Stata: Descriptive Analysis

Topics: Calculate means, medians, percentages, and 95% confidence intervals in Stata

The `tabulate` and `mean` commands are used for exploratory analysis, descriptive statistics, and when generating new variables.

1. Means, medians, and percentages

This lecture covers three descriptive statstics: means and medians which are used to summarize continuous data, and percentages which are used to summarize categorical data.

1a. Mean

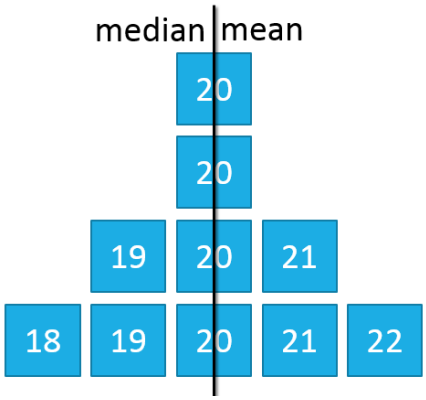
The mean, or average, is calcuated as the sum of values divided by the count of values. Means are used to describe variables that are normally distributed.

In words:

$$mean = \frac{sum\ (values)}{count\ (values)}$$

Formula:

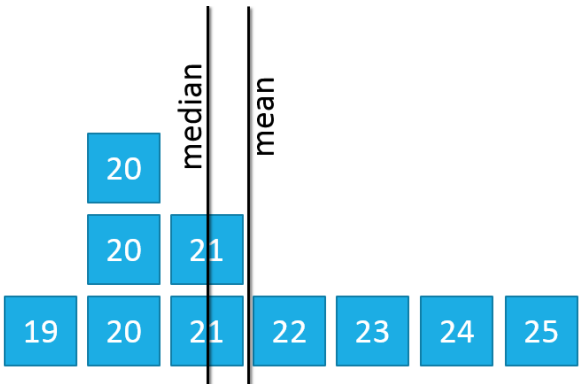
$$\bar{x} = \frac{\sum x_i}{N}$$



Normally distributed

Here is an example of a normally distrbuted variable; these are the ages of 10 university students. We see that their ages take on a bell shaped distrubtion. When the data are normally distrbuted, half of the data values are larger than the mean, and half of the data values are smaller than the mean.

1.b Medians



NOT normally distributed (skewed)

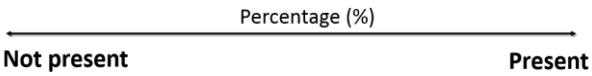
In variables that are not normally distrbuted, the values are skewed, or have one long tail. In non-normally distributed data, the mean is influenced by extreme values and therefore the median (the center value if you lined up all values from low to high) is a more useful descriptive statistic than the mean. If we lined up all values from lowest to highest, the median would be at the 50th percentile mark.

$$median = \frac{21+21}{2} = 21$$

If there are an odd number of values, the median will be the centeral number if you lined up the values from lowest ot hightest. If there is an even number of values, the median will be the average of the two central numbers.

1c. Percentages

Percentages are used to describe categorical data, and exist on a scale of 0-100. Percetages are not to be confused with proportions which exist on a scale of 0-1. The percentage is the proportion muliplied by 100.



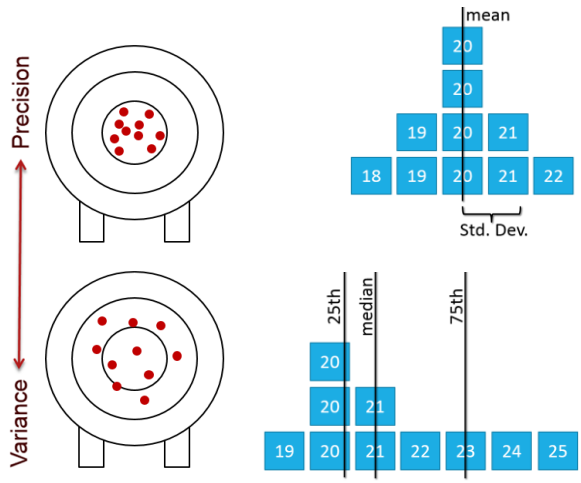
$$\% = \frac{\sum(present)}{N} * 100$$

Proportions are calculated from a binary response – a characteristic is present or not. A proportion is the mean of a binary response (it is the count of observations that have the characterstics of interest, divided by the total count of observations), and we multiply this by 100 to get the percentage.

General:	Categorical		
Specific:	Binary	Ordinal	Nominal
Example:	**Male** Yes (1) No (0)	**Education** None (0) Primary (1) Secondary (2) Higher (3)	**Marital** Married (1) Divorced (2) Widowed (3) Single (4)

Percentages are used to describe all categorical variables including binary, ordinal, and nominal variables. Binary values have two values, usually coded 0 and 1. Ordinal data are categories that follow a logical order like education level. And nominal data are categories that do not follow a logical order like marriage status. In variables that have more than two categories, we can think of each category as having a binary response and we calculate a percentage for each category.

2. Variance

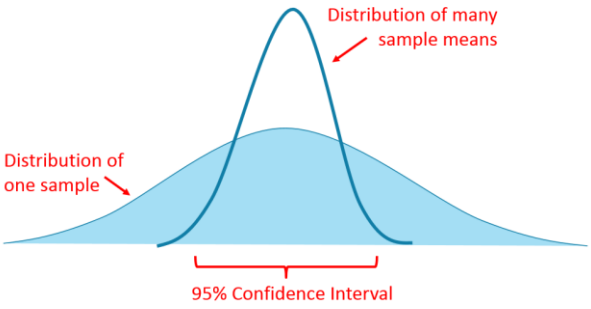


Variance is related to precision, or level of understanding. The variance describes the amount of spread around the mean of a variable.

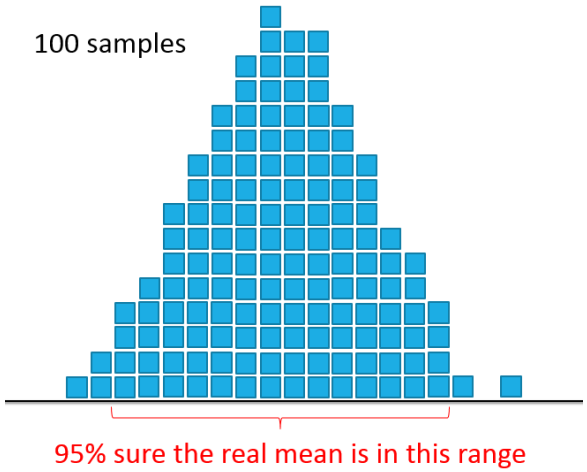
For normally distributed conintuous data, and for binary data, spread is summarized as a standard deviation around the mean.

For skewed continous data, spread is summarized as an interquartile range around a median. If you lined all values from smallest to largest, the interquartile range includes values between the 25th and 75th percentile.

2a. 95% Confidence Interval

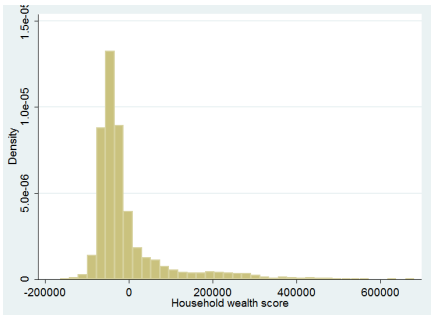


In survey data analysis, we usually do not present standard deviations or interquartile ranges; instead we present 95% confidence intervenals. While the confidence intervals are influenced by the spread of the data values in our one sample, it actually representes the spread of means in a hypothetical situation where we select multiple samples from the same population. We are concerned about this mean of sampled means because by chance any one sample might not perfectly represent the underlying population.

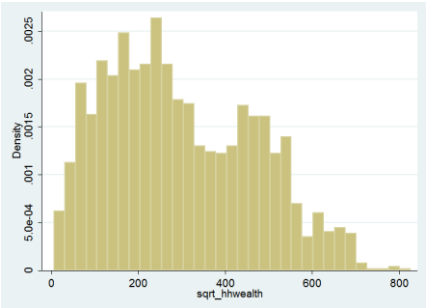


One way to think of the 95% confidence interval is to imagine that we drew 100 independent samples from the same population and made 100 different means. The mean in each sample would be slightly different due to chance. The 95% confidence internval would be the spread of the middle 95 means. The 95% confidence interval represents a range of values that we are almost certain contains the real mean in the population.

2b. How to analyze skewed continous values

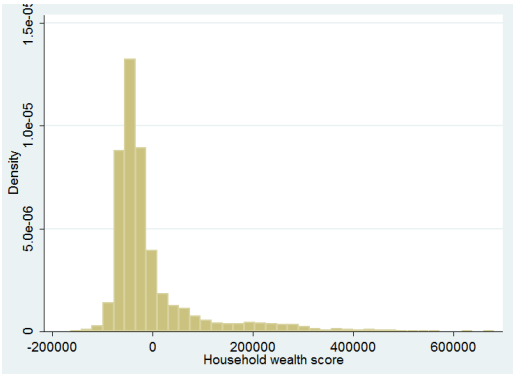



```
gen sqrt_hhwealth = sqrt(hhwealth)
hist sqrt_hhwealth
```



Stata calculates 95% confidnece intervals for means and percentages. Stata will assume that any continous variable follows a normal distribution. For continuous data that are NOT normally distributed, you should either:

(1) mathematically transform the values so they take on a normal distribution, for example by taking the square root, or




```
egen hhwealth_quintiles = cut(hhwealth), group(5)
svy: tab hhwealth_quintiles, percent ci
```

hhwealth_ quintiles	percentages	lb	ub
0	20.21	18.72	21.79
1	20.2	18.99	21.47
2	20.22	19.01	21.48
3	20.44	19.16	21.78
4	18.93	17.11	20.89
Total	100		

(2) cateogrize the values. Transformation of continous values is outside the scope of this course, and it is not typically done when analyzing categorical outcomes. I recommend that you categorize all values because it addresses any issues of skewed distributions, but it also simplifies interpretation which we will discuss further in the logistic regression videos.

3. Subpopulation

In non survey data, we can perform analyses in subpopulations by excluding (or dropping) observations that are not in the subpopulation. YOU CANNOT DROP OBSERVATIONS IN SURVEY DATA ANALYSIS, or your variance estimates will be biased.

In survey data analysis, we must analyze data according to the sample design. If we designed a sample to be representative of women age 15-49, but we are only analyzing women age 15-24, we still need to keep women age 25-49. The statistical reason is that Stata needs the total nubmer of observations in the dataset (the number of observations accoding to the sample design) in the denominator of the variance estimator. If we dropped some observations, the variance will be incorrect and could lead to incorrect answers to our research question.

```
* Survey Design
gen weight = v005/1000000
svyset [pweight=weight], psu(v021) strata(v023)

 * Kigali
 recode v023 (1/3 = 1 "Kigali")
        (4/30 = 0 "Not Kigali"), gen(kigali)

 *Mean (normal distribution), with subpopulation
 svy, subpop(kigali): mean mage
```

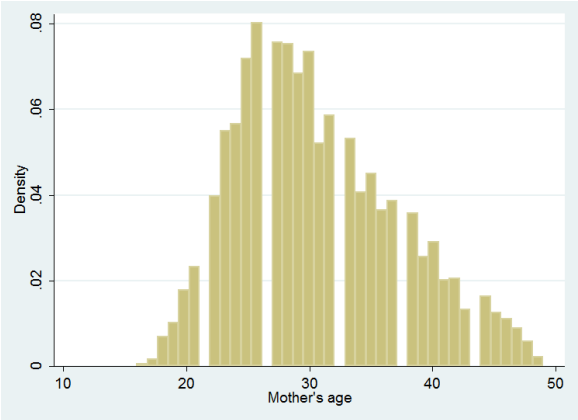
The correct way to perform a subpopulation analysis with survey data in Stata is to specify the subpopulation in a `svy:` statement. You create a new variable where 1 = in the subpopulation, and 0 = not in the subpopulation. This way Stata knows to use only the subpopulation data in the mean

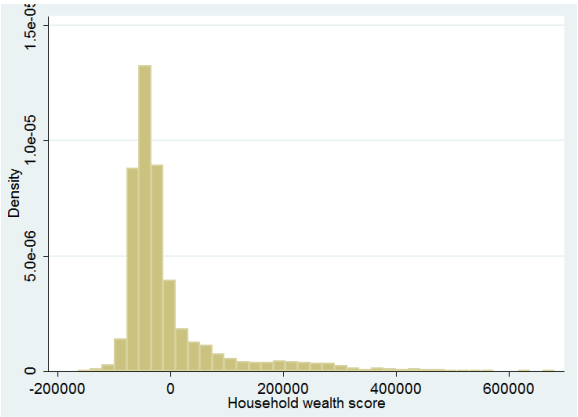
estimator, but can still count the total number of observations acording to the sample design for the variance estimator.

3a. Describe a continuous variable

Before performing descriptive analysis with survey data, we must specify the sample design in a `svyset` statement. Watch the Survey Stata Analysis video at www.populationsurveyanalysis.com to review this command. The svyset statement is absolutely essential before performing descriptive analysis with survey data.

```
*histogram to view distribution
histogram mage
histogram hhwealth
```





To describe a continuous variable in Stata, we start by looking at the shape of the distribution in a `histogram` so that we can decide whether to move forward with a mean or median calculation. We have the kid's recode file open. Let us compare two continuous variables – mother's age (mage) and household wealth score (hhscore) - so you can see the difference between a normal and skewed distribution.

Do you see how the distribution of mother's ages makes a nice bell shaped curve? This is what a normal distribution looks like. We can use either a mean or median to describe the center, or average, of this distribution. The mean is the conventional statistic used to describe a normally distributed variable.

Household wealth score, on the other hand, does not follow a normal distribution. These scores are calculated from a principle components analysis. The scores range from -163,000 to +680,000 – the scores themselves do not mean anything, the but relationship of the scores to each other are important. A household with the score of -2000 is wealthier than a household with a score of -20,000.

We see that the household wealth scores are skewed; the upper tail is much longer than the lower tail which means that there are a small number of people who are very rich, and a concentration of people who are poor. If we calculated mean household wealth score from this distribution, the mean score would be higher than 50% of the population because a few rich people are outliers and strongly influence the mean

upward. For skewed variables, either transform the values so they follow a normal distribution, or as I recommended earlier, categorize them.

```
* Mother's age
codebook v012
//check for missing coded as 9,
//or don't know coded as 8

*Household wealth score
codebook v191
//check for missing coded as 9,
//or don't know coded as 8
```

Before calculating the, we check the codebook for each variable to see how missing values and "don't know" responses are coded. Sometimes missing values are coded as 9 and don't know responses are coded as 8, and we should recode these as missing before performing descriptive analysis. Neither of these variables have missing or don't know categories, however, so we are okay to move forward.

```
*Mean (Mother's age)
svy: mean mage
```

```
.         svy: mean mage
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      30          Number of obs     =    9002
Number of PSUs   =     492          Population size   = 9136.94
                                    Design df         =     462

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        mage |   30.64164   .0956778      30.45362    30.82966
--------------------------------------------------------------
```

3a.1. Calculate the mean

Calculation of the mean is quite easy. Since we have specified the survey design above in a svyset statement, we can tell Stata to apply that survey design to our analysis with a `svy:` statement, then we type `mean` and the variable, in this case `mage`. The average age of mother's who have a child under five in Rwanda is 30.6 years rounded. We are 95% sure that the real mean age of mother's in the population is between 30.5 and 30.8 years.

```
*Mean (Mother's age, in Kigali)
svy, subpop(kigali): mean mage
```

```
.         svy, subpop(kigali): mean mage
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       3          Number of obs     =     990
Number of PSUs   =      60          Population size   = 871.623
                                    Subpop. no. obs   =     990
                                    Subpop. size      = 871.623
                                    Design df         =      57

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        mage |   29.60887   .2891831      29.02979    30.18795
--------------------------------------------------------------
Note: 27 strata omitted because they contain no subpopulation
      members.
```

3b. Describe a continuous variable with subpopulation

Describing a continuous variable in a subpopulation is easy. You need only add the subpopulation option to the `mean` command like this:

To calculate the mean and 95% confidence intervals in a subpopulation, write comma, `subpop()`, variable name after `svy` and before colon. So the mean age of mothers who have a child under five in Kigali (not all of Rwanda) is 29.6 years, with a 95% confidence interval of 29.0 and 30.2 years.

```
*Percentage (Cook with Charcoal)
svy: tabulate charcoal, percent ci
```

Cooking with charcoal	percentages	lb	ub
No	91.33	89.83	92.63
Yes	8.67	7.372	10.17
Total	100		

4a. Describe a categorical variable

We describe categorical variables with percentages using the `tabulate` statement. For example, to describe the percentage of children who live in households that cook with charcoal, we would type `svy: tabulate [variable name]`. To display percentages between 0 and 100 (rather than propotions between 0 and 1), we add the option `percent`, and to estimate the 95% confidence intervals we add the option `ci`. 8.7% of under-five households cook with charcoal in Rwanda.

```
*Percentage (Cook with Charcoal, in Kigali)
svy, subpop(kigali): tabulate charcoal, percent ci
```

Cooking with charcoal	percentages	lb	ub
No	35.34	25.31	46.85
Yes	64.66	53.15	74.69
Total	100		

4b. Describe a categorical variable in a subpopulation

To estimate a percentage in a subpopulation, simply add the `subpop` option after `svy` and before the colon. When we perform the same analysis in Kigali only, we find that 64.7% under-five households cook with charcoal.

The `mean` and `tabulate` commands allow you to describe the means and 95% confidence intervals of a single variable in the population, or a subpopulation.