Introduction to Statistics in STATA

Basic univariate descriptive statistics. Basics of bivariate inferential statistics.

Tomáš Oleš
Department of Economic Policy
Faculty of Economics and Finance

January 29, 2025

Agenda

- Obtain basic univariate descriptive statistics using STATA.
- Obtain basic bivariate inferential statistics using STATA.
- Perform simple bivariate analyses using STATA.

Why Do We Need Statistics?

Why Do We Need Statistics?

- "Facts are stubborn, but statistics are more pliable." - Mark Twain
- "Before the curse of statistics fell upon mankind we lived a happy, innocent life, full of merriment and go and informed by fairly good judgment." - Hilaire Belloc

What is Univariate Descriptive Statistics?

Univariate statistics describe and summarize a single variable in a dataset.
These include:

- Measures of Central Tendency: mean, median, mode.
- Measures of Dispersion: variance, standard deviation, range, interquartile range

... math is coming!

Univariate Descriptive Statistics

- Measures of Central Tendency:
  - Mean: $\bar{x} = \frac{\sum x_i}{n}$
  - Median: Middle value of ordered data
  - Mode: Most frequently occurring value
- Measures of Dispersion:
  - Variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
  - Standard Deviation: $s = \sqrt{s^2}$
  - Range: $\max(x) - \min(x)$
  - Interquartile Range: $Q_3 - Q_1$

Descriptive Statistics in Stata

```
/* load built-in dataset */
sysuse nlsw88, clear

/* basic summary statistics */
summarize wage hours

/* detailed statistics with percentiles */
summarize wage, detail
```

Mean and standard deviation provide insights into central tendency and dispersion. The ‘detail‘ option includes percentiles, skewness, and kurtosis.

Frequency and Percentiles

```
/* frequency table for categorical variables */
tabulate race
tabulate industry, sort

/* percentiles and quartiles */
centile wage, centile(25 50 75)
```

- ‘tabulate‘ provides counts and percentages.
- ‘centile‘ helps understand data distribution.

Visualizing Data: Histograms

```
/* histogram for wage */
histogram wage, bin(20) normal
```

- Helps visualize distribution.
- Can compare against a normal curve.
- Adjust ‘bin()‘ for granularity.

Skewness and Kurtosis

Skewness: Measures symmetry of distribution.

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3}{(n-1)s^3} \tag{1}$$

Kurtosis: Measures tail heaviness.

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{(n-1)s^4} \tag{2}$$

Skewness and Kurtosis in Stata

- Positive skew: Right-tailed distribution.
- Negative skew: Left-tailed distribution.
- Kurtosis > 3:Heavy tails (leptokurtic).
- Kurtosis < 3: Light tails (platykurtic).

```
/* test for skewness and kurtosis */
sktest wage
```

Boxplots for Outlier Detection

```
/* boxplot for wage */
graph box wage
```

- Median line represents the central tendency.
- Whiskers show variability.
- Outliers appear as individual points.

Descriptive Statistics and Graphs: workout1.dta dataset

```
/*back to using the workout1 dataset*/

use datasets/workout1,clear
encode v07, gen(v07_num) //turns v07 into numeric
/*shows frequency distributions*/
tab v07_num
fre v07_num
hist v07_num, discrete percent addlabel xlabel(1/2, valuelabel noticks)
graph pie, over(v07_num) plabel(_all percent)
```

```
/*open a Stata installed dataset*/
sysuse auto,clear

/*summary statistics*/
sum price
sum price, d
mean price
tabstat price weight length, stats(mean sd range count) by(foreign)
tabstat price weight length, stats(mean sd range count) by(foreign) col(stat
tab foreign rep78, sum(mpg)
```

Plotting

```
/*open a stata-installed dataset*/
sysuse nlsw88,clear
hist wage, frequency
replace race=. if race==3 //category 3 set to missing
graph box wage, by(race)
```

Bivariate Inferential Statistics

These include:

- Correlation
- t-test
- ANOVA
- Chi-squared test

... math is coming!

Correlation Analysis

Definition: Examines the relationship between two continuous variables.
Formula: Pearson Correlation Coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} \tag{3}$$

```
/*open Stata-installed data*/
sysuse nlsw88, clear
/*correlation analysis*/
pwcorr wage ttl_exp, star(0.05) obs
corr wage ttl_exp
```

Interpretation: A moderate positive/negative correlation between wage and experience (r $=$ 0.27, $p < 0.05$).

Independent t-test

Definition: Tests if the means of a variable differ between two independent groups.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{4}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{5}$$

```
/*independent t-test*/
ttest wage, by(collgrad)
ttest wage, by(collgrad) unequal
sdtest wage, by(collgrad)
```

Interpretation: The mean hourly wage of those with a college degree differs non-significantly/significantly from those without (t(2244) = –13, $p < 0.001$).

Analysis of Variance (ANOVA)

Definition: Tests for differences between more than two independent means.
Formula: F-ratio

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}} = \frac{MS_{between}}{MS_{within}} \tag{6}$$

```
/*anova*/
tab race, sum(wage)
anova wage race
pwcompare race, pveffects
```

Interpretation: There is a non-significant/significant difference in mean hourly wages across racial groups.

Chi-Squared Test

Definition: Tests the relationship between two categorical variables.
Formula: Chi-Squared Statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{7}$$

where:

- O_i is the observed frequency,
- E_i is the expected frequency.

```
/*chi-square test*/
tab union collgrad, col chi2
```

Interpretation: There is a non-significant/significant relationship between union membership and having a college degree.