Introduction to Statistics - Young Researchers Fellowship Program

Lecture 1 - Introduction to Statistics & Tabular Data Logic

Daniel Sánchez Pazmiño

Laboratorio de Investigación para el Desarrollo del Ecuador

September 2024

What is statistics?

What is statistics?

- A **methodology** for collecting, analyzing, interpreting, and presenting numerical information.
- A statistic is often referred to as a **numerical fact** or a **piece of data** which describes a particular characteristic of a group of individuals.
  - In the field of statistics, we typically don't refer to individual data points as statistics.
- In several fields, statistics is used as an aid to decision making under uncertainty.
- In a research context, statistics will be needed to understand phenomena, make predictions, and test hypotheses emerging from theory.
- *Statistics is the systematic investigation of the correspondence of theory with the real world*

Data in statistics

Data in statistics

- Because statistics is concerned with information, **data** is often the starting point of any statistical analysis.
- No clear definition of data can possibly satisfy everyone, but we can think of data as a collection of **facts** to be analyzed.
  - Data is **plural** for **datum**.
- A **dataset** is a collection of data points, which can be organized in a **table**, often about a specific topic, purpose, experiment, study, or context.

Broad types of data

- Typically, statistics makes a distinction between two broad types of data:

1. **Quantitative** data, which is numerical in nature, meaning it can be measured and expressed in numbers.
   - Discrete data: whole numbers (e.g., number of students in a class).
   - Continuous data: real numbers (e.g., height, weight).
2. **Categorical** or "qualitative" data, which is non-numerical in nature, meaning it cannot be directly measured or expressed in numbers.
   - Nominal data: categories without order (e.g., colors).
   - Ordinal data: categories with order (e.g., levels of satisfaction).

How a dataset might look like

- What type of data can be identified for each column in the dataset?

Types of datasets

- Datasets can be classified into different types based on the nature of the data they contain.

1. **Cross-sectional data**: data collected at a single point in time.
   - Example: a survey conducted in 2024.
2. **Time series data**: data collected over time at regular intervals, for a single entity
   - Example: monthly sales data from 2020 to 2024.
3. **Panel data**: data collected over time for multiple entities.
   - Example: monthly sales data from 2020 to 2024 for multiple companies.

- Often, we also hear about repeated or pooled cross-sectional data, which is a combination of cross-sectional and time series data; we observe multiple cross-sections at different points in time.

Tabular data logic

Tabular data logic

- A dataset is typically organized in a **table** with rows and columns.
- Datasets often collect characteristics of individuals or entities, which are typically referred to as **elements**.
  - Elements are not necessarily the observations in a dataset, elements are those entities or individuals for which we hold information.
- When data is *tidy*, a table structure typically allows for easy identification of the following elements:
  - We will talk more about **tidy data** concept in the R companion module.

1. **Variables**: columns in the table, which represent a characteristic of an element.
2. **Observations**: rows in the table, which represent a collection of variable values for a single element.

Elements are not observations

- Sometimes, elements in a dataset (for the SUPERCIAS dataset, companies) are not the same as observations.
- We may observe multiple observations for a single element, which is why we need to be clear about the distinction between elements and observations.
- When we observe multiple observations for a single element, we typically refer to this as **repeated measures** or **panel data**.
- It is in this context when it comes in handy to difference between long and wide format datasets.

Long vs. wide format

- Long vs. wide format refers to the way data is organized in a table.
- In the **long format**, each row represents a single observation, and each column represents a variable.
- In the **wide format**, each row typically represents a single element. Columns may represent variables, but also repeated measures or time points of the same variable.

Example of long vs. wide format - business creation per province

- Long format: each row represents a single observation (business creation per province per year).

Table 1: Long format business creation per province (SUPERCIAS)

Province	Year	Number of businesses created
ORELLANA	2013	82
AZUAY	2023	1267
IMBABURA	2012	97
COTOPAXI	2024	174
EL ORO	2010	176

- Notice that if a province creates businesses in multiple years, we will have multiple rows for the same province.

Example of long vs. wide format - business creation per province

- Wide format: each row represents a single element (province), and columns represent variables (business creation per year).

Table 2: Wide format business creation per province (SUPERCIAS)

Province	2013	2023	2012	2024	2010
ORELLANA	82	NA	NA	NA	NA
AZUAY	NA	1267	NA	NA	NA
IMBABURA	NA	NA	97	NA	NA
COTOPAXI	NA	NA	NA	174	NA
EL ORO	NA	NA	NA	NA	176

- We will never have multiple rows for the same province in the wide format.

Sources of data

Where do we get data for statistical analysis?

1. Experimental studies
2. Observational studies
3. Secondary sources (existing datasets)

Descriptive statistics vs. statistical inference

Descriptive statistics vs. statistical inference

- **Descriptive statistics** summarize and describe the main features of a dataset.
  - Descriptive statistics are used to describe the data as it is.
  - Examples: mean, median, mode, standard deviation, variance, etc.
  - Tables and visualizations are commonly used as well
- **Statistical inference** is the process of making predictions or inferences about a population based on a sample.
  - A population is the entire group of individuals or entities we are interested in.
  - A sample is a subset of the population.

Why even care about descriptive statistics?

- Sometimes we do have the entire population data, but we still use descriptive statistics.
  - E.g. a small population such as this group of students.
- Sometimes gathering data from the entire population is not feasible, so we use a sample to make inferences about the population.
- Descriptive statistics should be first understood well before moving to statistical inference.
  - We later use descriptive statistics to summarize the sample data and make inferences about the population based on the sample statistics
  - We will require an understanding of probability theory to make inferences about the population with the descriptive statistics of the sample.

Notation

- In statistics, we often use the following notation:
  - n: the number of observations in a dataset.
  - x_i: the i-th observation in a dataset.
- We often want to underscore the difference between a descriptive statistic calculated for a sample or a population.
  - For a sample, we use a lowercase letter to denote the statistic (e.g., $\bar{x}$ for the sample mean). We call this a sample statistic.
  - For a population, we use a Greek letter to denote the statistic (e.g., μ for the population mean). We call this a population parameter.
- In the rest of these slides, I will use the sample notation for simplicity. We will later discuss the difference between sample statistics and population parameters as we move to statistical inference.

Univariate descriptive statistics

Univariate data

- **Univariate data** refers to data that consists of a single variable or attribute.
- We are interested in summarizing single variables in a dataset.
- We'll later discuss bivariate descriptive statistics, which summarize the relationship between two variables.

Central tendency measures

Measures of central tendency are values that represent the center of a data set.

- **Mean**: the average of the data
- **Median**: the value that divides the data into two equal parts
- **Mode**: the value that appears most frequently

The mean $\bar{x}$

- The average or, more specifically, the **arithmetic mean** is the sum of all observations divided by the number of observations.
  - There are other types of means, such as the geometric mean, harmonic mean, etc.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \left(x_1 + x_2 + ... + x_n \right)$$

- The mean is sensitive to outliers, so use with caution.
  - Very large or very small values can make the mean less representative of a typical value in the dataset.

Software implementation for the mean

- Use the base R function `mean()` to calculate the mean of a variable.
- You may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.

```
# Calculate the mean of the capital suscrito variable

mean_capital_suscrito <- mean(supercias_raw$capital_suscrito,
                              na.rm = TRUE)

mean_capital_suscrito
```

```
[1] 136384.3
```

Median

- The median is the value that divides the data into two equal parts.
- You must first sort the data from smallest to largest.
  - If the number of observations is odd, the median is the middle value.
  - If the number of observations is even, the median is the average of the two middle values.
- The median is less sensitive to outliers than the mean.
  - Typically, income data is reported using the median because it is less affected by extreme values.

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right) & \text{if } n \text{ is even} \end{cases}$$

Software implementation for the median

- Use the base R function `median()` to calculate the median of a variable.
- You may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.

```
# Calculate the median of the capital suscrito variable

median_capital_suscrito <- median(supercias$capital_suscrito,
                                  na.rm = TRUE)

median_capital_suscrito
```

```
[1] 800
```

Mode

- The mode is the value that appears most frequently in the dataset.
- A dataset can have multiple modes if multiple values appear with the same frequency.
- The mode is not always defined, and it may not be unique.
- The mode is less commonly used than the mean and median.

Software implementation for the mode

- There is no built-in function in base R to calculate the mode.
- You can look for a function in a package or use the `table()` function to count the frequency of each value, then find the maximum frequency.

```
# Calculate the mode of the capital suscrito variable

mode_capital_suscrito <- table(supercias$capital_suscrito)
```

Software implementation for the mode

Var1	Freq
0.04	2
0.1	1
0.2	2
0.32	4
0.4	32
0.48	2

Dispersion or variability measures

- **Dispersion** or **variability** measures describe how spread out the data is, or how much the data points differ from their central tendency.
- Common measures of dispersion include:
  - **Range**: the difference between the maximum and minimum values.
  - **Variance**: the average of the squared differences from the mean.
  - **Standard deviation**: the square root of the variance.
  - **Variation coefficient**: the standard deviation divided by the mean.

Range

- The range is the difference between the maximum and minimum values in a dataset.

$$\text{Range} = \text{Max} - \text{Min}$$

- The range is sensitive to outliers, so use with caution.
- Use `diff(range(x))` in R to calculate the range of a variable.
  - Use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.
  - `range()` returns a vector with the minimum and maximum values, and `diff()` calculates the difference between them.

Variance and standard deviation

- The variance is the average of the squared differences from the mean.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right]$$

- The standard deviation is the square root of the variance.

$$s = \sqrt{s^2}$$

- The standard deviation is in the same units as the data, unlike the variance.
  - The variance is squared, so it cannot be directly interpreted; higher variance means more spread out data, but not much else.

Implementation in R

- Use the base R function `var()` to calculate the variance of a variable.
- Use the base R function `sd()` to calculate the standard deviation of a variable.
- You may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.

Implementation in R

```
# Calculate the variance of the capital suscrito variable

range(supercias$capital_suscrito, na.rm = TRUE)
```

```
[1] 4e-02 9e+08
```

```
diff(range(supercias$capital_suscrito, na.rm = TRUE))
```

```
[1] 9e+08
```

```
var(supercias$capital_suscrito, na.rm = TRUE)
```

```
[1] 1.844574e+13
```

```
sd(supercias$capital_suscrito, na.rm = TRUE)
```

```
[1] 4294851
```

Coefficient of variation

- The coefficient of variation is the standard deviation divided by the mean.

$$\text{CV} = \frac{s}{\bar{x}}$$

- The coefficient of variation is a relative measure of dispersion.
  - It is useful when comparing the variability of two variables with different units or scales.
- The coefficient of variation is expressed as a percentage or a proportion (decimal).
- No direct function in R, but you can calculate it manually by dividing the standard deviation by the mean.

Measures of position

- Measures of position describe the relative position of a data point within a dataset.
- Common measures of position include:
  - **Percentiles**: the value below which a given percentage of observations fall.
  - **Quartiles**: values that divide the data into four equal parts.
  - **Deciles**: values that divide the data into ten equal parts.
  - **Z-scores**: the number of standard deviations a data point is from the mean.

Percentiles

- Percentiles are values below which a given percentage p of observations fall.
- The p-th percentile is the value below which $p\%$ of the data fall.
- Again, we must first sort the data from smallest to largest before calculating percentiles. Then apply the formula:

$$\text{Percentile}_p = \left(\frac{p}{100}\right)(n+1)$$

- The median is the 50th percentile.

Quartiles and deciles

- **Quartiles** are values that divide the data into four equal parts.
  - The first quartile (Q1) is the 25th percentile.
  - The second quartile (Q2) is the median.
  - The third quartile (Q3) is the 75th percentile.
- **Deciles** are values that divide the data into ten equal parts.
  - The first decile (D1) is the 10th percentile.
  - The second decile (D2) is the 20th percentile.
  - And so on...
- The "five-number summary" includes the minimum, Q1, median, Q3, and maximum.
  - Provides a quick summary of the data distribution.

The interquartile range (IQR)

- The **interquartile range (IQR)** is the difference between the third and first quartiles.
  - Technically a measure of spread/variability, yet requires quartiles to calculate.

$$\text{IQR} = Q3 - Q1$$

- The IQR is less sensitive to outliers than the range.
  - It is a better measure of spread for datasets with very large or very small values.

Z-scores

- The Z-score is the number of standard deviations an observation is from the mean.
- The Z-score is calculated as:

$$Z = \frac{x - \bar{x}}{s}$$

- The Z-score tells us how many standard deviations a data point is from the mean.
  - A Z-score of 0 means the data point is at the mean.
  - A Z-score of 1 means the data point is one standard deviation above the mean.
- The Z-score is useful for identifying outliers in a dataset.

Outliers

- An **outlier** is an observation that is significantly different from other observations in the dataset.
- Outliers can be due to errors in data collection, measurement error, or a true anomaly in the data.
  - Either very large or very small values.
- Outliers can significantly affect the mean and standard deviation, making them less representative of the data.
- The Z-score is useful for identifying outliers in a dataset.
  - Typically, a Z-score greater than 3 or less than -3 is considered an outlier (we'll discuss this later).

R implementation for position measures

- Use the base R function `quantile()` to calculate percentiles, which will also calculate quartiles, deciles, etc.
- The `probs` argument specifies the percentiles to calculate.
- You may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.
- Use `IQR(x)` to calculate the interquartile range.

R implementation for position measures

```
# Calculate the 25th, 50th, and 75th percentiles of the capital

quantile(supercias$capital_suscrito, probs = c(0.25, 0.5, 0.75),
```

```
 25%  50%  75%
 400  800 1200
```

```
# Calculate the interquartile range of the capital suscrito vari

IQR(supercias$capital_suscrito, na.rm = TRUE)
```

```
[1] 800
```

R implementation for position measures

- To calculate the Z-score, we use the function `scale()`.
- The `center` argument specifies the value to center the data around (the mean by default).
- The `scale` argument specifies the value to scale the data by (the standard deviation by default).
- You may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.
- Extract a value you're interested in from the resulting object, or manually calculate the Z-score using the formula.

R implementation for position measures

```
# Calculate the Z-scores of the capital suscrito variable

z_scores <- scale(supercias$capital_suscrito)

head(z_scores)
```

```
             [,1]
[1,] -0.02053256
[2,] 16.84892490
[3,]  0.87314625
[4,] -0.01545671
[5,] -0.02802991
[6,]  0.08466318
```

Measures of distributional shape

- These descriptive statistics describe the shape of the distribution of the data.
- By distribution, we mean the way the data is spread out or distributed across the range of values.
- Common measures of distributional shape include:
  - **Skewness**: a measure of the asymmetry of the data.
  - **Kurtosis**: a measure of the "peakedness" of the data.

Skewness

- **Skewness** is a measure of the asymmetry of the data distribution.
- A distribution is **symmetric** if the left and right sides are mirror images of each other.
- A distribution is **positively skewed** if the right tail is longer than the left tail.
  - The mean is greater than the median.
- A distribution is **negatively skewed** if the left tail is longer than the right tail.
  - The mean is less than the median.

Skewnness

- The skewness of a dataset is calculated as:

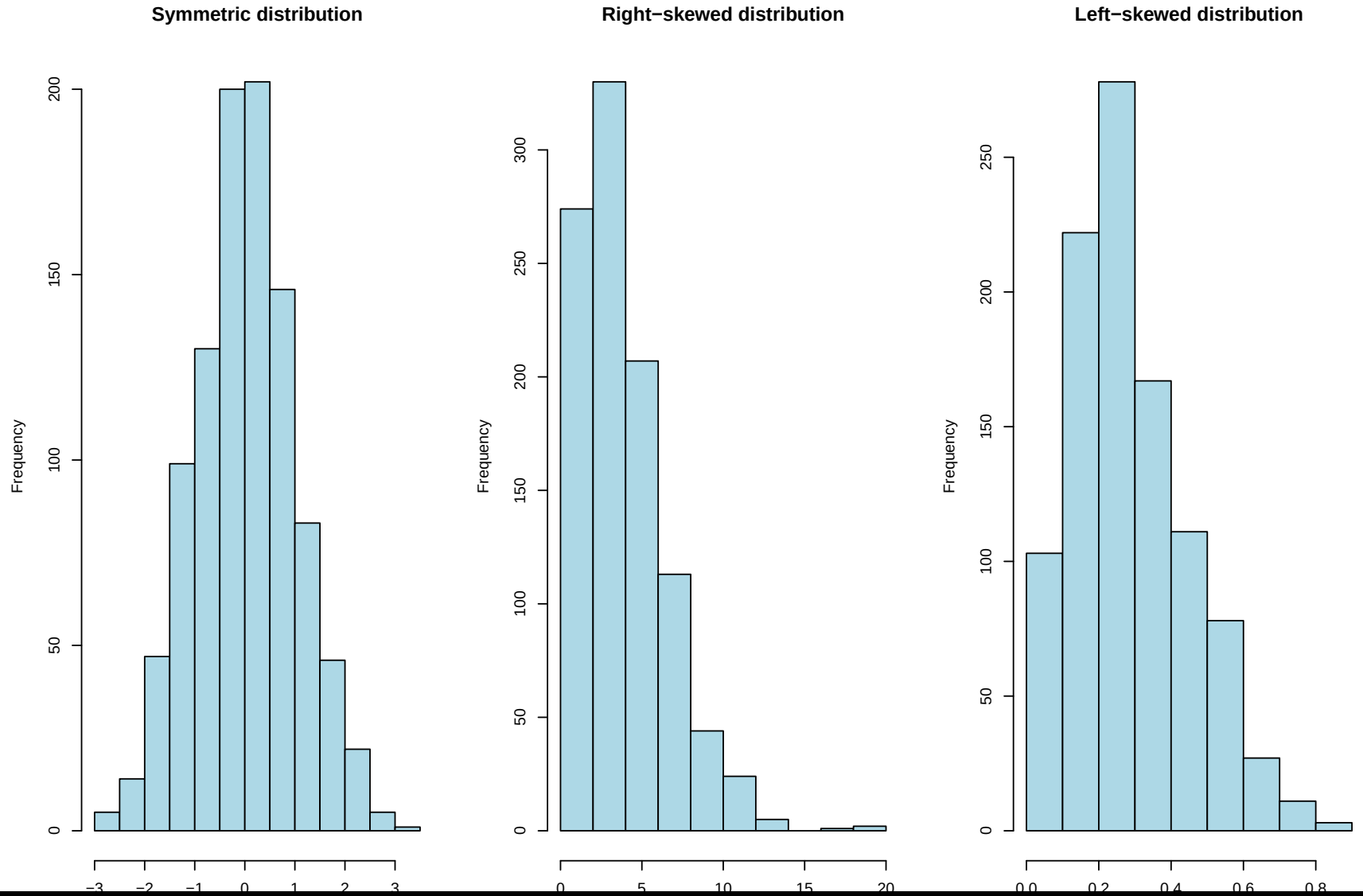
$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

- The skewness is 0 for a symmetric distribution.
  - Positive skewness indicates a right-skewed distribution.
  - Negative skewness indicates a left-skewed distribution.
- The skewness is not as commonly used as the mean, median, and standard deviation.

How skewed distributions look like

- **Symmetric distribution**: mean = median = mode.
- **Right-skewed distribution**: mean > median > mode.
- **Left-skewed distribution**: mean < median < mode.

How skewed distributions look like



Kurtosis

- **Kurtosis** is a measure of the "peakedness" of the data distribution.
- A distribution with high kurtosis has a sharp peak and fat tails.
  - This is called a **leptokurtic** distribution.
  - The tails are heavier than a normal distribution.
- A distribution with low kurtosis has a flat peak and thin tails.
  - This is called a **platykurtic** distribution.
  - The tails are lighter than a normal distribution.

Kurtosis

- The kurtosis of a dataset is calculated as:

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

- The kurtosis of a normal distribution is 3.
  - A distribution with kurtosis greater than 3 is leptokurtic.
  - A distribution with kurtosis less than 3 is platykurtic.

R implementation for shape measures

- There are several packages which provide functions to calculate skewness and kurtosis.
- The `e1071` package provides the `skewness()` and `kurtosis()` functions.
- The `moments` package provides the `skewness()` and `kurtosis()` functions.
- Base R does not have built-in functions to calculate skewness and kurtosis.
- Once again, you may need to use the `na.rm = TRUE` argument to remove missing values, depending on the dataset.

Stata implementations

- You can get many of these statistics in Stata using the `summarize` command.
  - `summarize` will give you the mean, median, and other statistics for all variables in the dataset.
  - You can use the `detail` option to get more detailed statistics, which include the range, IQR, among others.

```
summarize capital_suscrito, detail
```

Stata implementations

- For the coefficient of variation, you can calculate it manually by dividing the standard deviation by the mean, using the `egen` command to create a new variable.

```
egen m = mean(capital_suscrito)
egen s = sd(capital_suscrito)
gen cv = s / m
```

- Any generated variable can be viewed using the `list` command or by using the `browse` command.

Stata implementations

- To calculate percentiles, you can use the `pctile` command.
- The `pctile` command will calculate the specified percentiles for the variable.
- It requires a new variable name, followed by the variable you're getting the percentiles for, and the number of "divisions" you want.
- Include `return list` to get the results in a list.

```
pctile capital_suscrito_quart = capital_suscrito, nq(4)
return list
```

Stata implementations

- Generate a new variable using `gen` to calculate Z-scores for all observations.

```
generate zscore = capital_suscrito - r(mean)/ r(sd)
```

Stata implementations

- To calculate skewness and kurtosis, you can use the `sktest` command.
- The `sktest` command will calculate the skewness and kurtosis of the variable.

```
sktest capital_suscrito
```