Stata Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery
Chapter 8: Test of Hypotheses

Program Notes Outline

Note 8.1 – Testing a hypothesis about the mean assuming the variance is unknown
Note 8.2 – Testing the hypothesis about a population proportion
Note 8.3 – Correlation coefficients and their p-values
Note 8.4 – Testing the hypothesis of no difference in two population means assuming equal variances
Note 8.5 – Testing the hypothesis of no difference in two population means assuming unequal variances
Note 8.6 – Paired t-test
Note 8.7 – Testing a hypothesis about the difference of two proportions

Chapter 8 Formulas

Test	Test Statistic	Distribution of Test Statistic
One sample z-test	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	standard normal distribution
One sample t-test	$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t-distribution with n-1 degrees of freedom
Independent samples t-test assuming equal variances	$\frac{(\bar{x}_1-\bar{x}_2)-\Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	t-distribution distribution with degrees of freedom, df = n_1+n_2-2, and s_p = $\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$
Independent samples t-test assuming unequal variances	$\frac{(\bar{x}_1-\bar{x}_2)-\Delta_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	t-distribution distribution with degrees of freedom, df = $\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2-1)}}$
Paired t-test	$\frac{\bar{x}_d-\mu_d}{s_d/\sqrt{n}}$	t-distribution with n-1 degrees of freedom

Note 8.1 – Testing a hypothesis about the mean assuming the variance is unknown

The **ttest** command in Stata can be used to test a hypothesis about the mean when the population variance in unknown. As an example, we use the DIG200 data set to test the null hypothesis that the population mean is significantly different from 122.3 mmHg.

```
Stata commands:

ttest sysbp = 122.3
```

The Stata output provides the number of observations, the mean, standard error, standard deviation, a 95% confidence interval, the value of the t-statistic, and its degrees of freedom. The **level( )** option can be used to change the level of the confidence interval.

```
Stata output:

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   sysbp |     199    125.8241    1.288642    18.17853    123.2829    128.3653
------------------------------------------------------------------------------
    mean = mean(sysbp)                                            t =   2.7348
Ho: mean = 122.3                         degrees of freedom =      198

  Ha: mean < 122.3             Ha: mean != 122.3                 Ha: mean > 122.3
 Pr(T < t) = 0.9966         Pr(|T| > |t|) = 0.0068          Pr(T > t) = 0.0034
```

Stata also provides three p-values: two that are based on one-sided tests and one that is based on a two-sided test. The first p-value (on the left) is associated with the alternative hypothesis that the population mean is less than 122.3 mmHg. The second p-value (in the center) is associated with the alternative hypothesis that the population mean is not equal to 122.3 mmHg. Finally, the third p-value (on the right) is associated with the alternative hypothesis that the population mean is greater than 122.3 mmHg.

To obtain a 99% confidence interval, we use the **level( )** option as shown below:

```
Stata commands:

ttest sysbp = 122.3, level(99)
```

```
Stata output:

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
   sysbp |     199    125.8241    1.288642    18.17853    122.4725    129.1757
------------------------------------------------------------------------------
    mean = mean(sysbp)                                            t =   2.7348
Ho: mean = 122.3                         degrees of freedom =      198

  Ha: mean < 122.3             Ha: mean != 122.3                 Ha: mean > 122.3
 Pr(T < t) = 0.9966         Pr(|T| > |t|) = 0.0068          Pr(T > t) = 0.0034
```

Note 8.2 – Testing the hypothesis about a population proportion

The **ttest** command in Stata can be used to test a hypothesis about the population proportion. The test uses a t-statistic to test the null hypothesis that π equals π_0 instead of the *z*-statistic shown in the text. In addition, the estimated standard error uses *n*-1 in its denominator instead of *n* and a continuity correction is not used. Hence there will be slight differences between the test statistic provided by the **ttest** command and that shown in the text. Additionally, the p-value is calculated using the *t*-distribution, not the normal distribution. For large sample sizes, there will be little difference between the *t* and *z*-test statistics and their corresponding p-values. To illustrate how the **ttest** command is used in this situation, we refer to Example 8.4 and begin with the Stata commands that create the data. To create the data set, we started by creating 140 observations with a value of "1" and then replaced 54 of the observations with the value of "0" as shown in the Stata commands below:

```
Stata commands:

set obs 140
gen immun = 1
replace immun = 0 in 87/140
```

Using the **ttest** command below:

```
Stata commands:

ttest immun = 0.75
```

Stata provides the following output:

```
Stata output:

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   immun |     140    .6142857    .0412867    .4885114    .5326545     .695917
------------------------------------------------------------------------------
    mean = mean(immun)                                            t =  -3.2871
Ho: mean = 0.75                                 degrees of freedom =      139

   Ha: mean < 0.75             Ha: mean != 0.75                 Ha: mean > 0.75
 Pr(T < t) = 0.0006         Pr(|T| > |t|) = 0.0013          Pr(T > t) = 0.9994
```

To obtain results similar to those presented above, we can use the **prtest** command in Stata. The test statistic obtained with the **prtest** command is based on large-sample theory. Using the **prtest** command below:

```
Stata commands:

prtest immun = 0.75
```

```
Stata output:

One-sample test of proportion                 immun: Number of obs =      140
------------------------------------------------------------------------------
```

```
    Variable |       Mean   Std. Err.                      [95% Conf. Interval]
-------------+------------------------------------------------------------------
       immun |   .6142857    .041139                      .5336547    .6949167
------------------------------------------------------------------------------
    p = proportion(immun)                                         z =  -3.7084
Ho: p = 0.75

    Ha: p < 0.75                 Ha: p != 0.75                 Ha: p > 0.75
 Pr(Z < z) = 0.0001         Pr(|Z| > |z|) = 0.0002          Pr(Z > z) = 0.9999
```

Finally if the data set is relatively small, the **bitest** command in Stata can be used since it provides exact p-values. Using the **bitest** command below:

```
Stata commands:

bitest immun = 0.75
```

```
Stata output:

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
       immun |      140          86          105     0.75000      0.61429

  Pr(k >= 86)              = 0.999863  (one-sided test)
  Pr(k <= 86)              = 0.000271  (one-sided test)
  Pr(k <= 86 or k >= 123) = 0.000399  (two-sided test)
```

Note 8.3 – Correlation coefficients and their p-values

Stata uses the **correlate (corr)** command to calculate Pearson's correlation coefficient. Some details are already provided in the program notes for chapter 3. As an example, we use the data in Table 8.4 showing infant mortality rates from 1988 and correlate them with health expenditures as percentage of GDP from 1987. For complete details refer to Example 8.7 in the textbook.

The Stata commands below were used to compute Pearson's correlation coefficient.

```
Stata commands:

input infant_mortality health_expenditure
 4.8  6.8
 5.8  9.0
 6.1  7.4
 6.8  8.5
 6.8  7.7
 7.2  8.6
 7.5  8.2
 7.5  6.0
 7.8  8.6
 8.1  6.0
 8.1  7.1
 8.3  7.5
 8.7  7.1
 8.9  7.4
 9.0  6.1
 9.2  7.2
 9.3  6.9
10.0 11.2
10.8  6.9
11.0  5.3
13.1  6.4
end

corr infant_mortality health_expenditure
```

```
Stata output:
(obs=21)

             | infant~y health~e
-------------+------------------
infant_mor~y |   1.0000
health_exp~e |  -0.2429   1.0000
```

The **pwcorr** command which stands for pairwise correlations can also be used. The **pwcorr** command has an option **sig** which provides the p-value associated with the null hypothesis, Ho: ρ = 0. The Stata commands are shown below:

```
Stata commands:

pwcorr infant_mortality health_expenditure, sig
```

Stata output:

	infant~y	health~e
infant_mor~y	1.0000	
health_exp~e	-0.2429	1.0000
	0.2887	

Note 8.4 – Testing the hypothesis of no difference in two population means assuming equal variances

The **ttest** command is again used to test the hypothesis of no difference in two population means. The **ttest** command provides a test statistic and p value assuming that the population variances are equal. If we do not wish to assume equal variances, we can use the **unequal** option. In Example 8.9, we are assuming equal variances, and we use the **ttest** command with the **by( )** option to distinguish between the two groups being compared:

```
Stata commands:

ttest prop_from_fat, by(group)
```

```
Stata output:

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      19    .3526842    .0204989    .0893526    .3096176    .3957508
       1 |      14    .3292857    .0260267     .097383    .2730584     .385513
---------+--------------------------------------------------------------------
combined |      33    .3427576    .0160317    .0920951    .3101021    .3754131
---------+--------------------------------------------------------------------
    diff |            .0233985    .0326879               -.0432689    .0900659
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.7158
Ho: diff = 0                                     degrees of freedom =       31

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7603         Pr(|T| > |t|) = 0.4795          Pr(T > t) = 0.2397
```

Note 8.5 – Testing the hypothesis of no difference in two population means assuming unequal variances

As stated in Program Note 8.4, the **ttest** command is used to test the hypothesis of no difference between two population means. We now use the test statistic associated with the assumption that the variances are unequal. In Example 8.10, we refer to the data on ages of AML and ALL patients from Chapter 7 Table 7.7. We wish to test that the mean AML age is less than or equal to 5 years greater than the mean ALL age versus the alternative that the difference is greater than 5 years. The first 51 observations are from the AML patients and the last 20 are from the ALL patients. The Stata commands below refer to the hypothesis test of no difference between the mean age of AML and ALL patients.

```
Stata commands:

ttest age, by(dx_type) unequal
```

In the Stata output below, notice that Satterthwaite's degrees of freedom equals to 32.501 which is exactly equal to the degrees of freedom presented in the textbook. However, the test statistic calculated in Example 8.10 considers the alternative hypothesis of a 5-year difference between ages.

```
Stata output:

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      51    49.86275     2.31195    16.51063    45.21906    54.50643
       1 |      20       36.65     3.99096    17.84812    28.29682    45.00318
---------+--------------------------------------------------------------------
combined |      71    46.14085    2.113033    17.80473    41.92653    50.35516
---------+--------------------------------------------------------------------
    diff |            13.21275    4.612253                3.823571    22.60192
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.8647
Ho: diff = 0                     Satterthwaite's degrees of freedom =  32.5011

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9964         Pr(|T| > |t|) = 0.0073          Pr(T > t) = 0.0036
```

Note 8.6 – Paired t-test

The **ttest** command can be used to test for no difference in two dependent population means. In Example 8.11, we are given the mean difference between the diastolic blood pressure readings, and the standard deviation has been calculated. Therefore Stata's **ttesti** command, called the immediate form, is used where the first value after the command is the number of observations, the second value is mean difference, the third value is the standard deviation, and the fourth value is the mean difference under the null hypothesis. Here the **level** option, **level(90)**, indicates that a 90% confidence interval will be computed.

```
Stata commands:

ttesti 53 10.6 8.5 0, level(90)
```

```
Stata output:

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [90% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      53        10.6    1.167565         8.5    8.644692    12.55531
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   9.0787
Ho: mean = 0                                     degrees of freedom =       52

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

Note 8.7 – Testing a hypothesis about the difference of two proportions

The **ttest** command can be used to provide an approximate test statistic for the test of the equality of two population proportions. The estimate of the standard error of the difference differs slightly from that found using the binomial formula because of the division by $n - 1$ in **ttest** instead of n used by the binomial calculation. The reported p-value is also slightly off because the t-distribution instead of the normal distribution is used in its calculation. For large samples, these differences are small. Here we create the data used in Example 8.12. The data are the compliance status of 42 milk producers in the East and 50 milk producers in the Southwest. We will use the variable comply to indicate compliance status and the variable region to distinguish between the East and the Southwest.

```
Stata commands:

set obs 92
gen comply = 0
replace comply=1 in 1/12
replace comply=1 in 43/63

gen region=1
replace region=2 in 43/92
```

The variable comply equals "0" to indicate compliance and "1" to indicate non-compliance. The variable region equals "1" when referring to the East and equals "2" when referring to the Southwest. A t-test at the 0.01 significance level is conducted with the Stata commands below:

```
Stata commands:

ttest comply, by(region) level(99)
```

```
Stata output:

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      42    .2857143    .0705521      .45723    .0951402    .4762883
       2 |      50          .42    .0705084    .4985694     .231041     .608959
---------+--------------------------------------------------------------------
combined |      92    .3586957    .0502776    .4822457    .2264183     .490973
---------+--------------------------------------------------------------------
    diff |           -.1342857    .1005048               -.3987708    .1301993
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -1.3361
Ho: diff = 0                                     degrees of freedom =       90

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0924         Pr(|T| > |t|) = 0.1849          Pr(T > t) = 0.9076
```