Statistics 104 — Fall, 2004 — Assignment 7
Due Wednesday, November 24th, 2004.

Readings (Moore and McCabe)

- Sections 7.1 and 7.2 and Chapter 8.

Against All Odds videotape

The relevant tapes for this week are numbers 21 (Inference for one mean), 22 (Comparing two means), and 23 (Inference for proportions).

Written Assignment (Moore and McCabe)

- **MM:** 6.70, 6.76, 6.80, 6.86, 6.90, 7.20, 7.40, 7.68 (plus (d) Calculate a 95% confidence interval for the difference in the mean ratios), 7.86.

  When you do these problems, you should think about: (1) testing vs. confidence intervals, (2) z vs. t distribution, (3) one-sided vs. two-sided tests. For these problems, you may use Stata to calculate means and standard deviations, to look up p-values, and to check your results, but you should show how you calculate the test statistics yourself and not leave it up to Stata to carry out the tests and calculate the confidence intervals.

Additional problemss

1. The following are net returns to investment (given as %) for 8 domestic publishing houses: 6.8, 10.6, 8.1, 5.0, 6.9, 10.4, 3.1, 6.3

   The following are net returns to investment (given as %) for 8 European publishing houses: 7.7, 12.1, 11.4, 7.7, 6.7, 12.9, 2.7, 5.8

   Enter each data set into Stata and answer the following questions. You may use Stata to do the calculations.

   (a) Suppose each sample was chosen by simple random sampling from the population of publishing houses in the corresponding region. Do a statistical test to see whether the average returns to investment for domestic and European publishers are the same. State clearly what the null and alternative hypotheses are (in ordinary English, not just as an equation), the kind of test you used, why you used it, and what assumptions were required to use that test. State the p-value, and interpret the results.

   (b) You now find out that the European publishing houses are actually the overseas branches of the randomly-sampled domestic publishers in your sample. (The domestic publishers are listed in the same order as the corresponding European publishers.) What would now be your null and alternative hypotheses, what kind of test would you use and why, and what assumptions are required? State the p-value, and interpret the results.

(c) Briefly explain the differences between the p-values you just calculated. What caused them to be different? (A scatterplot might help to make this clear.)

Challenge Problem: Paired vs two-sample z-tests

Assume that you have n paired samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \sim N(\mu_x, \sigma)$ and $y_i \sim N(\mu_y, \sigma)$ and the correlation of x_i and y_i is ρ and assume that $\rho > 0$. Let $d_i = x_i - y_i$.

1. Show that $\bar{d} = \bar{x} - \bar{y}$.

2. Find the variance of d_i.

3. Find the standard error of $\bar{d}$.

4. Assume that x_i and y_i are independent. What is the standard error of $\bar{x} - \bar{y}$?

5. Which test examining $H_0 : \mu_x - \mu_y = 0$, will give the bigger z statistic? If $n = 20$, $\sigma = 5$, and $\rho = 0.5$, how much bigger will the bigger z statistic be?

6. If $\rho < 0$, will the same statistic in the previous part that gives the bigger value give the bigger value now?

Note: Since under the assumption for both test statistics, z will be standard normal for the paired and two-sample z test. Thus the statistic that gives the larger value will be more powerful. Thus the correlation in the paired case indicates whether pairing should be used, or whether two independent samples should be used instead.

Stata Hints

Calculating p-values: To calculate values of the cumulative distribution function (from which you can calculate p-values) you may use the `normprob` function in Stata. To calculate $P(Z < z)$, the probability that a standard normal will be less than or equal to a number z, use:

```
. generate npvalue=normprob(z)
```

(You have to put a number in for z. There is nothing special about the variable name `npvalue`, it is just used as an example. After you run the command, use the `list` command to print the results.) This gives the same values you would get from Table A, but not limited to the values included there.

You may compute a similar CDF value for the t distribution with the `tprob` function:

```
. generate tpvalue=tprob(df,t)
```

where now `df` is the degrees of freedom. Now `tpvalue` will contain $P(T_{\texttt{df}} > t)$.

There are similar functions for some other distributions we will run into later in the semester; `chiprob(df,x)` for the chi-square distribution, and `fprob(df1,df2,x)` for the F distribution.

You can also use Stata to calculate inverse CDFs (i.e. to get critical values, given the tail area), using functions `invnorm(p)` for the normal or `invt(df,p)` for the t distribution.

Note: If you do not have any data in you Stata session when you compute these values you must set the number of observations to be non-zero. If you are using Stata to compute single p-values, it suffices to use the following command before you start

```
. set obs 1
```

t confidence intervals: Stata computes a *confidence interval* for the mean of variable *varname* at level *level* (for example, 95 for a 95% interval) with the command

```
ci varname, level(level)
```

The ci function can also be used to calculate the exact confidence intervals for binomial proportions. The form of the command is

```
ci varname, level(level) binomial
```

where varname contains a 0/1 variable giving the failure/success indicators for each trial. If you already have the summary statistics, you can use the `cii` function. See `help ci` for more information. To get the large sample confidence interval for a binomial proportion, you need to use the `prtest` function. Again see `help prtest` for more information.

One- and two-sample tests on population means: Assume that there is a variable `v1`, and maybe another one, `v2`. With one column simply do

`. ttest v1=`μ_0

where μ_0 is replace by a number that corresponds to the null hypothesis value of the mean. For a two sample ttest you can do

```
. ttest v1=v2
```

Here the default is to do a paired t-test, assuming equal variances. You can modify it by

```
. ttest v1=v2, unpaired
```

or

```
. ttest v1=v2, unpaired unequal
```

One- and two-sample binomial tests: The large sample hypothesis tests based on the normal approximation to the binomial can be done with the `prtest` function. Usually the immediate form is more useful here. The forms for the one- and two-sample cases are

```
. prtest n #successes, count
```

and

```
. prtest n1 #successes1 n2 #successes2, count
```

Again for more information check `help prtest`.