

## Stat 111 Spring 2011 Week 6 Problem

### 1. Two sample $t$ tests.

- a) When comparing Normal samples from two populations with the same standard deviation, the pooled 2-sample  $t$  statistic satisfies the definition of a  $t$  random variable under a null hypothesis about the difference in means. This implies that  $t^2$  is an  $F$  random variable (under the null hypothesis). Demonstrate the two sample  $t$  test and CI for comparing the temperatures of samples of men and women.
- b) It is a strong assumption to say the variances for two (or more) samples are equal. In the 2-sample problem we can estimate the variances separately to construct an unpooled  $t$  statistic. Show that the unpooled 2-sample  $t$  statistic does not satisfy the definition of a  $t$  random variable, even when the data are exactly Normal.

It turns out the sampling distribution is very close to  $t$ , but with a smaller degrees of freedom than for the pooled test. There is a complicated formula for finding the best degrees of freedom value, but a conservative (i.e., smaller than needed) degrees of freedom value is the smaller of  $n_0 - 1$  and  $n_1 - 1$ . Justify this by imagining a matched-pairs  $t$  test for randomly paired individuals from the two independent samples.

Unfortunately there is not such a simple fix for the  $k$  sample problem. Standard ANOVA (i.e. multiple regression) software assumes equal sd's, so it is up to the user to check that assumption. Sometimes a transformation of the data (e.g., logarithms) can help balance the variability in different samples.

- c) The generalization of the pooled 2-sample  $t$  test is the One-way ANOVA  $F$  test. Suppose you have  $k$  samples of sizes  $n_j$ , with averages  $\bar{X}_j$  and standard deviations  $s_j$ . Call the overall average  $\bar{X}$  and the total sample size  $N$ . Then the  $F$  statistic for testing a null hypothesis of equal means is

$$F = \frac{\text{MSM}}{\text{MSE}}; \quad \text{MSM} = \sum_{j=1}^k \frac{n_j(\bar{X}_j - \bar{X})^2}{k-1}; \quad \text{MSE} = \sum_{j=1}^k \frac{(n_j - 1)s_j^2}{N - k}.$$

In the case where all  $n_j = n$ , show that the  $F$  statistic satisfies the definition of an  $F$  random variable when  $H_o$  is true. Also show that  $F = t^2$  when  $k = 2$ .

### 2. Log transform and Non-parametric tests.

Sometimes a  $t$  test is more appropriate for some transformation of the data. In other situations, we may decide not to assume a distribution at all.

- a) Download the file grant.txt from my Stat 111 web site. These are data for a random sample of NSF grants taken by a Swarthmore student in 1998. The first column is the dollar value of the grant, and the second column indicates the gender of the principal investigator on the proposal (0=female; 1 = male). Here are commands to read the data into R. Note that you will have to specify the path to your directory where the data file was saved.

```
x = scan("/Users/...specify path to directory.../grants.txt")
xmat = matrix(x, ncol=2, byrow=T)
y = xmat[,1]; male = xmat[,2]; N = length(y)
y0 = y[male==0]; y1 = y[male==1]; n0 = length(y0); n1 = length(y1)
```

This results in vectors  $y_0$  and  $y_1$  of lengths  $n_0$  and  $n_1$ , which report the grant values for females and males, respectively. You can use the R functions `mean`, `var` and `sd` to get summaries, and `hist` to display a histogram of the data. Carry out an unpooled 2-sample  $t$ -test to determine whether the mean grant amounts differ with the gender of the principal investigator. Comment on the appropriateness of the  $t$  procedures for these data.

- b) One way to assess the appropriateness of a  $t$  test is to “bootstrap” new data sets based on the original, and look at the distribution of the differences of bootstrapped sample averages. Use the following commands to simulate 1000 bootstrapped samples for the raw dollar values (assuming no difference for males and females):

```
simdiff = simse = simt =rep(0,1000)
for(iter in 1:1000){
  newy = sample(y, N, replace=T)
  newy0 = newy[male==0]; newy1 = newy[male==1]
  simdiff[iter] = mean(newy1)-mean(newy0)
  simse[iter] = sqrt(var(newy0)/n0 + var(newy1)/n1)
  simt[iter] = simdiff[iter]/simse[iter]}
```

Explain how this simulation approximates the null sampling distribution for the difference in averages, the standard error for the difference, and for the  $t$  statistic for this population. Make a histogram and normal quantile plot of the simulated differences. Note the distribution is close to Normal, despite the strong skew in the distribution of the individual values (the CLT works!). Plot the standard errors against the differences and show that the independence assumption appears to be violated. Regardless, the distribution of the  $t$  statistics appears close to Normal (or  $t$  with large degrees of freedom).

- c) Often taking logs of dollar values results in a distribution closer to Normal. Use the `log` function in R to create vectors  $x$ ,  $x_0$  and  $x_1$  containing the log dollar values. Repeat the procedures from part b for the log values, and show that the  $t$  assumptions are even closer to being met.
- d) Compare the interpretations of the  $t$  confidence intervals for the raw data and the log data. Explain why the  $p$ -values for the tests are so different, given that  $t$  procedure appears to be valid for the raw data. Hint: they test different hypotheses.
- e) There are many *non-parametric* tests that can be used for non-Normal data. A simple test is to replace the data values with their overall ranks (lumping together the two samples before ranking). Typing  $\mathbf{r} = \mathbf{rank}(\mathbf{y})$  yields a vector  $r$  with a 1 for the smallest  $y$  value, a 2 for the next smallest, etc. If there are repeated values, all receive the same average rank value. The ranks are reasonably well-behaved, so with larger samples a  $t$  test on the ranks is often appropriate. Carry out a 2-sample  $t$  test on the ranks and

compare your  $p$ -value to the test for the raw dollar values and for the logged dollar values.

- f) An alternative  $p$ -value for the test in c may be obtained via simulation. A **permutation test** estimates the probability of getting a  $t$ -statistic on the ranks as far from 0 as yours (under  $H_o$ ) by repeatedly permuting the ranks and recomputing the  $t$  statistic. The estimated  $p$ -value is then the proportion of simulated  $t$  statistics that are as extreme or more extreme than the observed value (computed for the actual ranks). In R, type **newr = sample(r)** to get a vector containing all of the same ranks, but in a scrambled order. For each iteration, generate a newr, break it into newr0 and newr1 (for females and males) and compute and save a  $t$  statistic. Generate 1000 new  $t$  values and estimate the  $p$ -value. Compare to the  $p$ -value in part d.

### 3. Hypergeometric, Binomial and Chi-square tests.

- a) I gave an experimental survey in Stat 1 asking students to agree or disagree with a statement that there is sometimes a need either for the use of “enhanced interrogation techniques” or for the use of “torture.” Surveys were distributed at random to 26 students, with 13 students receiving each version of the statement. Of the students asked about EIT, 6 of 13 agreed. Of those asked about “torture”, only 1 of 13 agreed. Is this good evidence that the wording of the question elicited systematically different responses? What is another explanation for the difference? Use the Fisher Exact test (based on the hypergeometric distribution) to quantify the evidence for a wording effect.
- b) For larger samples, a Normal approximation to the Binomial distribution may be used to test one or two probabilities. The previous presentation analyzed data for a random sample of  $n = 629$ , chosen from all NSF grants funded in 1997. Of these, 161 had female principal investigators. Carry out a test of whether NSF grants are equally likely to have a male or female PI (note that this is not the same as testing whether male or female PI’s are equally likely to be funded). Construct a 95% CI for the proportion of male PI’s.
- c) Of the sample grants with male PI’s, 383 were for at least \$100,000 dollars. Of the grants with female PI’s, 121 were for at least \$100,000 dollars. Test for equal proportions of such grants for male and female PI’s and find a CI for the difference in proportions.
- d) Using a Normal approximation for a 2-sample Binomial test is equivalent to the Chi-square test of independence. Describe the Chi-square statistic, and show that for the data in part c it evaluates to the square of your  $z$  statistic.
- e) There were 238 grants with male PI’s funded at over \$200,000, and 65 grants that large with female PI’s. Construct a 2x3 table of counts for grants with male and female PI’s in the three different dollar ranges. Carry out a Chi-square test on these data and explain your conclusions.

### Problems to turn in:

1. A woman has probability  $p$  of being pregnant and takes a pregnancy test that will be positive with probability 0.9 if she is pregnant and with probability 0.05 if she is not pregnant.
  - a) Let  $\theta = 1$  if the woman is pregnant and  $\theta = 0$  if not. If the test is positive, what is the likelihood function for  $\theta$ ? What is the  $p$ -value for a test of  $H_o : \theta = 0$  vs.  $H_a : \theta = 1$ ?
  - b) Suppose the woman thinks her probability of being pregnant is  $p = 0.15$ . What is the conditional probability that she is not pregnant, given a positive test result?
  - c) Explain why the conditional probability is more informative than the  $p$ -value in part a.
  - d) What value of  $p$  would make a positive test indecisive? That is, find  $p$  to make the posterior odds of no pregnancy to pregnancy, given a positive test, equal to 1.0. Also find the value of  $p$  that would make a negative test result indecisive.
  
2. I introduce hypothesis testing in Stat 11 and Stat 61 by passing around a box with a 6-sided die and having each student roll the die and keep track of the total of the rolls. What they don't notice is that I use a die with two 1's and no 6's. At the end, we use the Normal approximation to find the probability of getting an average roll as far from 3.5 as ours, and (usually) find that our average is improbably low.
  - a) With  $n = 20$  rolls, find the rejection region of a test of  $H_o : \mu = 3.5$  vs.  $H_a : \mu \neq 3.5$  for  $\alpha = 0.1$ . Use the Normal approximation.
  - b) What is the probability that we will reject  $H_o$  in the correct direction? That is, find the *power* of the test to detect the bias of this die at the  $\alpha = 0.1$  significance level (we don't consider it a success if you accidentally reject in the wrong direction).
  - c) How many rolls would we need to have power of at least 0.9 with  $\alpha = 0.01$ ?
  
3. Matched Pairs. In a *matched pairs t* test, two paired samples are collected. It would be inappropriate to treat these as two independent samples, but we can think of the differences for each pair as independent. One study of the effect of Friday the 13th recorded the numbers of hospital admissions for "accidents" at a British hospital on  $n = 6$  Friday the 13th and Friday the 6th pairs for the same month. The values for the 13th are 13, 12, 14, 10, 4, 12, with corresponding values for the 6th of 9, 6, 11, 11, 3, 5.
  - a) Estimate the standard deviation of the difference in averages and compare this to the value you would get treating these as two independent samples. Explain how the positive correlation between the samples leads to a smaller standard deviation for the differences.
  - b) Assuming the differences are a sample from a Normal distribution, carry out a test of no difference in the mean number of accidents on these two dates against a 1-sided alternative implying typically more accidents on the 13th. Report your  $p$ -value.
  - c) An alternative to the matched-pairs  $t$  test is the sign test. The sign test looks only at the sign (positive or negative) of each difference. There are various ways of dealing with ties (differences of 0) but these data don't have that problem. Define a parameter, state hypotheses and report a  $p$ -value for the sign test for these data.
  - e) Explain why the sign test may be preferable to the  $t$  test in some situations, despite its lower power.