Alex Pang

Csci 381 – Data Analytics.

Email :

Differences between database, and a data warehouse.

|  |  |  |
| --- | --- | --- |
| Dimension | Database | Data Warehouse |
| Purpose | Data stored in databases can be used for many purposes including day-today operations | Data stored in DW is cleansed data useful for reporting and analysis |
| Granularity -Data granularity is the level of detail considered in a model or decision making process or represented in an analysis report. The greater the granularity, the deeper the level of detail. | Highly granular data including all activity and transaction details | Lower granularity data; rolled up to certain key dimensions of interest |
| Complexity - S | Highly complex with dozens or hundreds of data files, linked through common data fields | Typically organized around a large fact tables, and many lookup tables |
| Size | Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size | Grows as data from operational databases is rolled-up and appended every day. Data is retained for long-term trend analyses |
| Architecural Choices | Relational, and object-orionted databases | Star schema, or snowflake schema |
| Data Access mechanisms | Primarily through high level languages such as SQL. Traditional programming access DB through Open DataBase Connectivity (ODBC) interfaces | Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools |

Datamining – is the art of science of discovering useful innovative patterns from mining data.

Data Mining techniques –

Decision trees – 70% of data mining is from decision trees. They are the most popular and important data mining technique. There are many popular algorithms to make a decision tree.

Regression - This is a well-understood technique from the field of statistics. The goal is to find a best fitting curve through the many data points.

Artificial neural networks – To predict the past futrew and more. This will be the future as computational power increases over time.

Association rule mining – Also called market basket analysis when used in retail industry. To find the correlation between what is purchased by whom what where and why.

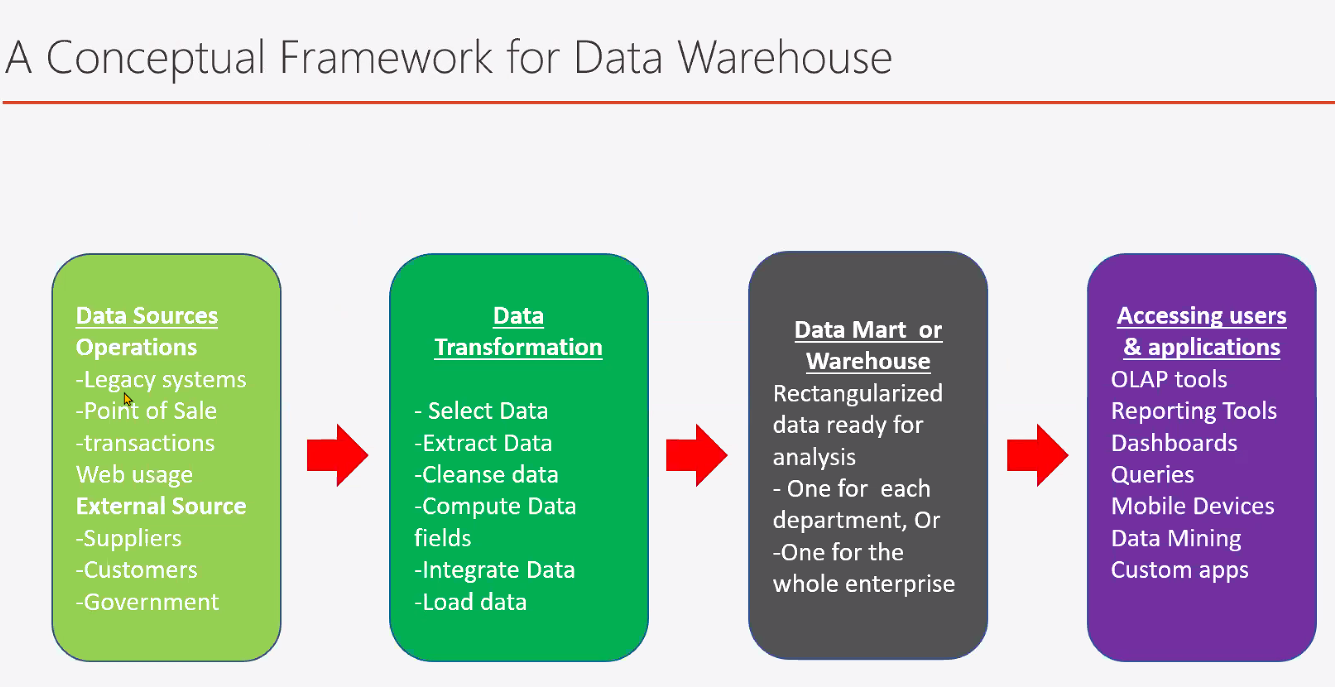
BIBM model – the goal of data analytics.

Data as a new asset class or natural resource.

Database vs data warehouse – OLTP (On-Linbe Trnasaction Processing) vs OLAP (On-Line Analytical Processing)

Oltp – Operations – Business strategy -> Business Processes -> Master data, Transactions.

OLAP – information – data analytics, data mining , decision making -> Business data warehouse.



ETL and Data Marts

Extraction, Transformaiton, and Loading.

* E – Extraction – Get the data
* T – Transforamtion – Make it useful data
* L – Loading – Save it to the warehouse

Data Marts (Sub-sets of the dw-

* Don’t mess with my data
* Keep it simple
* Something I missed.

1st normal form

2nd normal form

3rd normal form

* Look at the dummies guide for 1st to 3rd normal forms.

Python crash course

Import os;

Import getpass

Os.getuid

Data mining 101

Now we have the data in a datastorew, how are we going to useful information out?

SQL query => simple aggregation (mean) => simple statistics (Standard deviation) => Hypothesis testing => data mining => artificial intelligence

1. Make sure to know sql well ( master sql )
2. Low hanging fruits
3. Best bang for the bucks!

SQL exercises, practice, and solutions. <Https://w3resource.com/sql-exercies/>

Who will be responsible for collecting unstructed data and transform them into a database table?

Python stuff

Df = pd.read\_cv()

SQL notes

**Like** keyword searches through terms where it contains x in the name

So if your table looks like

Face gay

Face poop

And

Face man

If you do

Select \* from table

Where name like ‘face %’

It should return all names that start with the word face.

**Union** keyword, merges two queries together.

For example if your table is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **year** | **subject** | **winner** | **country** | **category** |
| 1970 | Physics | Hannes Alfven | Sweden | Scientist |
| 1970 | Physics | Louis Neel | France | Scientist |
| 1971 | Economics | Simon Kuznets | Russia | Economist |

To have a query together from the table and print everything out you would have to do

Select \* from table where year = 1970 union (select \* from table where year = 1971)

Know your statistics (terms we need to know well)

Mean, standard deviation

Distribution

Alw of large numbes

Statistical significane

Survival bias

Bias vs variance.

There are two kinds of statistics

1. Destriptive statistics vs inferential statistics
2. Descriptive statistics consists of organizing and summarizing data
3. Mean, Standard deviation, skew, quantile, and ranks.
4. Inferential Statistics (predictive statistics) consistsd of using data you have already collecting to form conclusions
5. Hypothesis testings
6. Estimiation (using sample mean to predict population mean)
7. The importance of correct sampling
8. The importantance for consideration in sampling
9. Systematic bias
10. Survival bias
11. Size of the samples (Cost)

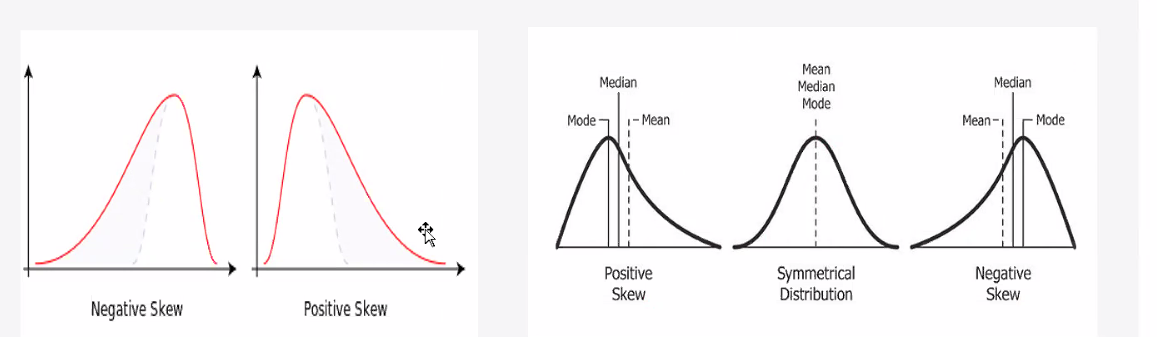
//Get the notes/slides for lecture on 2/17/2020

A lot of formulas and shit that are needed.

1. Desriptive Statistics
2. Central tendency refers to the measure used to determine the center of a distribution of data. Itg is used to find a single score that is most represnative of an entire data set.
3. Mean, Median, and Mode.
4. Mean is

Skew

1. **Skewness** is the degree of distortion from the symmetrical ebll curve, or the normal curve. It measures the lack of symmetry in data distribution. It differentiates extreme values in one versus the other tail. A symmetrical distribution will have a skewness of 0.



Kurtosis

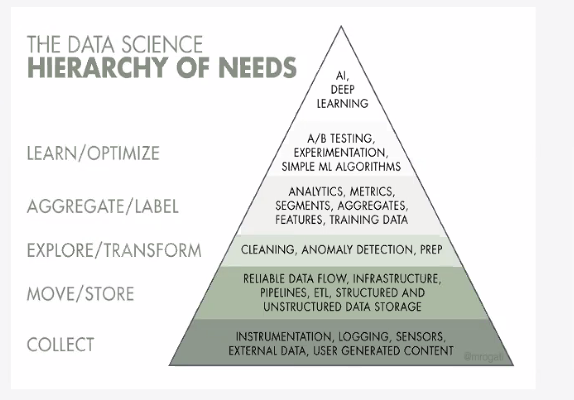
1. **Kurtosis**, on the other hand, refers to the pointedness of a peak or the tails in the distribution curve. The main difference between skewness and kurtosis is that the former talks of degree of symmetry, whereas the latter talks of the degree of peakedness (or tailedness) in the frequency distribution.
2. **Mesokurtic**: This distribution has a kurtosis statistic similar to that of the normal distribution. The standard normal distribution has a kurtosis of three.
3. **Leptokurtic (Kurtosis > 3):** tails are fatter, has more outliers. Peak is higher and sharper than mesokurtic
4. **Platykurtic: (Kurtosis < 3):** tails are thinner, has less outliers than the normal distribution. the peak is lower.

//note

In pandas the kurtosis definition is slightly different. Normal distribution has zero kurtosis. Leptokurtic kurtosis is > 0 and Platykurtic kurtosis is < 0.

Review quiz 2 due Monday. All I need to do is email him my answers.

When I submit the quiz submit as DA\_RQ\_2 as the subject.



Data analytics research process

1. Define your question or problem you want to solve

Make observations, collect the data

Identify possible important factors (features, and attributes to the data)

Test where the factors are important or not

Continue to find important factors

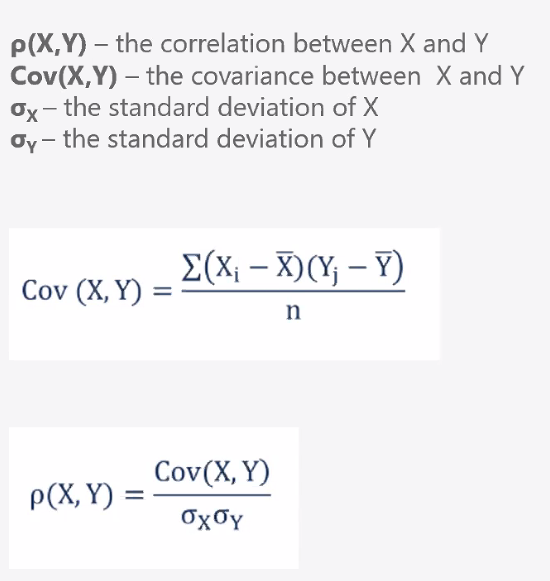
Until you feel like you got it.

Does my data have bias – The 5 types of bias in data analytics.

1. 1 confirmation bias – You have the conclusion before you have your data.
2. 2 selection bias -
3. 3 Outliers -
4. 4 Overfitting and underfitting – will come back to this important topic
5. 5 Confounding variabelen - not sure.

Remember – In pandas the kurtosis is set at 0 and not at 3. So if we use pandas notation we need to talk about leptokurtic, and platykurtic kurtosis is at >0 and <0 instead of >3 and <3.

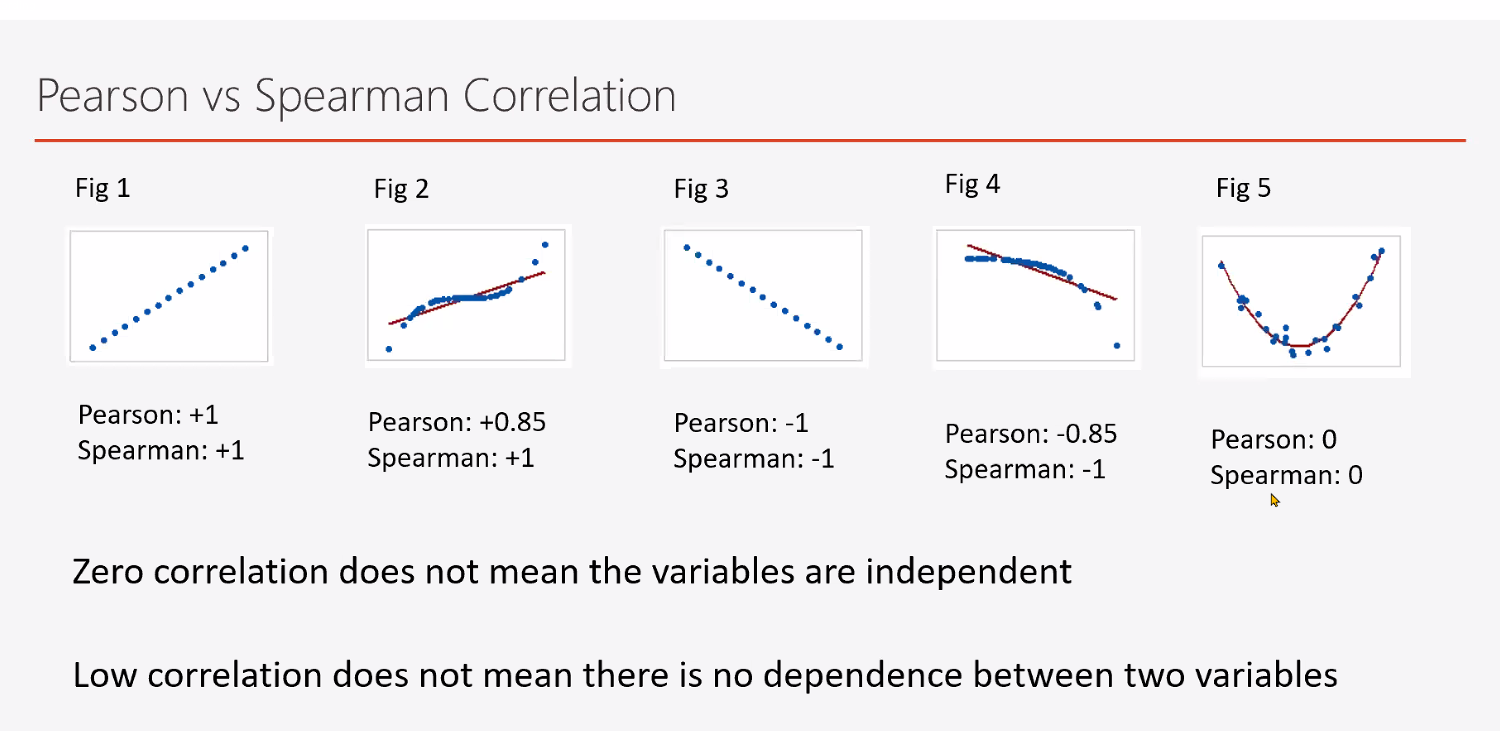
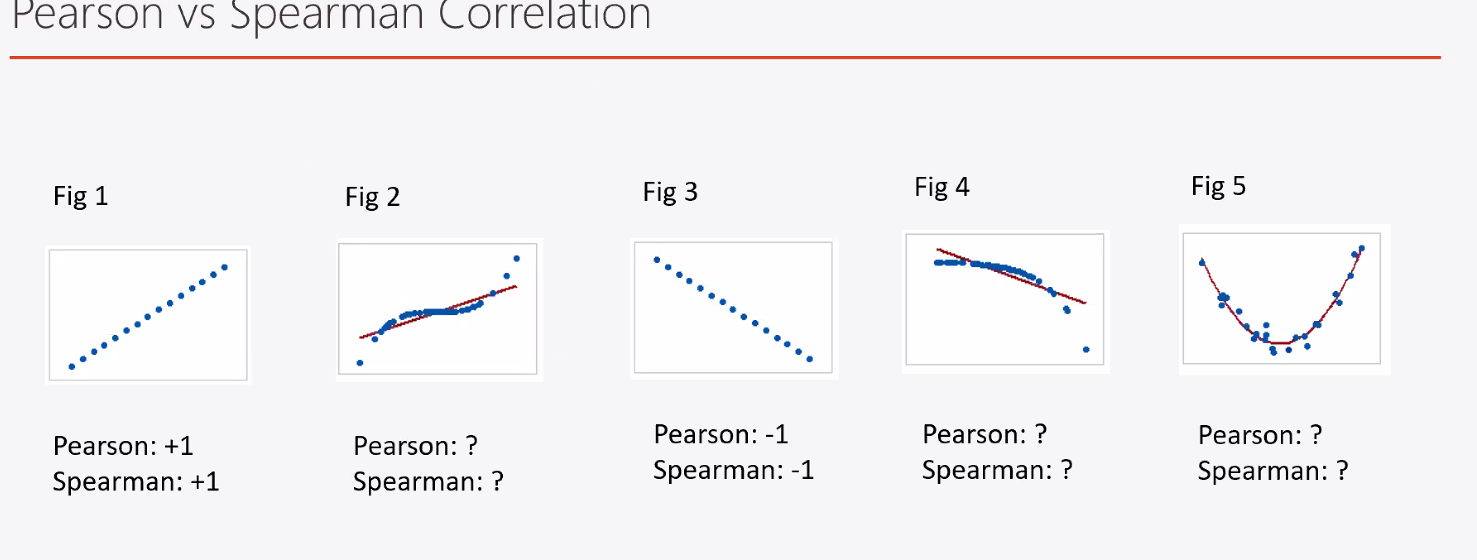
* Remember to watch the videos about the statistics.



1. Quickly finding out Covariance and Correlation between data
2. Covariance measures the linear relationship between two variables.
3. Positive covariance: Indicates that there are two variables that tend to move in the same direction.
4. Negative covariance: Reveals that two variables tend to move in inverse directions.
5. Covariance can range from negative infinity to positive infinity.
6. Correlation is the scaled measure of covariance. It is dimensionless. In other words, the correlation coefficient is always a pure value and not measured in any units.

Correlation is between -1 and +1.

1. **Pearson product moment correlation**
2. The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other.
3. For example, you might use a Pearson correlation to evaluate where a home price increase in a city is related to the unemployment rate in that area.
4. **Spearman rank-order correlation**
5. The spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.
6. Spearman correlation is often used for ordinal variables. For example, you might use a Spearman correlation to study how the order in which employees complete a test exercise is related to the months they have been employees.
7. In a scatterplot, Pearson Correlation coefficients measure linear relationship while Spearman is more concerned on whether the relationship is monotonic or not.



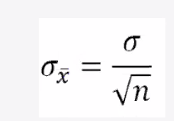
Correlation and Causaiton can lead to bad statistics

Such as MURDER RATES GOES UP WHEN ICE CREAM SALES GO UP

The rates of violent crime and murder have been known to jump when ice cream sales do. But presumably, buyiong ice cream doesn’t turn you into a killer.

But correlation is a good tool to identify, but can go horribly wrong.

1. **Inferential Statistics –** making estimations of the population from samples
2. **Parameters:** A characteristic that describes a population is called a parameter.
3. **Statistic:** A characteristic that describes a sample is called a statistic. Statistics are most often used to estimate the value of an unknown parameter.

Distribution of a sample mean

1. The central limit theorem: Independent of the actual distribution of the population, if we take a big enough sample size, when we repeat taking the sample task again and again, the distribution of the sample mean follows a normal distribution.
2. That is why we can often use the normal distribution behind hypothesis testing.

Tldr about statistics

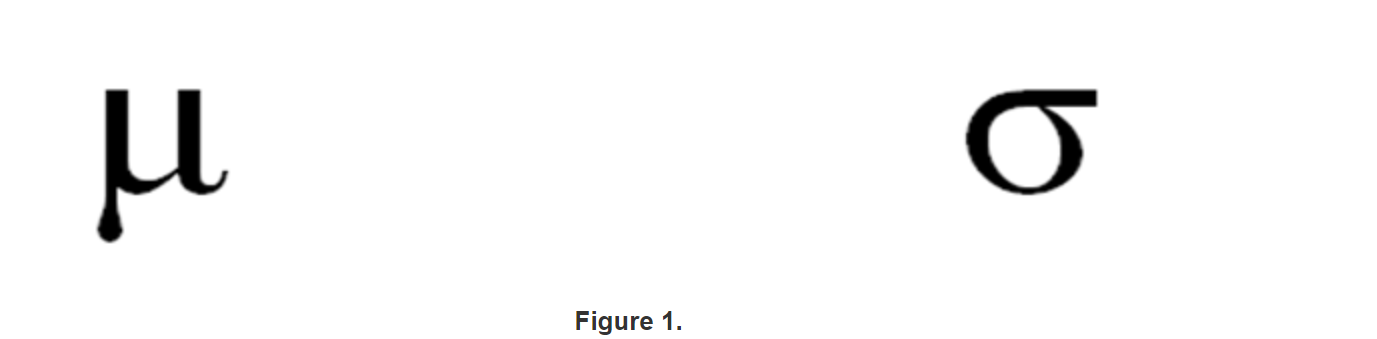
Sampling at N = x where x is the an amount you want to test every time. Is because the more random variance you have the better sampling it is theoretically with a big enough sample size.

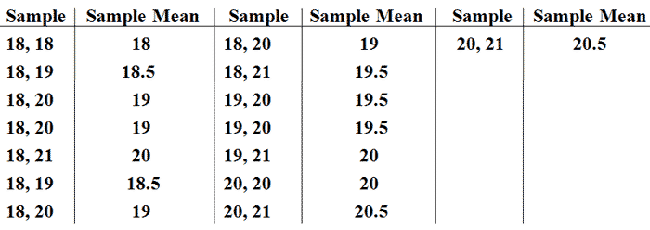
Mean = Average

Sample = the

Parameters are usually written as Greek letters. I’ve already taught you about two: population mean, and population standard deviation.

Mu (MEW) SIGMA

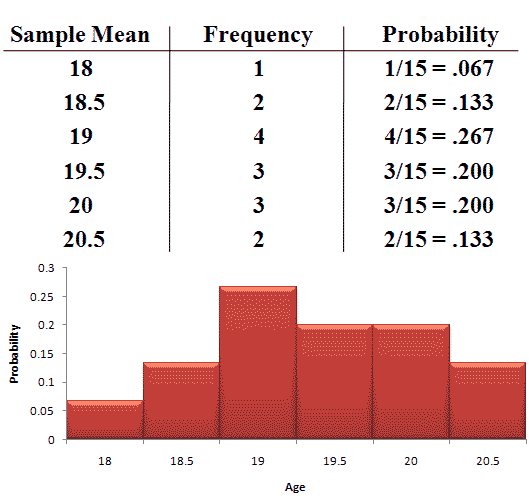




From this chart we can infer that sample means what’s being used.

Mean = average

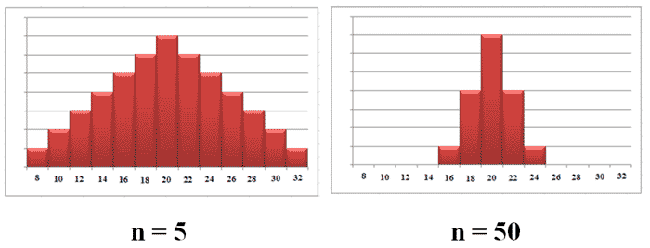
Sample mean = what’s being used that average



From the chart above we can ifner that

Frequency = ho wmany came

Probabiltity = the amount of frequency/(total count of the sample being used)



When you take a larger sampling where N is higher. The standard deviation grows smaller. For example this image

The standard deviation of the sampling distribution is also known as the Standard Error of the mean

SE = standard error of the sample

σ = sample standard deviation

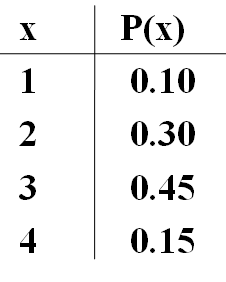
N = number of samples

Sample mean = [xbar](https://www.statisticshowto.com/wp-content/uploads/2009/09/xbar.bmp)

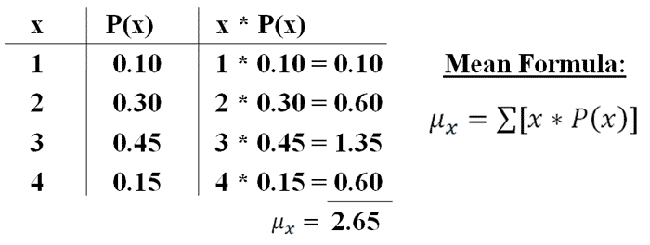
Mean, expected value of discrete random variables

Mean, and Expected value of Discrete random variables.

Below is the probability distribution for a golfer on a par 3 hole, where x = Number of Strokes to Complete Course.



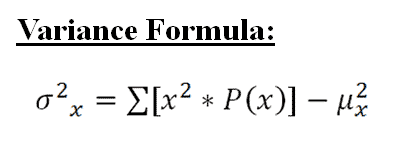
The mean can be calculated by multiplying each "x" by each "P(x)", then adding the resulting values together:



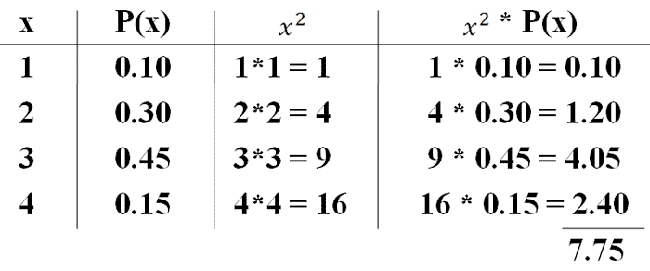
Here, the mean is 2.65

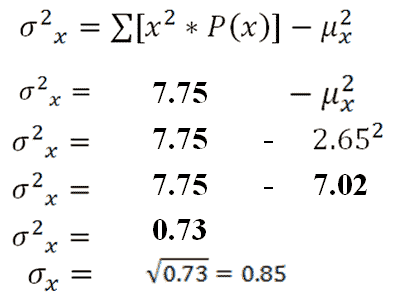
The mean we just calculated of 2.65 is an expected value. If we were to take a large enough sample of this golfers performance on par 3 holes, we expect his mean to approach 2.65.

This is a short example of the Law of Large Numbers.



To find the first part of the equation, we first square every "x". Then, we multiply each squared "x" by "P(x)". Last, we add together all resulting values.

 We find the first part of the equation to be 7.75. Now, we can plug in the rest of the values to get our answers:



The variance is 0.73, while the standard deviation is 0.85.

|  |
| --- |
| **The Central Limit Theorem** |
| The Central Limit Theorem states that regardless of the shape of the population distribution, the distribution of sample means will be approximately normal. |

From the central limit theorem, the following is true:

1. Population distributions that have no skew will lead to distributions of sample means that have no skew.

2. Population distributions that are skewed right will lead to distributions of sample means that have no skew.

3. Population distributions that are skewed left will lead to distributions of sample means that have no skew.

The distribution of sample means will become more normal as its sample size increases.

Good rule of thumb: sample distributions will usually be approximately normal if their sample size is n = 30 or larger.

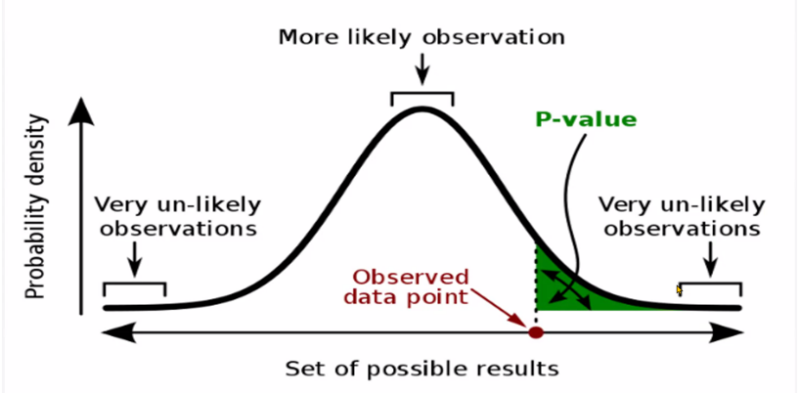
**Hypothesis testing**

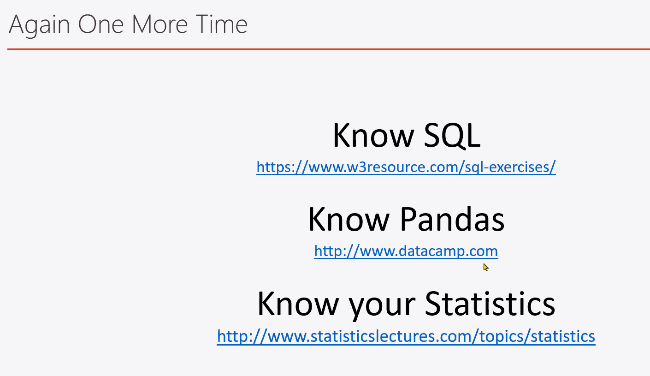
* Type 1 error (False positive, too excited to claim something non-existence)
* Type 2 error (False negative, failed to realize something real is going on)

**Hypothesis testing workflow**

1. Define null and alternative Hypothesis
2. State alpha
3. State Decision rule
4. Calculate test statistics
5. State results
6. State conclusion

(Find out how to do this shit)





Exploratory Data Analysis (Eda)

Before you build any sophisticated model, we need to do an EDA first.

An eda is the firs step into…

* Spotting mistakes and missing data
* Mapping out the underlying structure of the data
* Identifying the most important variables
* Listing anomalies and outliers
* Test a hypothesis / check assumptuions related to a specific model
* Establish a parsimonious model (One that can be used to explain the data with minimal predictor variables)
* Estimate parameters and figuring out the associate dconfidence intervals or margins of error

Most data scientists spend only 20% of their time on the actual data analysis and 80% of their time finding, cleaning, reorganizing huge amounts of data

* Duplicate data removed.
* Missing values need to be filled in (or handles)
* Data elements should be comparable (in similar units)
* Continuous values may need to be binned.
* Outlier data needs to be removed.
* Ensure dataset has no systematic biases for the phenomena under analysis.
* Be sure dataset has enough information density.

Deletion

* Pro: MOST EASY WAY AND NO AMBIGUITS
* Con: Can apply only if we have enough data, may introduce systematic bias

Imputation

* Use Mean, Median, or Mode.
  + Pro: Easy to understand, ok most of the time
  + Con: May introduce systematic bias

For time series data

* Use last observed data (forward fill) (df.fillna(method=’ffill’))
* Use latest available data (backwards fill) (df.fillna(method=’bfill’))

More advanced methods such as use nearest height

Linear Regression

Review the lecture from (3/1/2021)

This is a great review for jupyter/pandas/dataanalytics etc when actually doing work.

Quiz 4)

Question 1: Which of the following statements is/are true about Type-1 and Type-2 errors?

1. Type-1 error occurs when we rejects a null hypothesis when it is actually true

2. Type-2 error occurs when the prediction is positive while the actual case is negative

3. Type-2 error means the prediction is wrong from the actual case

1. A – Only 1

Question 2: If two variables depends on each other, their correlation cannot be zero because there are some dependence of one on the other (even though the effect may be small)

1. False. Correlation can be zero for variables that depend on each other

Question 3: What are the pro and cons of using deletion (ie dropping the data row) for fixing missing values?

1. Pro: It’s easy, and remove the ambiguity

Con: It may only be used when we have enough data otherwise, it may create systematic bias.