University of North Carolina
Chapel Hill

Soci708-001 Statistics for Sociologists

Fall 2009

Professor François Nielsen

Stata Commands for Module 7 – Inference for Distributions

For further information on any command in this handout, simply type `help` followed by the name of the command in Stata.

For confidence intervals, also see page 35 of the Stata and SAS Guide pdf (click on Documents in side bar; guide is linked under Software Documentation).

1 Statistical Functions in Stata

1.1 Normal Distribution Functions

The function `normal(z)` returns the cumulative standard normal distribution $P(Z \leq z)$.

```
. display normal(1.207)
.88628393
```

The function `invnormal(p)` returns z such that $P(Z \leq z) = p$.

```
. display invnormal(0.975)
1.959964
```

1.2 Student t Distribution Functions

The function `ttail(df, t)` returns the reverse cumulative (upper-tail) Student's t distribution for df degrees of freedom; given t it returns the probability $P(T > t)$.

```
. display ttail(7, 1.960)
.04540985
```

The function `invttail(df, p)` returns the inverse reverse cumulative (upper-tail) Student's t distribution for df degrees of freedom; given p it returns t such that $P(T > t) = p$.

```
. display invttail(7, 0.025)
2.3646243
```

1.3 Curve for Problem 7.113 p.481

For IPS6e Problem 7.113 p.481 – Degrees of freedom and confidence interval width.
This is how to draw the curve requested in this problem in Stata:

```
. twoway function y=invttail(x,0.025), range(2 100) yline(1.96)
```

This is how to do it in R:

```
> curve(qt(0.975,x),2,100,xlab="Degrees of freedom", ylab="t*")
> abline(1.96,0,col="blue")
```

2 Entering Data for Confidence Intervals and One-Sample Tests

There are several ways to enter data in Stata to calculate confidence intervals for the mean and one-sample test statistics. Here are three of them.

2.1 Method 1

A quick-and-dirty method from Andrew Ritchey. Take IPS6e Problem 7.24 p.442 as an example. You have to enter 20 observations. In Stata, first clear any data in memory. Then create a data frame with 20 observations and create a variable with all values missing:

```
. clear
. set obs 20
. gen mpg=.
```

Then go to the data editor (Data/Data Editor, or click the icon) and replace the missing values with the actual values. Then *close* the Data Editor (click on the ×). Your data are ready to use. See below.

2.2 Method 2

Another technique from Michele Easter. You want to enter 9 observations. After clearing any data in memory, we create a new variable called `var` and input the data. (Type `input var` and then enter the values. After the last value, type `end`.)

```
. clear
. input var

           var
  1. 3
  2. 3.2
  3. 3.2
  4. 3.3
  5. 2.9
  6. 3
  7. 3.1
  8. 3.1
  9. 3.4
 10. end
```

2.3 Method 3

Still another technique that is a time-saver for moderately long data sets from the textbook. Use the disk that comes with the text and navigate to the data sets, and go the the Excel folder. The data for Problem 7.24 p. 442 are under

Chapter 7. Find the file called `ex07_024.xls` and open it in Excel or another spreadsheet program. Then select and copy (Ctrl-C) the whole column of 20 observations including the header.

In Stata first clear any data in memory (clear) then open the Data Editor (Data/Data Editor in the menu). Then paste the data in the Data Editor (Ctrl-V). Then *close* the Data Editor (click on the ×). Your data are ready to use.

Say you want to list the data, calculate the mean and SD, and then calculate a 95% confidence interval. Some of the following commands might be useful.

First we calculate the CI "the hard way".

```
. list

     +------+
     |  mpg |
     |------|
  1. | 41.5 |
  2. | 50.7 |
  3. | 36.6 |

 ...

 20. | 43.3 |
     +------+

. stem mpg

Stem-and-leaf plot for mpg (MPG)

mpg rounded to nearest multiple of .1
plot in units of .1

  3** | 42
  3** | 66,73,73
  3** | 92
  4** | 15
  4** | 22,32,32,33,35
  4** | 43,46,50
  4** | 64,68,77
  4** | 80,84
  5** | 07

. * calculate mean and sd

. su mpg

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         mpg |        20       43.17    4.414939       34.2       50.7


. * calculate t* and margin of error m

. display invttail(19, 0.025)
2.0930241

. display 2.0930241*4.414939/sqrt(20)
2.0662551
```

```
. * lower bound of CI

. display  43.17 - 2.0662551
41.103745

. * upper bound

. display  43.17 + 2.0662551
45.236255
```

Now we do the CI the easy way.

```
. ci mpg

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
         mpg |         20       43.17    .9872104        41.10374    45.23626
```

3 One-Sample t-Test in Stata

The following two examples involve inputting data using the keyboard, but in general it is easier just to go into the Data Editor, or copy and paste from an already-entered Excel spreadsheet.

In the first command, we create a new variable called `var` and input the data. (Type `input var` and then enter the values. After the last value, type `end`.)

```
. input var

           var
  1. 3
  2. 3.2
  3. 3.2
  4. 3.3
  5. 2.9
  6. 3
  7. 3.1
  8. 3.1
  9. 3.4
 10. end
```

Check the mean by using the `summarize` command (a.k.a. `sum` or `su`).

```
. su var

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         var |         9    3.133333    .1581139        2.9        3.4
```

The mean for this sample is 3.133333. Now we would like to do a t-test to assess how likely it is that the true mean is greater than 3. Stata tests whether the $\mu > 3$, $\mu = 3$, or $\mu < 3$ at the same time.

```
. ttest var=3

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     var |       9    3.133333    .0527046    .1581139    3.011796     3.25487
------------------------------------------------------------------------------
    mean = mean(var)                                              t =   2.5298
Ho: mean = 3                                     degrees of freedom =        8

    Ha: mean < 3                 Ha: mean != 3                 Ha: mean > 3
 Pr(T < t) = 0.9824         Pr(|T| > |t|) = 0.0353          Pr(T > t) = 0.0176
```

Stata by default assigns a 95% confidence interval, but this can be changed using the option `level`. To tell Stata to use a 90% confidence interval, you would enter the command (output not shown):

```
. ttest var=3, level(90)
  ...
```

See Stata Help for more on the `ttest` command.

4 Matched Pairs in Stata

This example is very similar.

```
. input x

            x
  1. 9000
  2. 8000
  3. 6000
  4. 6000
  5. 8000
  6. 7000
  7. end

. ttest x=0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       6    7333.333    494.4132     1211.06    6062.404    8604.263
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  14.8324
Ho: mean = 0                                     degrees of freedom =        5

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

5 Two-Samples Problems in Stata

Now, using the dataset `woprops` (available from sidebar of course website, under Datasets), do a t-test to see whether State cabinets under Democratic

governors (demo=1) have a higher proportion of women (woprop) than State cabinets under Republican governors (demo=0).

The first approach (recommended in general) uses the option unequal, which causes Stata to carry out the t-test *without assuming equal variance of* woprop *in the two groups*. The degrees of freedom are calculated using Satterthwaite's approximation formula (1946).

```
. ttest woprop, by(demo) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      17    .1712353    .0199462    .0822401    .1289513    .2135193
       1 |      22    .2398636    .0287919    .1350461    .1799875    .2997397
---------+--------------------------------------------------------------------
combined |      39    .2099487    .0190242    .1188063    .1714362    .2484613
---------+--------------------------------------------------------------------
    diff |           -.0686283    .0350261               -.1397122    .0024555
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.9594
Ho: diff = 0                     Satterthwaite's degrees of freedom =  35.3172

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0290         Pr(|T| > |t|) = 0.0580          Pr(T > t) = 0.9710
```

The second approach assumes that the variances of woprop in the two groups compared are equal. It is the Stata default. In this case, the degrees of freedom are calculated as $(n_1 - 1) + (n_2 - 1)$. Except in special cases, however, this approach should not be used and *the option* unequal *should always be specified.*

```
. ttest woprop, by(demo)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      17    .1712353    .0199462    .0822401    .1289513    .2135193
       1 |      22    .2398636    .0287919    .1350461    .1799875    .2997397
---------+--------------------------------------------------------------------
combined |      39    .2099487    .0190242    .1188063    .1714362    .2484613
---------+--------------------------------------------------------------------
    diff |           -.0686283    .0372071                -.144017    .0067604
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.8445
Ho: diff = 0                                     degrees of freedom =       37

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0366         Pr(|T| > |t|) = 0.0731          Pr(T > t) = 0.9634
```

The command ttesti ("t-test immediate") can be used when one does not have the original data. Then the ns, means and sds for the two groups are input as part of the command. The next two commands (output not shown) give the same results as the two t-tests above. Note that ttesti can be used to check / replicate published results when the original data are not available.

```
. ttesti 17 .1712353  .0822401  22  .2398636 .1350461, unequal
  ...
. ttesti 17 .1712353  .0822401  22  .2398636 .1350461
  ...
```