Compilation of Study Questions from *StatPrimer* Units 11 - 18
12/13/01

Answers are linked to the Exercises pages.

(11.1) STUDY QUESTIONS
(A) What is the key distinction between an observational study and an experimental study?
(B) What does "double-blinded, randomized, controlled" mean?
(C) If statistics is *more* than just a compilation of computations methods, what then *is* it?
(D) Discuss problems associated with using "self controls."
(E) Discuss the difference intention-to-treat analysis and the more conventional method which considers only people who comply and complete treatments (so-called "efficacy analysis").
(F) Complete this sentence: According to Dallal, the 4 Basics of study design are . . .
(G) What is the difference between a longitudinal sample and cross-sectional sample?
(H) What is the main benefit of randomization?
(I) Write your own study question (not research question) here and submit it to me!

(12.1) STUDY QUESTIONS
(A) Differentiate between measurement error and processing error.
(B) List four different types of data entry errors.
(C) List four different methods that can be used to detect data entry errors.
(D) What is the function of an *EpiData* QES file?
(E) What extension is used to identify *EpiData* data files?
(F) List specific types of information you would include in a code book.
(G) Why should backup copies of data files be kept off site?
(H) What is the function of a code book?

(13.1) STUDY QUESTIONS
(A) Define "risk."
(B) The ratio of two incidence proportions is called a_____ _____.
(C) What symbol is used to represent the relative risk *estimator*?
(D) What symbol is used to represent the relative risk *parameter*?
(E) The incidence of hypertension in an exposed group is 10%. The incidence of hypertension in the unexposed group is 5%. What is the relative risk of hypertension associated with the exposure?
(F) The hypothetical sampling distributions of relative risks tend to be _____ normal.
(G) The uncorrected chi-square statistic introduced in *StatPrimer* unit 10 is attributed to Pearson ("Pearson chi-square statistic"). A different statistician adapted chi-square methods to account for fitting the continuous chi-square distribution to discrete data. What is the name of this statistician?
(H) Chi-square test statistics should <u>not</u> be used when an expected frequency is _____.
(I) What test is used to when an expected frequency in a 2-by-w table is less than 5?
(J) A relative risk is 9.0. What is the effect of the exposure in relative terms?
(K) A 95% CI for a *RR* is (1.1, 2.7). Is the association between the exposure and disease significant at $\alpha = .05$? Explain.

(14.1) STUDY QUESTIONS
(A) In what primary way does a case-control study differ from a cohort study?
(B) Fill in the blanks: Because subjects in case-control studies are selected based on their disease status, we can no longer estimate _____ directly. However, the _____ associated with an exposure can still be estimated through an odds ratio.
(C) What symbol is used to denote the odds ratio parameter? What symbol denotes the odds ratio estimator?
(D) The method used to calculate a confidence interval for an odds ratio that is presented in this chapter first has us

convert the point estimate for the odds ratio to a _____ scale.
(E) List the null hypothesis and alternative hypothesis addressed in this chapter.
(F) When is Fisher's test used in place of a chi-square test?
(G) In the 2-by-2 table used to summarize matched-pair case-control data, table cells *t* and *w* contain counts of _____ pairs, while cells *u* and *v* contain counts of _____ pairs.
(H) True or false? In matched case-control studies, information about concordant pairs is largely ignored.
(I) What is the name of the chi-square statistic used to test matched-pair data?

(15.1) STUDY QUESTIONS
(A) List a synonym for "variance."
(B) What symbol is used to denote the population variance? What symbol is used to denote the sample variance?
(C) Provide a synonym for "standard deviation."
(D) *When* will 95% of the data point be within *approx.* 2 standard deviations of the mean?
(E) What is the name of the rule that states "at least 75% of the values lie within 2 standard deviations of the mean."
(F) Name a good “distribution-free” ("nonparametric") measure of "spread."
(G) Describe a way in which the variability of a distribution can be shown graphically.
(H) The sum of squares is the sum square of deviation around the distribution’s _____.
(I) Express the sum of squares in terms of the sample variance.
(J) Name the primary test used to test for the inequality of two population variances (i.e., H_0: $\sigma^2_1 = \sigma^2_2$)?
(K) If you had a significant *F* ratio test, you would not want to calculate a pooled estimate of variance. Why?
(L) What does the standard error of the independent mean difference quantify?
(M) In pooling variances from two groups in which $n_1 = 11$ and $n_2 = 10$, df_1 = ____, df_2 = ____, and df = ____.

(16.1) REVIEW QUESTIONS
(A) What does *homoscedastic* mean?
(B) Why is analyzed in analysis of variance?
(C) What symbol is used to denote the mean of population *i*? What symbol is used to denote the estimator of the mean of population *i*?
(D) Provide a synonym for the variance between groups. Provide a synonym for the variance within groups.
(E) Why does one downwardly adjust the α level according to Bonferroni's method when doing *post-hoc* comparisons?
(F) Data set 1 has the following values: {90, 70, 50, 30, 10}. Data set 2 has the values {70, 60, 50, 40, 30}. Here is a side-by-side box-plot showing two groups:

```
Group 1  Group 2
      0|9|
       |8|
      0|7|0
       |6|0
      0|5|0
       |4|0
      0|3|0
       |2|
      0|1|
      (x10)
```

Compare the means and variances of the two groups.
(G) List the statistical assumptions behind ANOVA?
(H) *StatPrimer* suggests that statistical assumptions need not be met perfectly when conducting statistical tests. What are the bases of this recommendation?
(I) Write the *alternative* hypothesis tested by ANOVA in two different ways.
(J) We want to compare blood glucose levels (mmol/L) in male and female taxi drivers. What is the dependent variable in this analysis? What is the independent variable?

(K) The mean for all *N* subjects in an ANOVA combined is called the _____ mean.
(L) When would you consider using the Kruskal-Wallis test instead of ANOVA?
(M) State the null hypothesis tested by Levene's test.
(N) An ANOVA is preformed comparing 4 groups with 8 people per group. df_B = ? df_W = ?
(O) A positive Levene's test gives you pretty clear evidence that the _____ assumption has been violated.

(17.1) STUDY QUESTIONS
(A) Correlation coefficients quantify the direction and strength of an association between two continuous variables. How do you determine the direction of the correlation? How do you determine its strength?
(B) What symbol is used to denote the correlation coefficient in the sample?
(C) What symbol is used to denote the correlation coefficient in the population?
(D) A *t* statistic can be used to test a correlation coefficient for significance. This statistic is associated with ___ degrees of freedom.
(E) What is meant by "bivariate normality"?
(F) What is the range of values possible for *r*?
(G) The correlation between two variables in +.79. Interpret this correlation coefficient.

(18.1) STUDY QUESTIONS
(A) What is meant by a "functional dependency" between X and Y?
(B) How does an algebraic linear model differ from a statistical linear model?
(C) Linear relationships are characterized by a slope and intercept. What the slope of the model represent? What does the intercept of the model represent?
(D) What does a slope of 0 (zero) indicate?
(E) What is squared in a "least squares" line?
(F) Suppose the relationship between age in years (X) and height in inches (Y) in adolescents is modeled as "yhat" = 46 + 1.5x. Interpret the slope of this model.
(G) Using the model described in part F of this question, what is the expected height of a 10 year old?
(H) What is the value of the *t* percentile used to calculate a 95% confidence interval for a slope based on a sample of *n* = 25?
(I) What symbol is used to denote the slope in the sample? What symbol is used to denote the slope in the population?
(J) Under the null hypothesis of no linear relationship, the value of β is _____.
(K) A *t* statistic used to test a slope is associated with ___ degrees of freedom.
(L) Negative slopes suggest that as X increases, Y tends to _____.
(M) The normality and equal variance assumptions refer to the _____ of the model.