Notes

Tuesday, August 9, 2022 8:06 PM

Myroslava Sánchez Andrade A01730712 | 09/08/2022

Descriptive Statistics

Used to summarize information from raw data.
Most important descriptive statistics:

- Central tendency measures

Attempts to describe a data set with a single value which represents the middle or center of its distribution.
Main central tendency measures:

- Arithmetic mean

Average value of valid values of a variable X (assuming each value has the same importance), being X an attribute of a subject.

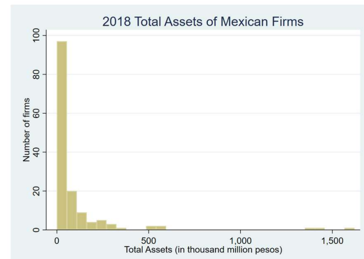
$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

- Median

The median of a variable is the its 50 percentile, mid-point of its values sorted in descending order. If two middle points, the median will be the arithmetic average of them.

When a variable **does not follow a probability distribution** close to normal distribution, the **best measure** for central tendency is the **median**.

The mean is sensible to extreme values.



Histogram skewed to the right

- Mode

Value that most appear in a variable (calculated for discrete variables).

- Dispersion measures

Used to measure how much on average the individual values of a variable change from the mean. Variance and standard deviation reflect variability in a distribution.

- Variance and the standard deviation

The variance of a variable X is the average of squared deviations (difference between the observed values of a variable and some other value) from each individual value X_i from its mean:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sigma_X^2$$

The variance is expressed in much larger units

Where:

X_i = Value i of the variable X

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = Arithmetic average of X

It is more used the sample variance (denominator n-1), instead of the population variance (denominator n). The *sample variance* is a more conservative value of the variance.

Rewriting the formula:

$$\mathrm{Var}(X) = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sigma_X^2$$

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\mathrm{SD}(X) = \frac{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{(n-1)}} = \sigma_X$$

The standard deviation is expressed in the same units as the original values

Data management

- Data transformations

- Return calculation

The return of a price is the % change of the price from one period (present period t) to the next (previous period t-1).

$$R_t = \frac{(price_t - price_{t-1})}{price_{t-1}} = \frac{price_t}{price_{t-1}} - 1$$

It is very recommended to calculate continuously compounded returns (cc returns) and cc returns instead of simple returns. Cc returns are calculated from the natural logarithm of prices.

- Natural logarithm

The natural logarithm of a number is the **exponent** that the number *e* (=2.71…) needs to be raised to get another number. The natural logarithm is the logarithm of base *e*.

Relation of *e* and the grow of financial amounts over time:
The general formula to get the final amount of an investment at the beginning of year x=1, for any interest rate R can be:

$$I_2 = I_1 * (1 + R)^1$$

The (1+R) is the growth factor of the investment.
But, if the interests are calculated each month, the investment would end up with a higher amount. The general formula would end up like this:

$$I_2 = I_1 * (\ 1 + \frac{R}{N})^{1*N}$$

A **continuously compounded** rate would give as a result the Euler constant for the growth factor.
We can generalize annual interest rate, so that e^R is the growth factor when interests are compunded every moment. On the other hand, when compounding every instant we use e^r . The relationship between the growth rate and an effective equivalent rate would be:

$$EffectiveRate = e^r - 1$$

- Continuously compounded returns

One way to calculate it is subtracting the current price(t) minus the log of the previous price (t-1). (Difference of the log of the price):

$$r_t = \log(price_t) - \log(price_{t-1})$$

Other way would be:

$$r_t = \log(\ \frac{price_t}{price_{t-1}})$$

Histogram

Illustrates how the values of a variable are distributed in its range of values (frequency plot). The most common values, least common values, the possible mean and standard deviation can be appreciated.

Probability Density Functions

- PDF of a discrete variable

Sum of probabilities of x to be equal to a specific value.

$$f(x) = P(X = x_i)$$

We can express the Cumulative Density Function (probability that x will take values less than or equal to x) as:

$$f(x) = \sum_{i=1}^{n} P(X = x_i)$$

- PDF of a continuous variable

Integration of the function f(x), where f(x) is the PDF. We calculate the probability of the continuous variable x to be within a specific range.

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{a}^{b} f(x), dx = P(a \le x \le b)$$

Normal Distribution Function

The most popular continuous PDF is the well-known "bell-shaped" normal distribution defined as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})}$$

Where u is the mean of the distribution and ʊ squared is the variance of the distribution. The only two parameters to be defined in order to know the behavior of the continuous random variable x are: the mean of x and the variance of x. The normal distribution is normal around u.

• For the range $(\mu - \sigma) <= x <= (\mu + \sigma)$, the area under the curve is approximately 68%

• For the range $(\mu - 2\sigma) <= x <= (\mu + 2\sigma)$, the area under the curve is approximately 95%

• For the range $(\mu - 3\sigma) <= x <= (\mu + 3\sigma)$, the area under the curve is approximately 99.7%.