2nd) Applied Data Science - Statistics

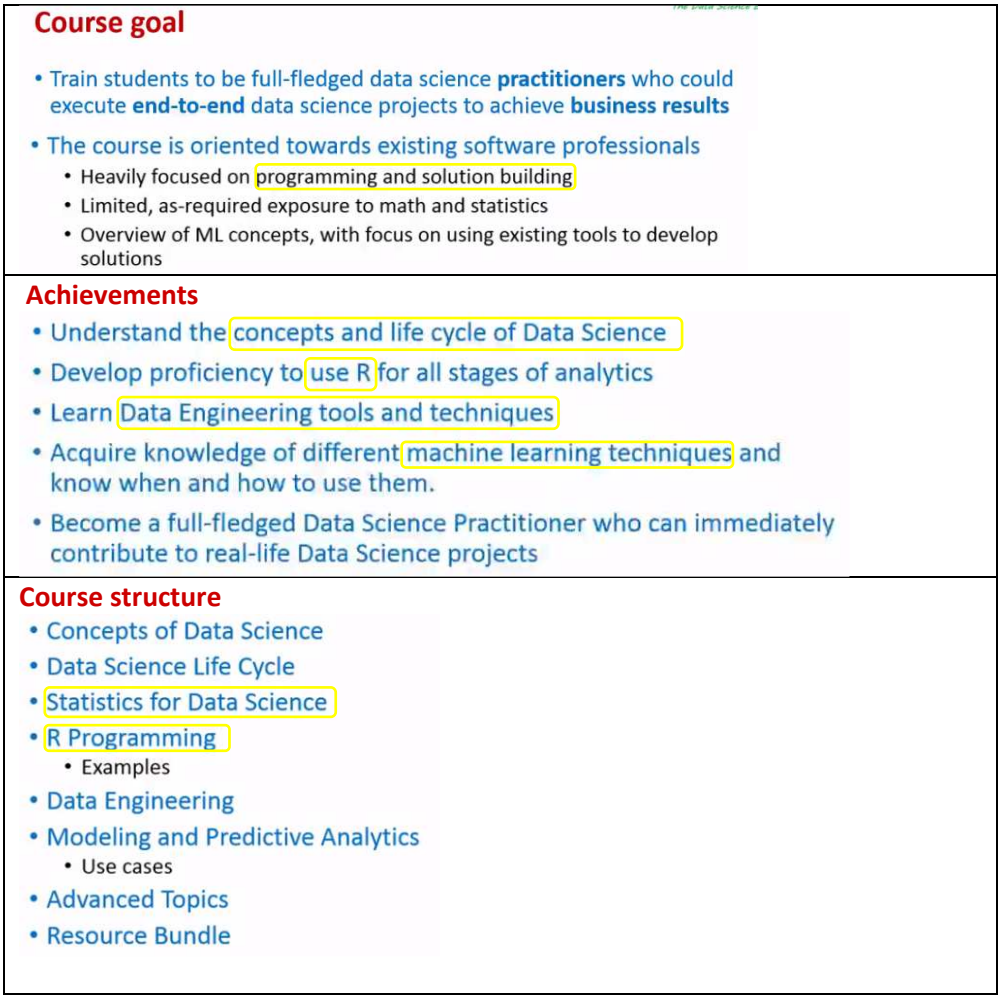
by Kumaran Ponnambalam F on Skillshare

This course focuses on the **Statistics** for Data Science. It goes through basic concepts of statistics that are required for performing data engineering and machine learning operations as a part of this series.

5 Lessons (1h 1m)

URL: https://www.skillshare.com/classes/Applied-Data-Science-2-Statistics/1621177206

1. About Applied Data Science Series 8:12



2. Types of Data 7:29

#4 Types of Data

- There are 4 types of data that you would deal with
- They differ in meaning and what operations you can do on them
- Types
  - Categorical or nominal
  - Ordinal
  - Interval
  - Ratio

Categorical (nominal)

- Represents categories or types
- Fixed list of values
- No implicit ordering or sequencing
- Examples :
  - Fruits : apples, oranges, grapes
  - Players : defender, mid-fielder, forward
  - Cars : sedan, coupe, SUV
  - **Gender:** Male, female
  - **Eye color:** Blue, green, brown
  - **Hair color:** Blonde, black, brown, grey, other
  - **Blood type:** O-, O+, A-, A+, B-, B+, AB-, AB+
  - **Political Preference:** Republican, Democrat, Independent
  - **Place you live:** City, suburbs, rural

Ordinal

- Represents categories
- But there is ordering among the values
- Represents a scale.
- Comparison possible (greater than, less than)
- Examples
  - Review Rating : Outstanding, Very Good, Good, Fair, Bad
  - Pain Levels: 1 – 10 (10 being the highest)
  - Student Grades : A, B, C, D, F
  - **Satisfaction:** Very unsatisfied, unsatisfied, neutral, satisfied, very satisfied
  - **Socioeconomic status:** Low income, medium income, high income
  - **Workplace status:** Entry Analyst, Analyst I, Analyst II, Lead Analyst
  - **Degree of pain:** Small amount of pain, medium amount of pain, high amount of pain

Interval

- Numeric Data
- Measurement where the difference is meaningful
- Represents time, distance, temperature etc.
- Addition, Subtraction possible
- Multiplication, division not possible
- Examples
  - Time of Day
  - Dates
  - Distance between two points
  - Temperature
- **Temperature:** Measured in Fahrenheit or Celcius
- **Credit Scores:** Measured from 300 to 850
- **SAT Scores:** Measured from 400 to 1,600

Ratio

- Numeric Data
- All arithmetic operations possible
- True Zero possible
- Examples
  - Weight
  - Speed
  - Amount
- **Height:** Can be measured in centimeters, inches, feet, etc. and cannot have a value below zero.
- **Weight:** Can be measured in kilograms, pounds, etc. and cannot have a value below zero.
- **Length:** Can be measured in centimeters, inches, feet, etc. and cannot have a value below zero.

Table of comparison

Operations	Nominal	Ordinal	Interval	Ratio
Discrete Values	Yes	Yes	Yes	Yes
Continuous Values	No	No	Yes	Yes
Frequency Distribution	Yes	Yes	Yes	Yes
Median and Percentiles	No	Yes	Yes	Yes
Add / Subtract	No	No	Yes	Yes
Multiply / Divide	No	No	No	Yes
Mean, Std. Deviation	No	No	Yes	Yes
Ratios	No	No	No	Yes
True Zero	No	No	No	Yes

+ ref :

https://blocnotes.iergo.fr/breve/nominales-ordinales-intervalles-et-ratios/

https://www.statology.org/tutorials/

3. Summary Statistics 16:10

Statistics for Data Science (the basics)

Summary statistics

- Describe a set of observations
- Observations have a number of data points; Summary statistics are used to characterize them
- Describe
  - Central Tendency
    - Mean, Median, Mode
  - Variation
    - Variance, Standard Deviation
  - Skew
    - Quartiles

Central Tendency (mean, median, mode)

- Mean : The average
  - Add all number and divide by their count
- Median: The middle value
  - Order the numbers and find the middle value
  - If the count is even, find average of the two middle values
- Mode: The most occurring value
  - The value that occurs most

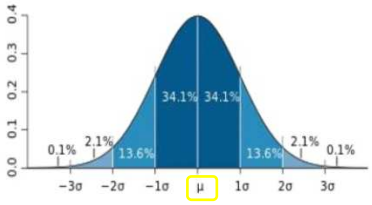
+ Range : diffrence between the lowest and highest values

Ex.

- Observations : 1, 3, 4, 5, 5, 7, 8, 9, 9, 9
- Count: 10
- Sum: 60
- Mean : Sum / Count = 60/10 = 6 μ
- Median : Middle Value = (5 + 7) / 2 = 6
- Mode: 9

+ Range = 8

The bell curve is commonly seen in statistics as a tool to understand **standard deviation**.



Variance & Standard deviation

- Describes how values are distributed around the mean
  - If most values are closer to mean, low variance
  - If significant differences in values, then high variance
- To compute
  - Find the mean
  - Square the differences from the mean
  - Sum of Squares
  - Divide by count

Values	Mean – Value	Square
4	0	0
6	-2	4
3	1	1
5	-1	1
2	2	4
Mean = 4		Sum=10
	Variance = 2	10 / 5
σ	Std. Dev = 1.41	$\sqrt{2}$

- Standard Deviation is Square Root of variance

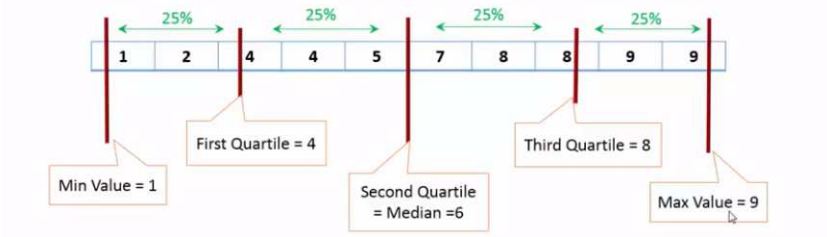
Variance tells us that how far away are the values from the mean.

A low Standard Deviation tells us that fewer numbers are far away from the mean.

A high standard deviation tells us that more numbers are far away from the mean.

Quartiles

- Describes the central tendency, distribution, range and skew in one set of measures
- Given a set of observations, we divide them into 4 equal sets.
- The boundaries form the quartiles



The Box Plot plots the 5-number summary of a variable: minimum, first quartile, median, third quartile and maximum.

Ex. of distribution and reading of a box plot

Min	1st	Median	3rd	Max	Comments
1	3	5	8	10	Evenly distributed
1	4	5	6	10	Most values closer to center
1	2	3	7	10	Skewed to the left
1	6	7	9	10	Skewed to the right

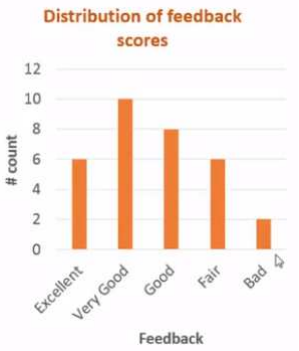
Outliers

- An Odd value occurring in a dataset
- Typically towards the min end or max end of the list
- Outliers tend to distort the summary statistics of a dataset
- Example
  - Observations : 1,2,4,5,20
  - Outlier: 20
  - With outlier, mean= 6.4, Std. Dev=6.94
  - Without outlier, mean= 3, Std. Dev=1.58

4. Statistical Distributions 19:05

Distributions (summarizing trends)

- Distributions show how data values are spread in a given observation set
- Distributions contain a set of bins
- Data is grouped in bins based on
  - Values (categorical, ordinal)
  - Value ranges (interval, ratio)

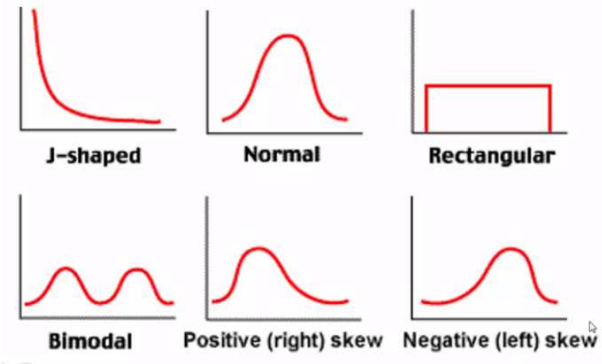


Ex. building a distribution with bins

4	7	3	2	6	9	8	2	5	2

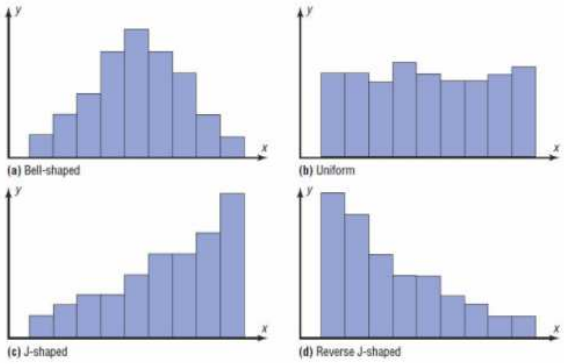
Bin	Values	Count
1-2	2,2,2	3
3-4	4, 3	2
5-6	6,5	2
7-8	7,8	2
9-10	9	1

Distribution shapes



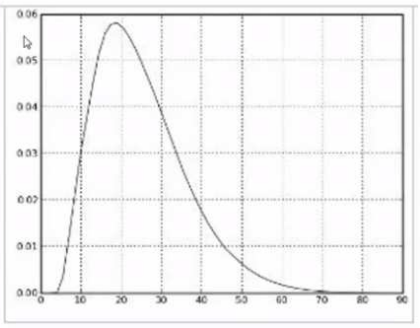
(b) Triangular

(i) Multimodal



Probability distributions

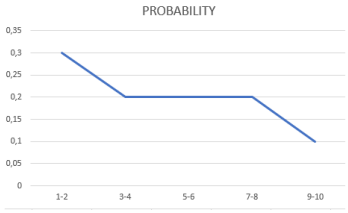
- Assigns a probability to each measurable subset of the possible outcomes of an experiment
- Each possible outcome (or range) plotted on the x-axis
- Probability (0 – 1) plotted on the y-axis
- Discrete or continuous



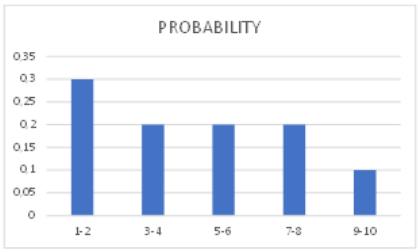
Ex. with

4	7	3	2	6	9	8	2	5	2

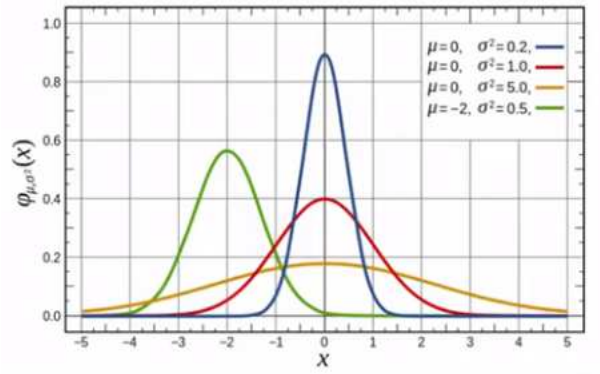
Bin	Values	Count	Probability
1-2	2,2,2	3	0.3
3-4	4, 3	2	0.2
5-6	6,5	2	0.2
7-8	7,8	2	0.2
9-10	9	1	0.1

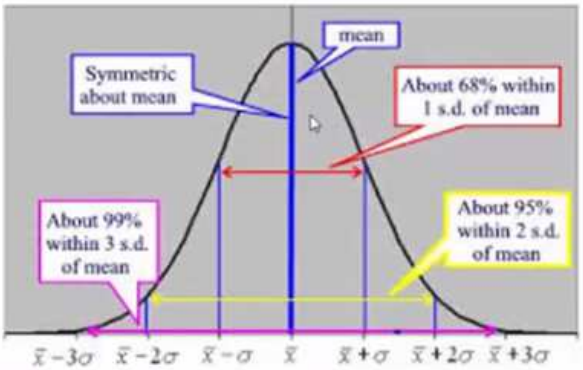


BIN	PROBABILITY
1-2	0,3
3-4	0,2
5-6	0,2
7-8	0,2
9-10	0,1

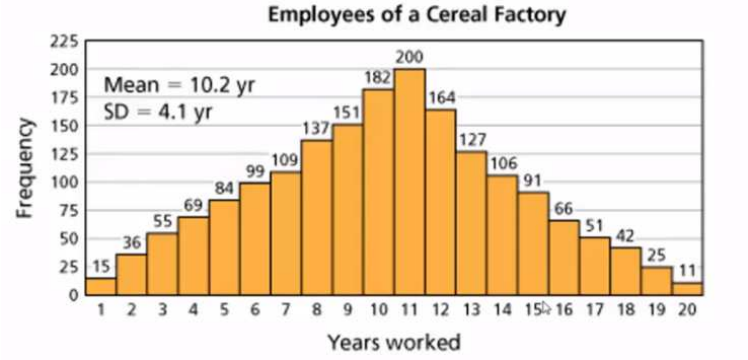


Normal distributions (Gaussian)





Ex. of gaussian distribution

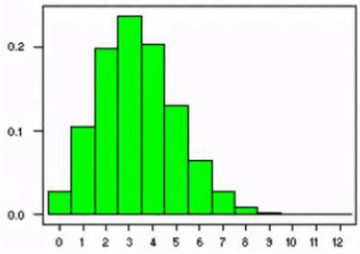


Ways to test a normal distribution: https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93

+ python sample codes…

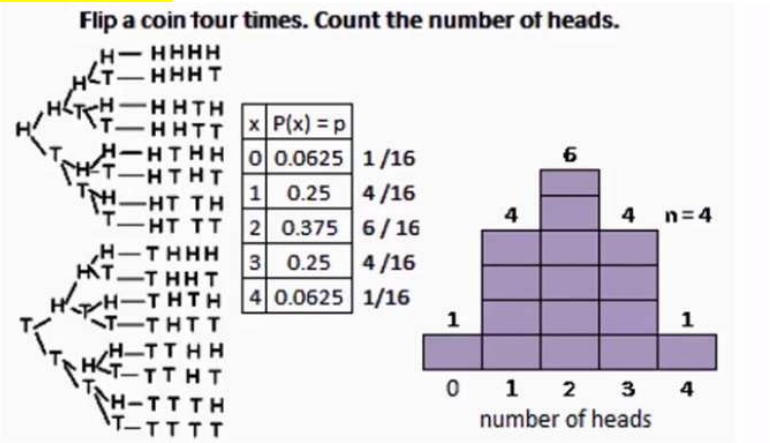
Binomial distribution

- Describes the probability of a Boolean outcome (Yes/ No)
- If
  - n is the number of trials
  - p is the probability of success
  - k is the number of successes
- Plots the probabilities of all values of k.



+ ref. https://fr.wikipedia.org/wiki/Loi_binomiale

Ex. pile ou face

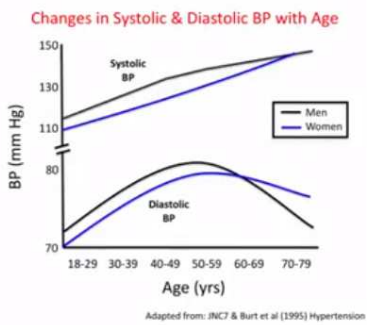


5. Statistics Correlations 10:09

Correlation (relationships)

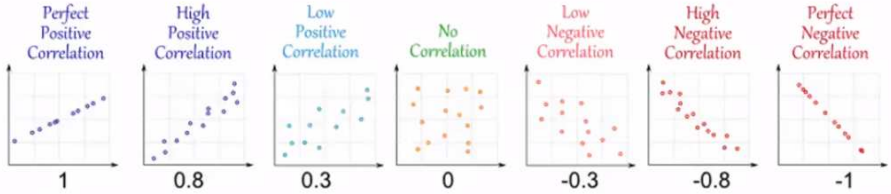
- Correlation : a mutual relationship or connection between two or more things
- Interdependence
- Correlation between 2 sets of data – how much does one change when the other changes
- The basis of data science

Example : Age and Blood Pressure



Measuring correlation

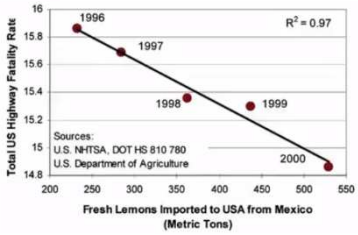
- Pearson's Correlation co-efficient
- Values range from -1 to +1



Causation vs. Correlation

- Causation : The reason for a change in value
- Correlation does not imply causation
- Correlation might be due to
  - Causation
  - Common cause
  - Incidental

- An analysis needed to establish why you see what you see



Ex. false correlation between two things