Revision Session - DS22

Inceptez

Statistics

Definitions :

- **Descriptive Statistics** - procedures used to organize and present data in a convenient, usable and communicable form
- **Mean** - Average value of a sample or population

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

N = number of items in the population

Sample Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n = number of items in the sample

- **Weighted mean** - Sum of a set of observations multiplied by their respective weights, divided by the sum of the weights

$$\bar{x}_w = \frac{\sum_{i=1}^{n}(w_i x_i)}{\sum_{i=1}^{n}(w_i)}$$

where as

$\bar{x}_w$ is the weighted mean variable

w_i is the allocated weighted vlue

x_i is the obsrved values

- **Median** - Value at the centre
- **Mode** - Value that occurs most

Statistics

- **Variance -** The average of square differences between observations and their mean

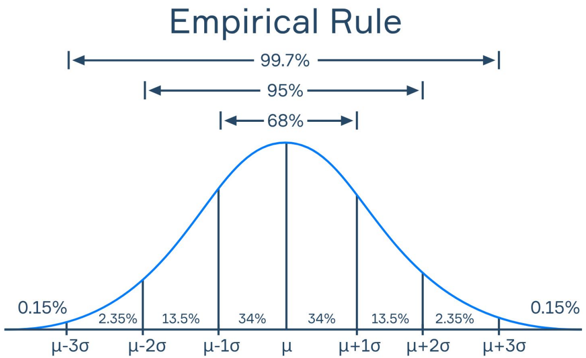
$$\sigma^2 = \sum (X_i - \bar{X})^2 / N$$

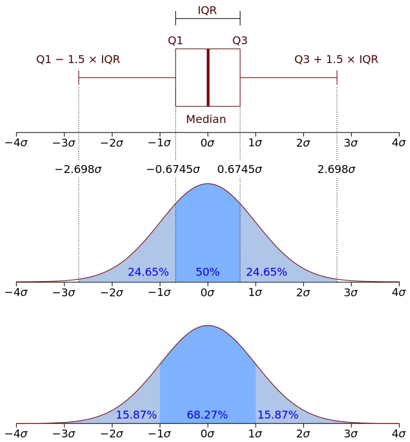
σ^2 = variance
X_i = the value of the ith element
$\bar{X}$ = the mean of X
N = the number of elements

- **Standard Deviation -** Square root of the variance

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Interpreting σ :

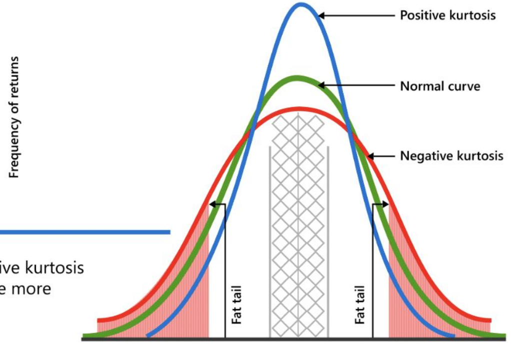




Statistics

The measure of symmentry :

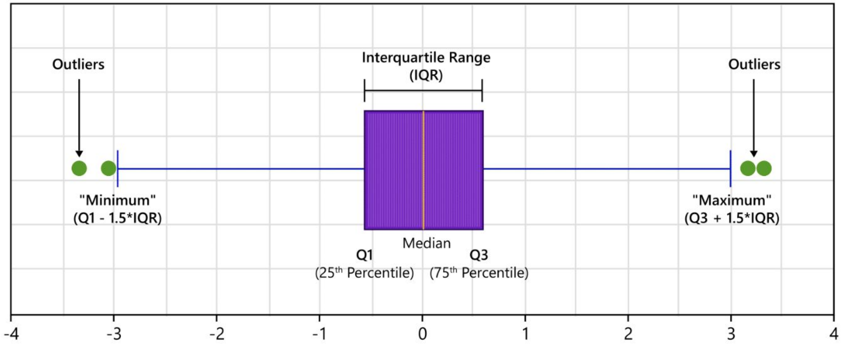
Skewness is the asymmetry of a distribution. A positively skewed distribution has a "tail" pulled in the positive direction. A negatively skewed distribution has a "tail" pulled in the negative direction. Most stock market returns are negatively skewed.



Normal not always the norm

Kurtosis refers to how peaked the curve is:steeper means positive kurtosis and fatter means negative kurtosis. Fat tails occur when there are more outsize returns on the downside or upside,or both, than the normal curve suggests.

Box-and-whisker plot : A graphic that summarizes the data using the median and quartiles, and displays outliers. Good for comparing several groups of data.



Statistics

Correlation :

When there is some relationship between two things

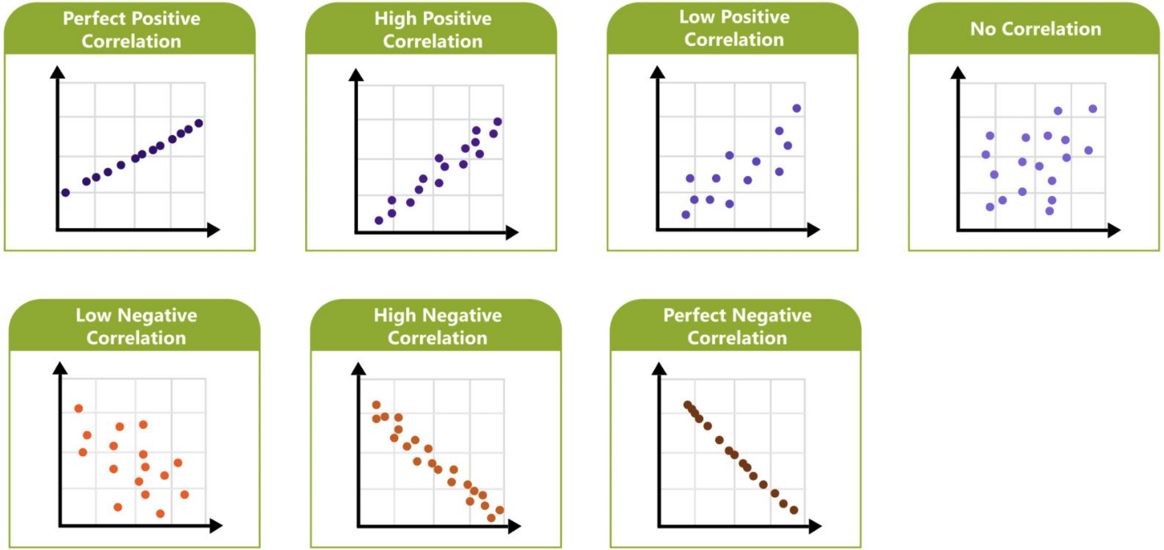
- Correlation always take values between-1 and 1

  -1 is a perfect negative correlation, which means as one thing gets bigger the other thing gets smaller

  0 is no correlation at all, basically is no relationship between these things

  1 is a perfect positive correlation, which means that when one thing gets bigger so does the other

- The closer the correlation value is -1 to 1, the tighter (more linear) the relationship will be on a scatter plot (see below on Pearson's coefficient)



Statistics

Probabilities... (Chance) :

How likely something (an event) is to happen

Kind of Probabilities :

- **Conditional Probabilities** - Probability of an event happening based on whether or not something else happened
- **Joint Probabilities** - Probability of two events happening at the same time
- **Unconditional Probabilities** - Are just the summation of all probabilities

$$\text{Probabbility} = \frac{\textit{How many times event happened}}{\textit{Total Outcomes}}$$

Kind of Events :

- **Mutually Exclusive** - Events that can't happen at same time
- **Non-Mutually Exclusive** - Events that can happen at the same time
- **Independent** - When an event's probability isn't affected by anything else happening or not happening(e.g. a coin toss isn't affected by previous coin toss)
- **Dependent** - Events whose probabilities change based on each other happening or not happening

Cumulative Distribution Function

$$F_X(x) = \mathbb{P}(X \leq x)$$

Cumulative Distribution Function

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

$$\int_{-\infty}^{\infty} f_X(t)\,dt = 1$$

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Statistics

Probability Distributions :

Poisson Distribution :

notation	$Poisson\,(\lambda)$
cdf	$e^{-\lambda}\sum_{i=0}^{k}\frac{\lambda^i}{i!}$
pmf	$\frac{\lambda^k}{k!}\cdot e^{-\lambda}$ for $k\in\mathbb{N}$
expectation	λ
variance	λ
mgf	$\exp\left(\lambda\left(e^t-1\right)\right)$
ind. sum	$\sum_{i=1}^{n}X_i\sim Poisson\left(\sum_{i=1}^{n}\lambda_i\right)$

Story - the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event

Binomial Distribution

notation	$N\,(0,1)$
cdf	$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt$
pdf	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
expectation	$\frac{1}{\lambda}$
variance	$\frac{1}{\lambda^2}$
mgf	$\exp\left(\frac{t^2}{2}\right)$

story: normal distribution with $\mu=0$ and $\sigma=1$.

Story - the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p

Normal Distribution :

notation	$N\,(\mu,\sigma^2)$
pdf	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$
expectation	μ
variance	σ^2
mgf	$\exp\left(\mu t+\frac{1}{2}\sigma^2t^2\right)$
ind. sum	$\sum_{i=1}^{n}X_i\sim N\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\sigma_i^2\right)$

Story - describes data that cluster around the mean

Standard Normal Distribtion :

notation	$N\,(0,1)$
cdf	$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt$
pdf	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
expectation	$\frac{1}{\lambda}$
variance	$\frac{1}{\lambda^2}$
mgf	$\exp\left(\frac{t^2}{2}\right)$

story: normal distribution with $\mu=0$ and $\sigma=1$.

Story - normal distribution with μ = 0 and σ = 1

Tests

1. Tests for mean
    a. Z - Test
    b. T - Test
2. Tests for Variance
    a. Chi-Square
    b. F - Test

ML

		ALGORITHM	DESCRIPTION	APPLICATIONS	ADVANTAGES	DISADVANTAGES
Supervised Learning	Linear Models	**Linear Regression**	A simple algorithm that models a linear relationship between inputs and a continuous numerical output variable	USE CASES 1. Stock price prediction 2. Predicting housing prices 3. Predicting customer lifetime value	1. Explainable method 2. Interpretable results by its output coefficients 3. Faster to train than other machine learning models	1. Assumes linearity between inputs and output 2. Sensitive to outliers 3. Can underfit with small, high-dimensional data
		Logistic Regression	A simple algorithm that models a linear relationship between inputs and a categorical output (1 or 0)	USE CASES 1. Credit risk score prediction 2. Customer churn prediction	1. Interpretable and explainable 2. Less prone to overfitting when using regularization 3. Applicable for multi-class predictions	1. Assumes linearity between inputs and outputs 2. Can overfit with small, high-dimensional data
		Ridge Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients closer to zero. Can be used for classification or regression	USE CASES 1. Predictive maintenance for automobiles 2. Sales revenue prediction	1. Less prone to overfitting 2. Best suited where data suffer from multicollinearity 3. Explainable & interpretable	1. All the predictors are kept in the final model 2. Doesn't perform feature selection
		Lasso Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients to zero. Can be used for classification or regression	USE CASES 1. Predicting housing prices 2. Predicting clinical outcomes based on health data	1. Less prone to overfitting 2. Can handle high-dimensional data 3. No need for feature selection	1. Can lead to poor interpretability as it can keep highly correlated variables
	Tree-Based Models	**Decision Tree**	Decision Tree models make decision rules on the features to produce predictions. It can be used for classification or regression	USE CASES 1. Customer churn prediction 2. Credit score modeling 3. Disease prediction	1. Explainable and interpretable 2. Can handle missing values	1. Prone to overfitting 2. Sensitive to outliers
		Random Forests	An ensemble learning method that combines the output of multiple decision trees	USE CASES 1. Credit score modeling 2. Predicting housing prices	1. Reduces overfitting 2. Higher accuracy compared to other models	1. Training complexity can be high 2. Not very interpretable
		Gradient Boosting Regression	Gradient Boosting Regression employs boosting to make predictive models from an ensemble of weak predictive learners	USE CASES 1. Predicting car emissions 2. Predicting ride hailing fare amount	1. Better accuracy compared to other regression models 2. It can handle multicollinearity 3. It can handle non-linear relationships	1. Sensitive to outliers and can therefore cause overfitting 2. Computationally expensive and has high complexity
		XGBoost	Gradient Boosting algorithm that is efficient & flexible. Can be used for both classification and regression tasks	USE CASES 1. Churn prediction 2. Claims processing in insurance	1. Provides accurate results 2. Captures non linear relationships	1. Hyperparameter tuning can be complex 2. Does not perform well on sparse datasets
		LightGBM Regressor	A gradient boosting framework that is designed to be more efficient than other implementations	USE CASES 1. Predicting flight time for airlines 2. Predicting cholesterol levels based on health data	1. Can handle large amounts of data 2. Computational efficient & fast training speed 3. Low memory usage	1. Can overfit due to leaf-wise splitting and high sensitivity 2. Hyperparameter tuning can be complex