you were to use backwards elimination, which term would you remove first?
- Describe how you would proceed using backwards elimination to arrive at a final model.