How to perform a Descriptive statistics using the summarize command in Stata?

Authored by

stats writer

June 29, 2024

The descriptive statistics for the variables in the dataset are a set of numerical measures that summarize the characteristics and patterns of the data. These statistics are displayed in the Stata Annotated Output using the "summarize" command, which provides a comprehensive summary of the data's central tendency, dispersion, and shape. This includes measures such as the mean, median, standard deviation, and interquartile range. The Stata Annotated Output also displays the number of observations, missing values, and unique values for each variable, giving a complete overview of the data's distribution and variability. This information is important for understanding the data and identifying any potential outliers or unusual patterns. Overall, the descriptive statistics provided by the "summarize" command in Stata are essential for initial data exploration and can inform further data analysis and decision-making.

Descriptive statistics using the summarize command | Stata Annotated Output

This page shows an example of getting descriptive statistics using the summarize command with footnotes explaining the output. In the first example, we get the descriptive statistics for a 0/1 (dummy) variable called female. This variable is coded 1 if the student was female, and 0 otherwise. In the second example, we get the descriptive statistics for a continuous variable called write, which was the score students received on a writing test. We use the detail option to get additional information, including

percentiles, skewness and
kurtosis. You do not have to use the detail option with all
continuous variables.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2
(highschool and beyond (200 cases))

summarize female

Variablea| Obsb Meanc Std. Dev.d Mine Maxf
----------+---
female | 200 .545 .4992205 0 1

a. Variable - This column indicates which variable is being
described. You can list more than one variable after the summarize
command; when you do, you will see each variable on its own line of the output.

b. Obs - This column tells you the number of observations (or
cases) that were valid (i.e., not missing) for that

variable. If you had
200 observations in your data set, but you had 10 missing values for the
variable female, then the number in this column would be 190.

c. Mean - This is the mean of the variable. In this case, our variable female ranges from 0 to 1 (the min and max values), so the mean is
actually the proportion of observations coded as 1.

d. Std. Dev. - This is the standard deviation of the variable. This gives information regarding the spread of the distribution
of the variable.

summarize write, detail

writing score

Percentiles Smallesti
1%e 31 31
5% 35.5 31
10% 39 31 Obsb 200

25%f 45.5 31 Sum of Wgt.k 200

50%g 54 Meanc 52.775
Largestj Std. Dev.d 9.478586
75%h 60 67
90% 65 67 Variancel 89.84359
95% 65 67 Skewnessm -.4784158
99% 67 67 Kurtosisn 2.238527

e. 1% - This is the first percentile. Percentiles are calculated by ordering the values of a variable from lowest to highest, and then finding the value that corresponds to whatever percent you are interested in, in this case, 1%. Hence, 1% of the values of the variable write are equal to or less than 31.

f. 25% - This is the 25th percentile, also known as the first quartile.

g. 50% - This is the 50th percentile, also known as the median. If you order the values of the variable from lowest to highest,

the median would be the value exactly in the middle. In other words, half
of the values would be below the median, and half would be above. This is
a good measure of central tendency if the variable has outliers.

h. 75% - This is the 75th percentile, also known as the third quartile.

i. Smallest - This is a list of the four smallest values of the variable. In this example, the four smallest values are all 31.

j. Largest - This is a list of the four largest values of the variable. In this example, the four largest values are all 67.

b. Obs - This column tells you the number of observations (or
cases) that were valid (i.e., not missing) for that variable. If you had
200 observations in your data set, but you had 10 missing values for the
variable female, then the number in this column would

be 190.

k. Sum of Wgt. - This is the sum of the weights. In Stata, you can use different kinds of weights on your data. By default, each case (i.e., subject) is given a weight of 1. When this default is used, the sum of the weights will equal the number of observations.

c. Mean - This is the arithmetic mean across the observations. It is the most widely used measure of central tendency. It is commonly called the average. The mean is sensitive to extremely large or small values.

d. Std. Dev. - This is the standard deviation of the variable. This gives information regarding the spread of the distribution of the variable.

l. Variance - This is the standard deviation squared (i.e., raised to the second power). It is also a measure of spread of the

distribution.

m. Skewness - Skewness measures the degree and direction of
asymmetry. A symmetric distribution such as a normal distribution has a
skewness of 0, and a distribution that is skewed to the left, e.g., when the
mean is less than the median, has a negative skewness.

n. Kurtosis - Kurtosis is a measure of the heaviness of the
tails of a distribution. A normal distribution has a kurtosis of 3. Heavy tailed
distributions will have kurtosis greater than 3 and light tailed distributions
will have
kurtosis less than 3.