UMES Interdisciplinary Research Center

Statistical Analysis Using STATA

Dr. Abaidoo, Rexford

3/14/23

Introductory and Advanced Research Analysis Training

Statistical Analysis Using STATA

1. **Descriptive Statistics**
   a. *des*
   b. *br*
   c. *su, sum, su, de*
   d. *ci means mpg, level(90)*
   e. *ci variances mpg, level (99)*
   f. *using the by command: by foreign: su mpg*

2. **Correlation Analysis**

   a. *cor, :generate simple correlation table*
   b. *pwcor. Sig : Generate correlation with significance levels*

Fundamentals of Hypothesis Testing

- **Hypothesis Test for variance:**

- **One sample test for variance**

- *sdtest mpg=5*

- **Two sample Test for equality of variance**

- *sdtest mpg=weight*

Hypothesis test for the Mean

a. Single Variable estimation

i. One-tailed test: Right

Test the hypothesis that the average mpg is greater than 25.

- *H0 : $\mu \leq 25$*
- *HA : $\mu > 25$*

ttest mpg = 25

ii. One-tailed test: Left

Test the hypothesis that the average weight is less than 3500.

- *H0 : $\mu \geq 3500$*
- *HA : $\mu < 3500$*

ttest weight = 3500

iii. Two-tailed test

Test the hypothesis that the average headroom is different than 3.5.

- *H0 : μ =3.5*
- *HA : μ 6 ≠ 3.5*

ttest headroom = 3.5

Two Variable Hypothesis test estimation- Paired vs Unpaired t-test

Paired t test - A paired (samples) t-test is used when you have two related observations

ttest Montana = North Dakota, level(95)

Unpaired t test (equal and unequal variance)

ttest mpg = gear ratio, unpaired level(95)

ttest mpg = gear ration, unpaired unequal level(95)

test with independent categorical variables using the *by* command– With equal or unequal variance

ttest mpg , by(foreign) level(99) unequal

ttest mpg , by(foreign) level(99)

Chi-square test: for association between two categorical variables

tab schtyp female, chi2

Data: use https://stats.idre.ucla.edu/stat/stata/notes/hsb2

relationship between two categorical variables

Hypothesis test Using Summary statistics

T-test from Summary Statistics

Example:

Assume that $\sigma_1 = \sigma_2$

ttesti N1 Mean1 SD1 N2 Mean2 SD2, CI Level.

ttesti ……., level (99)

Do Not Assume $\sigma_1 = \sigma_2$

ttesti ……., level(99) unequal

Example 1: Specific Motors of Detroit has developed a new automobile known as the M car. 24 M cars and 28 J cars from Japan were read test to compare miles per gallon performance. The sample statistics are shown below:

	M Cars	J Cars
Sample size	24	28
sample mean	29.8	27.3
sample std dev	2.56	1.81

Can we conclude, using 0.05 level of significance, that the mpg performance of M cars is greater than the mpg performance of J cars?

ttesti 24 29.8 2.56 28 27.3 1.81, level (95) unequal

Example 2 Par Inc is a manufacturing of golf equipment and has developed a new golf ball that has been designed to provide extra distance. A sample of par golf balls was compared with a sample of golf balls made by Rap Ltd. The same information are provided below.

	Par Inc	Rap Ltd
Sample size	120 balls	80 balls
sample mean	275 yards	258 yards

- The sample standard deviation for the two firms are 15 for Par Inc and 20 for Rap LTD
- Can we conclude, using 0.01 level of significance, that the mean driving distance of Par In is greater than the mean distance of Rap Ltd

ANOVA – Using the anova command

Data: use https://stats.idre.ucla.edu/stat/stata/notes/hsb2

One-way ANOVA

A one-way analysis of variance (ANOVA) is used when you have a categorical independent variable (with two or more categories)

For example, we can test whether the mean of **write** differs between the three program types (**prog**). The command for this test would be:

anova write prog

Factorial ANOVA

A factorial ANOVA has two or more categorical independent variables (either with or without the interactions) and a single normally distributed interval dependent variable

anova write female ses

Hypothesis testing for the Proportion

Summary Statistics Approach (su)

1. Two groups, A and B, each consist of 100 randomly assigned people who have a disease. One serum is given to Group A and a different serum is given to Group B; otherwise, the two groups are treated identically. It is found that in groups A and B, 75 and 65 people, respectively, recover from the disease. (a) Test the hypothesis that the serums differ in their effectiveness using $\alpha = .05$.

prtesti 100 .75 100 .65, level(95)

(i.e. N1 p1 N2 p2, desired CI level)

2. Out of 80 household in Salisbury surveyed, 25 indicated that they participated in a neighborhood watch program. In Delmar, 70 household were surveyed and 20 claim to be involved in such a program. Test if there is evidence to conclude that there is a significant difference in the proportion of residence participating in neighborhood watch program between the two cities

Survey Data and Experimental Data Analysis

- **Convert string into numeric**

- **Su**
- **Cor**
- **Hypothesis testing**
- **Practical demonstration**

Simple and Multiple Regression Analysis

Creating and index:

Pac v1 v2 v3
Predict pc1 pc2 pc3
Predict luxury
Practical Demonstration

a. Using the regression command in stata

reg price mpg weight……

Robustness Test after regression

i. check for linearity – rvfplot
ii. check for multicollinearity -vif
iii. check for heteroskedasticity – hettest
iv. check for normality - swilk mpg or hist mpg, frequency normal, sfrancia mpg, sktest mpg

b. Creating Interaction or Moderating Effects

i. (a continuous variable) = c
ii. (a 0-1 dummy variable) = i
iii. Example: c.age#i.male
iv. Use "##" instead of "#" to add full interactions,

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2

reg science i.female##c.read

c. Translating Results into publishable format

i. eststo
ii. esttab, se ar2 r2 scalars(F)

Data Visualization in Stata

Single Variable Graphs:

- hist mpg, percent
- hist mpg, discrete
- hist mpg, frequency normal color(blue)
- hist mpg, discrete frequency normal
- hist mpg, discrete bcolor(blue) frequency normal
- kdensity mpg
- kdensity mpg, normal
- kdensity price if mpg>=20
- kdensity price if mpg>=20, normal
- kdensity price if mpg>=20, normal color(red)
- hist mpg, kdensity by (foreign)

Multiple Variable Graphs:

- *tw mpg weight or scatter mpg weight*
- *tw lfit mpg weight*
- *tw (scatter mpg weight)(lfit mpg weight)*
- *scatterplot with CI. tw lfitci price mpg*
- *tw line price mpg*. (make sure the data is sorted first by using the sort command)

  note: can also just use line ….
- *tw con mpg price, sort (price)*
- *tw bar price mpg* Making bar graphs
- *graph tw line mpg price, sort(price)*
- *graph tw line mpg price, sort(price) scheme(s1mono)*
- *graph pie, over (foreign) or graph pie, over(xxxx) sort descending*
- *graph bar mpg rep78 length*
- *graph matrix mpg price*