Part Two: Practice Questions; Answer Key

Answers in arial font.

1 After struggling for weeks to stay awake in your 1pm Biostatistics class, you and your classmates surmise that perhaps the subject is not so boring after all, but that people tend to fall asleep after lunch as their blood sugar decreases. You decide to test the null hypothesis that eating lunch does not affect a person's ability to stay awake. ($H_0: \mu_1 = \mu_2$, where μ_1 is mean time awake without eating lunch, and μ_2 is mean time awake after eating lunch) For six weeks, you skip lunch before every other class. On days that you do eat before class, you eat a lunch consisting of the same variety of apple and a peanut butter and jelly sandwich made of the same ingredients. During each class, you record the length of time between the beginning of class and when you begin to feel drowsy. After six weeks [$n_1 = 6$ (lunch) and $n_2 = 6$ (no lunch)], you obtain a t-value of 2.4, which exceeds the critical value $t_{0.05(2),10} = 2.228$. You show your analysis to your study partner, but she declares that your test is invalid. She is correct. Explain why.

Your data are pseudoreplicated, because you measured responses of the same person (yourself) multiple times and treated those measurements as independent observations. In this case, $n_1 = n_2 = 1$, so $\nu = n_1 + n_2 - 2 = 0$. Not valid to compare with $t_{0.05(2),10}$.

2 Human use of Federal public lands in the US exceeds 500 million visitor days annually. The US population currently numbers slightly more than 298 million. Ignoring foreign visitors, mean visitation of Federal public lands by US residents is about 1.67 days per capita. Using members of this class as a sample, determine the probability that the mean visitation rate of WWU/Huxley environmental science students is the same as the national mean visitation rate.

Answer using one-sample *t*-test, with $H_0 : \mu = 1.67$. With $n = 20$, critical value: $t_{\alpha(2),n-1}$

3 As discussed on the first day of class, many questions in environmental science can be posed as "Is there a difference between sampled populations?", "Is there an effect of one variable on another?", or "Is there an association between two variables?" Using statistical tests in this course, answers to these questions depend on (1) the magnitude of the difference, effect, or association, and (2) sample variability. For each of the following tests, identify the variable representing each of these quantities. Example: two-sample t-test, magnitude of difference = $\overline{X}_1 - \overline{X}_2$; sample variability = $s_{\overline{X}_1 - \overline{X}_2}$

	Magnitude	Variability
a) Paired-sample t-test	$\overline{d}$	$s_{\overline{d}}$
b) Single factor ANOVA	groups MS	error MS
c) Tukey multiple comparison test	$\overline{X}_A - \overline{X}_B$	SE = $\sqrt{errorMS / n}$
d) Simple linear regression	reg.MS or $b - \beta_0$	residual MS or s_b

4 In 1999 the American Society for the Testing of Materials (ASTM) announced a change in standards for evaluating chemical toxicity. Previously, toxicity was determined using analysis of variance and associated multiple comparisons tests, in which test subjects were exposed to one of several concentrations of the compound of interest, including zero concentration. Results were used to determine concentrations that caused (statistically) significant effects and concentrations that had no detectable effect. The new standard uses regression analysis to determine dose-response relationships and to determine the threshold concentration below which no effects are observed. Which approach provides more statistical power? Why?

Regression provides greater statistical power because it can be used to determine the threshold concentration precisely. Conclusions of analysis of variance are restricted to the exposure levels studied. If the sampling regime does not include the actual threshold exposure, analysis of variance (and associated multiple comparisons test) will identify the nearest (greater) exposure as the threshold. Hence, analysis of variance would miss effects at exposures between the actual threshold and identified (sampled) value. Overlooking these effects is a type II error, and reduces statistical power.

For the following two questions (5 and 6), describe the following.
(a) The appropriate statistical analysis to perform, including number of tails, fixed or random effects, and parametric or nonparametric tests.
(b) State any assumptions necessary in using the appropriate statistical analysis.
(c) State the null hypothesis or hypotheses to be tested.
(d) State the criterion for rejection of the null hypothesis or hypotheses.

5 Research question: Is there an association between the depth of tidepools and number of microhabitats within them? Data on these two characteristics were recorded for 100 randomly selected tidepools.

a) Simple linear correlation, two tailed, parametric
b) Samples drawn from bivariate normal distribution.
X & *Y* sampled at random from normally distributed populations.
c) $\rho = 0$
d) *t*-test: $|t| \geq t_{\alpha(2),\nu}$ where $\nu = n - 2 = 98$ *F*-test: $F \geq F_{\alpha(2),\nu,\nu}$, i.e., $F \geq F_{\alpha(2),98,98}$

6 Research question: Do textbooks required for natural science courses cost more than texts required for social science courses? Prices charged for required texts were recorded from samples of ten courses each selected from upper and lower division courses in the natural sciences and social sciences (20 courses total).

a) Two-sample *t*-test, one-tailed, parametric
b) Normally distributed residuals
c) $\mu_{nat.sci.} \leq \mu_{socialsci.}$
d) If $\alpha = 0.05$,
$t_{calc} \geq t_{\alpha(1),\nu}$
$t_{calc} \geq t_{0.05(1),18} = 1.734$

Alternatively, use ANOVA
(2-factor ANOVA: subject X division)
(could do as single factor w/ 4 samples)

7 Research question: Does DDE in the tissues of birds cause them to lay thin eggs? You decide to address this question using simple linear regression, comparing eggshell thickness with DDE concentration in tissue samples from 100 thrushes. You obtain the following result.

a (intercept) = 30.0 s_a = 3.45
b (slope) = -2.8 s_b = 0.96

Source of variation	DF	SS	MS
Total	99	875.44	
Regression	1	224.31	224.31
Residual	98	651.13	6.644

a) Present a complete statistical analysis to test the hypothesis that there is no relationship between eggshell thickness in thrushes and DDE concentration in their tissues (i.e., that the slope of the regression line equals zero).

Note regression ANOVA table completed above.

t-test: (suppose $\alpha = 0.05$)

$$t = \frac{b - \beta_0}{s_b} = \frac{-2.8 - 0.0}{0.96} = -2.917$$

e.g., for $\alpha = 0.05$, $t_{0.05(1),98} = 1.661$

t_calc > t_crit, $0.0025 > P > 0.001$ => reject H_o

Conclude that DDE causes thinning in (thrush) eggshells, at rate of -2.8 units per unit of DDE.

Note: because *Ho* is one-tailed, must use *t*-test, above. For your review, below is how you could test the analogous two-tailed hypothesis using an *F*-test.

F-test:

$$F = \frac{\text{regression MS}}{\text{residual MS}} = 33.76$$

e.g., for $\alpha = 0.05$, $F_{\alpha(1),1,n-2} = F_{0.05(1),1,98} = 3.95$ (use 90 DF because 98 not in table => 90)

$F_{\text{calc}} > F_{\text{crit}}$, $P << 0.0005$ => reject H_o

b) Write the equation that relates tissue DDE concentration (X) with eggshell thickness (Y).

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

c) What fraction of total variation in initial eggshell thickness is explained by regression results above?

$$r^2 = \frac{\text{regression SS}}{\text{total SS}} = 224.31 / 875.44 = 0.256$$

d) Predict the thickness of shells laid by a thrush whose tissues contain DDE at 4.4 ppm.

$$Y_i = 30.0 - 2.8 X_i$$
$$Y_i = 30.0 - 2.8 * 4.4 = 17.7$$

8 Customer bagging preferences were surveyed at the Community Food Coop and at Sehome Haggen. Bagging preferences ("paper or plastic?") were determined for 100 customers selected randomly at each store, with the following results. Test the null hypothesis that bagging preferences are independent of store. Use a significance level of 0.05.

Bag Preference	Community Food Coop		Sehome Haggen		R_i
	Observed	Expected	Observed	Expected	
paper	30	25	20	25	50
plastic	10	40	70	40	80
use own bag	40	21	2	21	42
no bag	20	14	8	14	28

Expected Frequencies, $\hat{f}_{ij} = \dfrac{(R_i)(C_j)}{n}$ C1 = 100, C2 = 100, $n = 200$

DF = $(r - 1)(c - 1) = 3$

$$\chi^2 = \sum\sum \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$$

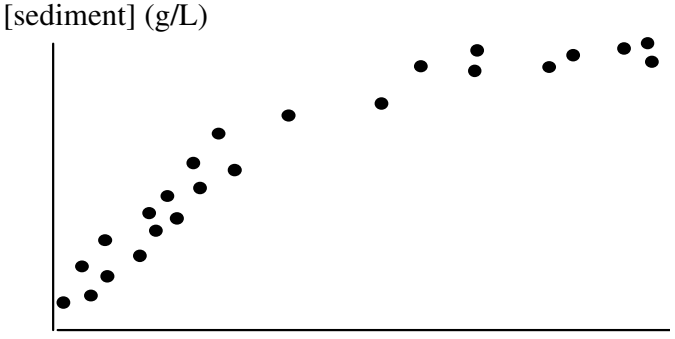
$\chi^2_{calc} = 86.52$ $\chi^2_{0.05,3} = 7.815$

$P < 0.001$

Reject H_0

Conclude that customers at Coop and Haggen differ in bagging preferences.
(bagging preferences are not independent of store)

9 You are studying the effect of rainfall on sediment load in high order streams in the Mt. Baker-Snoqualmie National Forest. Before conducting a regression analysis, you plot your data and obtain the following scatterplot.



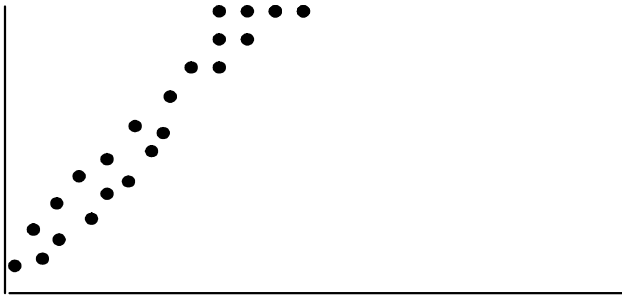
a) Which assumption of simple linear regression analysis do these data violate? How can you tell?

Linear relationship between *X* and *Y*. Can tell from curvilinear shape or plot of residuals after trying simple linear regression (negative residuals at low and high values of *X*, positive residuals at intermediate *X*).

b) How could you perform regression analysis using these data, but without violating the assumption identified in part (a)? For full credit, be explicit in your description (i.e., include any equations you would use), and draw a new graph showing data used in your regression.

Transform predictor (independent) variable (ppt), using square root transformation $X' = \sqrt{X + ½}$, or logarithmic transformation $X' = \log(X + 1)$.

[sediment] (g/L)



Log(ppt)

10 Research question: how is elk migration down to private land in autumn affected by date, start of hunting season, and number of hunters? The Colorado Division of Wildlife conducted an experiment in which hunting season dates and numbers of hunters allowed in experimental areas were manipulated. Elk migration in those areas was measured, and a set of models were fit to data on elk movement, with the following results.

Model No.	Model variables	*K*	AIC_c	ΔAIC_c	*w*
1	date, area, # hunters, huntseason, 2-way interactions	11	5835.37	0	0.441
2	date, area, # hunters, huntseason, 2- and 3-way interactions	15	5835.52	0.15	0.409
3	date, area, # hunters, huntseason, 2-, 3-, and 4-way interactions	16	5837.53	2.16	0.150
4	date, year, area, huntseason, 2-way interactions	19	5847.17	11.80	0.001
5	date, year, area, huntseason, 2- and 3-way interactions	29	5863.74	28.37	0.000
6	date, year, area, huntseason, 2-, 3, and 4-way interactions	32	5869.81	34.44	0.000

a) Complete the table above, by filling in values for ΔAIC_c and *w*.

b) Which model performs best according to Information criteria?

Model 1, because it has lowest AIC_c score.

c) What is the probability that the model identified in (b) really is the best among the models considered?

$W_1 = 0.441$ or 44.1 % probability

d) What is the confidence set for the best model, among the models considered?

Method 1, 95%: Confidence set: [1, 2, 3] Method 2, for $\Delta \leq 2$: Confidence set: [1, 2]
Method 3, with $C = 20$: Confidence set: [1, 2, 3].

11 Is the allocation of Greenways money among geographic regions of Bellingham unfair relative to the number of people living in those regions? Treat the following information (Greenways expenditures from 1990 and 1997 levies; population data from 2000 census) as sample data to address the research question.

Use Chi-squared goodness of fit test.

Region	% population	% Greenways $$	$(f_i - f\hat{})^2/f\hat{}$
Lower Wh. Cr. – Boulevard Pk. Corridor	9	3	4.000
NW Bellingham & Urban Fringe	26	27	0.036
SW Bellingham	30	41	4.033
SE Bellingham & Urban Fringe	10	7	0.900
NE Urban Fringe	2	3	0.500
East & NE Bellingham	23	19	0.696

$\chi^2_{calc} = 10.17$ $\chi^2_{0.05,5} = 11.07$ $0.10 > P > 0.05$

In what form would you need this information to address the question more appropriately?

Greenways expenditures as frequencies (actual $$ spent)

12 Do benzene and dioxin impact survival of fathead minnows to the same degree?
Survival rates of 50 randomly selected fathead minnows were recorded after exposure to a range of benzene concentrations, and survival rates of 50 other randomly selected fathead minnows were recorded after exposure to a range of dioxin concentrations. Simple linear regression analysis produced the following results (some values omitted intentionally). Use these results to answer the research question.

Benzene Exposure (fraction alive / ppm)

a (intercept) = 0.98 s_a = 0.02
b (slope) = − 0.12 s_b = 0.033

Source of variation	DF	SS	MS
Total		3.7	
Regression		2.5	
Residual		1.2	
Benzene $(X_i - \overline{X})$		22.6	

Dioxin Exposure (fraction alive / ppm)

a (intercept) = 0.98 s_a = 0.02
b (slope) = − 0.22 s_b = 0.036

Source of variation	DF	SS	MS
Total		4.1	
Regression		2.7	
Residual		1.4	
Dioxin $(X_i - \overline{X})$		22.6	

Null hypothesis: H_0: $\beta_1 = \beta_2$

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}} \qquad s_{b_1 - b_2} = \sqrt{\frac{(s^2_{Y \cdot X})_p}{\left(\sum x^2\right)_1} + \frac{(s^2_{Y \cdot X})_p}{\left(\sum x^2\right)_2}}$$

$$(s^2_{Y \cdot X})_p = \frac{(\text{residual SS})_1 + (\text{residual SS})_2}{(\text{residual DF})_1 + (\text{residual DF})_2}$$

$$t = \frac{-0.12 - (-0.22)}{0.049} = 2.04$$

$t_{0.05(2),96} = 1.985$ $0.05 > P > 0.02$

If $\alpha = 0.05$, reject H_0
Conclude that dioxin decreases survival more than does benzene.