Statistics Intuition Cheatsheet

Yahya Almardeny
almardeny@gmail.com

September 2019

1 Abstract

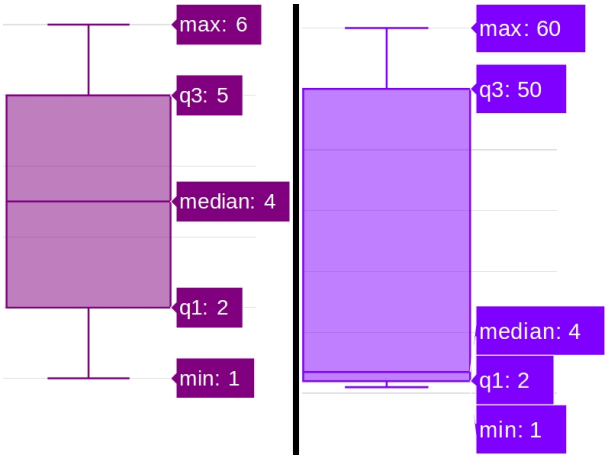
The purpose of this sheet is to provide a quick recap about the intuition of the fundamental concepts in statistics and/or its applications.
It assumes that you are familiar with the formulas of the basic statistical tests.
This sheet is not about how to calculate tests or solve statistical issues.

2 Descriptive Statistics

- **Mean (μ)**: The Arithmetic Average of Data Values. It is highly susceptible to extreme values (i.e. outliers).
  - *Example 1 - No Outliers*:
    $Data = [1, 2, 3, 4, 5] \implies \mu = 3$ which makes sense since 3 is within the range [1-5] (i.e. good representative of the data center).
  - *Example 2 - With Outliers*:
    $Data = [1, 2, 3, 4, 5, 50] \implies \mu = 10.83$ which is not a perfect representative of the center of data since 10.83 is outside the range [1-5] that contains the majority of data points, rather the mean moved towards the value 50!.

- **Median**: The Middle Number in an Ordered Array. It is **not** affected so much by outliers.
  - *Example 1 - No Outliers*:
    $Data = [1, 2, 3, 4, 5] \implies median = 3$ note how it is the same as the mean value above because the distribution of the data points is perfectly symmetric and has zero skewness.
  - *Example 2 - With Outliers*:
    $Data = [1, 2, 3, 4, 5, 50] \implies median = 3.5$ which is still a good representative of the center of data unlike the mean that we got above 10.83!.

- **Quartiles**: Values that divide an ordered list of numbers into quarters Q1, Q2 (the median), and Q3. It is mainly used to find Interquartile (see next) and in the Box and Whisker Plot which can describe the center and spread of the distribution visually in a five-number summary. Ponder the following Box plots (On left: Data=[1,2,3,4,5,6], On right: Data=[1,2,3,4,50,60] which is skewed):



- **Interquartile (IQR)**: It is Q3 - Q1. However, it is not used that much. It can be used to reflect the degree of the outlierness cutoff.
  - *Example*:
    Consider the following 3 data samples:
    $Data_1 = [1, 2, 3, 4, 5] \quad Q1 = 2, \quad Q3 = 5 \implies IQR_1 = 3$
    $Data_2 = [1, 2, 3, 4, 50] \quad Q1 = 2, \quad Q3 = 16 \implies IQR_2 = 14$
    $Data_3 = [1, 2, 3, 4, 100] \quad Q1 = 2, \quad Q3 = 28 \implies IQR_3 = 26$
    note how $IQR_3 > IQR_2$ because $Data_3$ contains more extreme element than $Data_2$: $100 > 50$.

- **Range**: It is the difference between the largest and smallest values in a dataset. It provides an indication of statistical dispersion and it is most useful in representing the dispersion of small data sets. It can be used as a reference to trigger a warning if readings are not within the acceptable range. However, it is very sensitive to outliers that could lead to a misleading result.

- **Standard Deviation (σ)**: Describes how Data Points Vary/Spread around the Mean (i.e. the variability of data). It is highly affected by outliers and it is not a perfect measure for finance

risk (see later). It has very important uses, such as in Variance, Z-Score and Coefficient of Variation (CV) (see later).

- **Variance (σ^2)**: The σ Squared (i.e. the average distance from the mean squared). Its importance (over σ) comes from its mathematical properties. For example: Let X, Y be two independent random variables. Then mathematically, we can find the Variance of the two variables like this: $\sigma^2(X \pm Y) = \sigma^2(X) + \sigma^2(Y)$. However, σ fails to find: $\sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)} = \sqrt{\sigma(X)^2 + \sigma(Y)^2}$ which is not $\sigma(X) + \sigma(Y)$.

- **Z-Score**: (a.k.a standard score). It is the number of standard deviations from the mean a data point is. In other words, the Z-score of a data point, is the distance between that point and the mean normalized by the standard deviation: $Z = \frac{x-\mu}{\sigma}$

  - *Its Importance - Example*:
    Let X, Y be two samples of students' grades in Maths from two different schools. Let $\mu_x = 70$, $\sigma_x = 15$, $\mu_y = 70$, $\sigma_x = 5$. Now if you know that Sarah from School X scored 80 in Maths, and Omar from School Y scored 75 in Maths. Whose grade (i.e. rank) is better in their school? You might think at the beginning it is Sarah. However, to do the comparison, we need to compare Apples to Apples, and since $\sigma_x \neq \sigma_y \implies$ we normalize by converting their grades to Z-Scores:
    $Z_{sarah} = \frac{80-70}{15} = 0.66$ and $Z_{omar} = \frac{75-70}{5} = 1$.
    It turned out that Omar is farer from the mean in his school compared to Sarah. Thus Omar's rank is higher in his school compared to Sarah in her school!.

- **Finance Risk and Coefficient of Variation (CV)**: The finance risk concerns about the danger or possibility that investors will lose money. If we take the return as an example, the finance risk would concern about the variation of return values over the years. Thus, the higher the standard deviation, the higher the risk (i.e. the return is not so consistent and it fluctuates a lot). However, σ is not the best measure since it doesn't take the mean of the returns into account as it is shown in the following example.

  - *Example*:
    Suppose that you have two datasets about the yearly rate of return of two Stocks A and B during the past 5 years:
    $Stock_A = [22\%, 15\%, 20\%, 26\%, 15\%]$
    $Stock_B = [8\%, 10\%, 12\%, 14\%, 16\%]$
    Calculations show that $\sigma_{stock_A} = 4.27$ and $\sigma_{stock_B} = 3.16$ which indicates that $Stock_A$ is more risky than $Stock_B$. Nevertheless, $CV_{stock_A} = 24.09$ and $CV_{stock_B} = 26.35$ which indicates that $Stock_B$ is more risky than $Stock_A$, and that's more reliable result. That's because the Coefficient of Variation $CV = \frac{\sigma}{\mu}$ adjusts for the size of the project and puts the standard deviation into context. Also, it is safer when you have huge difference in the means and want to compare their variations (you can think about it as a "relative σ" or a "normalized σ").

- **Linear Transformation**: Intuitively, when we apply linear transformation on a dataset, we basically change the center and the spread of data by Adding or/and Subtracting or/and Multiplying by or/and Dividing by a constant.
  Note that Linear Transformation does not change the relationship between the variables (e.g. the correlation still the same), rather the mean is shifted and the variance is scaled.
  One example you've already seen is the Z-Score where we normalized by σ to make new data has mean 0 and variance 1.
  Linear transformation helps to get rid of skewness in data, for example, Logarithm Transformation reduces right skewness, where Square Transformation reduces Left Skewness.

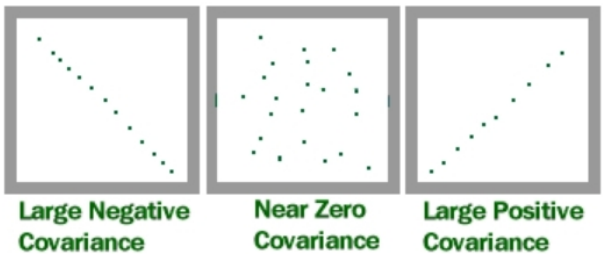
- **Empirical Rule**: There is a trend found in data that strictly has a mound shape. That trend called the Empirical Rule (or the 68–95–99.7 rule), which says: 68.27% of the data falls within 1 standard deviation. 95.45% falls within 2σ and 99.73% falls within 3σ from

the mean. That's useful when you know only the mean and the standard deviation.

- **Chebyshev's Rule**: It is an alternative to the Empirical Rule when you do not have a bell-shaped data (i.e. data is skewed).
  The rule says: at least 0% of data is within 1σ, at least 75% of data is within 2σ, at least 89% of data is within 3σ.

- **Skewness**: It measures the degree of asymmetry exhibited by data.
  As a rule of thumb: Negative Values = Skewed Left, Positive Values = Skewed Right, If $|skewness| < 0.8 \implies$ Don't need to transform data.

- **Covariance**: It measures the joint variability of two random variables. Intuitively, it is similar to Variance, but this describes how two variables vary together. Covariance reflects the trend (relationship) between the two variables (Positive, Negative, or No Relationship) which has enormous applications.



- **Covariance Matrix**: Intuitively, it is a matrix that generalizes the notion of variance to multiple dimensions. It calculates (and therefor find if there's a relationship) for every combination of variables. It has a lot of applications, such in Principle Component Analysis, Cholesky Decomposition and fixing Model Overfitting (we look at the features that have similar / very high covariance and we remove one of each since they are redundant - see next example).

  – *Example of Redundant Variables*:

$$Let \quad S = \begin{bmatrix} A & B & C \\ 1 & 3 & 6 \\ 4 & -4 & -8 \\ 10 & 1 & 2 \\ 5 & 20 & 40 \\ 7 & 28 & 56 \end{bmatrix} \in \mathbb{R}^3$$

be a dataset. Here we have 3 features (i.e. variable per column, name them A, B and C from left to right respectively). At a glance, you can see that variable C is always double the value of B (which is redundant in modeling). Now, suppose that we have a really big dataset, it will be impossible to analyze it by just looking at it. Thus, we can use the Covariance Matrix which reflects how feature C is related to feature B (it has high covariance highlighted in red):

$$Cov.\ Matrix = \begin{bmatrix} & A & B & C \\ A & 11.3 & 8.45 & 16.9 \\ B & 8.45 & 187.3 & 374.6 \\ C & 16.9 & 374.6 & 749.2 \end{bmatrix}$$

- **Correlation**: Very similar to Covariance. Correlation measures the strength of the dependence or association between two variables and its direction. You can think about it as a "custom" relationship and it has a lot of applications, for example but not limited to:

  – Finding the influence of the price on the quantity of the product.

  – Checking for Multicollinearity between features before applying Multiple Linear Regression or when calculating the variance inflation factor (VIF)).

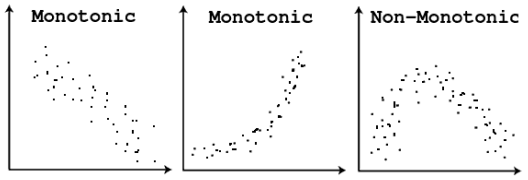
  Correlation ranges only between -1 and 1 (unlike Covariance) and it has 3 main types:

  – *Pearson r Correlation*: The most popular coefficient. It measures the strength of a linear relationship between two variables (both variables should be normally distributed and have homoscedasticity).
  According to Cohen's Standards, its *absolute* value has overall 3 categories:

    * Weak: up to 0.29
    * Medium: 0.3 to 0.49
    * Strong: 0.5 and above

  – *Spearman Rank Correlation*: It measures the strength of a monotonic relationship

between two variables. A monotonic relationship is where one variable increases, the other increases or decreases but not linearly. Ponder the following graph:



– *Kendall Rank Correlation*: It represents a probability; that is the difference between the probability that the two variables are in the same order versus the probability that the two variables are in different orders.
It is very similar to Spearman Rank Correlation, because both are for Rank Correlation and the interpretations of both coefficients are very similar and thus invariably lead to the same inferences. However, Kendall Correlation is recommended to be used with smaller sample sizes since P-values are more accurate and less sensitive to error.

3 Probability

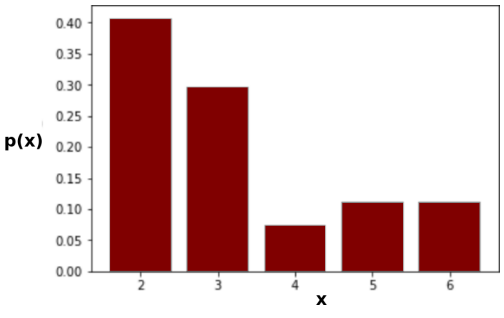
- **Definition**: Provides measures for reasoning the likelihood that an event will occur in an experiment (measure of uncertainty).

- **Rule**: The probability of each outcome s of any Event, should satisfy:
  - within the range [0 - 1]: $0 \leq P(s) \leq 1$.
  - all probabilities should sum up to 1 $\Sigma P(s_i) = 1$
  - the complement rule: $P(s) = 1 - P(\bar{s})$: the probability of s to happen equals the 1 - the probability of s doesn't happen.

- **Probability of Events:**
  1. **Marginal probability**: Simply $P(A)$ which means unconditional probability.
  2. **Addition Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ where $\cup$ means OR, $\cap$ means AND. Note how we are adding up probabilities because we want the probability of A OR B to happen, and that should intuitively increase the overall chances.
  3. **Joint (Compound) Probability Rule**:
     (a) Independent Events: $P(A \cap B) = P(A) \times P(B)$.
     (b) Dependent Events: $P(A \cap B) = P(A) \times P(B|A)$. where $P(B|A)$ reads: Probability of B given A had happened.

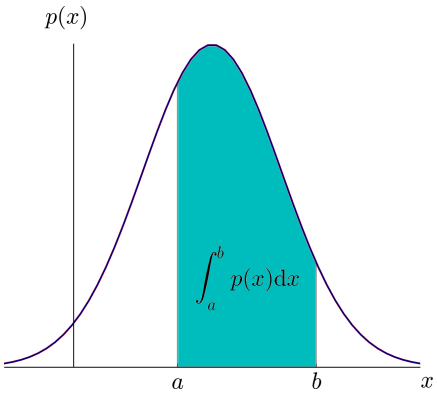
     Note how we multiply here, so the result will be lesser than any of P(A) and P(B) because we are working with numbers between 0 and 1. That makes sense because intuitively the overall chances should decrease since we want both A AND B to happen.
  4. **Conditional Probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  5. **Bayes Theorem**: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$. It gives the probability of A based on prior knowledge of B.

- **Probability Distributions:**
  1. **Probability Mass Function (PMF)**: Gives the probability a discrete random variable X takes on the value x: $P(X = x)$.



  2. **Probability Density Function (PDF)**: Gives the probability a continuous random variable X takes on the value x: $P(X =$

x) which means how dense the probability of X near x, and it will be the definite Integrals between two points.



3. **Cumulative Density Function (CDF)**: Gives the probability a random variable X is less than or equals the value x: $P(X \leq x)$. It is cumulative because it adds the probabilities up to x.
   (a) For Discrete R.V: It is the summation of previous probabilities $\Sigma p(x_i)$.
   (b) For Continuous R.V: It is the integral $\int_{-\infty}^{x} p(x)dx$

- **Expected Value of Random Variable:** Intuitively, it is the long-run average value of repetitions of the same experiment it represents (i.e. what outcome to expect on long run, you can think about it as the "mean")
  1. **For a Discrete R.V**: It is the average of R.V values based on their associated probabilities $E(X) = \Sigma(x_i \times p(x_i))$.
  2. **For a Continuous R.V**: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ where $f(x)$ is probability density function.

- **The Law of Large Numbers:** Let X be R.V of a population where its expected value (i.e Mean) is $E(X)$. Let $\overline{X_n} = \frac{x_1+x_2+...x_n}{n}$ be the mean of n samples $\in X$. As $n \to \infty$: $\overline{X_n} \to E(X)$. That means as we do more experiments, the practical outcomes become closer and closer to the outcomes of the probability theory.

- **Statistical Distributions**[1]
  1. **Binomial Distribution**:
     - How many success in finite number of trials
     - Keywords: Fixed and Independent Trials.
     - *10% Rule*: we assume independence (where it should be applied) if the sample $\leq 10\%$ of the population.
     - The more trials we do, the more Binomial becomes Normal distribution.
  2. **Bernoulli Distribution**: Simply a Binomial Distribution with just one trial.
  3. **Geometric Distribution**:
     - Keywords: How many trials until success?.
     - It is similar to Binomial dist. but here we do not have fixed number of trials because we do not know ahead how many trials until we get the desired outcome.
  4. **Poisson Distribution**:
     - Keywords: discrete probability dist. that predicts independent rare events that occur with a known constant rate λ.
     - Usually used instead of Binomial dist. if the number of trials is very high (fixed interval of time) and the probability (occurrence) of each is relatively low.
  5. **Normal Distribution (a.k.a Gaussian)**:
     - Keywords: continuous probability dist. that has the notation $X \sim \mathcal{N}(\mu, \sigma^2)$.
     - Symmetric and has a bell shape where most of the values are near to the mean.
     - Follows the 68–95–99.7 rule (a.k.a empirical rule).

[1] It is a big long topic, the purpose here is to create your map of knowledge, so you can associate each distribution with a few words. Only the most popular distributions included here.

6. **Uniform Distribution**:
   - Keywords: It is a probability dist. that has constant probability.
   - It can be Discrete Uniform Dist. or Continuous Uniform Dist.
7. **Chi-Square Distribution**:
   - Special case of the Gamma distribution.
   - Continuous distribution of a sum of the squares of k independent standard normal deviates (where k is known as degree of freedom). A standard normal deviate is a random sample from the standard normal distribution.
8. **Student's t-Distribution**:
   - Keywords: Used to estimate population parameters when the sample size is small and standard deviation is unknown.
   - Some of its main application is for assessing the statistical significance and confidence interval between two samples, and in linear regression analysis.

4 Inferential Statistics

- **Hypothesis**: A Claim that we want to test.
- **Null Hypothesis** H_0: The Default Hypothesis that is currently accepted.
- **Alternative Hypothesis** H_a: a.k.a "Research" Hypothesis. It is a new proposed hypothesis that involves the claim to be tested by using *Test Statistics*. Note that H_0 and H_a are opposite mathematically.
- **Possible Outcomes of Hypothesis Test**: 1. Reject Null Hypothesis. 2. Fail to Reject Null Hypothesis.
- **Test Statistics**: A test on a new sample data to get the new parameter.
- **Statistically Significant**: Where to draw a line to make a decision about the Null Hypothesis.
- **Intuitive Example**: Let $\mu = 5g$ be the weights mean of n chocolate bars a machine produces in a certain factory. Now a few years later, suppose a worker on this machine weighed randomly one bar and the weight was 20g! (very suspicious). The worker claimed that the machine is no longer producing the correct weight. In this example:
  - The Null hypothesis is: the weights mean was and still 5g.
  - The Alternative hypothesis is: the machine is no longer producing bars with correct weights.
  - The Test Statistics is: we sample n new chocolate bars and we calculate the new mean.
  - To judge the new results, we use concrete ways (i.e. boundaries) to decide if the test statistics is accepted (i.e. significant). Those ways are: The Confidence Interval (CI) and the Probability Value (P-value).
- **Confidence Interval**: If we draw one sample from a big population, the sample parameter (e.g. mean) most likely will be different from the real actual population parameter, simply because it's just a sample that doesn't necessarily represent all objects.
  A confidence interval is a range of values from drawing different samples that likely contains the True Value (i.e. the unknown value) of the Population Parameter.
  - *Example*: A confidence interval of weights mean on samples of the above chocolate bars = [8 to 10] indicates that the real weights mean of the population is likely to be between 8 and 10. Thus, the CI provides meaningful estimates because it produces ranges that usually contain the true value of the parameter.
  - *Hypothesis Decision Making*: Since the CI indicates the precision of the parameter estimate, it can be used to judge if the hypothesis test result is statistically significant. In the chocolate bars example above, if we sample many bars and check their

weights at different times randomly and found that the CI = [8 to 10] (which excludes the default=5g and doesn't overlap with it), we can then reject H_0 and say that the worker is right and the machine is no longer working properly.

- **Margin of Error**: is the range of values below and above the sample statistic in a confidence interval. It is usually used with CI to reflect the fact that there is room for error.
  - *Example*: Suppose we have one sample of n men heights, where $\mu = 175cm$ and $\sigma = 20cm$. The CI will equal $\mu \pm$ Margin of Error. Where Margin of Error $= Z\frac{\sigma}{\sqrt{n}}$. Where we look up the critical value Z-score from the table (also we can use T-score instead).

- **Level of Confidence** C: How confident we are in our decision on the Hypothesis. For example, saying that the above Confidence Interval = [8 to 10] comes from 20 chocolate bars samples at 95% Level of Confidence, means we are 95% sure that 19 out of 20 samples in our experiment contain correct measurement. Thus we were 95% right in rejecting H_0.

- **Level of Significance** α: Simply $1 - C$. So if level of confidence is 95%, *alpha* will be 5% (5% is the most used value in one tail test, where 2.5% for two tails).

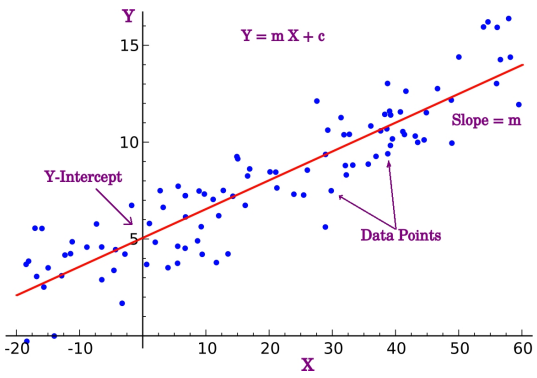
- **P-value**: How likely the result occurred by chance alone. It is the probability of obtaining a sample more extreme that the ones observed in your data (i.e. in H_a), assuming the Null Hypothesis is True (i.e. under H_0). "More Extreme" depends on the direction of the Test Tails (to left or to right). If P-value $<$ level of Significance $\alpha \implies$ Reject H_0. Otherwise, we fail to reject H_0.
  - *Intuitive Example*: What do "more extreme" and "under the null hypothesis" mean? Take the chocolate bars example above. The worker claims that the weights mean is no longer 5g (the H_a) because he sampled one bar and it weighed 20g!. The P-value here checks if it is probable to have 20g under H_0. In other words, P-value checks whether this 20g value was already considered in the calculation in the default hypothesis (the original one). That is, under the default distribution where $\mu = 5g$, if the probability of having a bar weighs 20g or more (in case of only one tail test) is high (i.e. P-value $> \alpha$), that means the original hypothesis did consider and include such extreme numbers and more (no rejecting of H_0). On the other hand, if P-value $< \alpha$, that means the probability of having such extreme values under the null hypothesis is very small and the new results are the valid ones (reject null hypothesis).

5 Applied Statistics

- **Linear Regression**: It is a common type of predictive analysis where we find the best regression estimator to explain the relationship between one dependent variable (a.k.a response) and one or *more* independent variables (a.k.a predictors). The *simplest* form of a Linear Regression is between two variables (dependent and independent) and defined by the line equation:

$$y = m \times X + c$$

where y is the response, X is the predictor and m and c are known as Thetas (or Weights). In this particular equation they are the slope of the line and the y-intercept respectively.



The ultimate purpose is the find the best line (i.e. the best fit) between X and Y.
It should be mentioned that when we have multi-independent variables, the regression is then called *Multiple Linear Regression*.
As a rule of thumb, if the correlation coefficient is reasonably large (positive or negative), the next step would be to fit the regression line which best models the data in order to help to make prediction on Y given X.

- R^2 **Coefficient**: Related to Linear Regression, it is a coefficient of determination which reflects the fit goodness of the Line:

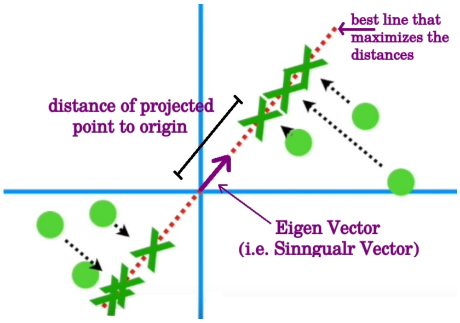
$$Variance\ _{Y_{actual}} \times R^2 = Variance\ _{Y_{predicted}}$$

So intuitively, the more R^2 is closer to 1, the more Actual Y and Predicted Y have same variance (i.e. same spread). That is the actual variance is fully *explained* by the regression line.

We prefer R^2 over the correlation coefficients because it can quantify a linear relationship that is more complicated than a *straight line*

- **Adjusted** R^2: Useful over R^2 because it decreases when we add more junk data and only increases if we have meaningful (i.e. useful) data.

- **Principle Component Analysis - PCA**: It is used mainly to reduce the dimensions of multivariate data (i.e. data that has more than one feature). It calculates for the importance of the features by using the Singular Value Decomposition (SVD) as follows:
  - First it centers all points around the origin, so the new mean will be zero (this shift won't ruin the relative relation between the points but it will make the calculations easier).
  - Then for each feature (i.e. dimension), it finds the best fit of a line made by projecting the points of that feature; the best line should maximize the distances from the projected points to the origin.
  - The line is called PCA1 for the first feature and PCA2 for the second and so on..
  - The Unit Vector on that line (unit vector has the same direction but its length is one) gives us what is known as Eigen Vector (a.k.a Singular Vector). This vector is the Eigen Vector of the Covariance Matrix and since Eigen Vector by definition doesn't rotate and only scaled if multiplied by a rotation matrix, and since the rotation matrix here is the Covariance Matrix of the features, the Eigen Vector will have the most spread of the feature points.



  - The cocktail recipe of how many we need from each feature according to the line and its slope called the Linear Combination.
  - The sum of the squared distances of the projected pointed on the line of each feature is called the Eigen Values, that's the PCA score (i.e. importance).
  - As a result, the PCA that has the most variance around, it is to be the most important.
  - To visualize PCA, check out youtu.be/FgakZw6K1QQ

- **Linear Discriminant Analysis - LDA**:
  - PCA reduces the dimensions by focusing on the importance of the features (the feature with most variation).
  - LDA also reduces dimensions but it focuses on maximizing the separation among known categories.
  - It maximizes the distances between the means of the features and minimizes the variances of them.

- Suppose we have two features, the purpose is:

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{ideally\ large}{ideally\ small}$$

- LDA accounts for the most variation between categories.

- **Analysis of Variance - ANOVA**: It is a statistical method to compare the means between two or more groups (like t-test but with a lot of groups / samples). ANOVA has different types, such as:
  - *One-Way ANOVA*: one factor with at least two independent levels.
  - *Repeated Measures ANOVA*: one factor with at least two dependent levels.
  - *Factorial ANOVA*: two or more factors and levels can be either dependent or independent or both (mixed).

  Assumptions on samples distribution:
  - are normally distributed
  - independent of errors
  - no outliers
  - homogeneity of variance

  ANOVA is useful to show the interactions effects to test hypothesis.

6 Miscellaneous

- **Standard Deviation v.s Standard Error**:

σ	std. error
How much the measurement spread out around the mean	The standard deviation of the mean (i.e. the mean of the means of many samples)
Quantifies the variation within a set of measurements	Quantifies the variations in the means from multiple sets of measurements (tells how the mean is distributed)

- **Bootstrap**: Select a lot of samples randomly from original dataset allowing duplicates (i.e. with replacement) then calculate the mean of each sample.

- **Classification Performance Evaluation**: Given the following abbreviations: Number of True Positives, True Negatives, False Positives, False Negatives is TP, TN, FP, FN respectively, the most important metrics for evaluating classification performance are:
  - **Sensitivity**:

$$Sensitivity = \frac{TP}{TP + FN}$$

    Intuitively speaking, if we have a 100% sensitive model, that means it did not miss any TP, in other words, there were no FN. However, here is a risk of having a lot of FP.
  - **Specificity**:

$$Specificity = \frac{TN}{TN + FP}$$

    Intuitively speaking, if we have 100% specific model, that means it did not miss any TN, in other words, there were no FP, however, there is a risk of having a lot of FN.
  - **Precision**:

$$Precision = \frac{TP}{TP + FP}$$

    Intuitively speaking, if we have a 100% precise model, that means it could catch all TP and there were no FP.
  - **Recall**: It is the *Sensitivity*, but usually this term is used more by the Machine Learning Engineer whereas the Sensitivity term is more used by the Statisticians.
  - **As a Rule of Thumb**: if the cost of having FN is high, we want to increase the model sensitivity (i.e. recall). For instance, in fraud detection or sick patient detection, we don't want to label/predict a fraudulent transaction (TP)

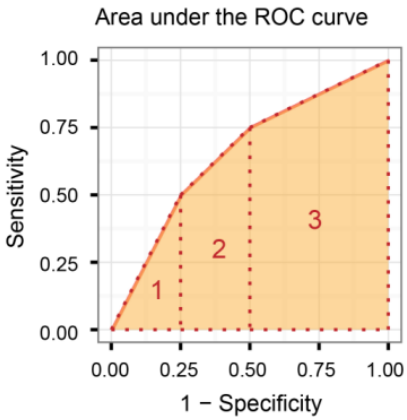
as non-fraudulent (FN). Also, we don't want to label/predict a contagious sick patient (TP) as not sick (FN). That is because the consequences will be worse than a False Positive (incorrectly labeling a a harmless transaction as fraudulent or a non-contagious patient as contagious).
On the other hand, if the cost of having FP is high, then we want to increase the model specificity and precision. For instance, in email Spam detection, we don't want to label/predict a non-Spam email (TN) as Spam (FP). Whereas failing to label a Spam email as Spam (FN) is less catastrophic.

- **F1 Score**: It's given by the following formula:

$$F_1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 Score keeps a balance between Precision and Recall. We use it if there is uneven class distribution, as precision and recall may give misleading results if used each one alone. In other words, F1 Score is a comparison indicator between Precision and Recall Values.

- **Area Under the Receiver Operating Characteristic Curve (AUROC)**:
It compares the overall (i.e. at different thresholds) Sensitivity vs (1-Specificity); that is the True Positive Rate vs False Positive Rate. Thus, the bigger the area under the curve, the greater the distinction between TP and TN.
- **AUROC vs F1 Score**: In general, AUROC is for many different levels of thresholds and thus it shows the overall performance since it has many F1-score values.



On the other hand, F1-score is applicable for any particular point on the AUROC. You may think of it as a measure of precision and recall at a particular threshold value whereas AUROC is the whole area under the ROC curve, and for F1-score to be high, both precision and recall should be high. Consequently, when you have a data imbalance between positive and negative samples, you should always use F1-score because AUROC averages over all possible thresholds.