Training on Statistical Tools for Research: Stata

Survey Data Analysis in Stata

Md Monowar Hossain

Institute of Statistical Research and Training (ISRT)
University of Dhaka
mmhossain1@isrt.ac.bd

September 27, 2025

One sample t-test

- A one sample t-test allows us to test whether a population mean (of a normally distributed interval variable) significantly differs from a hypothesized value.
- For example, using the **birthwt.dta** data file, say we wish to test whether the average birth weight of newborns differs significantly from 2500 gm.

```
use birthwt.dta, clear
ttest bwt=2500
```

One-sample t test

One-sample t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf. interval]	
bwt	**189**	**2944.587**	**53.04254**	**729.2143**	**2839.952**	**3049.222**

```
    mean = mean(bwt)                                          t =   8.3817
H0: mean = 2500                              Degrees of freedom =      188

   Ha: mean < 2500            Ha: mean != 2500             Ha: mean > 2500
 Pr(T < t) = 1.0000       Pr(|T| > |t|) = 0.0000         Pr(T > t) = 0.0000
```

- The mean of the variable bwt for this particular sample is 2944.587, which is significantly different from the test value of 2500.
- We would conclude that the mean birth weight is significantly higher than 2500.

Two independent samples t-test

- An independent samples t-test is used when you want to compare the means of a normally distributed interval dependent variable for two independent groups.
- For example, we may wish to test whether the mean birth weight is the same with smokers and nonsmokers.
- The test variable is **bwt** and group variable is **smoke**.

```
ttest bwt,by(smoke)
```

Two-sample t test with equal variances

Two-sample t test with equal variances

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
no	**115**	**3055.696**	**70.18559**	**752.6566**	**2916.659**	**3194.733**
yes	**74**	**2771.919**	**76.681**	**659.6349**	**2619.094**	**2924.744**
Combined	**189**	**2944.587**	**53.04254**	**729.2143**	**2839.952**	**3049.222**
diff		**283.7767**	**106.9688**		**72.75612**	**494.7973**

```
    diff = mean(no) - mean(yes)                              t =   2.6529
H0: diff = 0                               Degrees of freedom =      187

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9957         Pr(|T| > |t|) = 0.0087          Pr(T > t) = 0.0043
```

- The results indicate that there is a significant difference between the means with smokers and nonsmokers, since the p-value is 0.0087.
- More specifically, the mean birth weight of babies of nonsmokers is significantly higher than those of smokers.

One-way ANOVA

- A one-way analysis of variance (ANOVA) is used when you have a categorical independent variable (with two or more categories) and a normally distributed interval dependent variable and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable.
- For example, we may wish to test whether the mean birth weight differs among the three races.

```
oneway bwt race
```

Analysis of Variance

Analysis of variance

Source	SS	df	MS	F	Prob > F
Between groups	**5015725.25**	**2**	**2507862.63**	**4.91**	**0.0083**
Within groups	**94953930.6**	**186**	**510505.003**		
Total	**99969655.8**	**188**	**531753.488**		

Bartlett's equal-variances test: chi2(**2**) = **0.6595** Prob>chi2 = **0.719**

- As indicated in the ANOVA table, the mean of birth weight differs significantly among the levels of races. That is, the means of birth weight across the races are not the same.

Paired t-test

- A paired (samples) t-test is used when you have two related observations (i.e. two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another.
- Assume that twenty subjects participated in an experiment to study the effectiveness of a certain diet, combined with a program of exercise, in reducing serum cholesterol levels. Data are recorded on the serum cholesterol levels for the 10 subjects at the beginning of the program (before) and at the end of the program (after), and is available in the data file **diet**.

- The question to be answered is: Do the data provide sufficient evidence for us to conclude that the diet-exercise program is effective in reducing serum cholesterol levels?

```
use diet.dta, clear
ttest after=before
```

Paired t test

Paired t test

Variable	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
after	10	226.8	9.219303	29.154	205.9445	247.6555
before	10	244.6	10.77775	34.08225	220.219	268.981
diff	10	-17.8	4.762819	15.06136	-28.57425	-7.025755

mean(diff) = mean(**after** - **before**) t = **-3.7373**

H0: mean(diff) = 0 Degrees of freedom = **9**

Ha: mean(diff) < 0 Ha: mean(diff) != 0 Ha: mean(diff) > 0

Pr(T < t) = **0.0023** Pr(|T| > |t|) = **0.0046** Pr(T > t) = **0.9977**

- These results indicate that the diet-exercise program is significantly effective in reducing serum cholesterol levels.

Test for single proportion

- A one sample proportion test allows us to test whether the proportion of successes on a two-level categorical dependent variable significantly differs from a hypothesized value.
- For example, we may wish to test whether the proportion of mothers with low birth weight babies differs significantly from 0.40.
- The **prtest** command assumes that the variables it will act on are binary (0/1) variables and the proportion of interest is the proportion of 1's.

```
prtest low=0.4
```

One-sample test of proportion

```
One-sample test of proportion                  Number of obs    =       189

    Variable |       Mean   Std. err.                [95% conf. interval]
-------------+------------------------------------------------------------
         low |   .3121693   .0337058                  .2461071    .3782315

    p = proportion(low)                                          z = -2.4647
H0: p = 0.4

     Ha: p < 0.4                 Ha: p != 0.4                 Ha: p > 0.4
 Pr(Z < z) = 0.0069         Pr(|Z| > |z|) = 0.0137        Pr(Z > z) = 0.9931
```

- The results indicate that the proportion of mothers with low birth weight babies is significantly lower than the hypothesized value of 40%.

Test for two proportions

```
prtest ht=smoke
```

Two-sample test of proportions

```
Two-sample test of proportions                   ht: Number of obs =       189
                                              smoke: Number of obs =       189

    Variable |       Mean   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          ht |   .0634921   .0177372                      .0287278    .0982563
       smoke |   .3915344   .0355036                      .3219487    .4611201
-------------+----------------------------------------------------------------
        diff |  -.3280423   .0396877                     -.4058287   -.2502559
             |  under H0:   .0431254   -7.61   0.000
------------------------------------------------------------------------------
        diff = prop(ht) - prop(smoke)                             z =  -7.6067
    H0: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0000         Pr(|Z| > |z|) = 0.0000          Pr(Z > z) = 1.0000
```

- The results indicate that the proportion of mothers with hypertension is significantly lower than the proportion of mothers who are smokers.

Equivalence of two proportions test and the chi-square test of independence

```
tab low smoke, chi2      /* test of independence */
prtest smoke, by(low)    /* compares between the proportion
                            and proportion of smokers (in lo
```

- If you square the z value then you will obtain the chi-square value.
- p-values for both the tests are the same.

chi-square test

low birth weight	smoking status during pregnancy no	yes	Total
no	86	44	130
yes	29	30	59
Total	115	74	189

Pearson chi2(1) = 4.9237 Pr = 0.026

two proportions test

```
Two-sample test of proportions                     no: Number of obs =       130
                                                  yes: Number of obs =        59
------------------------------------------------------------------------------
       Group |       Mean   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
          no |   .3384615   .0415012                      .2571207    .4198024
         yes |   .5084746   .0650851                      .3809101     .636039
-------------+----------------------------------------------------------------
        diff |   -.170013   .0771908                     -.3213042   -.0187219
             |  under H0:   .0766189   -2.22   0.026
------------------------------------------------------------------------------
        diff = prop(no) - prop(yes)                               z =  -2.2189
    H0: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0132         Pr(|Z| > |z|) = 0.0265          Pr(Z > z) = 0.9868
```

Chi-square goodness of fit test

- A chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions.
- For example, suppose we believe that the women under different races are of equal proportions.
- We want to test whether the observed proportions from our sample differ significantly from the hypothesized equality of proportions.

```
findit csgof
csgof race
```

```
+---------------------------------------+
| race    expperc    expfreq    obsfreq    |
|---------------------------------------|
| white  33.33333  63         96         |
| black  33.33333  63         26         |
| other  33.33333  63         67         |
+---------------------------------------+
chisq(2) = 39.27, p = 0
```

- Since the p-value is very small, we can reject the null hypothesis that proportion of women is the same for all the races.

With explicit proportions

```
csgof race, expperc(33.3,33.3,33.3)
```

```
+-------------------------------------+
| race   expperc   expfreq   obsfreq  |
|-------------------------------------|
| white  33.3      62.937    96       |
| black  33.3      62.937    26       |
| other  33.3      62.937    67       |
+-------------------------------------+
chisq(2) = 39.31, p = 0
```

With hypothesized proportions

```
csgof race, expperc(50,20,30)
```

```
+---------------------------------------+
| race   expperc   expfreq   obsfreq   |
|---------------------------------------|
| white  50        94.5      96        |
| black  20        37.8      26        |
| other  30        56.7      67        |
+---------------------------------------+
chisq(2) = 5.58, p = .0615
```

- Note that we cannot reject the null hypothesis at 5% level of significance.

THANK YOU