Stat 135, Fall 2006 **A. Adhikari**

Summary of material in the second half of the term

ONE SAMPLE

CATEGORICAL DATA. The case of two categories was extensively covered before the midterm. So think in terms of more than two categories.

Questions: Does the sample come from a particular multinomial model? If the data have been cross-classified according to two variables, are the variables independent?

Answers: The first question can be answered by the χ^2 test of goodness of fit. If the model is not completely specified, you have to first estimate its parameters using maximum likelihood; 8.2, 8.5, 9.5. The second question is answered by the χ^2 test of independence; 13.4.

Question about estimation: Can you construct confidence intervals for the proportions in the cells? For the difference between the proportions in two cells?

Answer: Yes, specially if you remember how to use covariance; HW7.7.

Question, if the data come in matched pairs: If each pair is either 00, 01, 10, or 11, is the underlying probability of 1 the same for treatment and control?

Answer: McNemar's test, 13.5.

QUANTITATIVE DATA: ONE VARIABLE Parametric methods of inference for the population mean (and SD, in the case of a normal population) were extensively covered before the midterm.

Nonparametric method – questions: What is the median of the underlying population? Is the distribution symmetric about a specified point?

Answers: The first question can be answered by the sign test; Ex9.24. Its typical use is if the data come in matched pairs, to decide whether or not there is a 50% chance that the control value is greater than the treatment. The second can be answered by the Wilcoxon signed rank test; 11.3.2. The typical use is with paired comparisons, to decide whether or not the pre-treatment distribution is the same as the post-treatment distribution.

QUANTITATIVE DATA: MORE THAN ONE VARIABLE

The matched pairs situation is when a single sample provides observations on two variables **which can be measured in the same units**, and you are interested in comparing the distributions of the two variables. We have gone over this extensively, e.g. reducing the problem to a single sample of differences (see above).

Frequently we are interested in the relation between two different variables x and y, e.g. weight and height.

Normal theory – main question: Is y a linear function of the x's, apart from random noise?

Answer when there is one predictor: Simple linear regression; 14.2. Specifically:

The regression line as a least-squares line; lecture.

Properties of the regression estimates and residuals, including sum of squares decomposition and relation to r; HW9, 14.2.3, Ex14.10-12.

Inference for the parameters; lecture handout, 14.2.1, Ex14.13-14.

Assessing regressions, residual plots; HW10, 14.2.2.

Answer when there is more than one predictor: Multiple linear regression; 14.3-5, lecture, HW10. Specific issues: the F-test for the fit (can also be used for simple regression), the definition of r^2, collinearity and variable selection, how to plot the residuals, polynomial regression, using indicator variables to see differences between subgroups.

TWO INDEPENDENT SAMPLES

CATEGORICAL DATA

The case of two 0-1 samples was extensively covered before the midterm; with large random samples you can construct normal confidence intervals for the difference between the two population proportions of 1's. You can use a z-test for the equality of the proportions; this involves using the pooled estimate of the common p to estimate the SE for the difference, under the null hypothesis.

Question: Do the two samples come from the same underlying multinomial distribution?

Answer: χ^2 test for equality of distributions (also known as homogeneity); 13.3. This is identical to the χ^2 test for independence, which can be thought of as a test for the equality of conditional distributions.

QUANTITATIVE DATA

Parametric methods for comparing the two underlying population means were extensively covered before the midterm.

Nonparametric method – question: Do the two samples come from the same underlying distribution?

Answer: Wilcoxon rank-sum (Mann-Whitney) test; 11.2.3.

THREE OR MORE INDEPENDENT SAMPLES

CATEGORICAL DATA

Question: Do all the samples come from the same underlying multinomial distribution?

Answer: χ^2 test of homogeneity; 13.3. Notice that this can be used to decide whether three or more binomial samples all have the same underlying p.

QUANTITATIVE DATA

Normal theory – question: Assuming that the samples come from normal distributions with the same variance, do they also all have the same mean?

Answer: One-way ANOVA; 12.2. Model, sum of squares decompostion, F-test; 12.2.1.

Nonparametric method – question: Do all the samples come from the same underlying distribution?

Answer: Kruskal-Wallis test; 12.2.3.

GENERAL REMINDER. Keep in mind these easy basics.

1. If you are using the same data to construct more than one confidence interval or to perform more than one test, you can use the Bonferroni method to control the "simultaneous error" probability; 11.4.8, 12.2.2.2.

2. Suppose you have constructed a confidence interval for a parameter τ, and suppose g is a continuous monotone function. You can construct a confidence interval of the same level for $g(\tau)$ by simply applying g to each end of the confidence interval for τ. Typical examples are $g(x) = \sqrt{x}$ and $g(x) = 1/x$.

BAYESIAN METHODS

Now the parameter is being thought of as a random variable. Your opinion about the uncertainty in the random parameter is quantified by your **prior** probability distribution for the parameter, before data are gathered. The main question is: how do the data change your opinion about the parameter? Bayes' Rule implies that your **posterior** distribution of the parameter, given the data, is proportional to the product of your prior and the likelihood. In some nice situations, you can find a family of **conjugate priors**, which give rise to posterior distributions in the same family.