Part One: Review and Advice

Identifying the Appropriate Statistical Analysis

The following questions may help you select an appropriate test. They are listed in arbitrary order.

1 How many samples and/or treatments do you have?
2 Do measurements within a sample represent true replication, or are they sub-samples?
3 Judging from sample frequency distributions, are the populations likely to be normally distributed? More appropriately, are the residuals (ε in a linear model) normally distributed? (parametric vs. non-parametric test)
4 Are the variances in your samples similar, or are they very different? (parametric vs. non-parametric test, or transformation prior to analysis)
5 Do you need to compare means, variances, or proportions?
6 If there are two samples, are their measurements paired? Is each measurement in one sample clearly related to one and only one measurement in the other sample? (paired vs. two-sample test)
7 Are you comparing means of three or more samples? (ANOVA) If so, how many factors influence each sample? (one: single-factor ANOVA; two: two-way ANOVA; three or more: take a course in experimental design)
8 Are factors in an ANOVA fixed or random?
9 Do you need to determine a quantitative relationship between two or more sampled variables?
10 If the answer to 9 is yes, are you interested in testing for a cause-effect relationship between the variables or simply an association? (regression or correlation)
11 Are all the hypotheses plausible? Avoid or revise null hypotheses that can be rejected a priori.

ANOVA Review

Given a description of a sampling design, you should be able to answer the following questions.

1 Is ANOVA the appropriate kind of analysis to test the hypothesis(es)? Why?
2 If ANOVA is appropriate, is (are) the factor(s) fixed or random?
3 Can you construct an ANOVA table, including sources of variation, SS, DF, and MS?
4 Can you determine the correct ratio of MS to test a particular hypothesis?
5 If you are given a partial ANOVA table and a study description, can you complete the table to test ANOVA hypotheses?
6 If you do a multiple comparisons test after concluding from ANOVA that the groups differ, in which order should you compare the various pairs of groups?
7 How would you interpret ambiguous results in a multiple comparisons test? (e.g., test concludes that groups 1 and 3 differ from each other, but that cannot conclude that group 2 differs from either group 1 or group 3)
8 After concluding from a multiple comparisons test that population means differ, can you calculate the 95% confidence interval for each mean? (and 90% CI, 99% CI, etc)

Goodness of Fit and Independence Review

Given a description of a sampling design you should be able to answer the following questions.

1 Chi-square tests are used to analyze both goodness of fit and independence of factors. What is the difference?
2 Given hypothesized proportions for a set of categories, and frequencies of sample data observed in each category, can you calculate the χ^2 statistic?
3 How would you test H_0: the sample was from a population with a set of proportions, p_1, p_2, p_3, … ?
4 If a sample is drawn from a population affected by two factors, how would you determine the frequencies of sample data that would be expected if the factors acted independently?
5 How would you test H_0: the factors are independent?

Regression and Correlation Review

Given a sampling design for two variables, you should be able to answer the following questions.

1 Should correlation or regression analysis be used? Why?
2 In regression analysis, what is meant by the hypothesis H_0: $\beta = 0$?
3 Given a regression ANOVA table with sources of variation and SS, can you test the hypothesis H_0: $\beta = 0$?
4 Given the above ANOVA table, can you calculate r^2? What does this quantity mean?
5 How would you test the hypothesis H_0: $\beta = 0$ using a t-test?
6 When should you evaluate β with a one-tailed test?
7 Given calculated values of a and b (which are estimates of α and β), what is the equation of the line that relates Y to X?
8 When should you transform data prior to regression analysis? How can you determine if a transformation was effective?
9 Given calculated values for $\sum x^2$, $\sum y^2$, and $\sum xy$, can you calculate the correlation coefficient? What does this coefficient mean?
10 How would you test the hypothesis $\rho = 0$?
11 What information do you need to compare the slopes of two regression lines?

Model Selection

Given a research question and multiple scientific hypotheses that potentially answer the question, you should be able to complete the following steps or answer the following questions.

1 Is each hypothesis plausible? Why/not?
2 Translate each scientific hypothesis into a model for the response variable (data).
3 Is the number of models reasonable? (More than two, small enough to avoid spurious results)
4 Does each model represent its hypothesis well?
5 Can you fit each model to the data, and determine the residual error of the fit? Consider both least squares and likelihood methods.
6 Can you calculate AIC scores for each model?
7 Do your AIC scores include the correct value for K, the number of parameters estimated?
8 Should you use AIC or AIC_c to compare models?
9 Given a set of AIC or AIC_c scores, can you calculate an Akaike weight for each model?
10 Which model performs best according to Information criteria?
11 What is the confidence set for the best model, among the models considered?

Multi-model Inference

Given results of model selection mentioned above, you should be able to complete the following steps or answer the following questions.

1 Given a set of models that estimate the same parameter, use the entire set to generate a weighted estimate of that parameter. How much does the weighted estimate differ from the estimate generated by the single best model?
2 Given a set of models that predict data on a response variable using subsets of predictor variables, determine the order of importance of the independent variables.

Part Two: Practice Questions

1 After struggling for weeks to stay awake in your 1pm Biostatistics class, you and your classmates surmise perhaps the subject is not so boring after all, but 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.

2 Human use of Federal public lands in the US exceeds 500 million visitor days annually. The US population currently numbers about 300 million. Ignoring foreign visitors, mean visitation of Federal public lands by US residents is about 1.67 days per capita. Using this class as a sample, determine the probability that the mean visitation rate of WWU students is the same as the national mean.

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(s) representing each of these quantities. Example: two-sample t-test, magnitude of difference = $\overline{X}_1 - \overline{X}_2$; sample variability = $s_{\overline{X}_1 - \overline{X}_2}$

a) Paired-sample t-test
b) Single factor ANOVA
c) Tukey multiple comparison test
d) Simple linear regression

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?

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 criteria 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.

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 (40 courses total).

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		224.31	
Residual		651.13	

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).

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

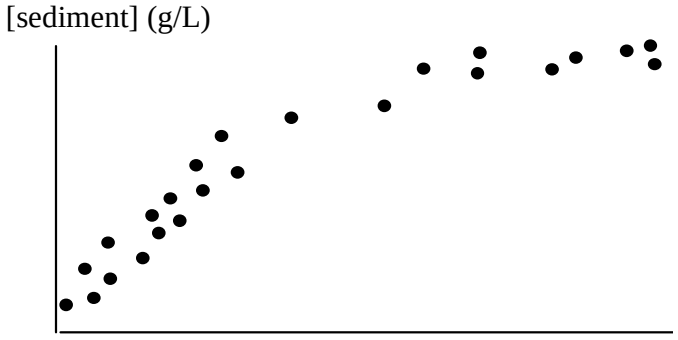
c) What fraction of total variation in initial eggshell thickness is explained by regression results above?

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

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
paper	30	20
plastic	10	70
use own bag	40	2
no bag	20	8

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?

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.

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, hunt season, 2-way interactions	11	5835.37		
2	date, area, # hunters, hunt season, 2- and 3-way interactions	15	5835.52		
3	date, area, # hunters, hunt season, 2-, 3-, and 4-way interactions	16	5837.53		
4	date, year, area, hunt season, 2-way interactions	19	5847.17		
5	date, year, area, hunt season, 2- and 3-way interactions	29	5863.74		
6	date, year, area, hunt season, 2-, 3, and 4-way interactions	32	5869.81		

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

b) Which model performs best according to Information criteria?

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

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

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.

Region	% population	% Greenways $$
Lower Wh. Cr. – Boulevard Pk. Corridor	9	3
NW Bellingham & Urban Fringe	26	27
SW Bellingham	30	41
SE Bellingham & Urban Fringe	10	7
NE Urban Fringe	2	3
East & NE Bellingham	23	19

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

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	