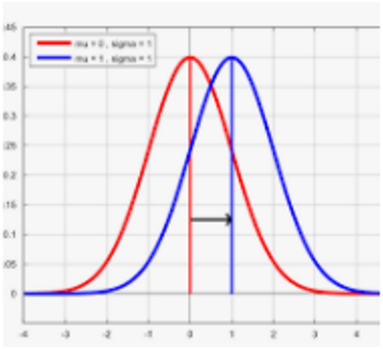
Inferential Statistics

Wednesday, May 26, 2021 4:49 PM

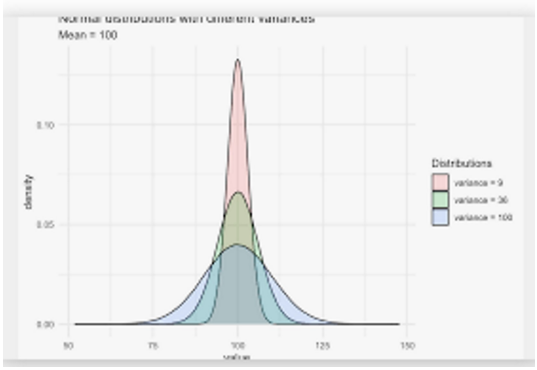
Features of Normal Distribution :

1. Symmetric Distribution
2. Bell Shaped Curve
3. X ~ N (μ , σ2)

- Change mean and keep std. constant :

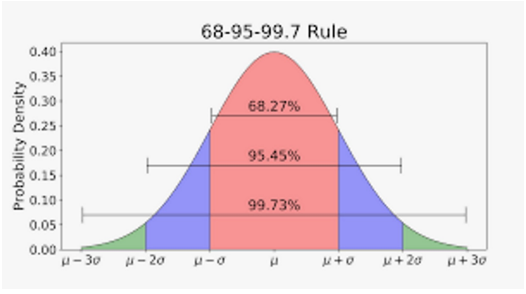


- Change std. and keep mean constant :



4. Gaussian Distribution

5.68-95-99.7 rule :

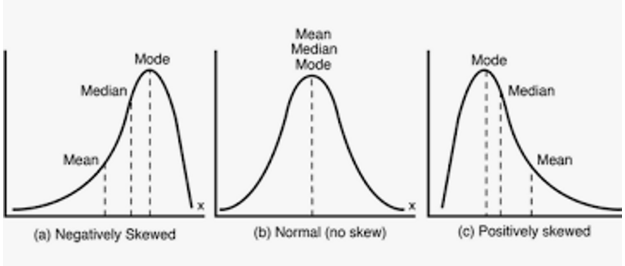


Skewness

In statistics, skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. In other words, skewness tells you the amount and direction of skew (departure from horizontal symmetry). The skewness value can be positive or negative, or even undefined. If skewness is 0, the data are perfectly symmetrical, although it is quite unlikely for real-world data. As a general rule of thumb:

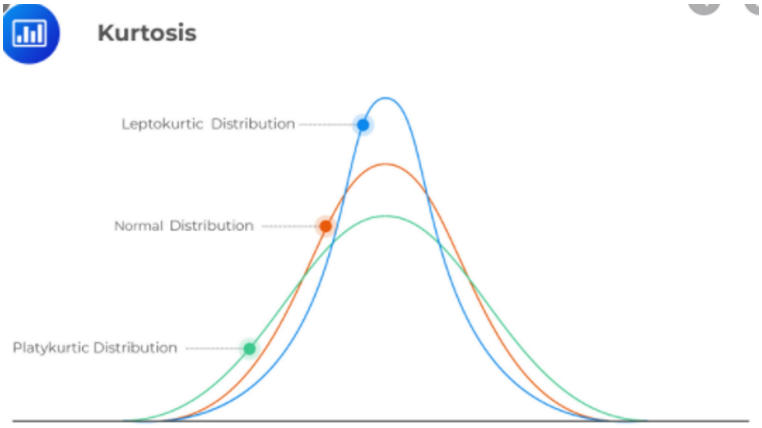
- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.

- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.



Kurtosis

Kurtosis tells you the height and sharpness of the central peak, relative to that of a standard bell curve.



- If kurtosis = +ve, it means distribution has peak.
- If kurtosis = -ve, it means distribution has flat.

Standard normal distribution

The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1. Any normal distribution can be standardized by converting its values into z-scores. The step is called Standardization step.

Z-score :

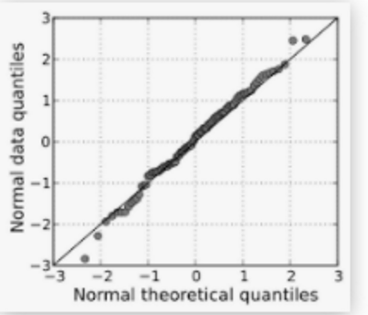
It means how much std. point away is a data point from the mean. Its formula is :

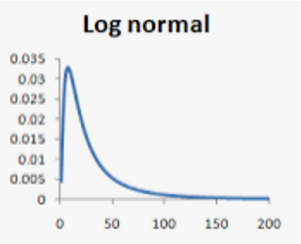
$$z = (x - \mu) / \sigma$$

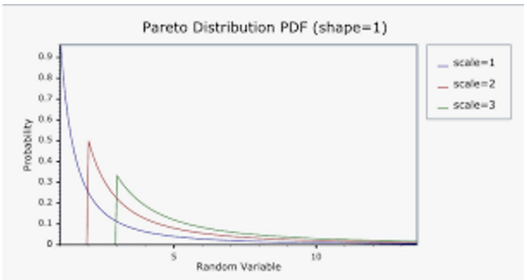
Test of Normality :

- QQ Plot :

  A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. We need two columns : theoretical Quantity and Observed Quantity.







SciPy's stats package provides a function called Boxcox for performing box-cox power transformation that takes in original non-normal data as input and returns fitted data along with the lambda value that was used to fit the non-normal distribution to normal distribution.

Population and Sample Data

A population data set contains all members of a specified group (the entire list of possible data values). A sample data set contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken.

Table 1: Mean and Standard Deviation				
Measure Name	Symbol for Population	Symbol for Sample	Computation for Population	Computation for Sample
Mean	μ	$\overline{x}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Standard Deviation	σ	s_x	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$ or the equivalent form $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x^2) - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N}}$	$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$ or the equivalent form $s_x = \sqrt{\frac{\sum_{i=1}^{n} (x^2) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$

Note: The (n-1) in sample std. is for removing any biasness in the data

What Is a Sampling Error?

A sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data. As a result, the results found in the sample do not represent the results that would be obtained from the entire population.
Sampling is an analysis performed by selecting a number of observations from a larger population. The method of selection can produce both sampling errors and non-sampling errors.

Types of Sampling :

1. Volunteer Sampling :
2. Convenience Sampling :
3. Uniform Random Sampling :

Central Limit Theorem (apply only when sigma pop. present)

The mean of the sampling distribution:

The mean of the sampling distribution of a sample mean $\bar{x}$ is equal to the population mean:

$$\mu_{\bar{x}} = \mu$$

The standard deviation of the sampling distribution:

$$\sigma_x^2 = \frac{\sigma^2}{n} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

also called standard error.
where n : no of data points in individual sample

- The mean of sampling distribution is always follows normal dist.
- The mean of sampling dist. Is always a point estimation but its very close.

- $\bar{X} \sim N(\mu_{\bar{x}}, \frac{\sigma^2}{n})$
- It should follow 68-95-99.7% rule, it means
  - $\mu_{pop} = [\mu_{\bar{x}} - 2\frac{\sigma}{\sqrt{n}}, \mu_{\bar{x}} + 2\frac{\sigma}{\sqrt{n}}]$ with 95 % confidence
  - Z* - {1,2,3}
  - The +/- region is called margin of error.
- We cannot use point estimate. We have to use confidence interval using std. because it can contains biasness due to presence of outliers. That's why we can come up with C.I. with some y% confidence.

T-score vs. z-score

- Has a sample size below 30,
- Has an unknown population standard deviation.

Like z-scores, t-scores are also a conversion of individual scores into a standard form. However, t-scores are used **when you don't know the population standard deviation**; You make an estimate by using your sample.

T = (X – μ) / [s/√(n)].

Where:

- s is the standard deviation of the sample.

Formula used in C.I. when T-score is used,
Where v : N-1 which is degree of freedom
and α : significance level and
α/2 : critical value

$$\bar{x} - (t_{1-\alpha/2,v})\frac{S}{\sqrt{N}}$$

$$\bar{x} + (t_{1-\alpha/2,v})\frac{S}{\sqrt{N}}$$

Understanding the Hypothesis Testing

Step - 1:
Alternate Hypothesis (Bold claim): $H_1 \Rightarrow >, <, \neq$
Null Hypothesis (Status Quo): $H_0 \Rightarrow \leq, \geq, =$

Step - 2:

- Collect a sample of size n
- Compute the mean from this sample $\bar{x}$

Step - 3: Compute Test Statistic:

- If population variance is known

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt[2]{n}}$$

- If population variance is unknown

$$t = \frac{\bar{x} - \mu}{s / \sqrt[2]{n}}$$

Step - 4: Decide significance level α. Lower α means you need stronger evidence to reject Null Hypothesis.

Step - 5.1: Apply decision rule:

- If test statistic is z-score -
  - Two tailed z-test:

    $$|z|>z_{\frac{\alpha}{2}} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

  - Right tailed z-test:

    $$z>z_{\alpha} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

  - Left tailed z-test:

    $$z<-z_{\alpha} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

- If test statistic is t-score
  - Two tailed t-test:

    $$|t|>t_{n-1,\ \frac{\alpha}{2}} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

  - Right tailed t-test:

    $$t>t_{n-1,\ \alpha} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

  - Left tailed t-test:

    $$t<t_{n-1,\ \alpha} \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

Step - 5.2: Compute p-value $P(Test\ Statistics|H_0)$

- For two tailed test:

  $$p\ value = 2*(1.0 - cdf(test\ statistic))$$

- For one tailed test:

  $$p\ value = (1.0 - cdf(test\ statistic))$$

  Now,

  $$if(p\ value<\alpha) \Rightarrow Accept\ H_1 or\ Reject\ H_0$$

CDF: The **cumulative distribution function** (**CDF**) of a real-valued random variable , or just **distribution function** of X , evaluated at x, is the probability that will take a value less than or equal to x .

$$F_X(x) = \mathrm{P}(X \leq x) \quad \text{(Eq.1)}$$

- Why do we need to standardize data?

  Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format. Standardized values are useful for tracking data that isn't easy to compare otherwise