Data and Statistics

DSLA COURSE

ROHIT PADEBETTU

Ultimate Goal of a Data Scientist

Is not to build models!

Using **Machine Learning**
to extract insights from **data**
to **solve problems**.

Problem-Solving Critical Thinking

- *Many mistakes boil down to one of the gotchas we'll cover today*
- *Gotchas are the building blocks of critical thinking*
- *Critical thinking is key in many jobs*
- *Critical thinking is key to divergent thinking*
- *Problem-solvers must be critical thinkers, therefore…*

Data Quality Issues

Data scientists spend 60% of their time cleaning data!

Incomplete

- lacking attribute values
- lacking certain attributes of interest
- containing only aggregate data

Noisy

- Containing Errors
  - Salary = -$1000
- Containing Outliers

Inconsistent

- Containing discrepancies in codes or names
  - Rating was 1-5 then A,B,C
- Discrepancies in duplicate records

Data Quality

Reliable Models are built on top of High Quality Data

Quality Metrics

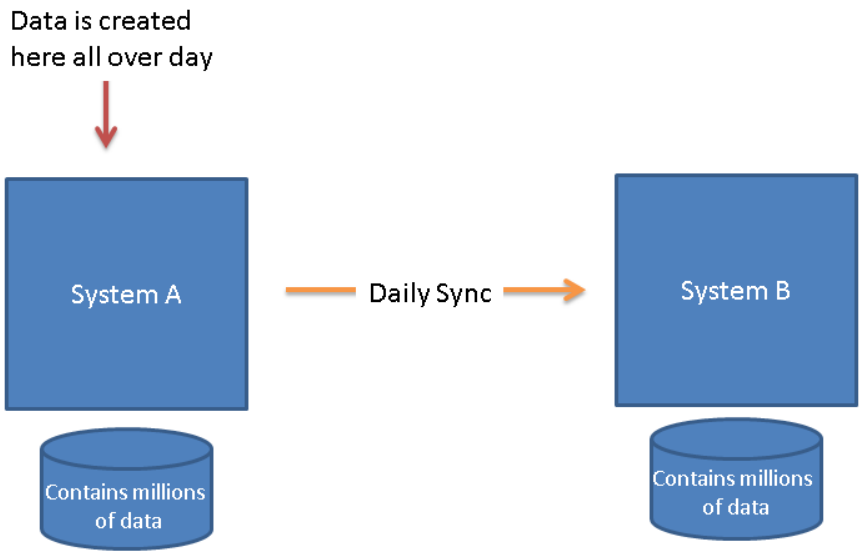
Accuracy (Reconciles)	• % data loads where data reconciles • # accuracy incidents
Consistency (Match Source)	• % data loads where data matches source • # consistency incidents
Timeliness (Right Time)	• % data loads delivered on-time • # timeliness incidents
Integrity (Right Rules)	• % load with Appropriate Business Rules Applied • # integrity incidents
Validity (Right Data)	• % loads with appropriate date range • # validity incidents
Completeness (No Noise)	• % records without noise (missing data) • # noise incidents

Potential Quality Metrics

Accessibility	• % of Critical Data Fields provided
Uniqueness	• % total where duplicate records exist
Compliance	• # of regulatory noncompliance data issues with HIPAA, PHI
Efficiency	• Avg. time taken for data quality issues to be resolved

Data Consistency

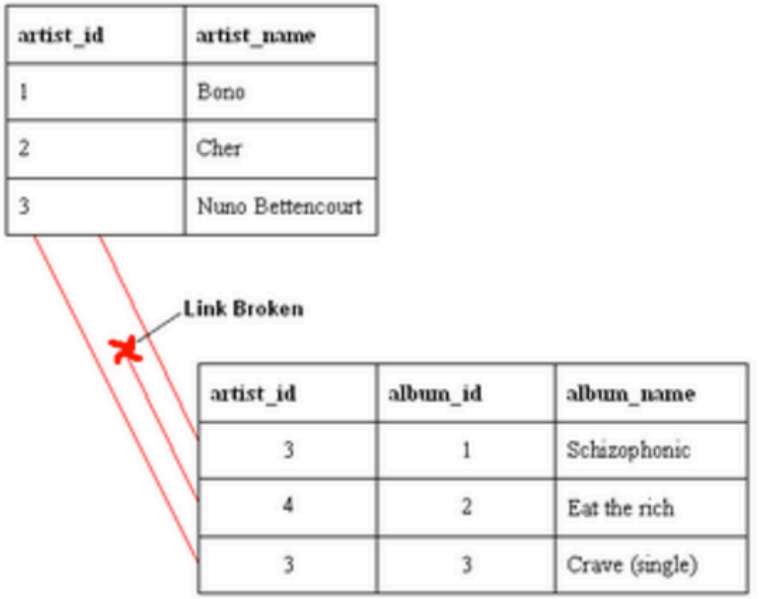
- *Refers to the usability of data*
- *Requires measures to prevent*
  - ✓ *Corrupted files*
  - ✓ *Incomplete records*
  - ✓ *Incomplete transactions (such as due to a system failure)*



Data Referential Integrity

Data Integrity refers to accuracy & reliability

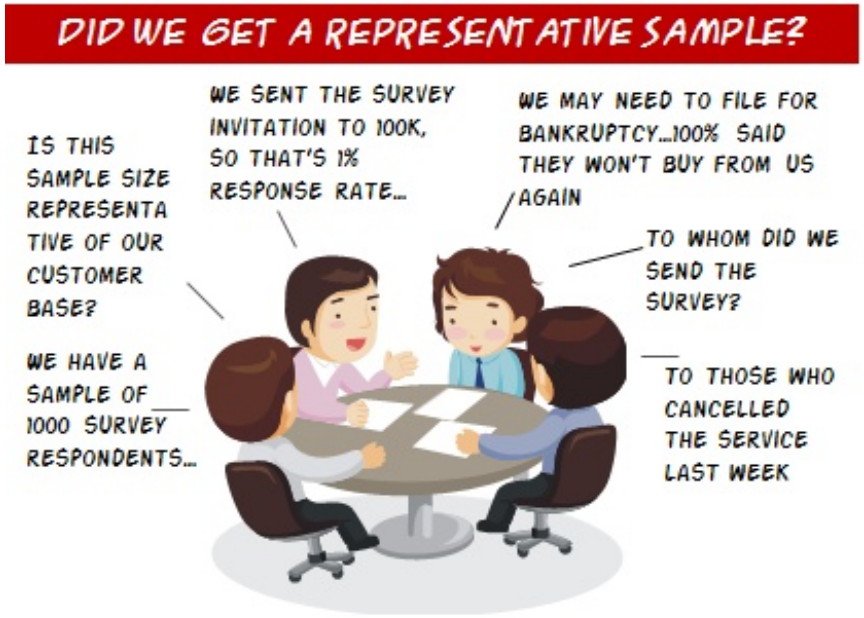
If it's not accurate, it might as well not exist.



Sampling Bias

Usually a result of systematic error in sampling

We will find the average height of Americans based on a sample of NBA players.



A large sample size is NOT the same as a representative sample
© Relevant Insights, LLC

Fixes

- *Get better, more Representative Samples*
- *Aggregate Data with new factors*

Confirmation Bias

Preferential search for evidence that confirms our conclusion

Where
- Data Collection
- Analysis
- Interpretation

Why
- Consciously
- Unconsciously

How
- Overfitting
- Cherry Picking
- Groupthink

Handling Incomplete Data

There are two types of people in this world:

1) Those who can extrapolate from incomplete data

Difficulties

May introduce bias

Or

Not make sense

Fixes

Imputation *Average/KNN*

Or

Build a model to predict it

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

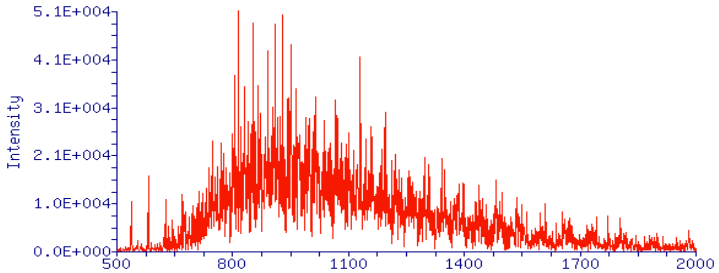
Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

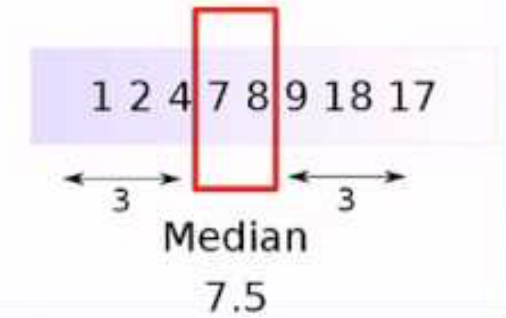
Handling Noisy Data

Random error or variance in measure attribute

Fixes

- *Binning followed by smoothening within bins*
- *Regression Smoothening*
- *Clustering to remove outliers*
- *Human Inspection !!*



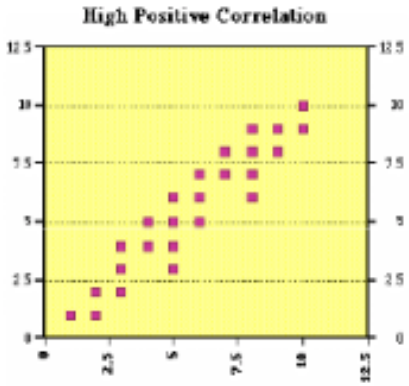


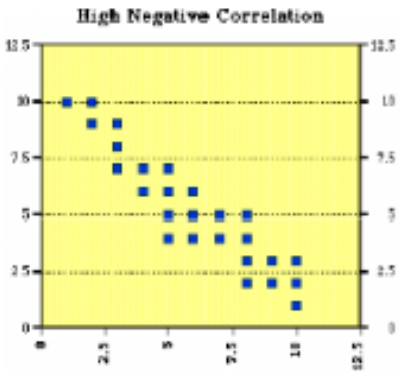
Handling Redundant Data

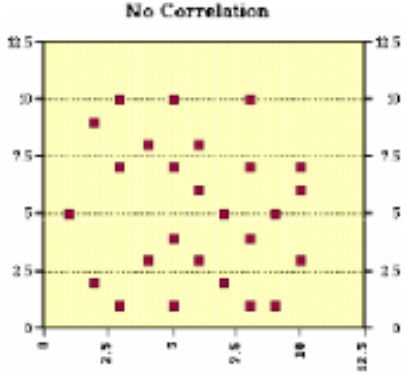
Duplicate or Equivalent Data

Difficulties

Can be source of overfitting







Fixes

Correlation Analysis of Variables

Handling Imbalanced Class Data

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

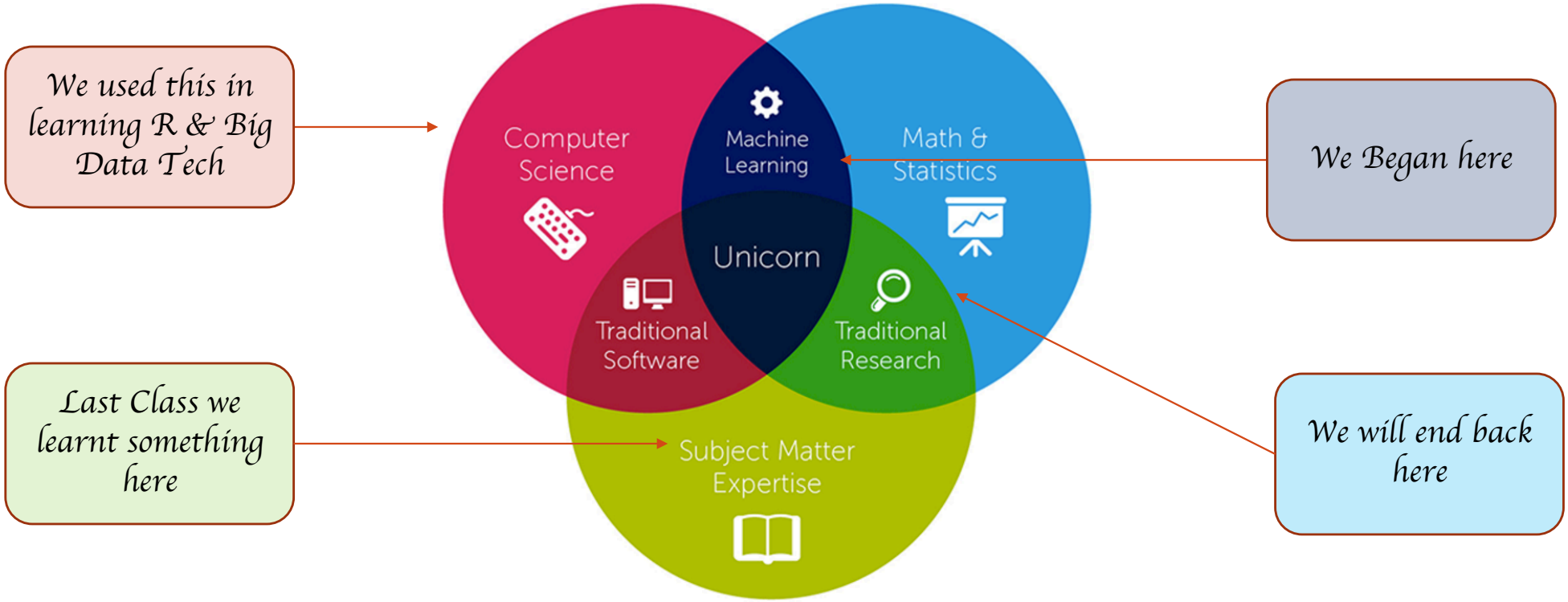
Remember this from LDA discussion?
97.2% Accuracy – not bad huh!!

Accuracy Paradox!

Fixes

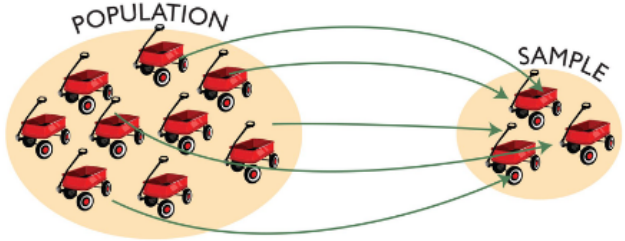
- *Stratified Sampling*
- *Weighting*
- *Consider Precision/ Recall/F1, ROC Curve in addition to Accuracy*
- *Oversampling minor class*
- *Under-sampling major class*
- *Algorithm Choice*

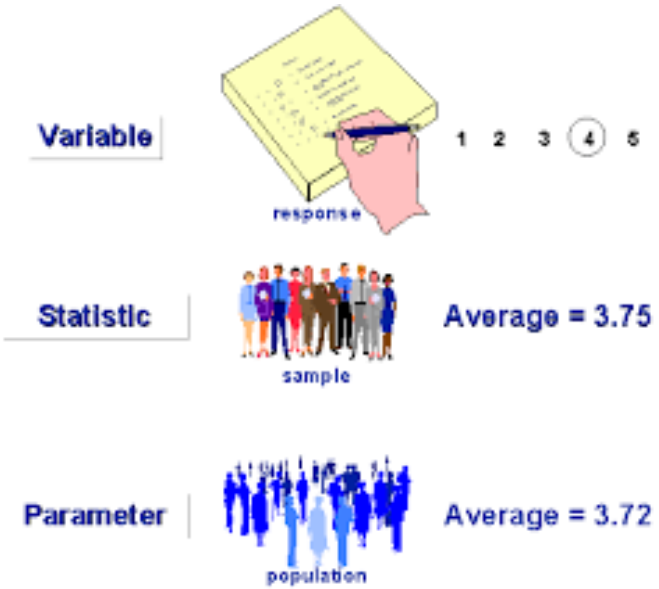
Who is a Data Scientist?

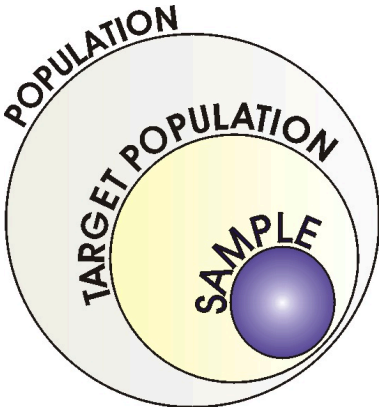


Sample vs Population

Remember: *We usually analyze samples not populations*





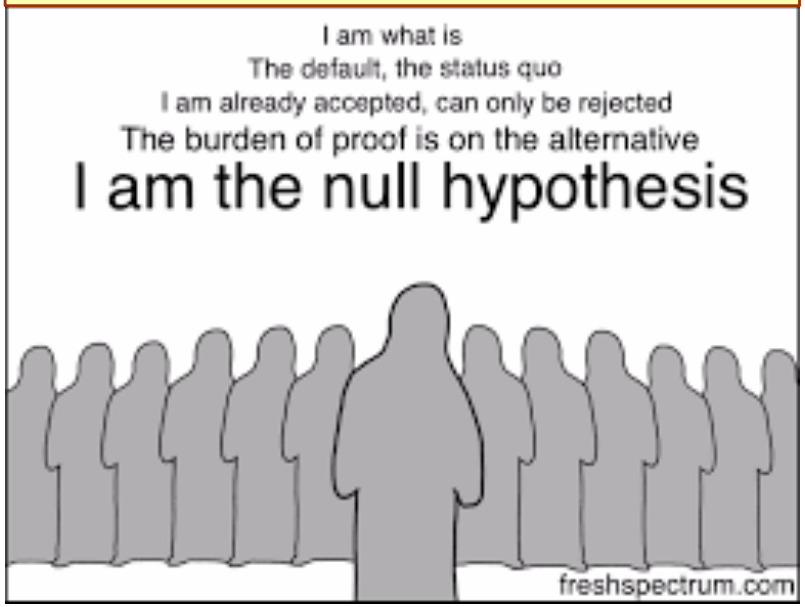


Assumptions

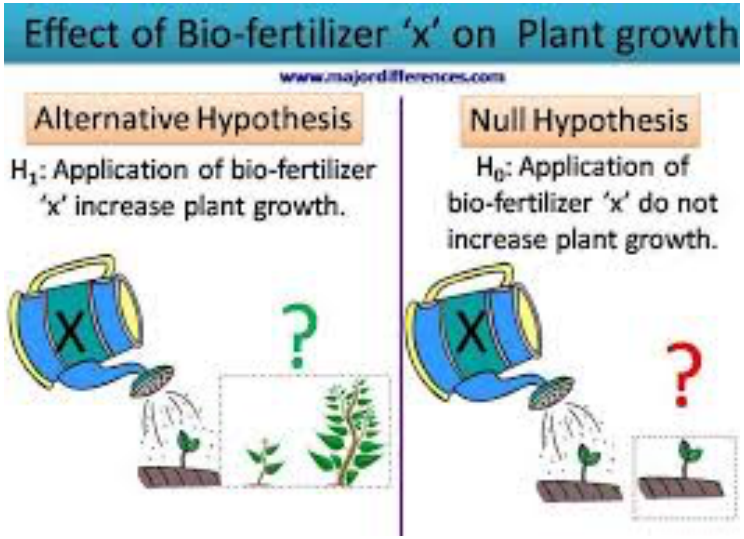
For inferential statistics to hold, sample should be representative of the population

Language of Statistics – Null Hypothesis

Null Hypothesis



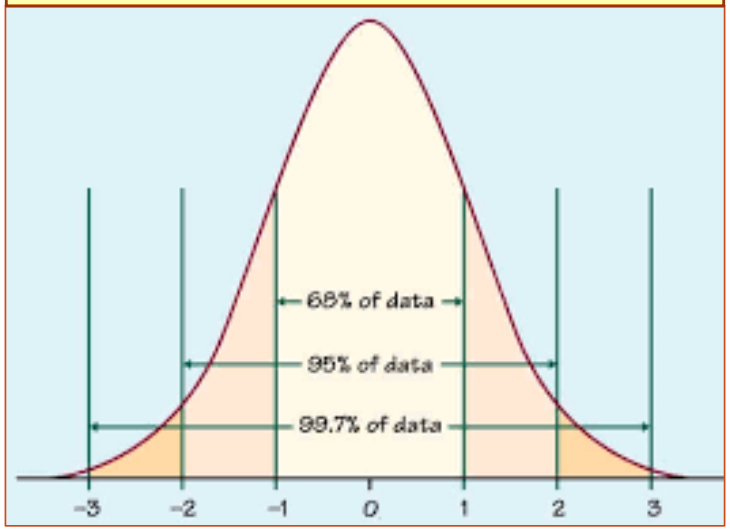
It is usually the opposite of what you want to prove



Language of Statistics- Z Score

Z Score calculates how many standard deviations an element is from the mean

Z Scores



$$z = \frac{x - \mu}{\sigma}$$

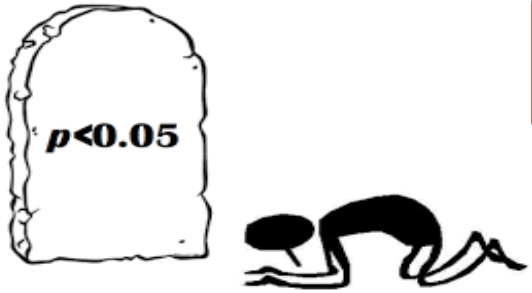
μ = Mean

σ = Standard Deviation

Higgs Boson : 5 sigma event

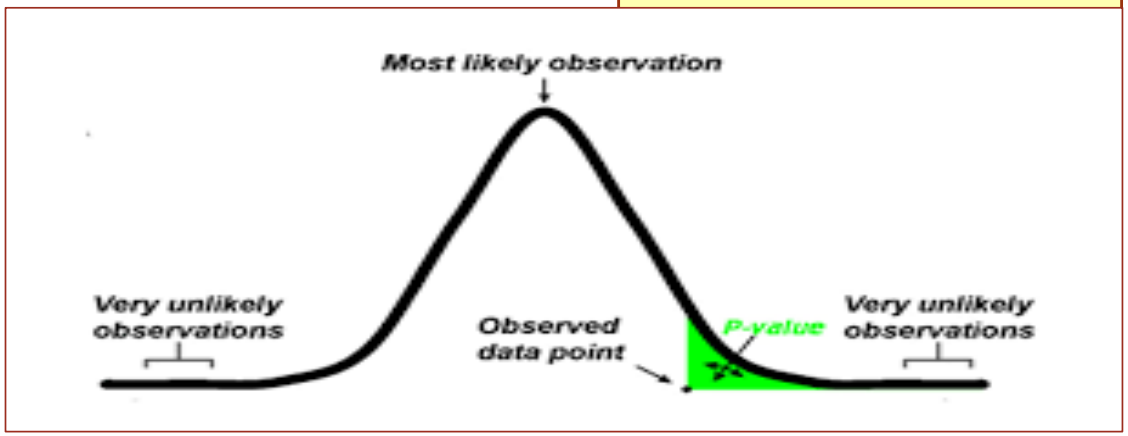
1 in 3.5 Million chance of data occurring by chance rather than from the Higgs Boson truly existing

Language of Statistics – P Value



It is the probability of observed event or more extreme event being due to chance

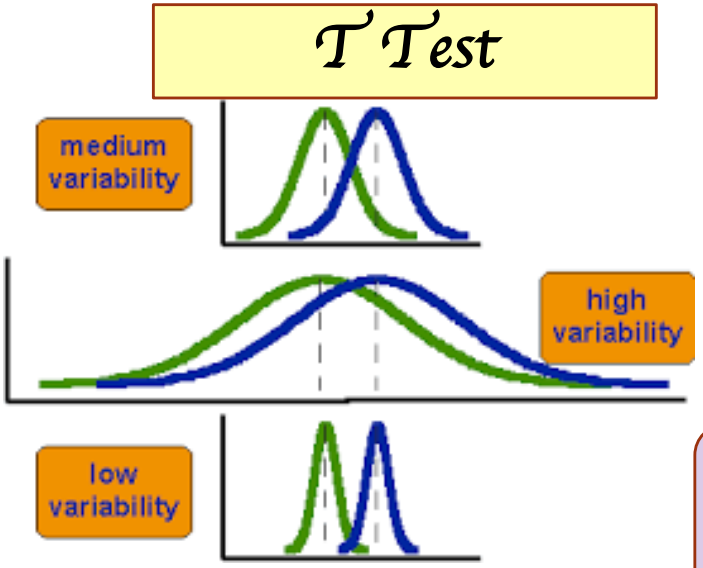
P Values



Values of *p*	Inference
$p > 0.10$	No evidence against the null hypothesis.
$0.05 < p < 0.10$	Weak evidence against the null hypothesis
$0.01 < p < 0.05$	Moderate evidence against the null hypothesis
$0.05 < p < 0.001$	Good evidence against null hypothesis.
$0.001 < p < 0.01$	Strong evidence against the null hypothesis
$p < 0.001$	Very strong evidence against the null hypothesis

Language of Statistics – T test

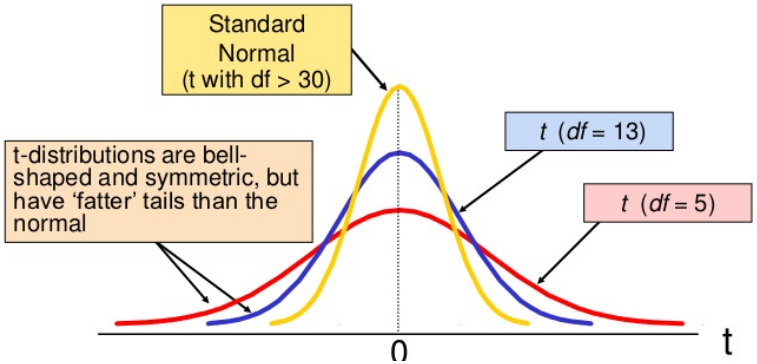
Compares two samples to see if there is a "significant" difference between them



$$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

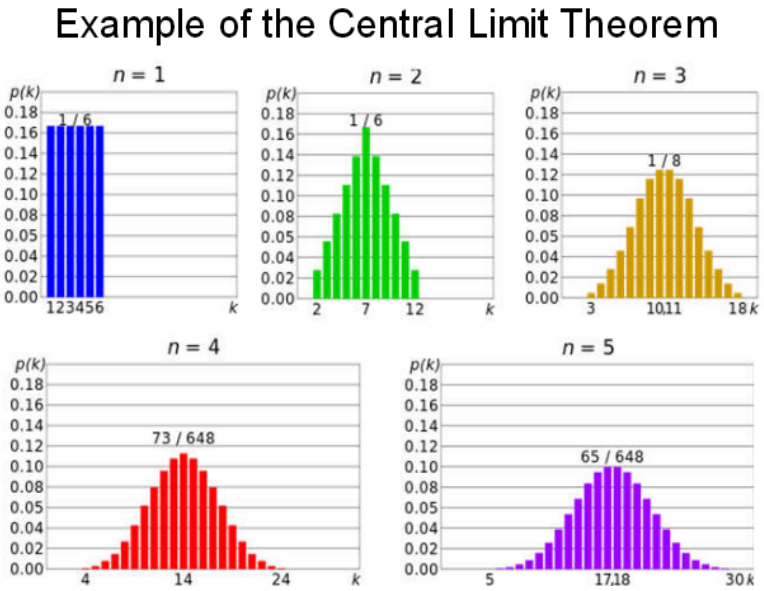
Null Hypothesis is always both samples are same mean

t ⟶ Normal as n increases



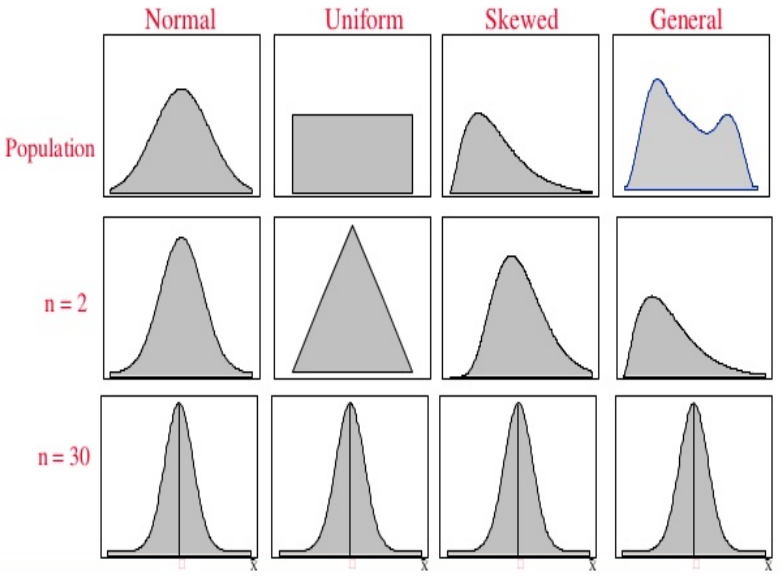
Central Limit theorem

Means of samples drawn from a population, following any distribution, is normally distributed



In Machine Learning this is the basis for

- Cross Validation
- Bagging
- Averaging results of Models

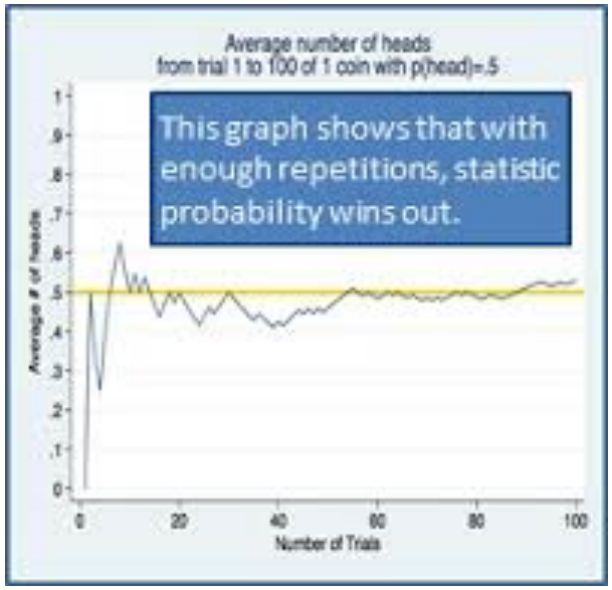


Law of Large Numbers

As sample size grows, it's average will get closer to that of the population

Number of Tosses	Number of Heads	Probability of Heads
4	1	25%
100	64	64%
1000	582	58.2%
10,000	4989	49.89%

Larger number of observations in a sample "might" make it representative of the population

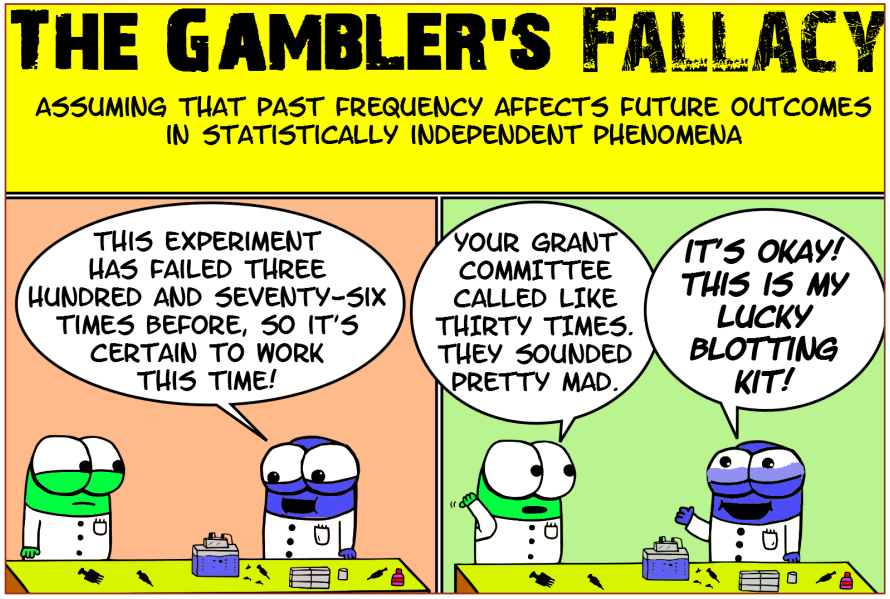


Law of Small Numbers

This is a Myth! There is no law of small numbers

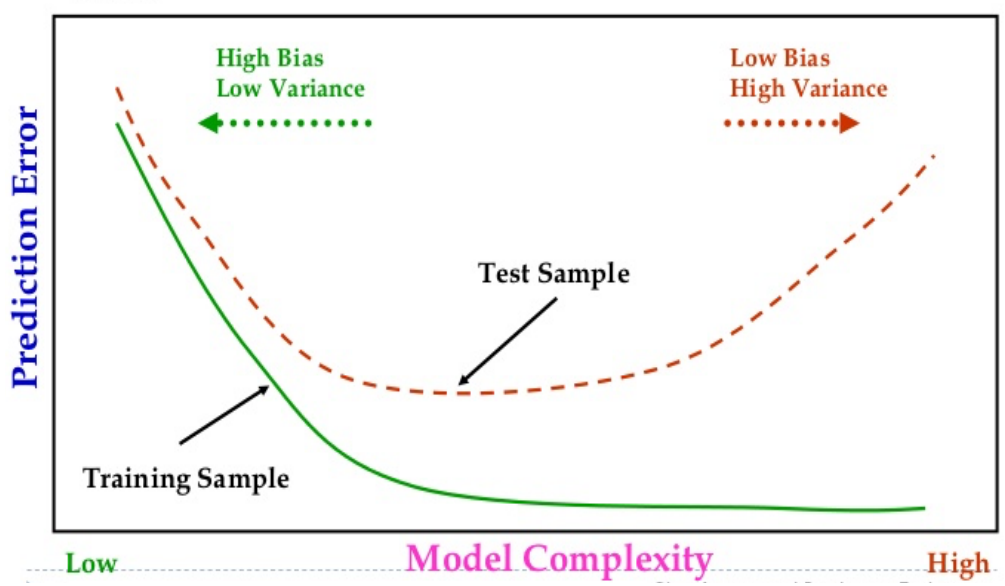
Gambler's Fallacy

General rules of probability hold in the long run. In short run events are random



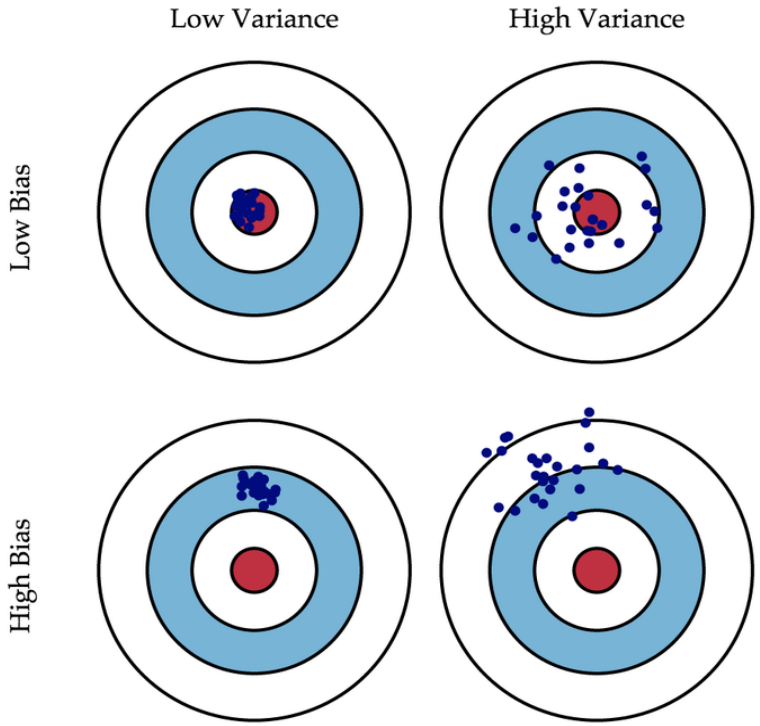
Bias Variance Problem

Under fitting leads to overly simplified models which haven't learned all the patterns in the data.



Overfitting leads to increased model complexity which leads to a variance problem as the model learns the noise in the data.

Bias Variance Problem



To Reduce Variance

- Get more data
- Reduce Features
- Regularization
- Pruning
- Ensemble Models

To Reduce Bias

- Obtain more features
- More data – but it doesn't help beyond a point

The key is pay attention to difference between training set error and test/validation set error

Regularization

To prevent overfitting, this adds a "Penalty Term" to the loss function

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

L1 Norm LASSO Regularization

LASSO can be used to do feature selection

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

L2 Norm Ridge Regularization

Last Slide

DSLA
DATA SCIENCE LEARNING ACADEMY

Hope you enjoyed as much as I did !