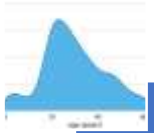
Descriptive Statistics

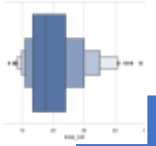
Data Visualization Basics

Data Visualization help us understand the data in an easy way. Through pictorial representation



Distribution Plot

- Dist-Plot
- Joint Plot
- Rug Plot
- Pair Plot



Categorical Plots

- Bar Plot
- Count Plot
- Box Plot
- Violin Plot

Matrix Plot

- Heat Map
- Cluster Map

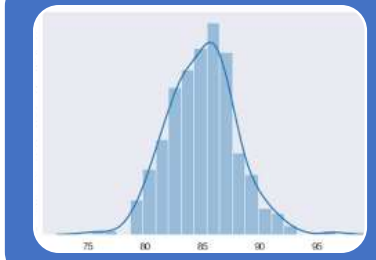
Advanced Plot

- Strip Plot
- Swarm Plot

Miscellaneous

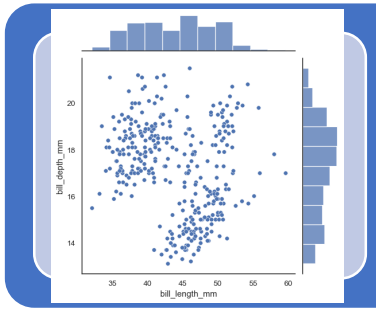
- Facet Grid
- Regression Plots

Distribution Plots



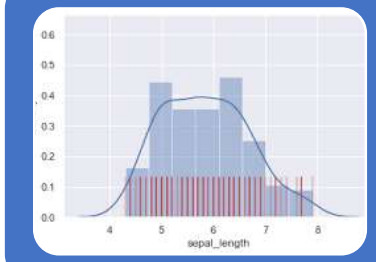
Dist-Plot

- Dist plot gives us the histogram of the selected continuous variable.
- We can change the number of bins
- Mostly used for **univariate analysis**



Joint Plot

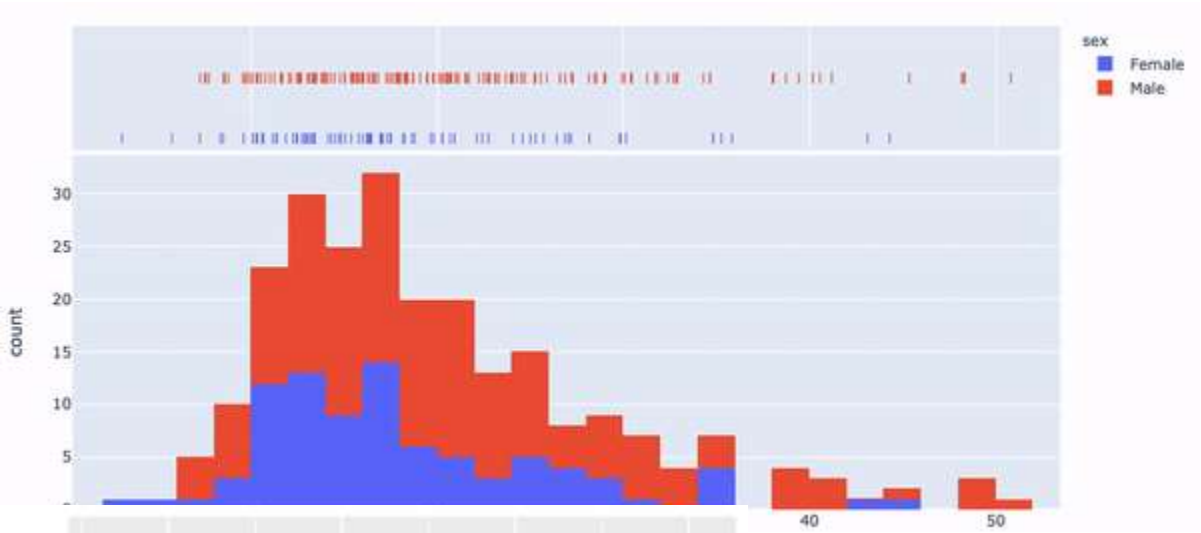
- It is the combination of the distplot of two variables.
- We obtain a scatter plot between the variable to reflecting their linear relationship.
- Mostly used for **bivariate analysis**

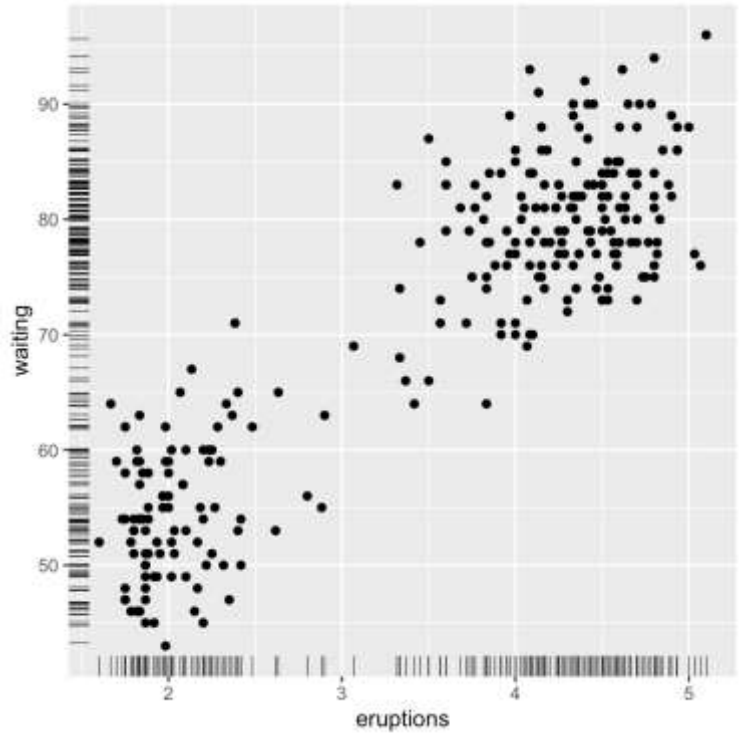


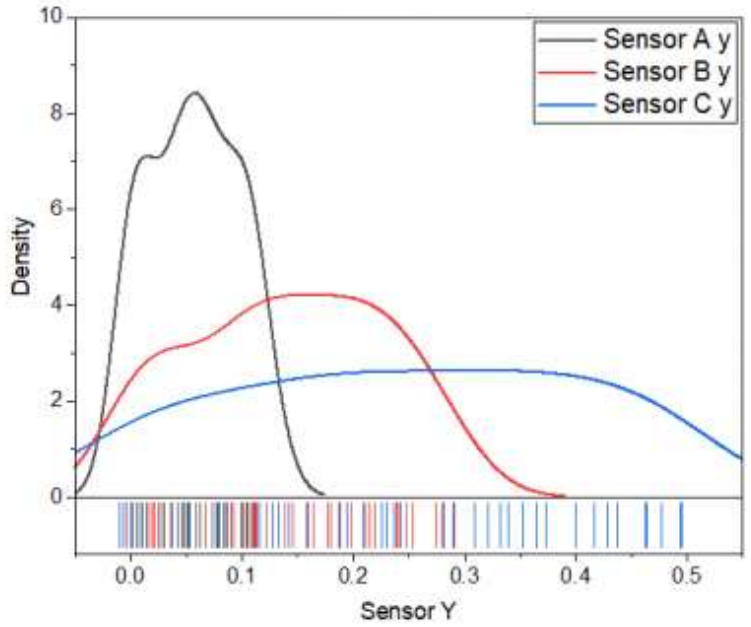
Rug Plot

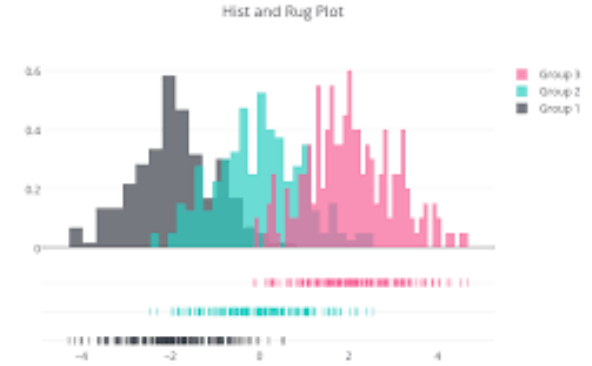
- It draws a dash mark instead of a uniform distribution as in distplot.
- Mostly used for **univariate analysis.**

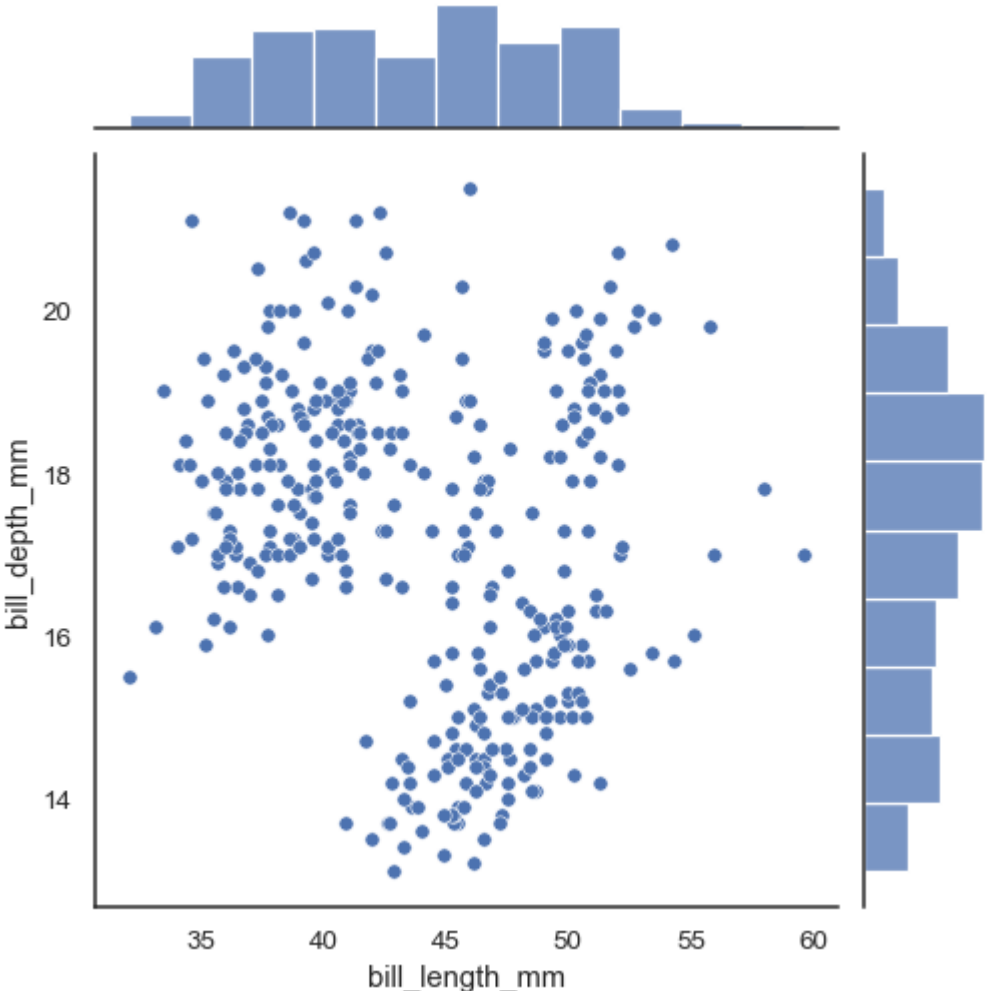
More Examples







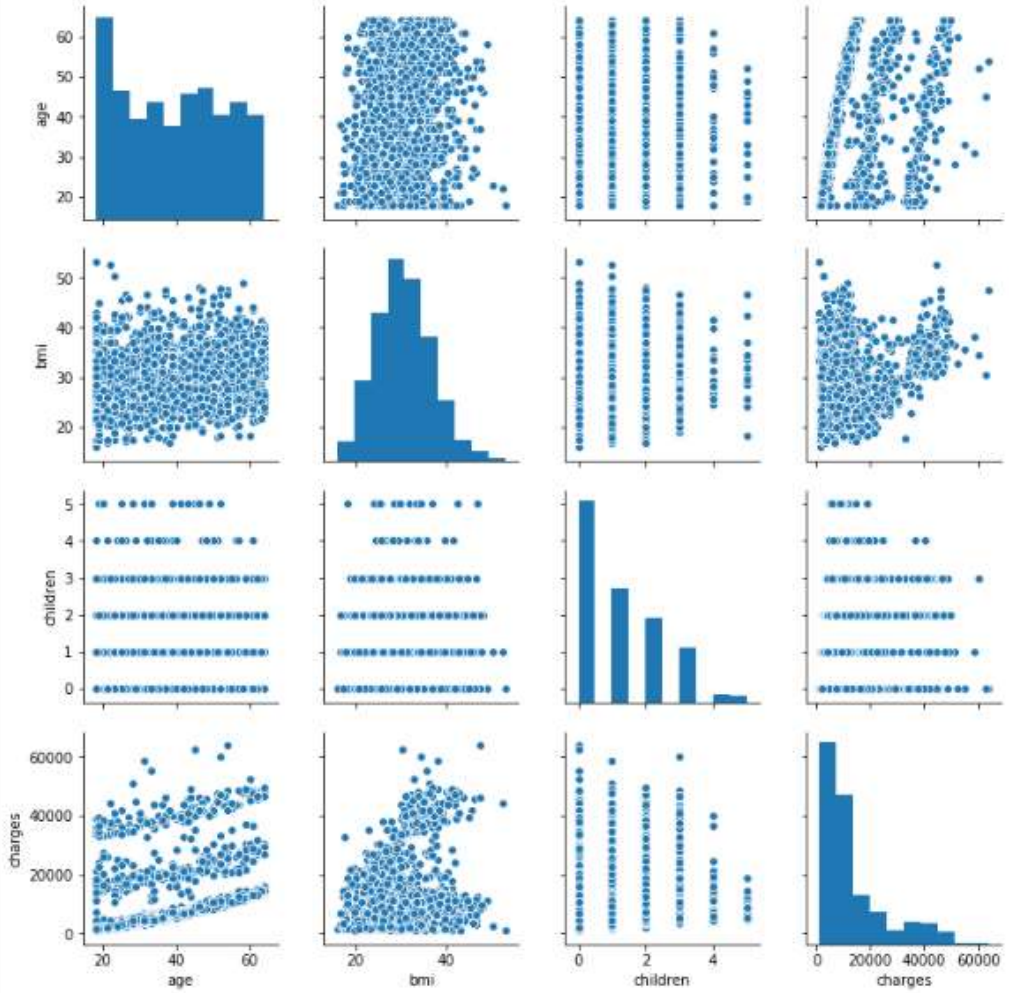


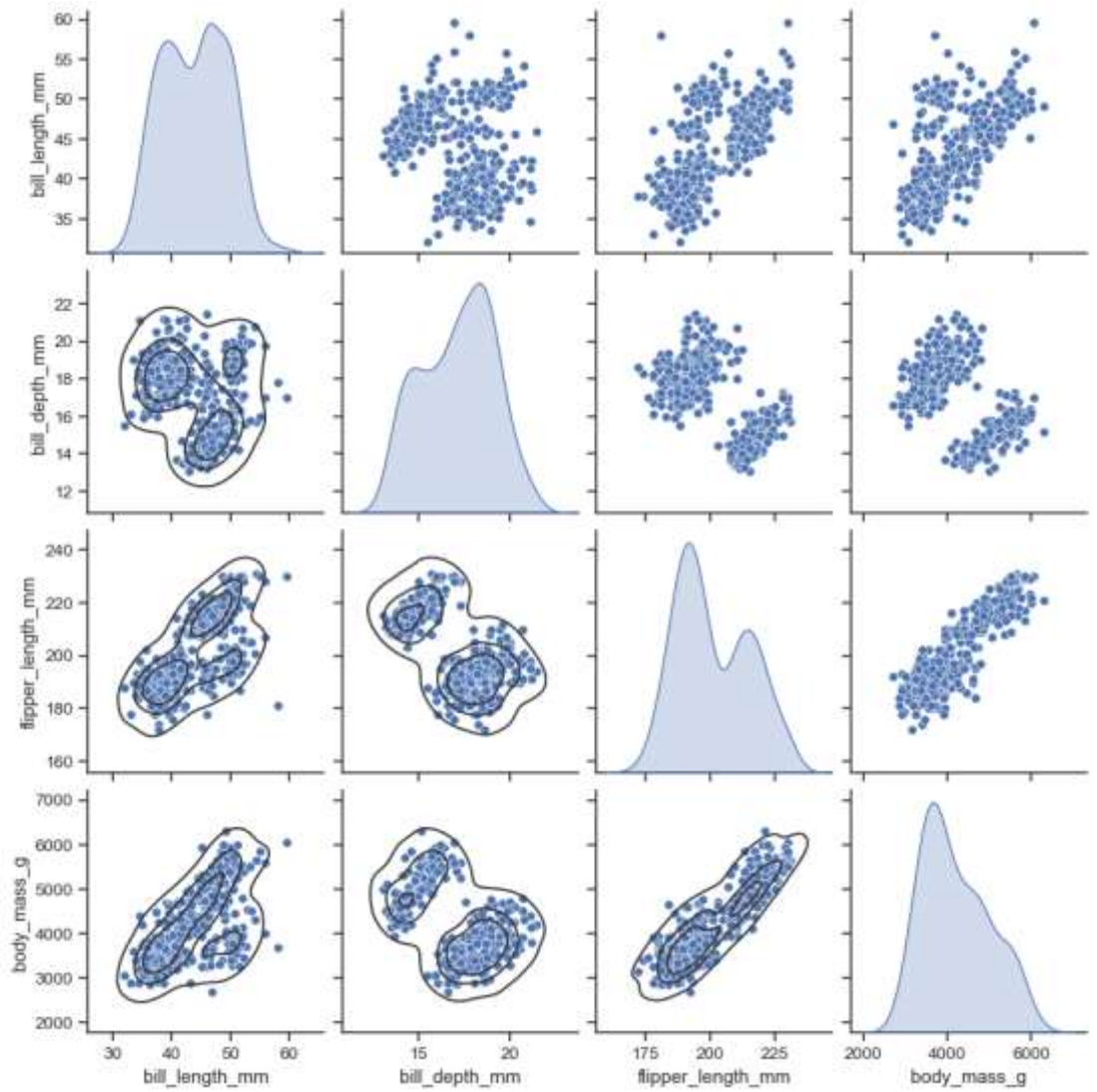


Pair plot

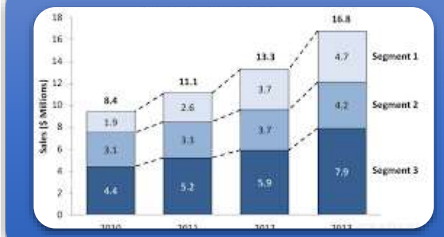
It takes all the numerical attributes of the data and plot pairwise **scatter plot for two different variables** and **histograms from the same variables**.

```
In [4]: sns.pairplot(df)
Out[4]: <seaborn.axisgrid.PairGrid at 0x1afb1b64ba8>
```



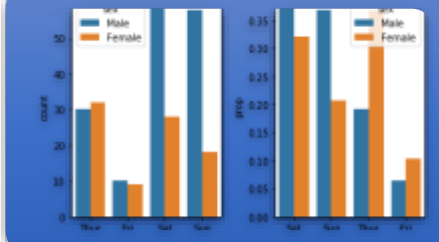


Categorical Plots



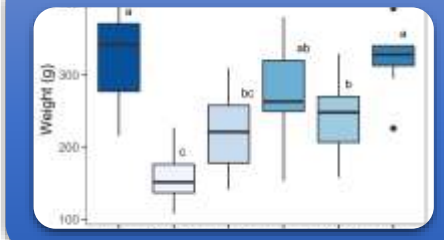
Bar Plot

- On the x-axis, we have a categorical variable and on the y-axis, we have a continuous variable.
- **Bivariate analysis.**



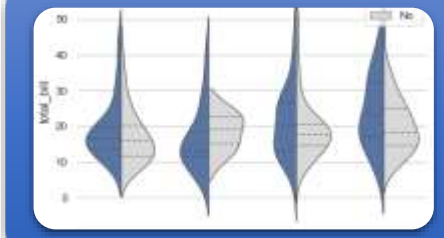
Count Plot

- It counts the number of occurrences of categorical variables.
- **Univariate analysis**.



Box Plot

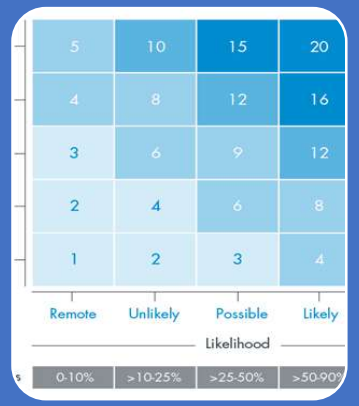
- It is a **5 point summary plot**. It gives the information about the maximum, minimum, mean, first quartile, and third quartile of a continuous variable. Also, it equips us with knowledge of outliers.



Violin Plot

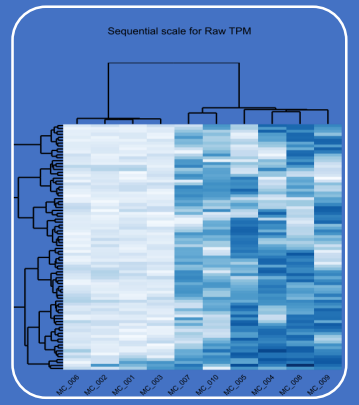
- It is similar to the Box plot, but it gives supplementary information about the distribution too.

Matrix Plots



Heat Map

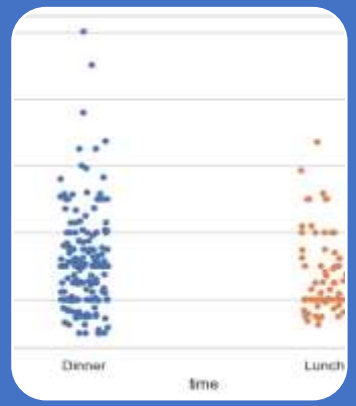
- A **data visualization** technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



Cluster Map

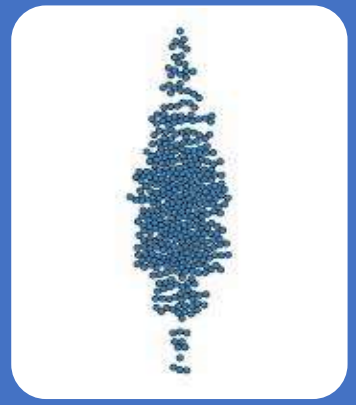
- a matrix data and want to group some features according to their similarity, cluster maps can assist us.
- Cluster maps use **Hierarchical clustering** to form different clusters. What is **Dendrogram??**
- sns.clustermap(df, figsize=(10,8), annot=True);

Advanced Plots



Strip Plot

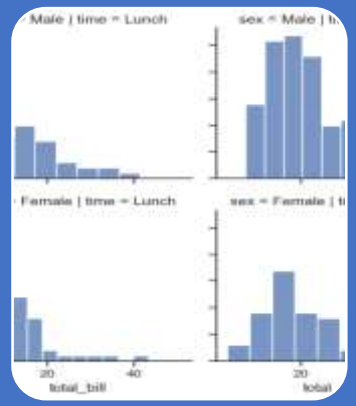
- It’s a plot between a continuous variable and a categorical variable.
- plots as a scatter plot but supplementarily uses categorical encodings of the categorical variable.



Swarm Plot

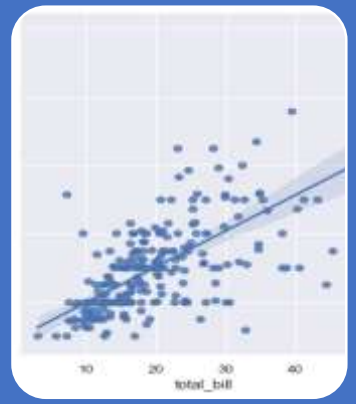
- It is the combination of a strip plot and a violin plot
- Along with the number of data points, it also provides their respective distribution.

Miscellaneous Plots



Grid: (*Facet Grid)*

- **facet_grid() forms a matrix of panels defined by row and column faceting variables**. It is most useful when you have two discrete variables, and all combinations of the variables exist in the data. If you have only one variable with many levels, try facet_wrap() .

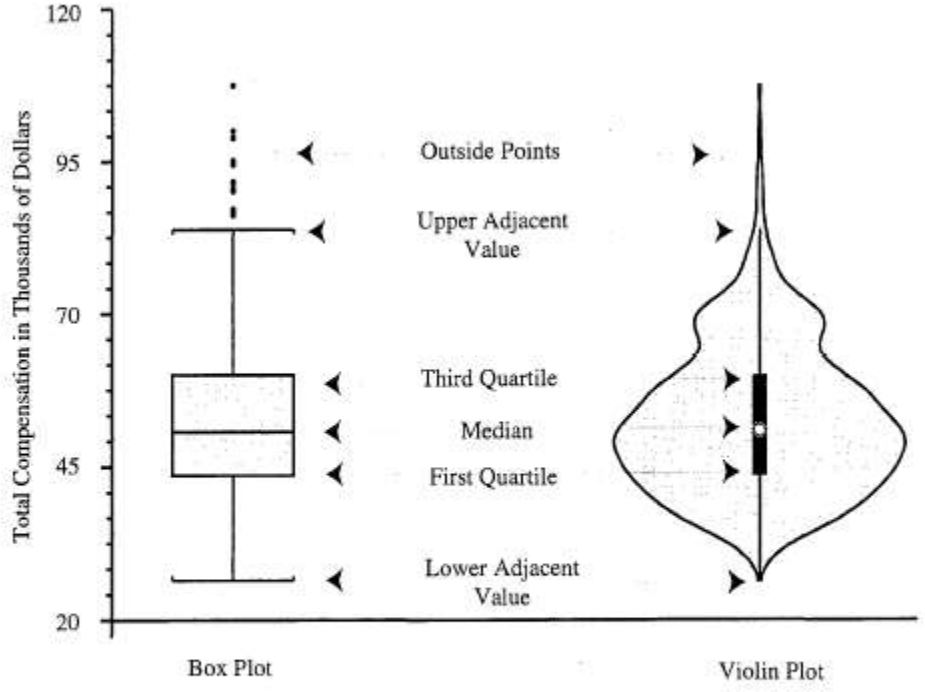


Regression Plot

- More advanced statistical plot that provides a scatter plot along with a linear fitting on the data.

Categorical Plots Example

The advantage of a **violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot**. On the other hand, the ***boxplot more clearly shows the outliers in the data.***



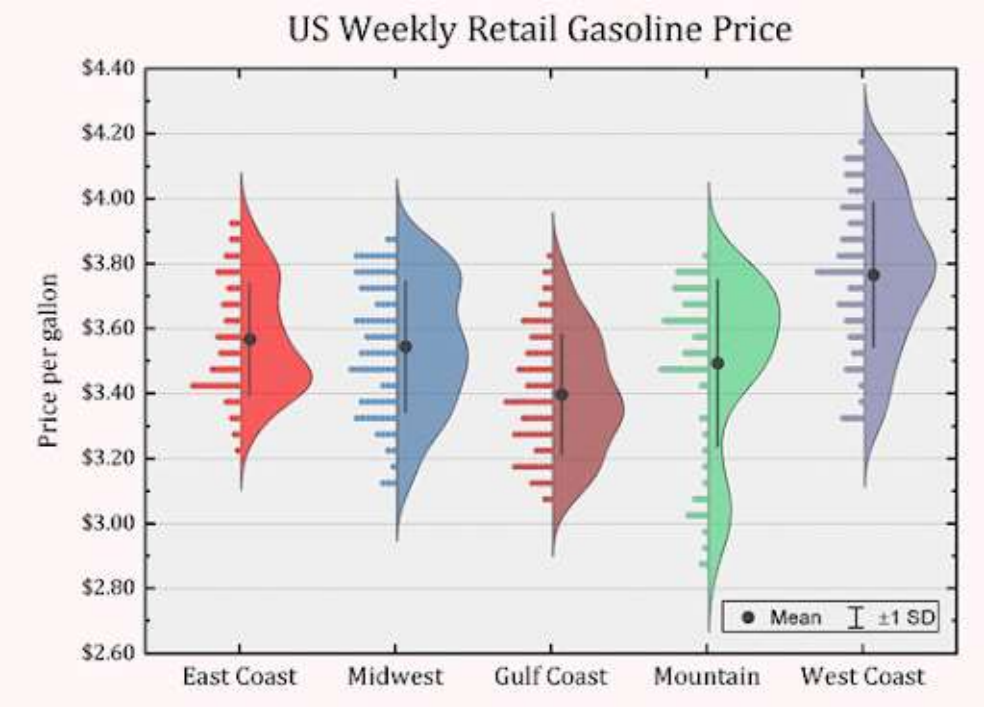
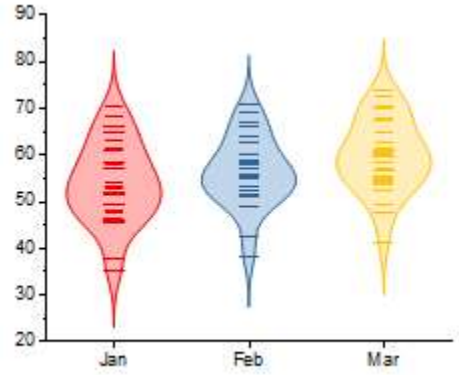
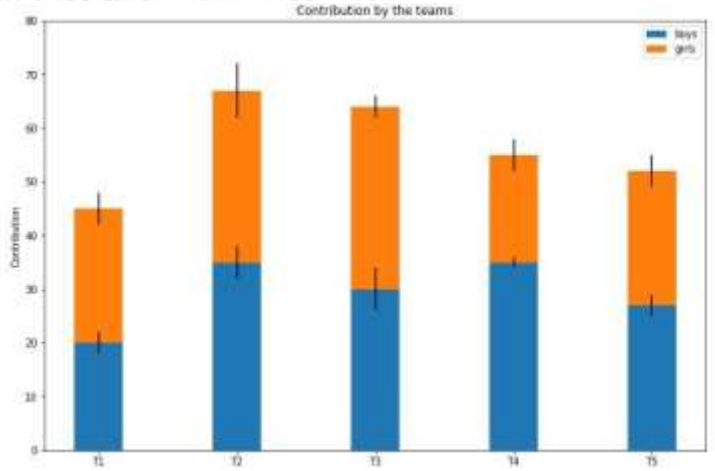


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

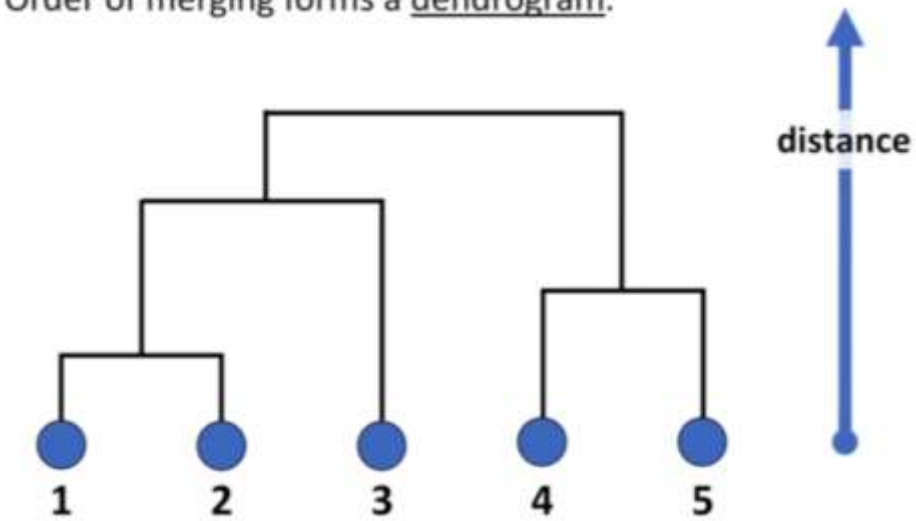




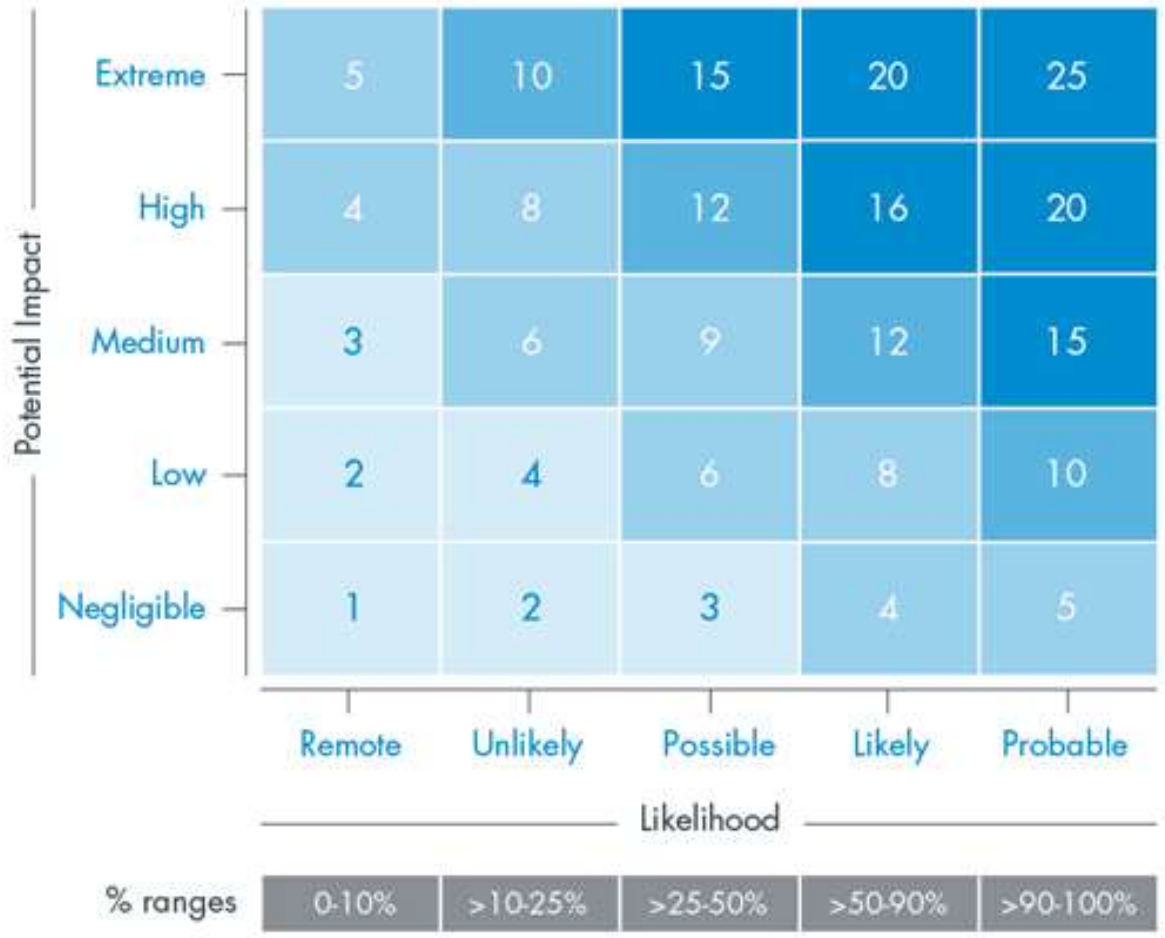
DENDROGRAM

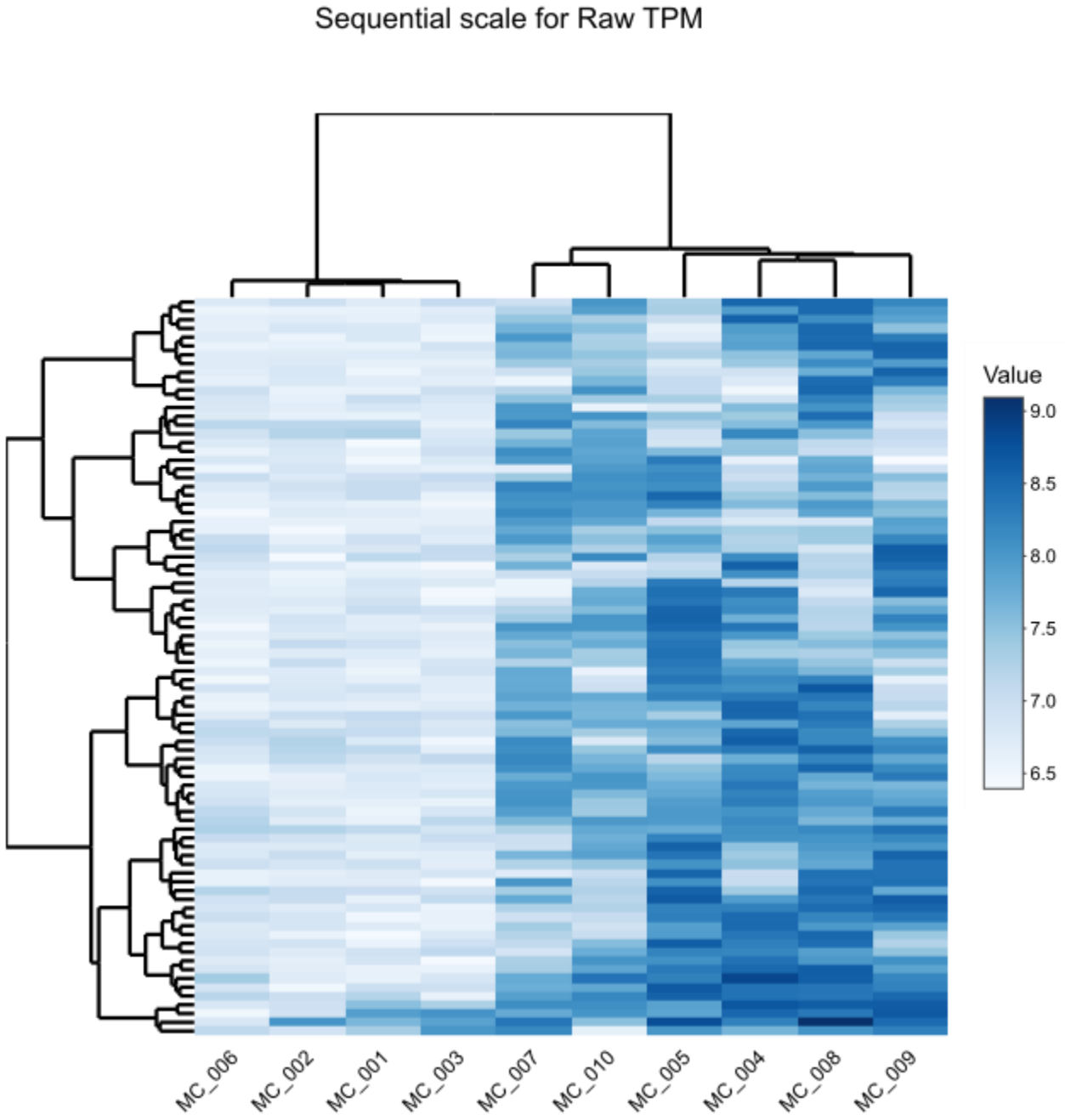
- By default **Euclidean distance** is used (Cosine , Manhattan distance are alternative).
- Closer the height of dendrogram, ***closer is the cluster.***

Order of merging forms a dendrogram.



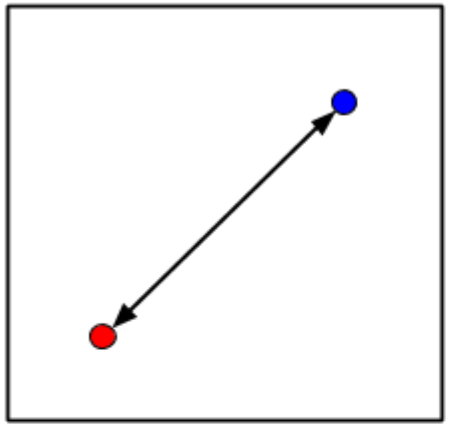
Matrix Plots Example



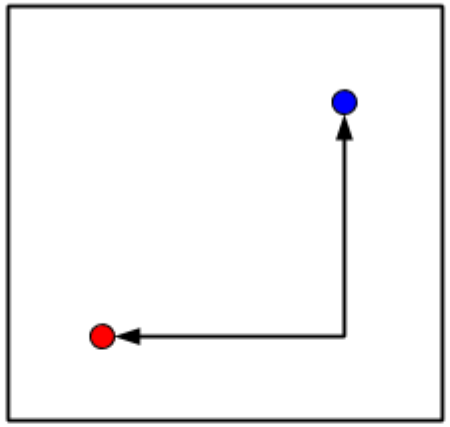


Methods to measure distance between 2 points

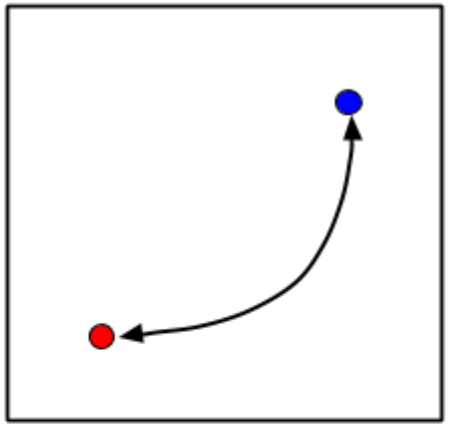
Euclidean



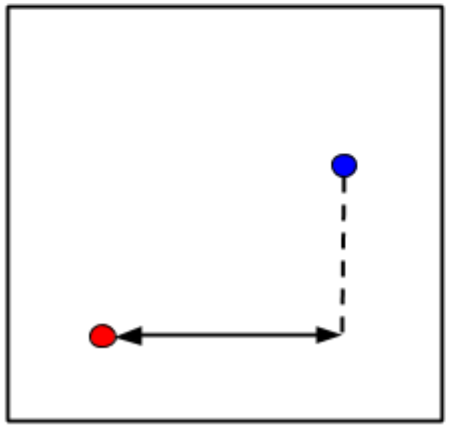
Manhattan



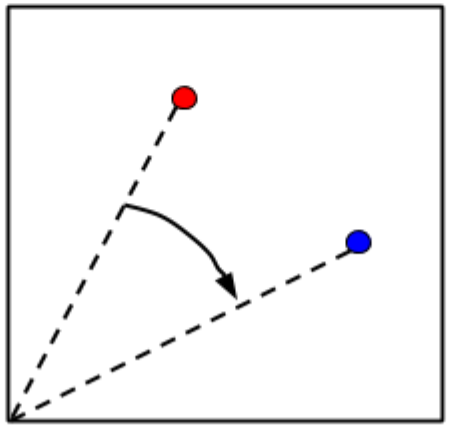
Minkowski



Chebychev

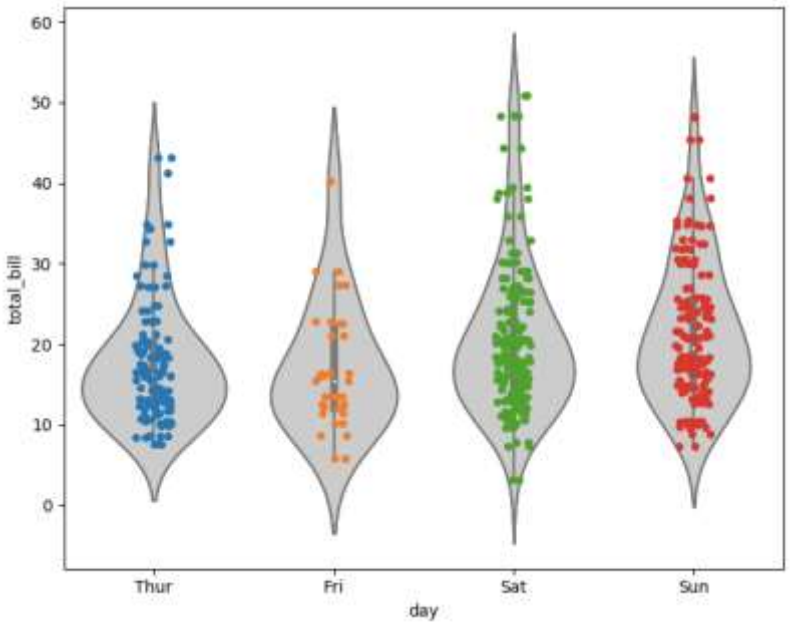


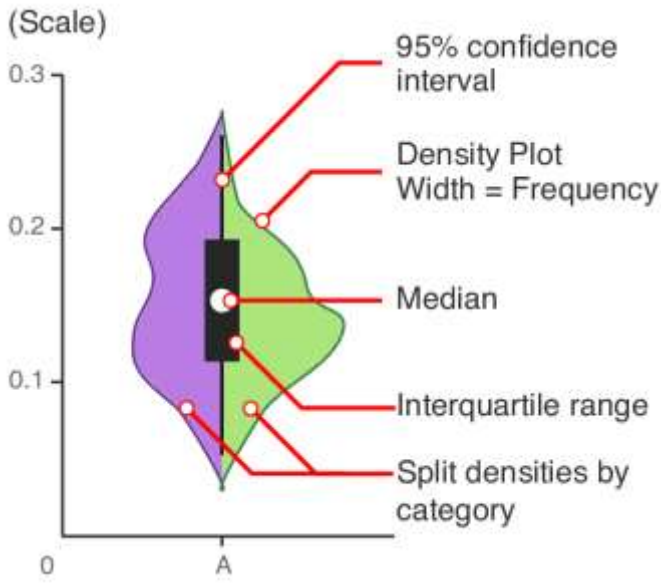
Cosine Similarity

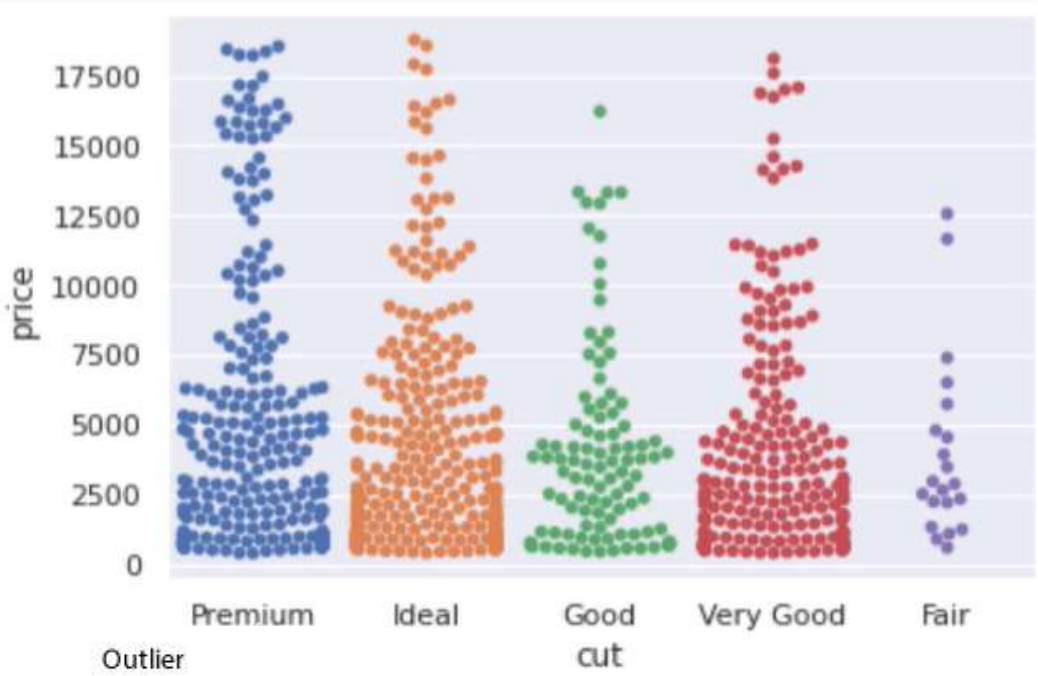


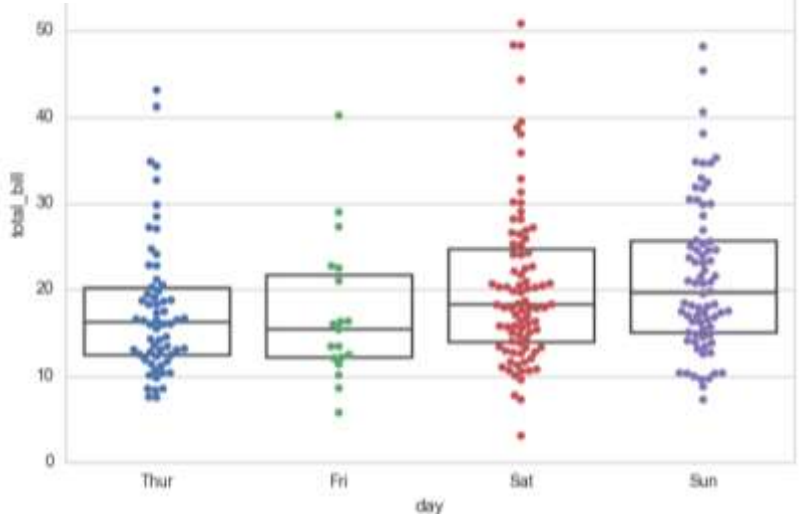
Hamming

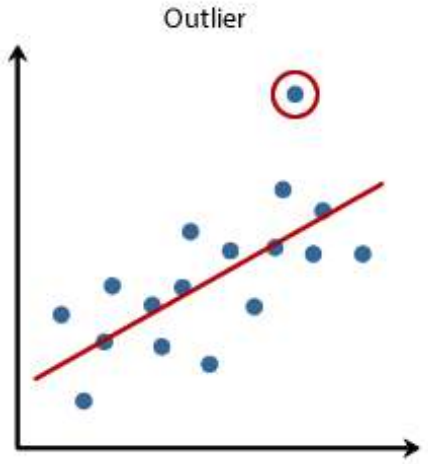
Other Plot Example

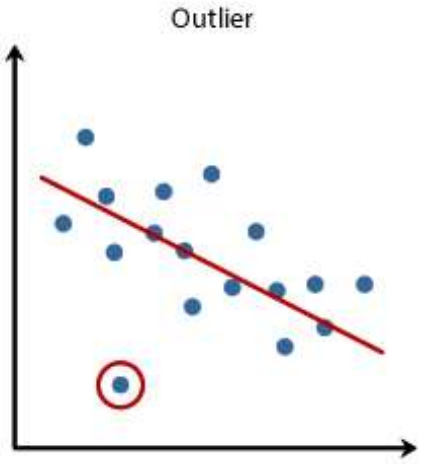


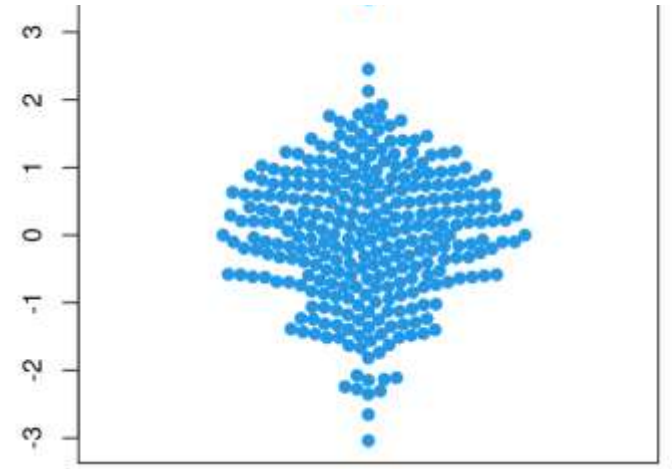


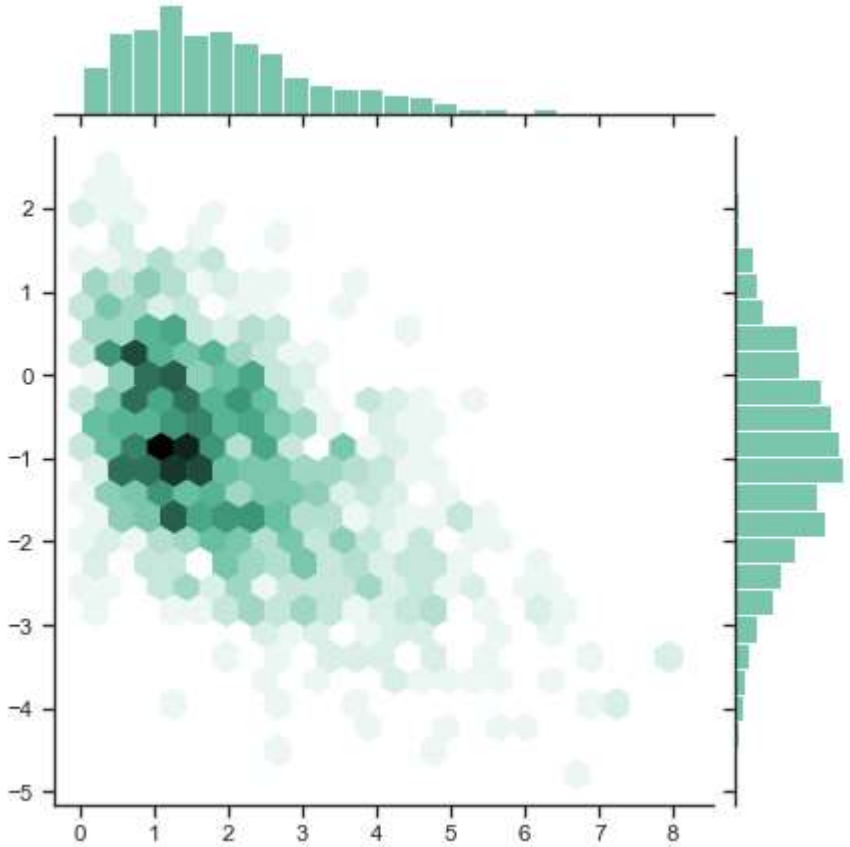


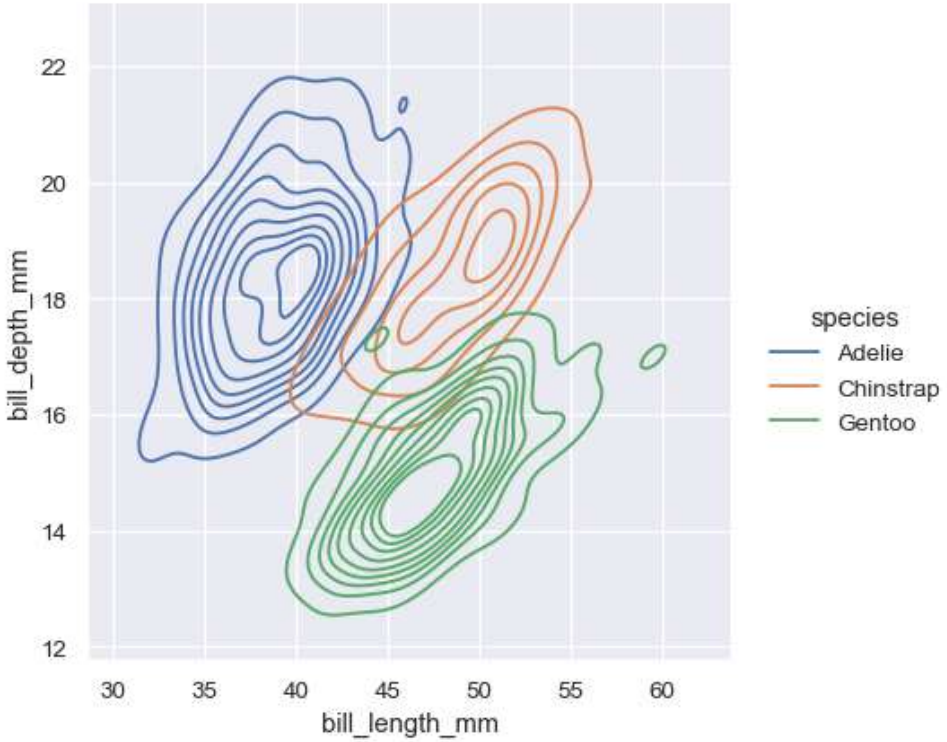












sns.displot(penguins, x='boll_length_mm, y= 'bill_depth_mm' , hue='species', kind= 'kde')

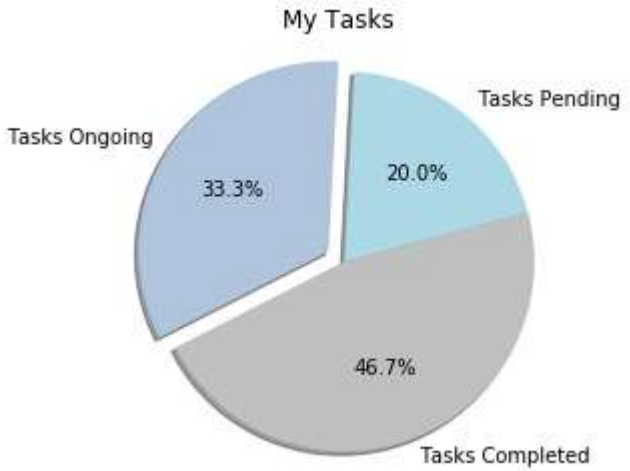
Contour plots:

A plot showing the density of two numeric variables like a topographical map

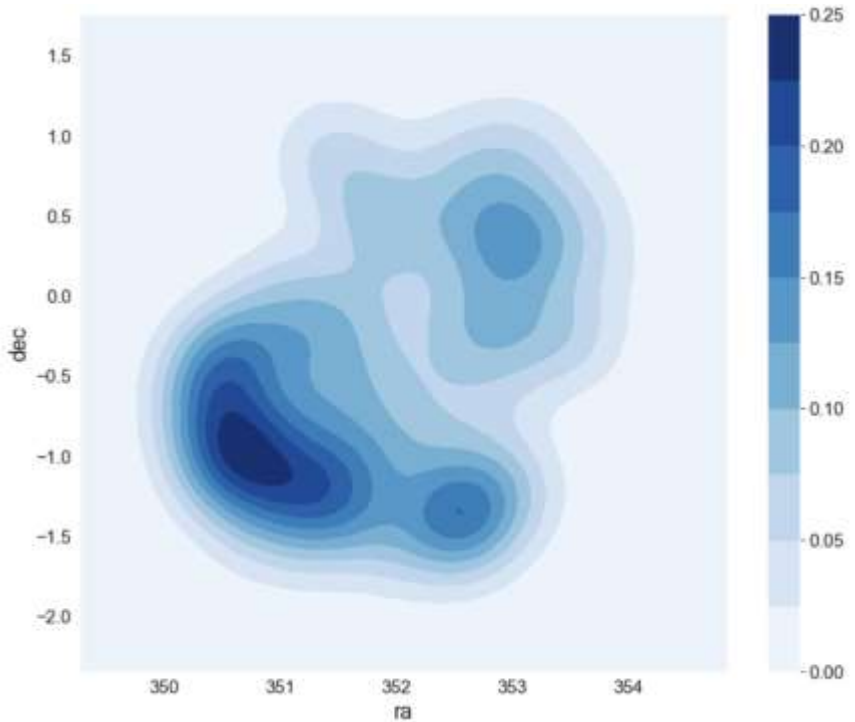
Hexagonal binning:

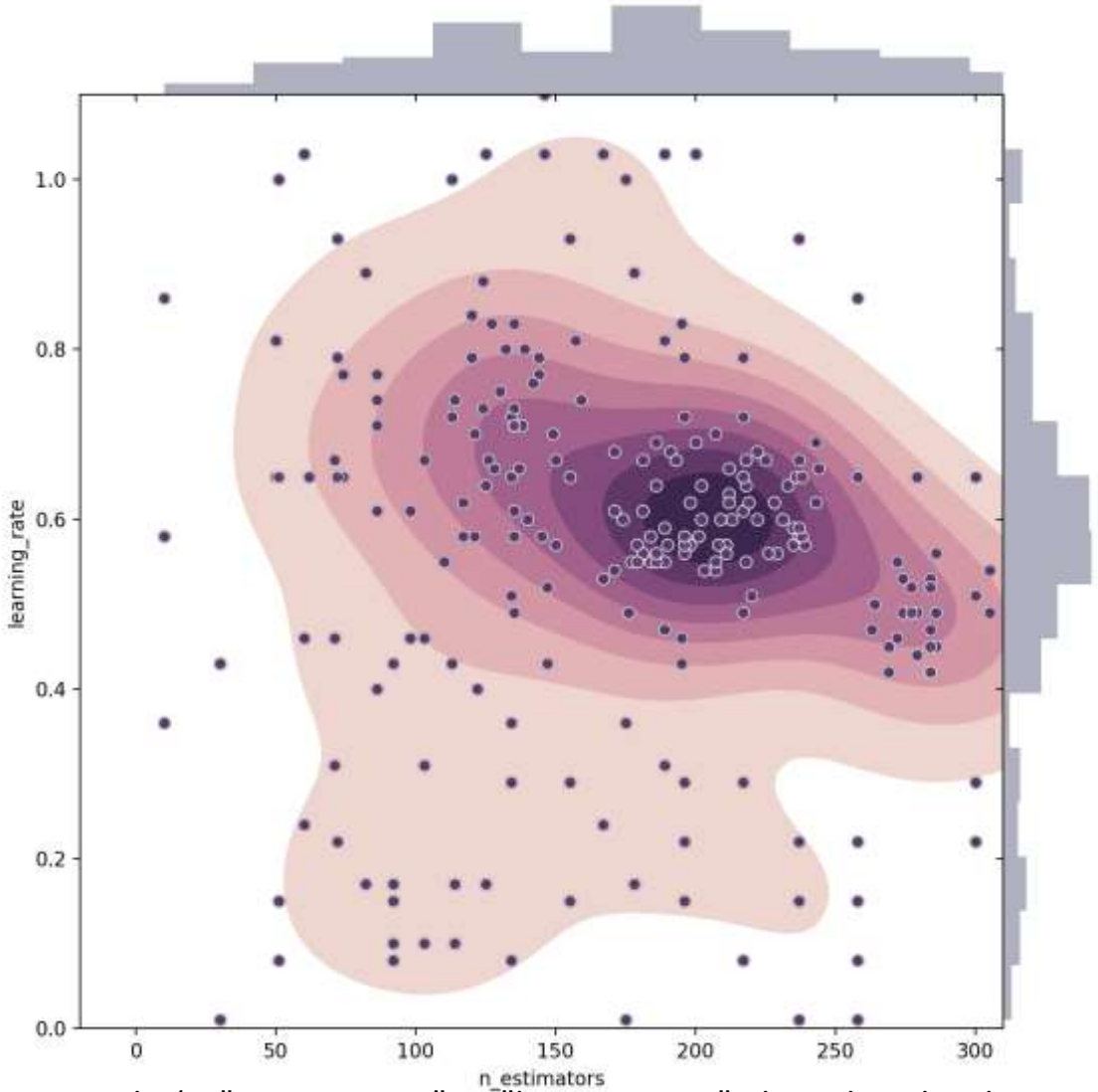
A plot of two numeric variables with the records binned into hexagons.

sns.jointplot(x=x, y=y, kind= 'hex', color='#4CB391')

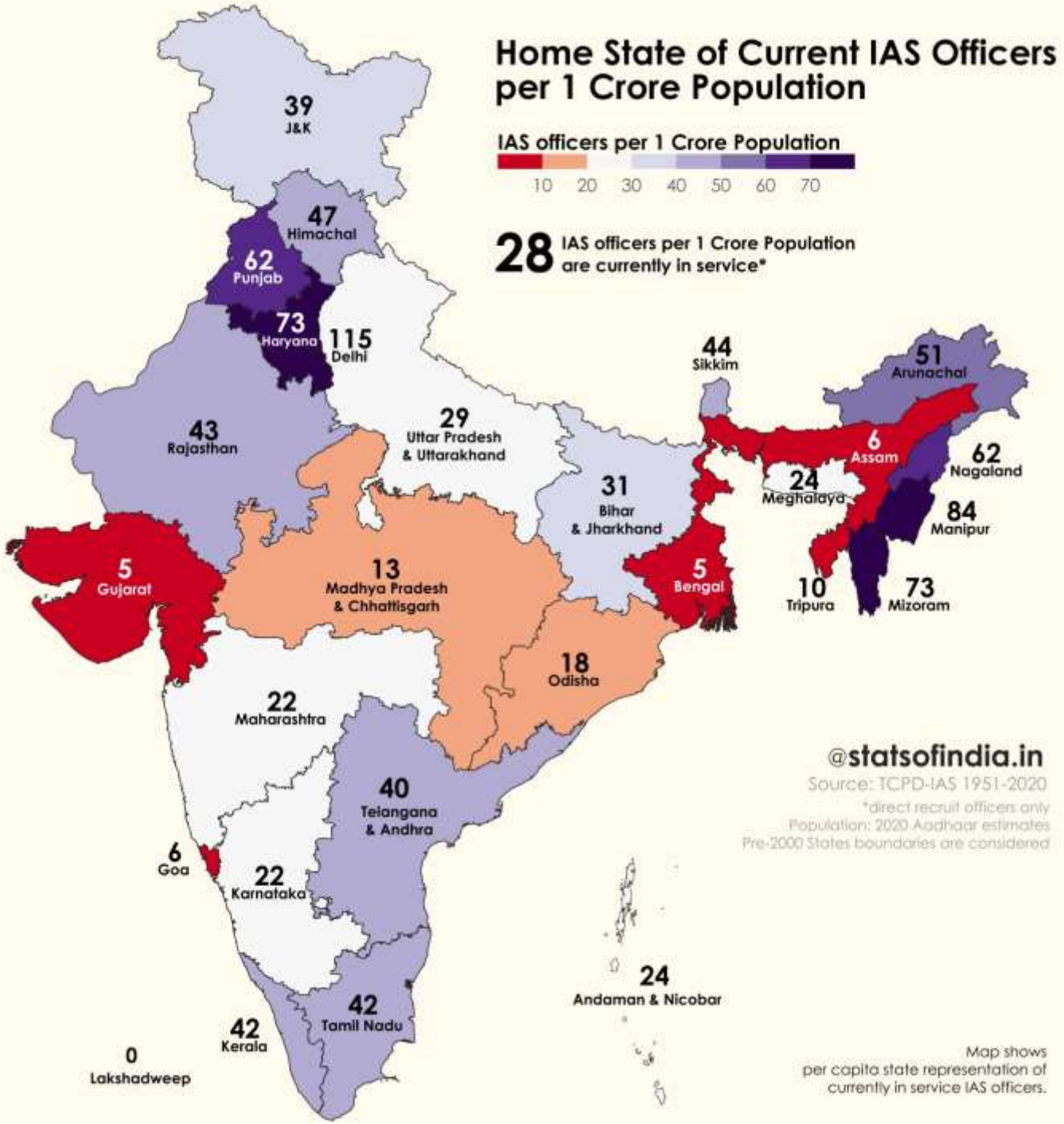


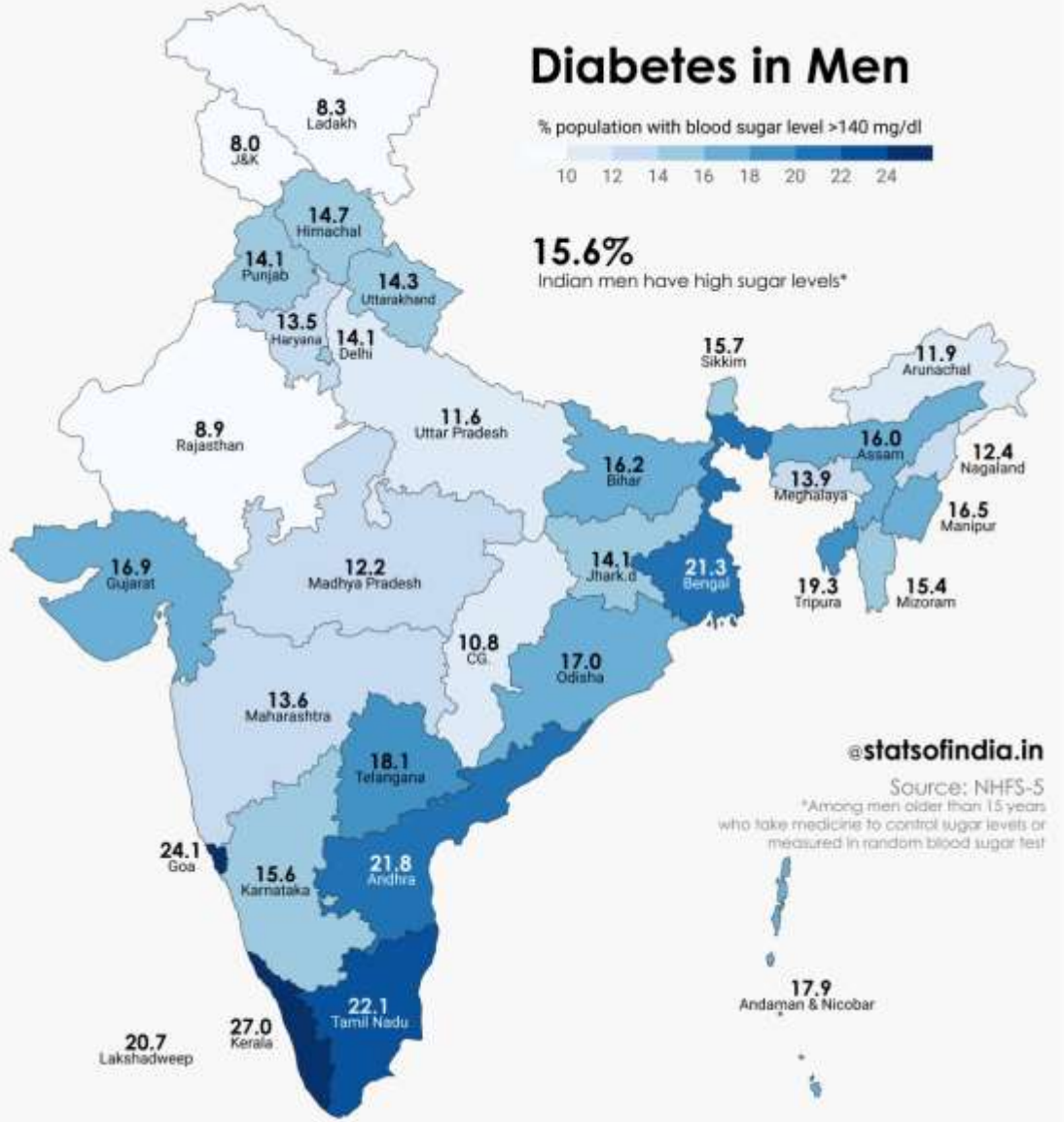
Pie charts:
The frequency or proportion for each category plotted as wedges in a pie.

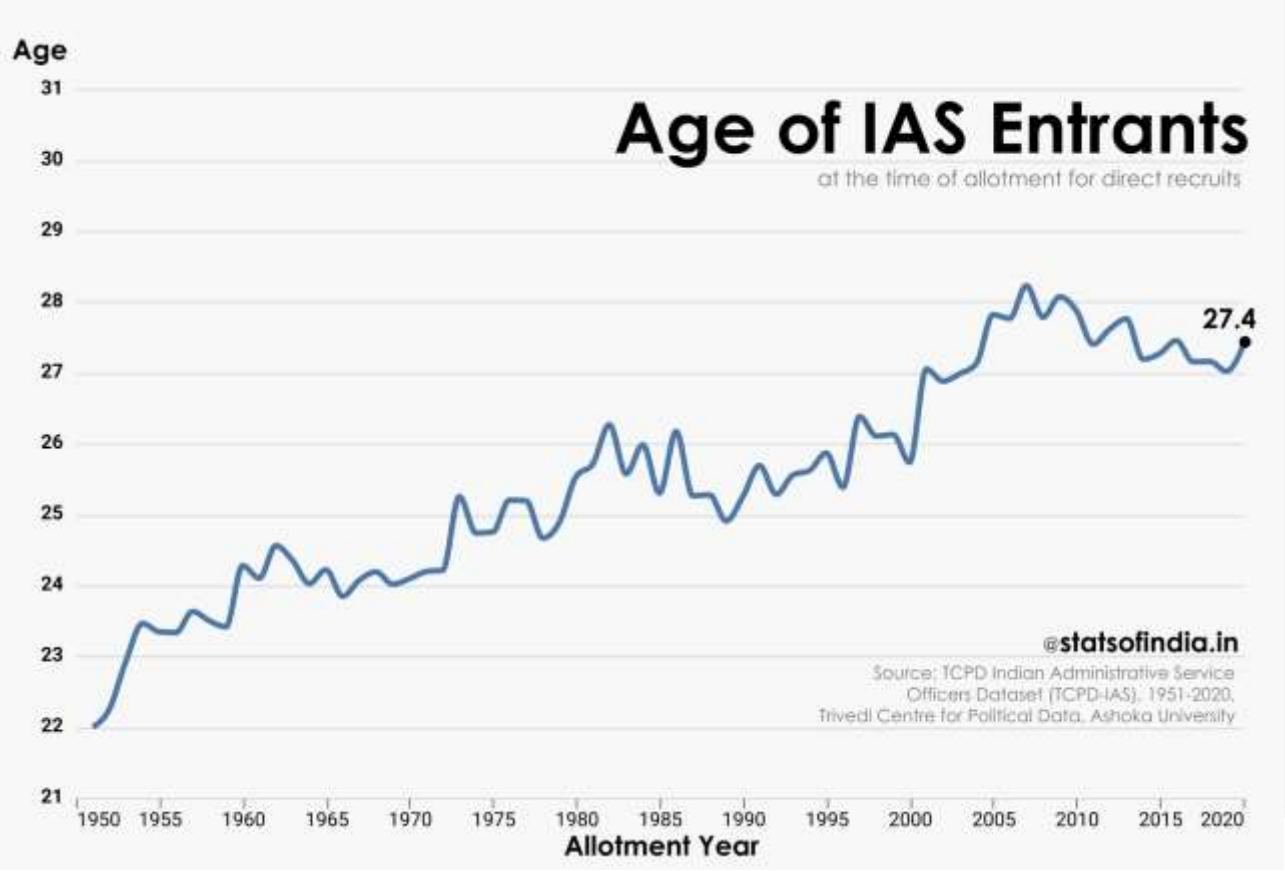


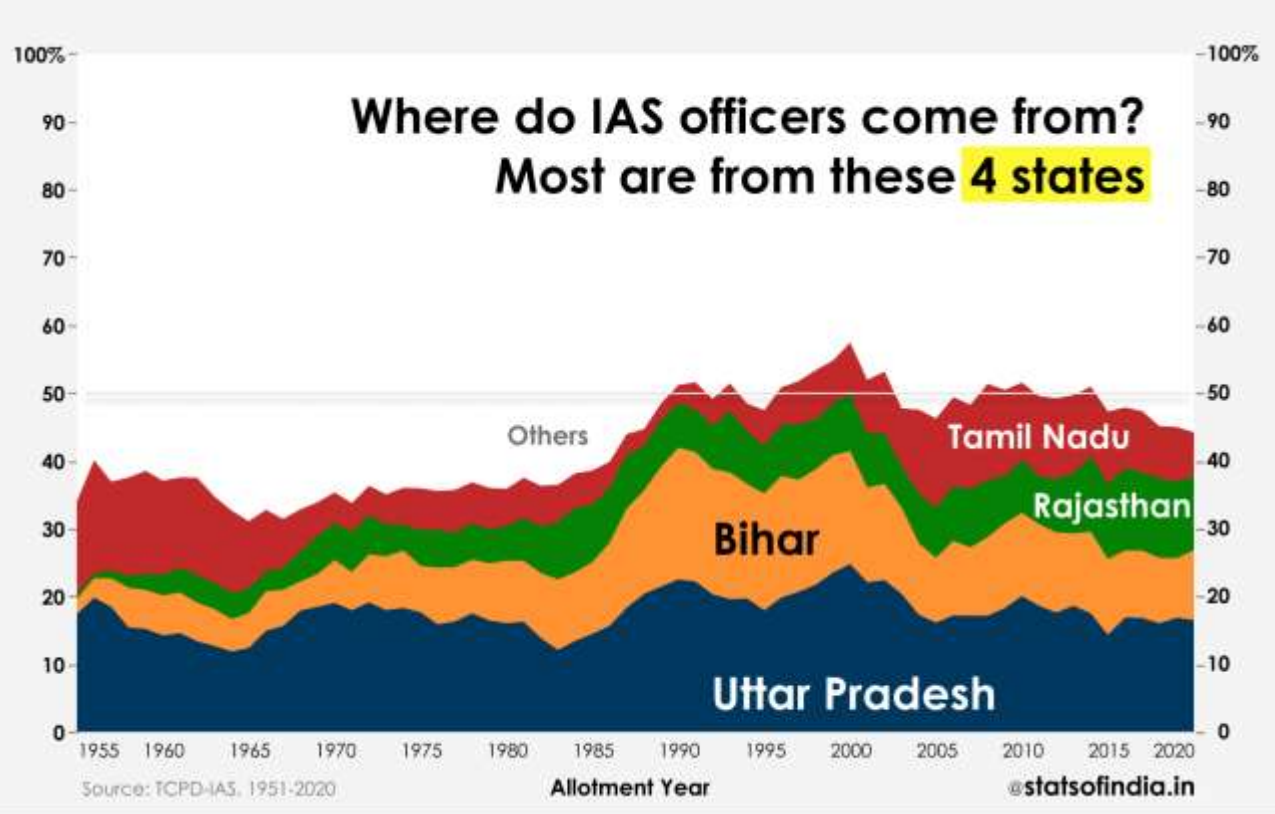



```
sns.jointplot(x="n_estimators", y="learning_rate", data=data, height=8,
ratio=10, space=0, color="#383C65")\
                .plot_joint(sns.kdeplot, zorder=0, shade=True,
shade_lowest=False, cmap=sns.cubehelix_palette(light=1, as_cmap=True),
legend=True, cbar=False, cbar_kws={})
```


Home State of Current IAS Officers per 1 Crore Population
IAS officers per 1 Crore Population
10 20 30 40 50 60 70
28 IAS officers per 1 Crore Population are currently in service*
39 J&K
47 Himachal
62 Punjab
73 Haryana
115 Delhi
43 Rajasthan
29 Uttar Pradesh & Uttarakhand
31 Bihar & Jharkhand
44 Sikkim
51 Arunachal
6 Assam
62 Nagaland
24 Meghalaya
84 Manipur
10 Tripura
73 Mizoram
5 Bengal
5 Gujarat
13 Madhya Pradesh & Chhattisgarh
18 Odisha
22 Maharashtra
40 Telangana & Andhra
6 Goa
22 Karnataka
42 Tamil Nadu
42 Kerala
0 Lakshadweep
24 Andaman & Nicobar
@statsofindia.in
Source: TCPD-IAS 1951-2020
*direct recruit officers only
Population: 2020 Aadhaar estimates
Pre-2000 States boundaries are considered
Map shows per capita state representation of currently in service IAS officers.


Diabetes in Men
% population with blood sugar level >140 mg/dl
10 12 14 16 18 20 22 24
15.6%
Indian men have high sugar levels*
8.3 Ladakh
8.0 J&K
14.7 Himachal
14.1 Punjab
14.3 Uttarakhand
13.5 Haryana
14.1 Delhi
15.7 Sikkim
11.9 Arunachal
8.9 Rajasthan
11.6 Uttar Pradesh
16.2 Bihar
16.0 Assam
12.4 Nagaland
13.9 Meghalaya
16.5 Manipur
16.9 Gujarat
12.2 Madhya Pradesh
14.1 Jhark.d
21.3 Bengal
19.3 Tripura
15.4 Mizoram
10.8 CG
17.0 Odisha
13.6 Maharashtra
18.1 Telangana
24.1 Goa
15.6 Karnataka
21.8 Andhra
22.1 Tamil Nadu
27.0 Kerala
20.7 Lakshadweep
17.9 Andaman & Nicobar
@statsofindia.in
Source: NHFS-5
*Among men older than 15 years who take medicine to control sugar levels or measured in random blood sugar test

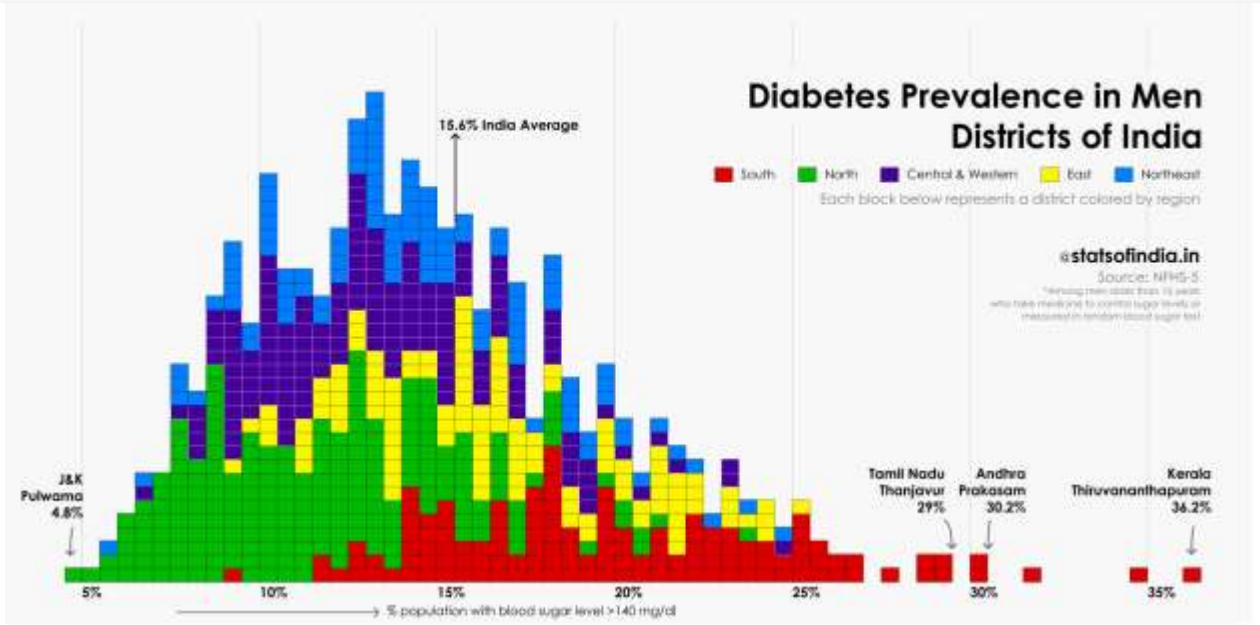




Pratap Vardhan
@PratapVardhan

Making sense of the world, one data point at a time. • 👨‍💻 Data Science @KhanAcademy • 🛠️ @Stats_of_India 📈 • pratapvardhan.com

Credits



Note:

- Boxplots and violin plots allow you to plot a numeric variable against a categorical variable.

  - This received alot of attention initially as it underlies the machinery of hypothesis tests and confidence intervals.
  - but, since formal hypothesis tests and confidence intervals play a small role in data science, and the bootstrap is available in any case, the central limit theorem is not so central in the practice of data science.

Descriptive Statistics

What Are Descriptive Statistics?

- Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the ***entire population*** or a ***sample of a population.***
- Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).
- **Measures of central tendency describe the center of a data set.**
  - ✓ **Measures of central tendency** include the mean, median, and mode.
- **Measures of variability or spread describe the dispersion of data within the set.**
  - ✓ **Measures of variability** include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.

Basics of Statistics

- ***Sample:*** A subset from a larger data set.
- ***Population:*** The larger data set or idea of a data set.
- ***N (n):*** The size of the population (sample).
- ***Random sampling:*** Drawing elements into a sample at random.
- ***Stratified sampling:*** Dividing the population into strata and randomly sampling from each strata.
- ***Simple random sample:*** The sample that results from random sampling without stratifying the population.
- ***Sample bias:*** A sample that misrepresents the population
- ***Bias:*** Systematic error.
- ***Data snooping:*** Extensive hunting through data in search of something interesting.
- ***Vast search effect:*** Bias or nonreproducibility resulting from repeated data modeling, or modeling data with large
- numbers of predictor variables.

PARAMETER

A number that describes
the data from a population

STATISTIC

A number that describes
the data from a sample

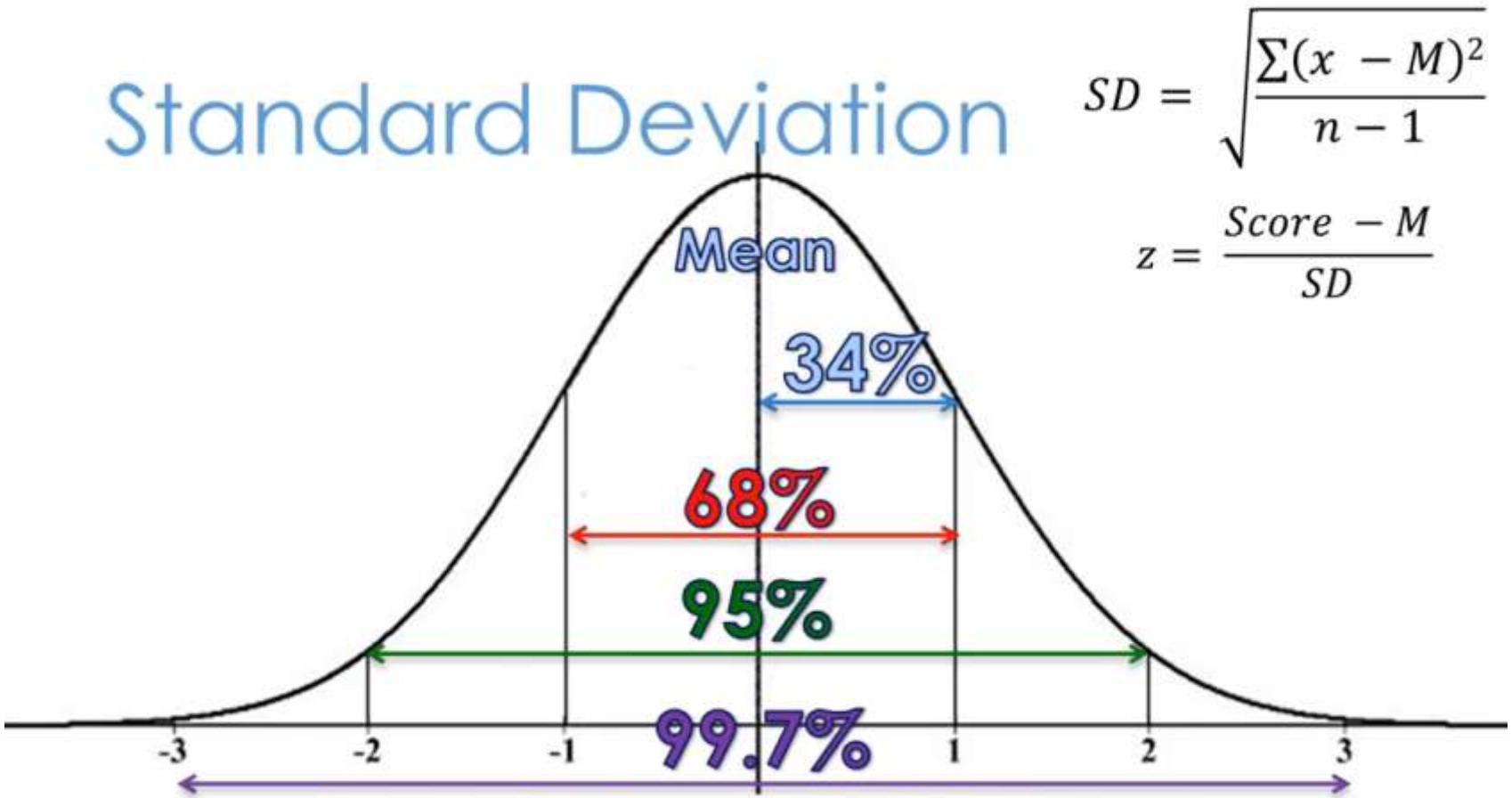
	SAMPLE	POPULATION
MEAN	$\bar{x}$	μ
STANDARD DEVIATION	s	σ
	STATISTIC	PARAMETER

- *Mean* ***(***average): The sum of all values divided by the number of values.
- *Weighted mean* ***(***weighted average***):*** The sum of all values times a weight divided by the sum of the weights.
- *Median* A.K.A *50th percentile****:*** The value such that one-half of the data lies above and below.
- *Weighted median****:*** The value such that one-half of the sum of the weights lies above and below the sorted data.
- *Trimmed mean (*truncated mean): The average of all values after dropping a fixed number of extreme values.
- *Mode****:*** The most commonly occurring category or value in a data set.
- *Robust (*resistant*)****:*** Not sensitive to extreme values.
- *Outlier (*extreme value): A data value that is very different from most of the data.

Standard Deviation: Measure how much variation exists in a distribution.

Low STD means value are closer to mean

High STD means values are spread out over a large range.



$$SD = \sqrt{\frac{\sum(x - M)^2}{n - 1}}$$

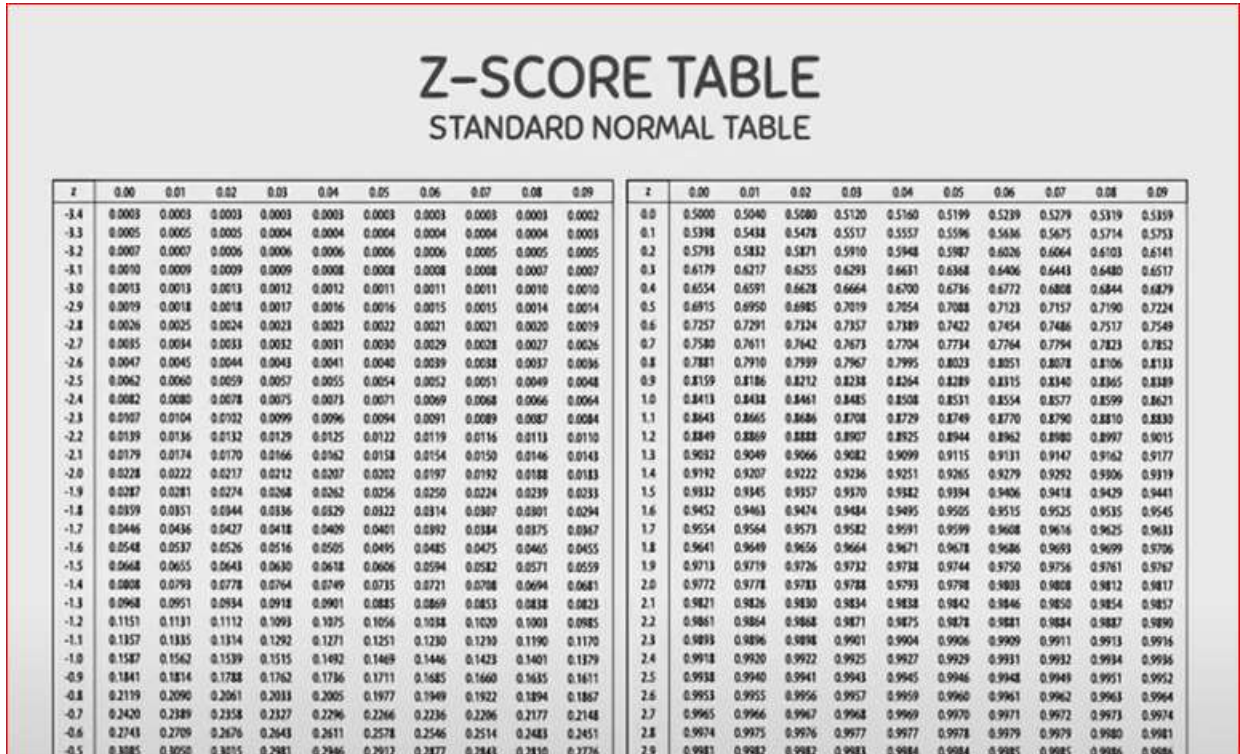
$$z = \frac{Score - M}{SD}$$

Z Scores

- **A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values.**
- Z-score is measured in terms of standard deviations from the mean.
- If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
- A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
- Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

~~ Investopedia.com

Note: Thus Z-Score above or below +3 and -3 respectively is considered outlier (58-95-99.7 Rule)



Z-SCORE TABLE

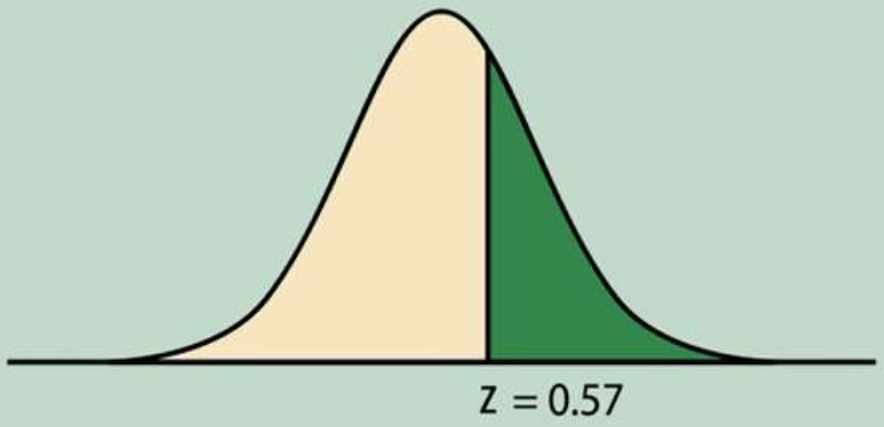
STANDARD NORMAL TABLE

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

$1 - AREA_{LEFT} = AREA_{RIGHT}$

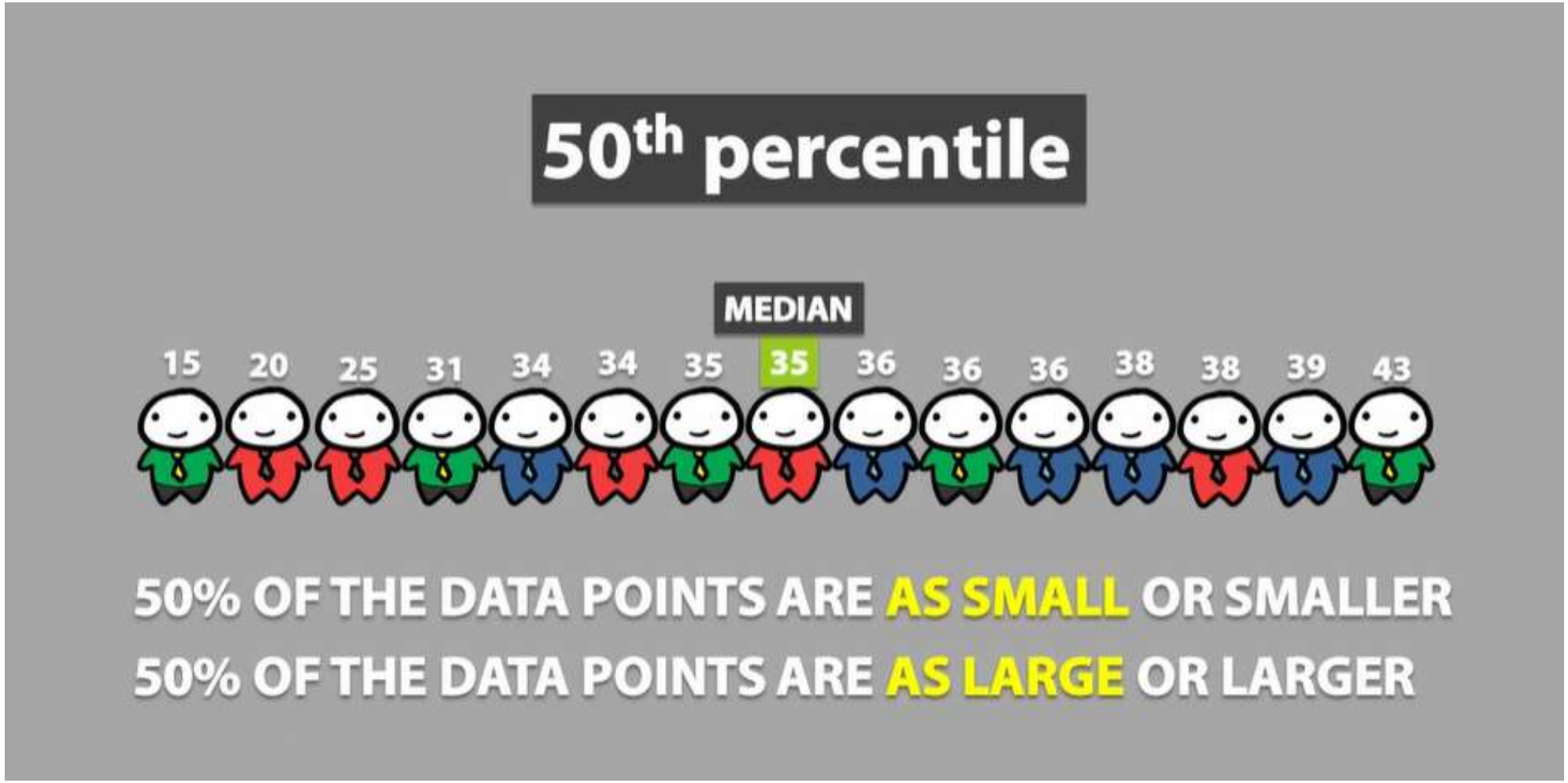
$1 - 0.7157 = 0.2843$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6631	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

Percentiles

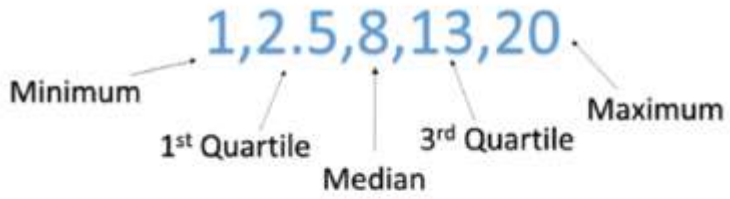
A percentile (or a centile) is **a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall**.



5 Number Summary

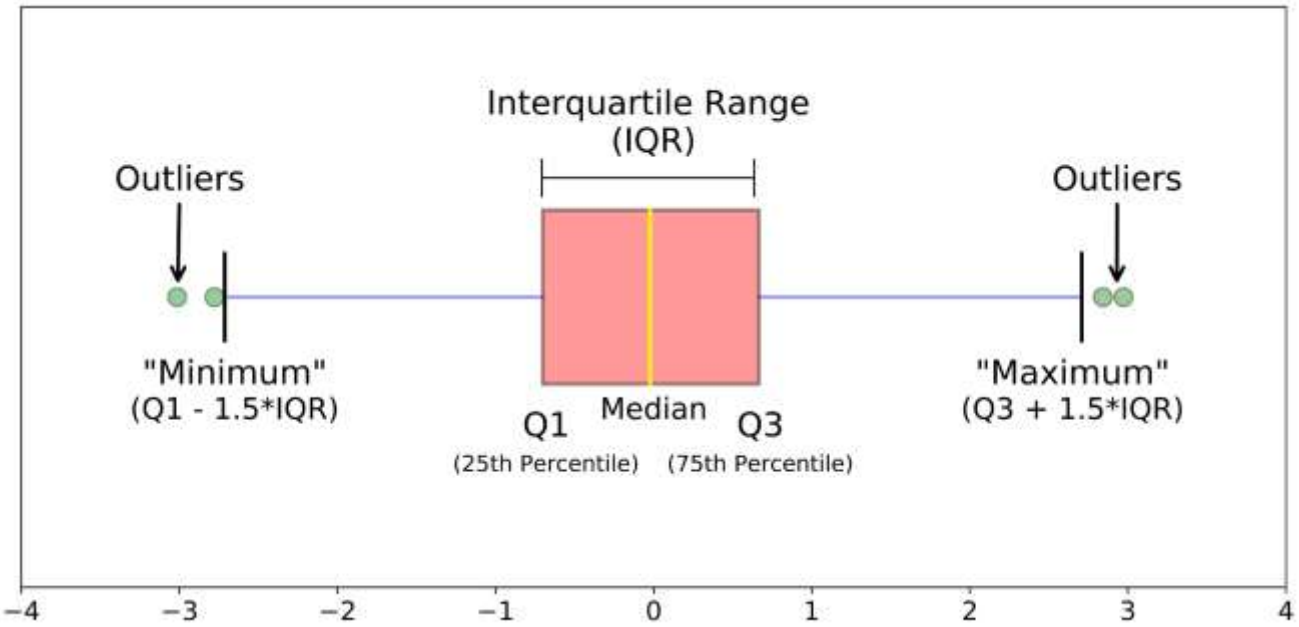
Five Number Summary For Data Set:

1,2,3,4,5,11,11,12,14,20,20



A five-number summary simply consists of the smallest data value, the first quartile, the median, the third quartile, and the largest data value. A box plot is a graphical device based on a five-number summary

What are Whisker??



In other words 5 no summary gives us way to describe a distribution using only 5 Numbers.

df.describe()

Note: Median of 1st half (till Q2) is called 1st Quartile i.e Q1 similarly Median of 2nd half (Q2 to end) is called 3rd Quartile i.e Q3

Outlier : Data Value < Q1 - 1.5(IQR) or Data Value > Q3 + 1.5(IQR)

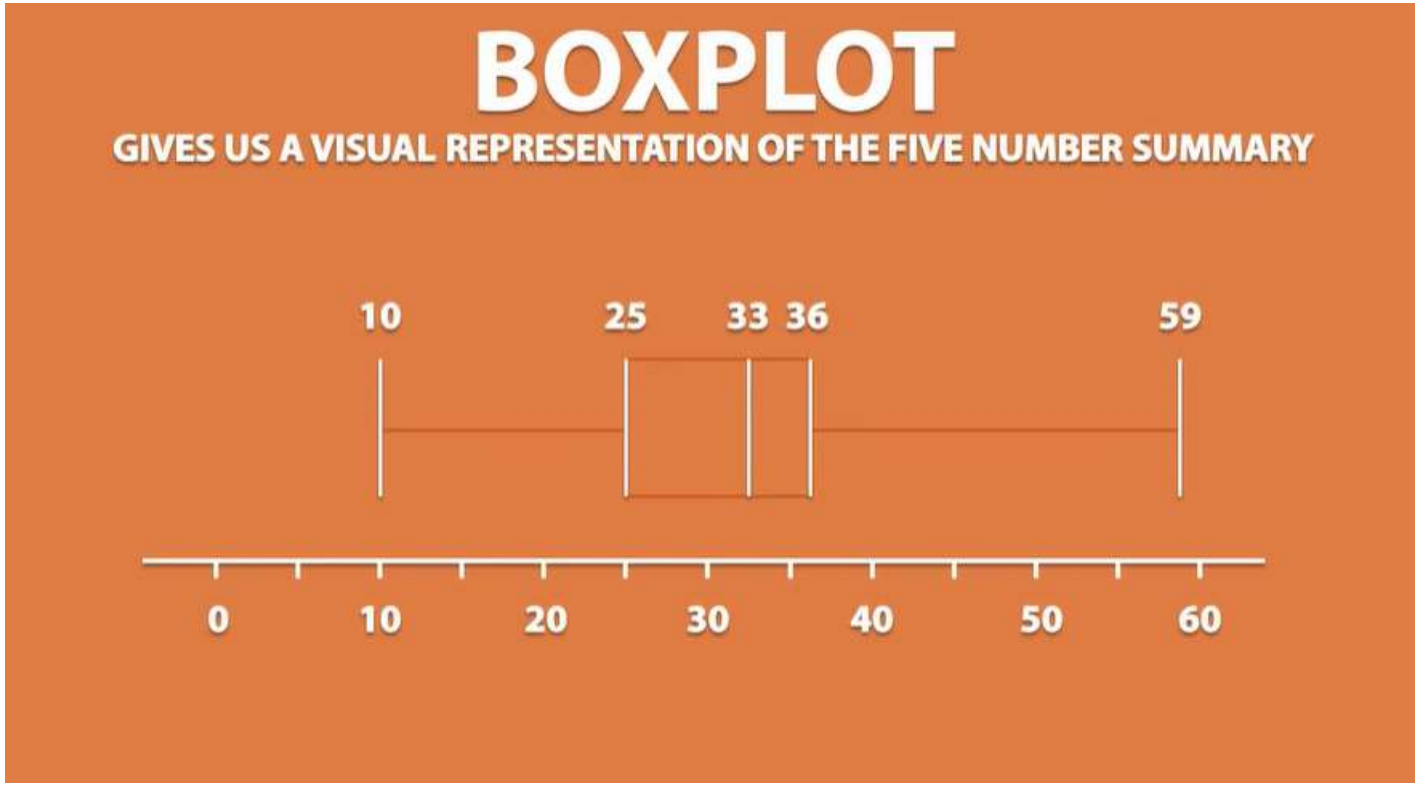
Question

Lets determine 5 Number Summary of: ie. Min , Q1, Median, Q3, Max

10 , 11, 12, 25, 25 ,27, 31, 33, 34 ,34, 35, 36, 43, 50, 59

Now calculate Outlier

Hint: calculate
Q1- 1.5(IQR)
&
Q3 + 1.5(IQR)



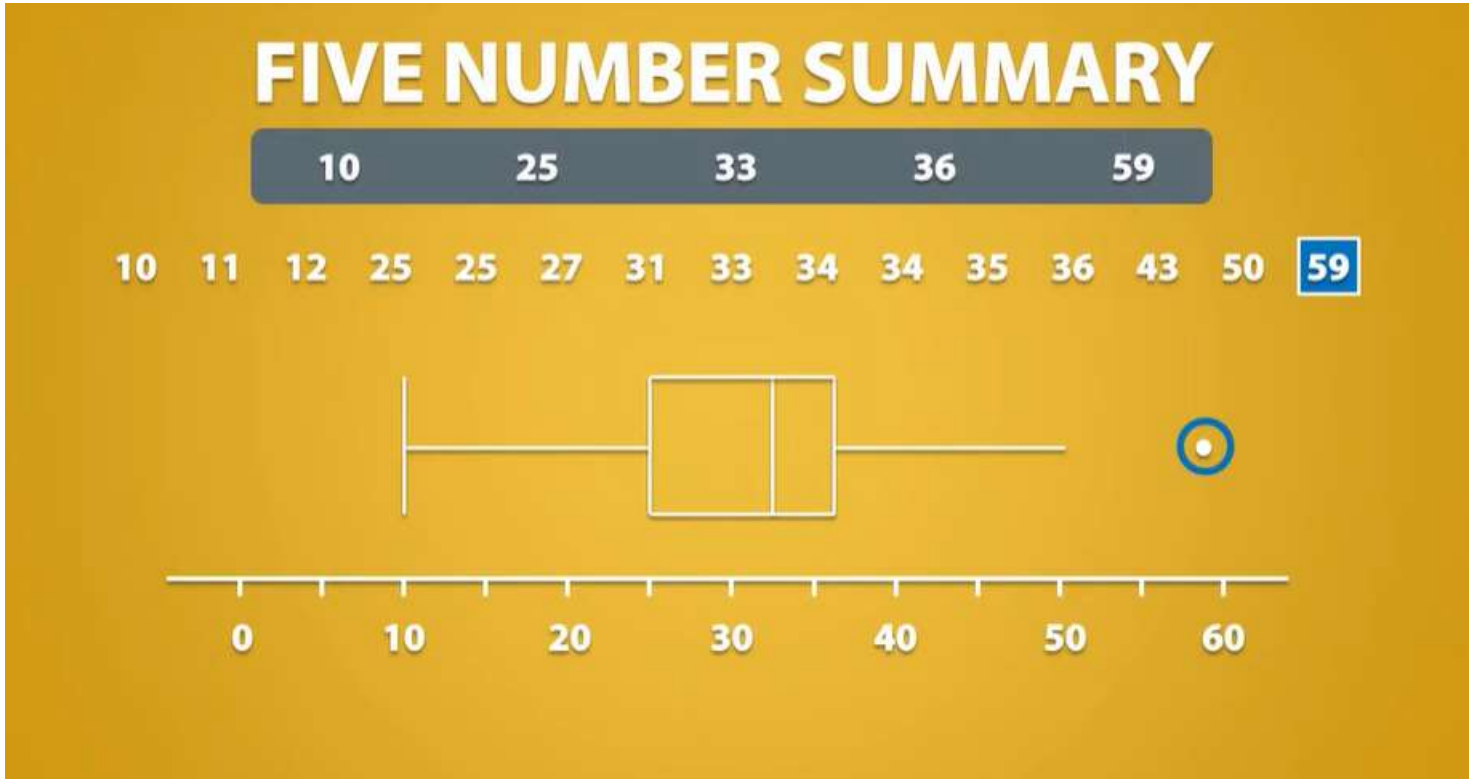
Question

Lets determine 5 Number Summary of: ie. Min , Q1, Median, Q3, Max

10 , 11, 12, 25, 25 ,27, 31, 33, 34 ,34, 35, 36, 43, 50, 59

Now calculate Outlier

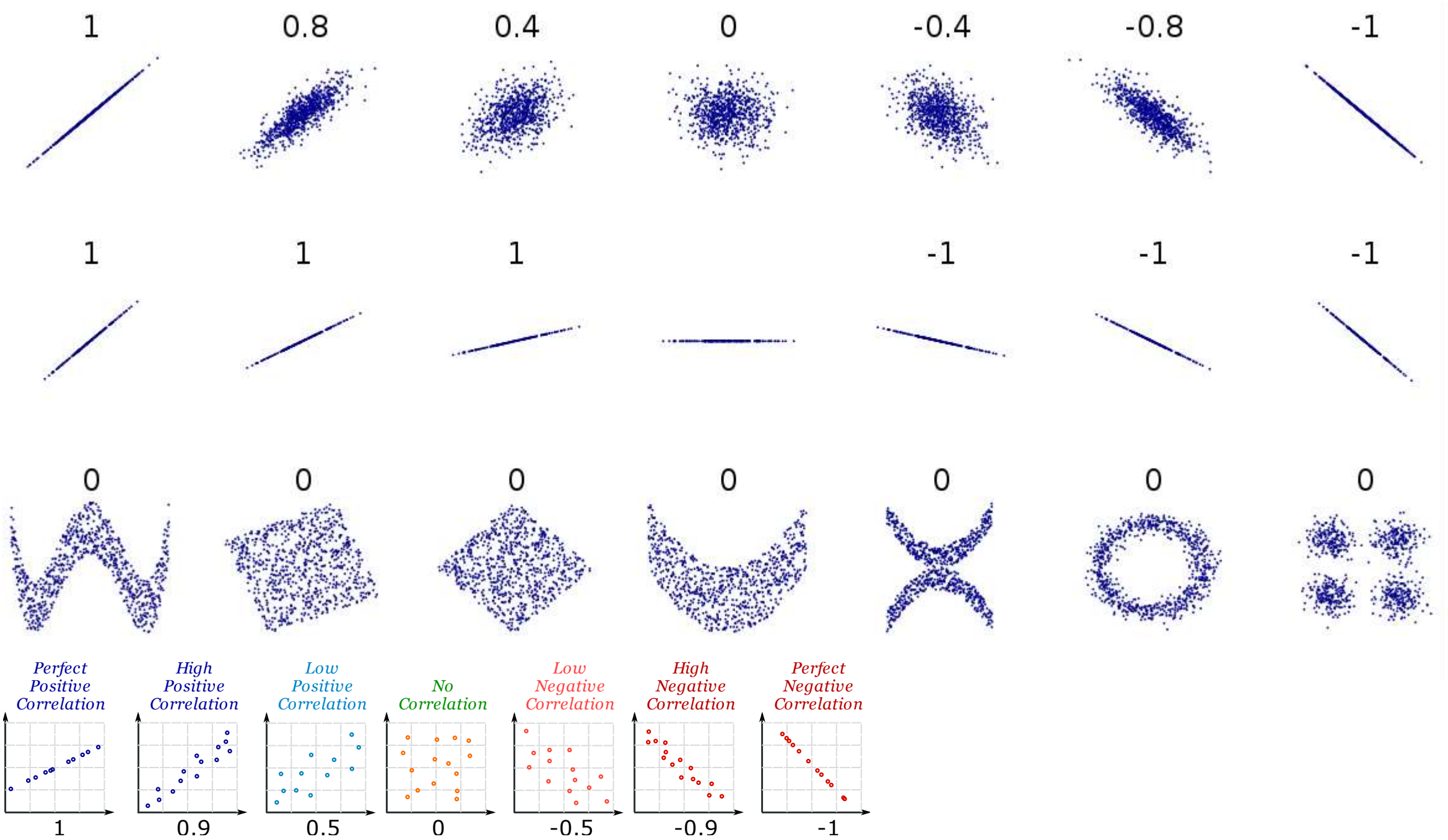
Hint: calculate
Q1- 1.5(IQR)
&
Q3 + 1.5(IQR)



Correlation

- Correlation is a statistic that measures the degree to which two variables move in relation to each other.
- Correlation measures association, but doesn't show if x causes y or vice versa—or if the association is caused by a third factor.

- The correlation coefficient is a standardized metric so that it always ranges from −1 (perfect negative correlation) to +1 (perfect positive correlation).
- A correlation coefficient of 0 indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0
Perfect Positive Correlation
High Positive Correlation
Low Positive Correlation
No Correlation
Low Negative Correlation
High Negative Correlation
Perfect Negative Correlation
1
0.9
0.5
0
-0.5
-0.9
-1

Covariance	Correlation
Covariance is a measure to indicate the extent to which two random variables change in tandem.	Correlation is a measure used to represent how strongly two random variables are related to each other.
Covariance is nothing but a measure of correlation.	Correlation refers to the scaled form of covariance.
Covariance indicates the direction of the linear relationship between variables.	Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.
Covariance can vary between -∞ and +∞	Correlation ranges between -1 and +1
Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed.	Correlation is not influenced by the change in scale.
Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary.	Correlation of two dependent variables measures the proportion of how much on average these variables vary w.r.t one another.
Covariance is zero in case of independent variables (if one variable moves and the other doesn't) because then the variables do not necessarily move together.	Independent movements do not contribute to the total correlation. Therefore, completely independent variables have a zero correlation.

Continuous: Data that can take on any value in an interval.
Synonyms: interval, float, numeric

Discrete: Data that can take on only integer values, such as counts.
Synonyms: integer, count

Categorical: Data that can take on only a specific set of values representing a set of possible categories.
Synonyms: enums, enumerated, factors, nominal, polychotomous

Binary: A special case of categorical data with just two categories of values (0/1, true/false).
Synonyms: dichotomous, logical, indicator, Boolean

Ordinal: Categorical data that has an explicit ordering.

PDF Probability Density Function

- A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

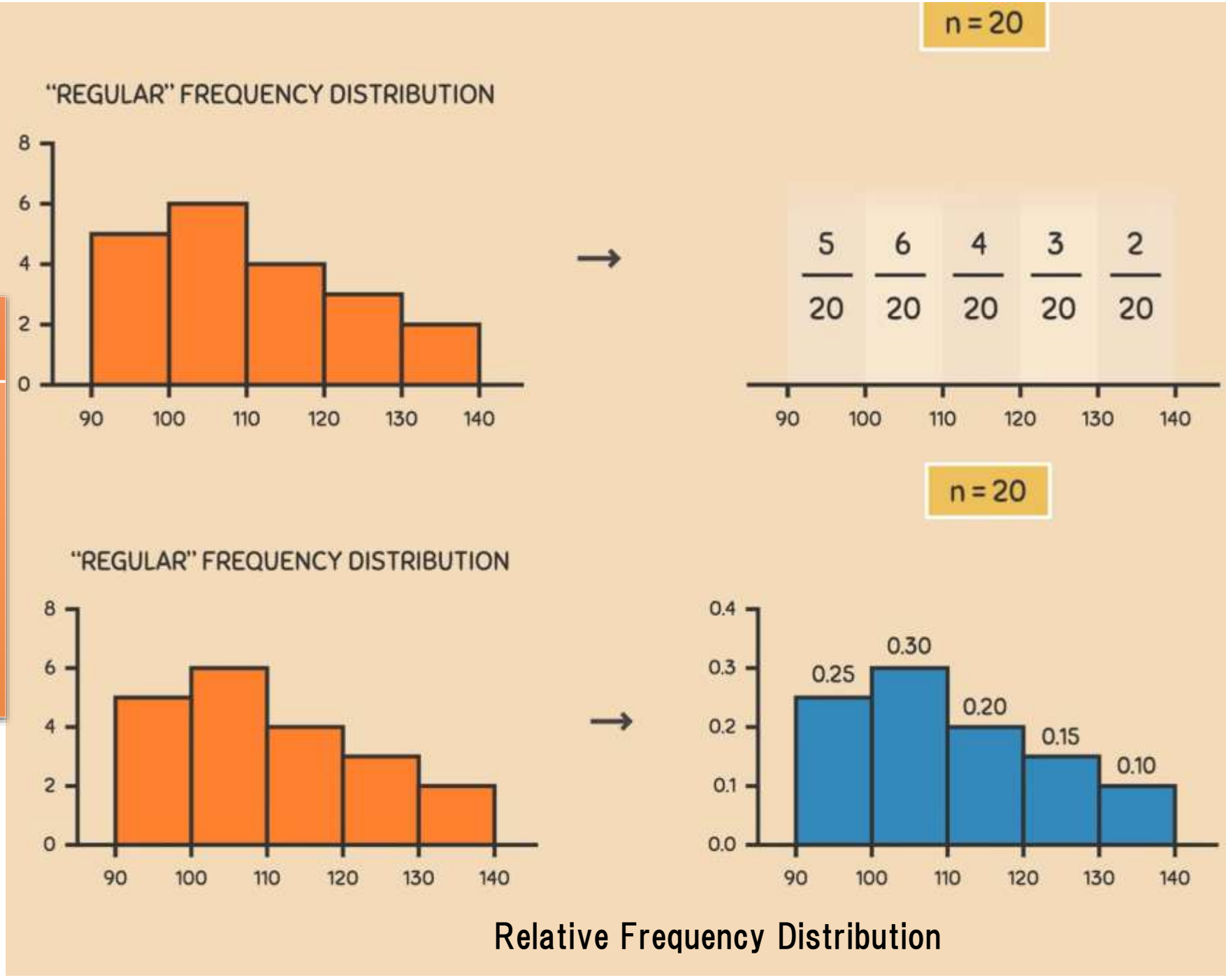
Likelihood

- Likelihood deals with fitting models given some known data ("what is the likelihood that this coin is/isn't rigged given that I just flipped heads six times in a row?"

KDE Kernel Density Estimate

- KDE allows us to **estimate probability density function** from our finite dataset.
- **Bandwidth:** (*bw=5)* larger the bandwidth smoother is the KDE.
- Smoother the KDE more data it looses. [sns.kdeplot(df.height , bw=12)]
- By default it is **scott** or we can switch to **silberman** these are methods to estimate appropriate bandwidth for your dataset.

Frequency Distribution	Probability Distribution
It records how often an event occurs. It is based on actual observations	It records the likelihood that an event is to occur. It is based on theoretical assumption of what should happen



Relative Frequency Distribution

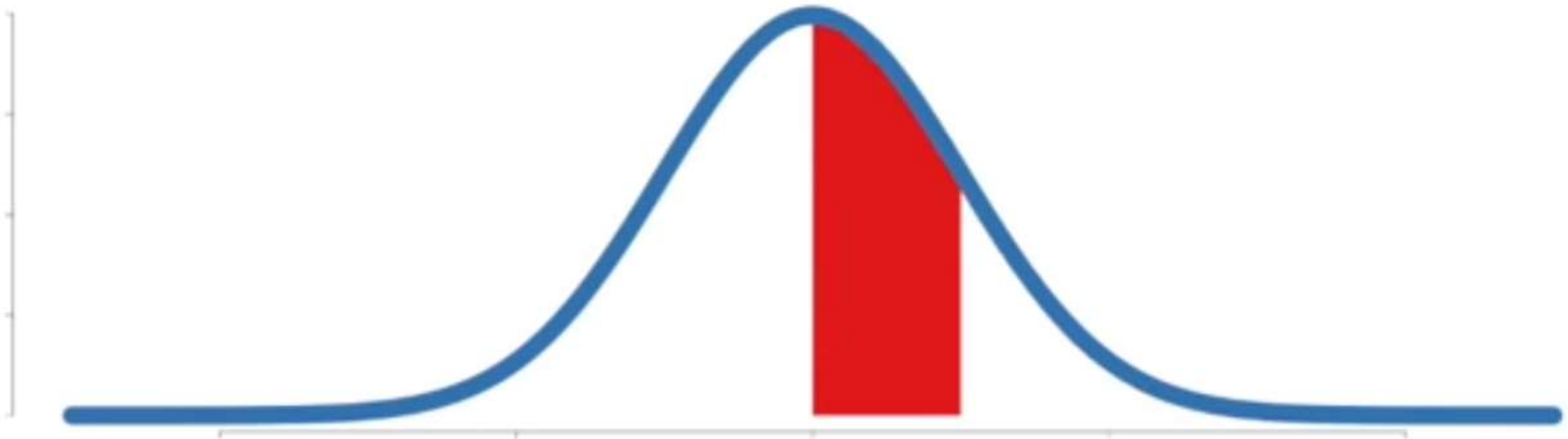
Histogram vs Density Curve

1. Gives us an idealized picture of population/ dataset without considering **irregularities** and **outliers**. Thus gives us a great overall picture of actual distribution and its tendencies.
2. Picture of histogram depends on how many interval we have, more intervals we have the better representation (distribution) of data. But with density curve we are not limited by the number of intervals that are and we can actually have infinite amount of intervals.
3. Smooth curve is generally easier to work with than a histogram especially when we are working with very large population.

In summary…

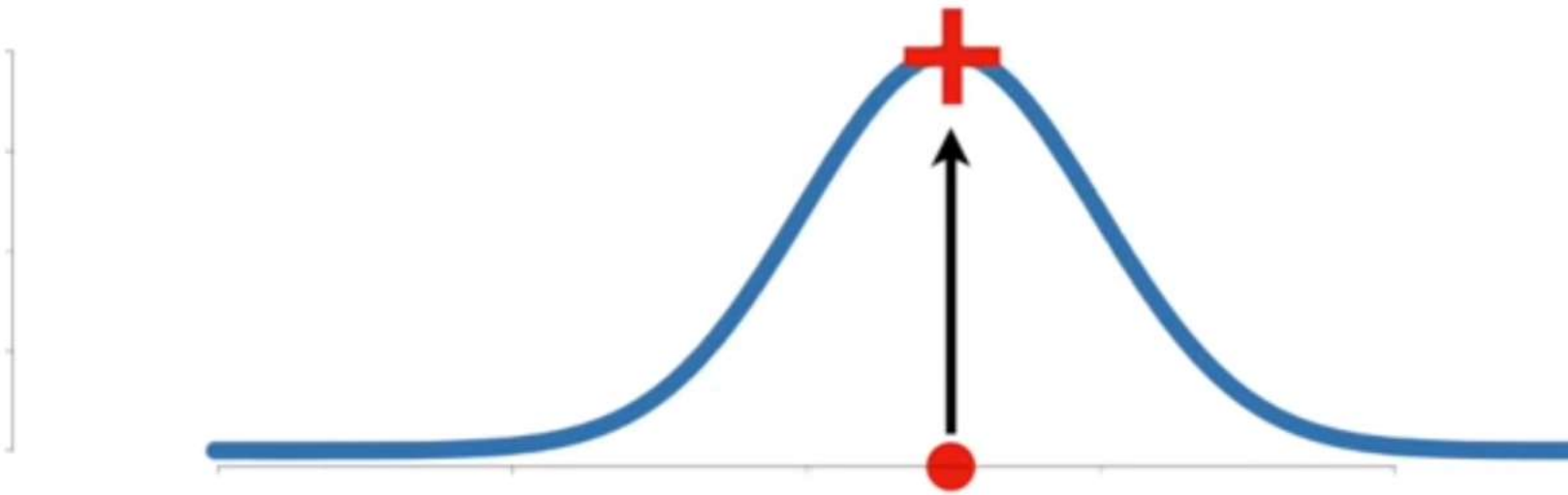
Probabilities are the areas under a fixed distribution…

pr(data | distribution)



Likelihoods are the y-axis values for fixed data points with distributions that can be moved…

L(distribution | data)



Probability Vs Likelihood

1. Probability" refers to the percentage of chances of foreseen outcomes based on parameters of values.
"likelihood" refers to the possibility of occurrences with different sets of parameter values that may lead to a sound conclusion.

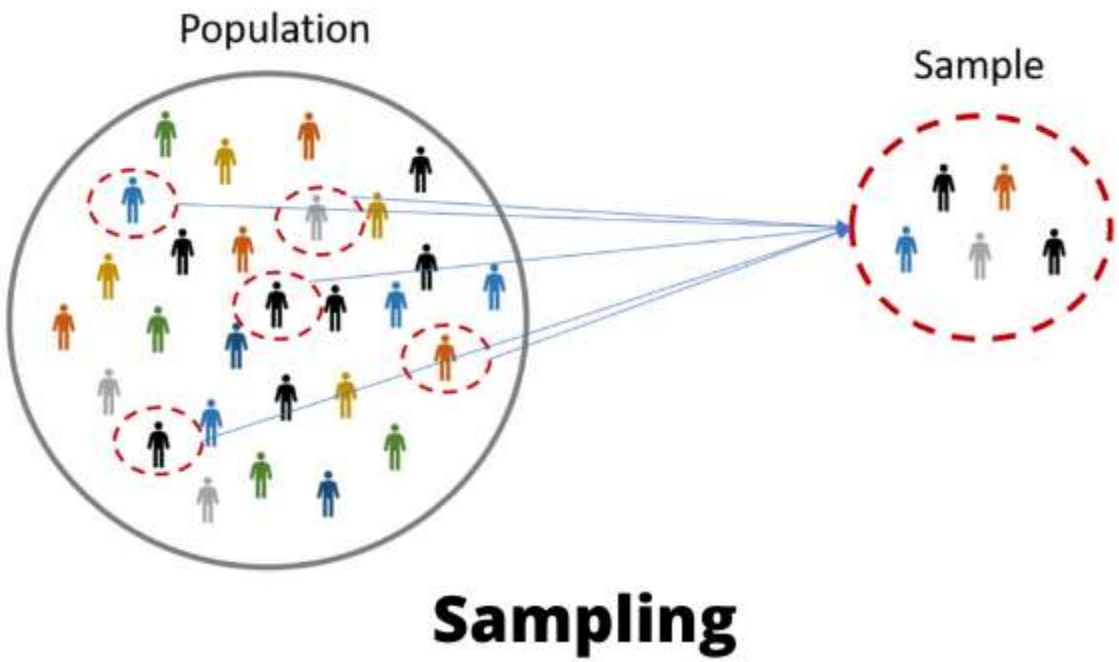
2. "Probability" and "likelihood" can be both used to express a prediction and odds of occurrences.

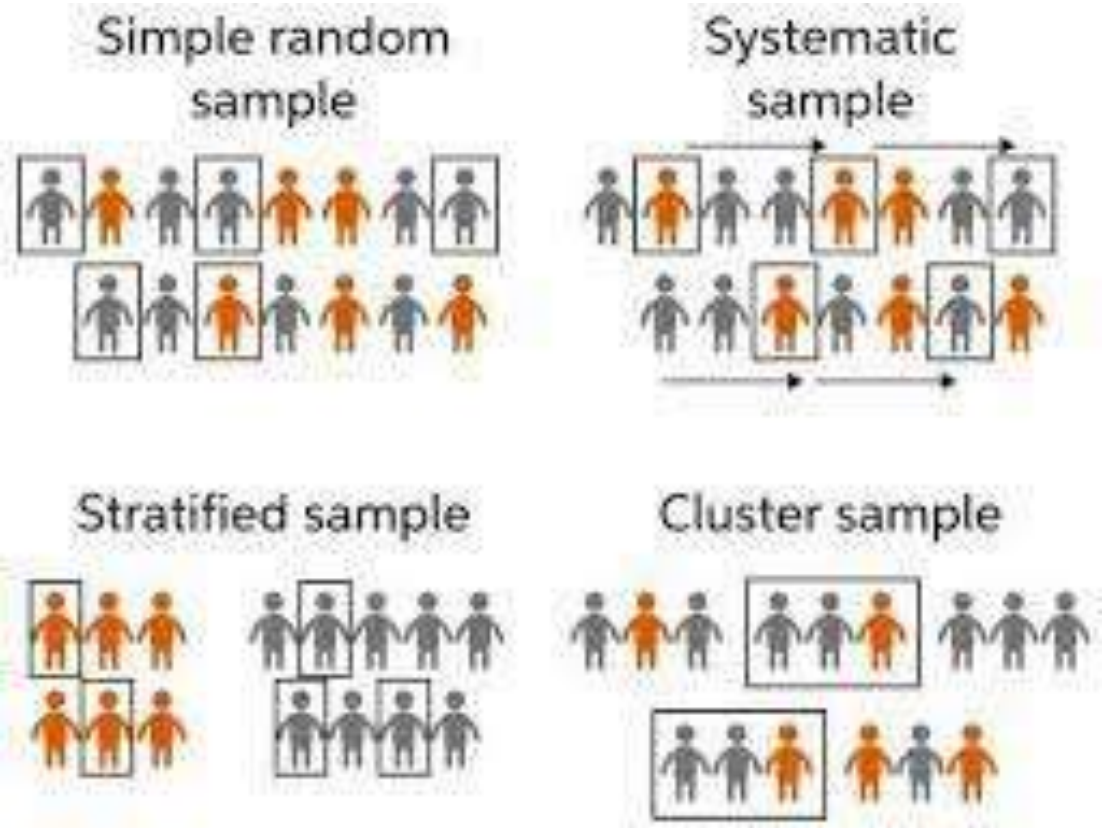
3. "Probability" refers to a "chance" while likelihood refers to a "possibility."

4. A probability follows clear parameters and computations while a likelihood is based merely on observed factors.

Sampling

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.



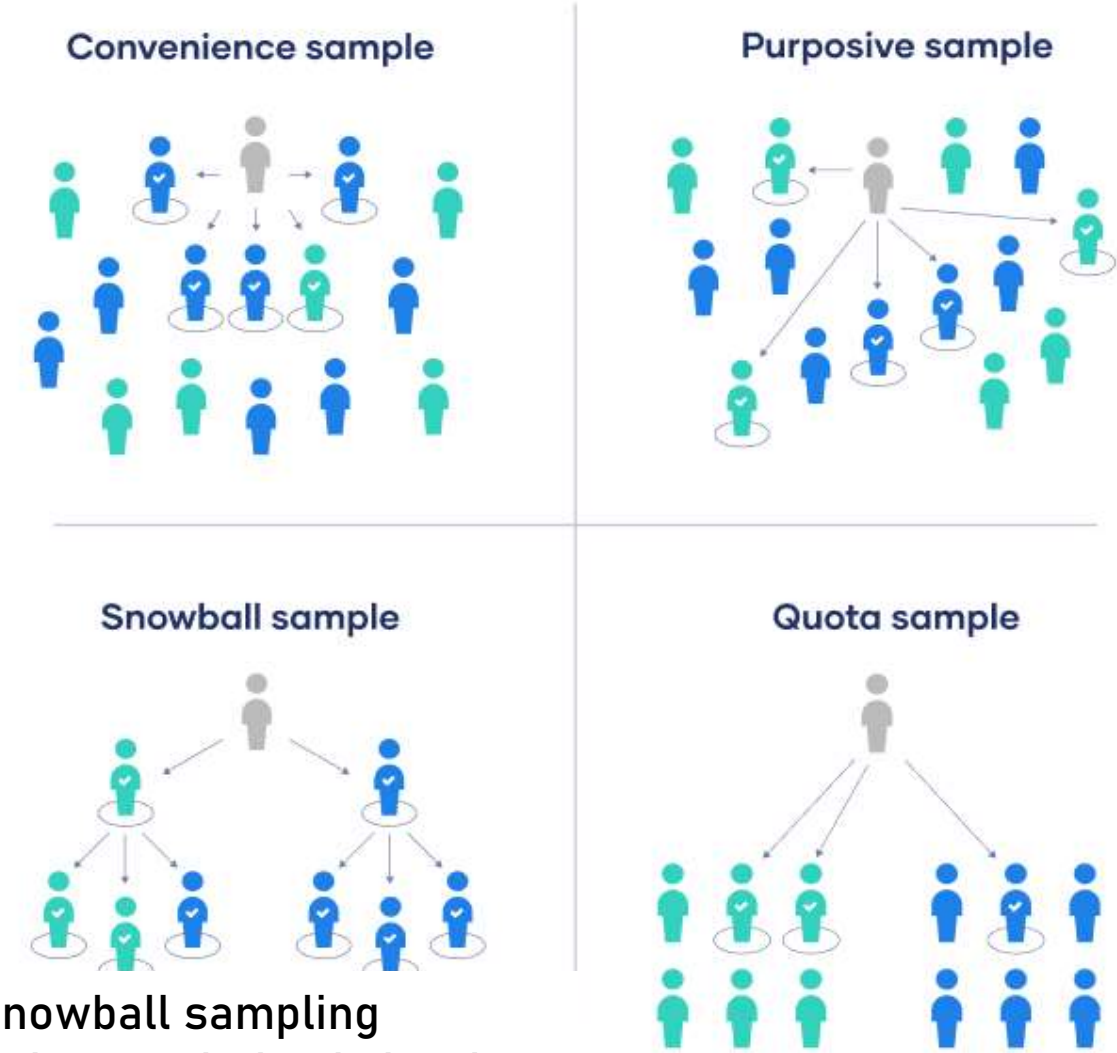


- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Non- Probability Sampling

Convenience sampling
A convenience sample simply includes the individuals who happen to be most accessible to the researcher.
Purposive sampling
This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.
Voluntary response sampling
Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).
Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.



Snowball sampling
If the population is hard to access, snowball sampling can be used to recruit participants via other participants.

a sample taken from a stratified population by sampling until a pre-assigned quota in each stratum is represented.

Probability sampling methods

1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.

Example

You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

Example

All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

3. Stratified sampling

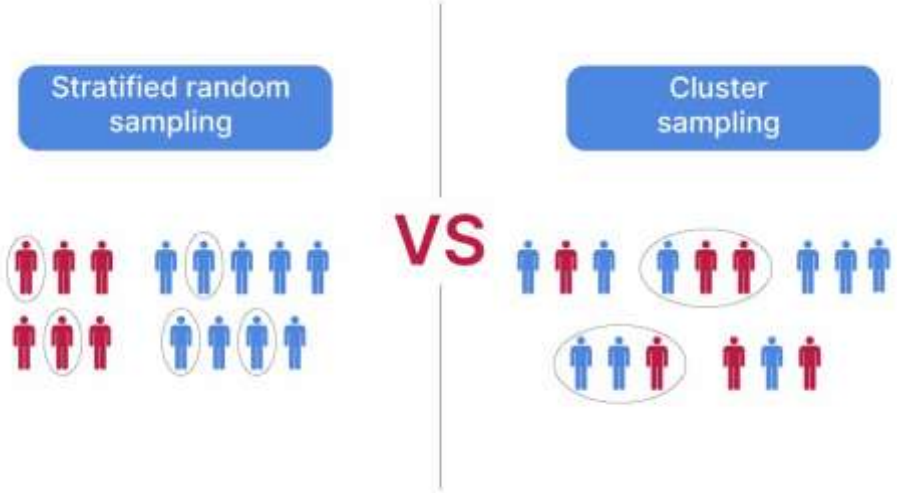
Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

To use this sampling method, you divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).

Based on the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.

Example

The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.



4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. *Instead of sampling individuals from each subgroup, you randomly select entire subgroups*.

If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above. This is called *multistage sampling.*

This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.

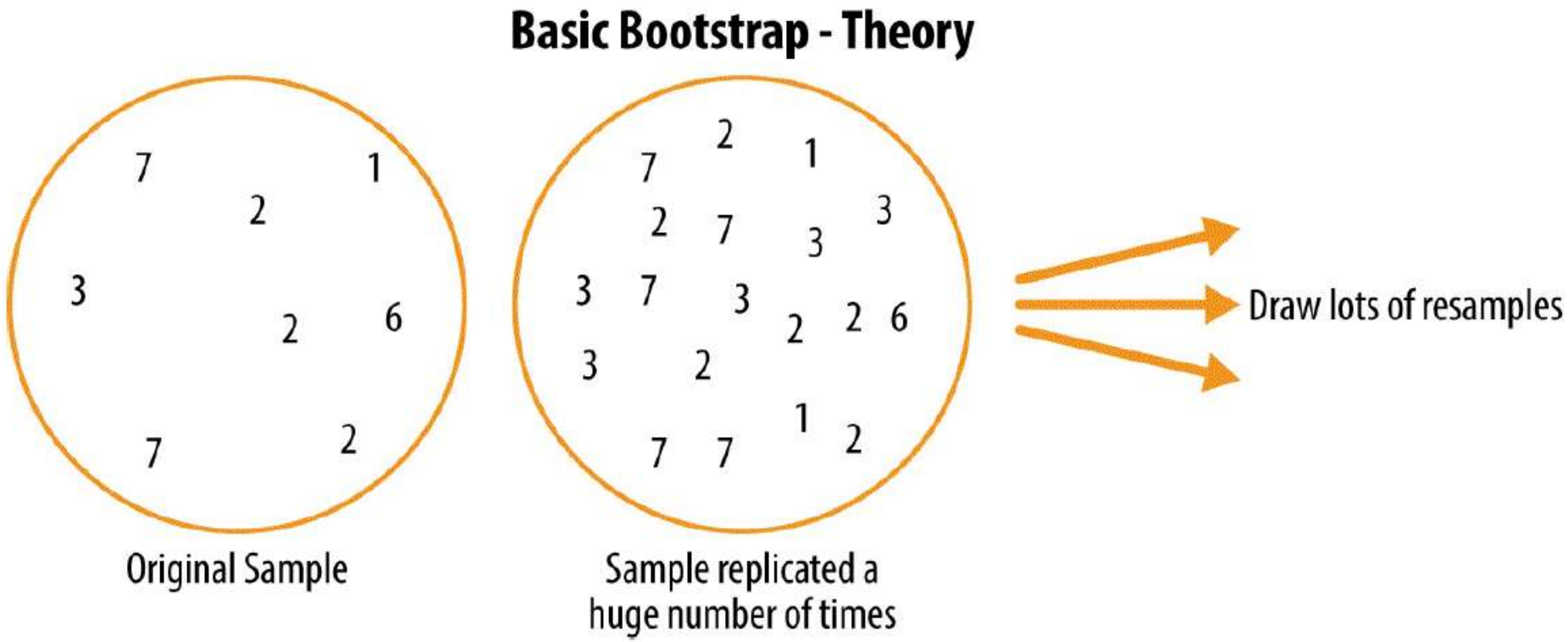
Example

The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Bootstrap

Bootstrap sample: A sample taken with replacement from an observed data set. It does not necessarily involve any assumptions about the data or the sample statistic being normally distributed.

Resampling: The process of taking repeated samples from observed data; includes both bootstrap and permutation (shuffling) procedures.



We simply replace each observation after each draw.

The algorithm for a bootstrap resampling of the mean is as follows, for a sample of size *n*:

1. Draw a sample value, record, replace it.
2. Repeat *n* times.
3. Record the mean of the *n* resampled values.
4. Repeat steps 1–3 *R* times.
5. Use the *R* results to:
    a. Calculate their standard deviation (this estimates sample mean standard error).
    b. Produce a histogram or boxplot.
    c. Find a confidence interval.

R, the number of iterations of the bootstrap, is set somewhat arbitrarily. The more iterations you do, the more accurate the estimate of the standard error, or the confidence interval. The result from this procedure is a bootstrap

Bootstrap can be used with multivariate data, with classification and regression trees running multiple trees on bootstrap samples and then averaging their predictions (or, with classification, taking a majority vote) generally performs better than using a single tree. This process is called *bagging* (short for "bootstrap aggregating": see "Bagging and the Random Forest").

The bootstrap met with considerable skepticism when it was first introduced; it had the aura to many of spinning gold from straw. This skepticism stemmed from a misunderstanding of the bootstrap's purpose.

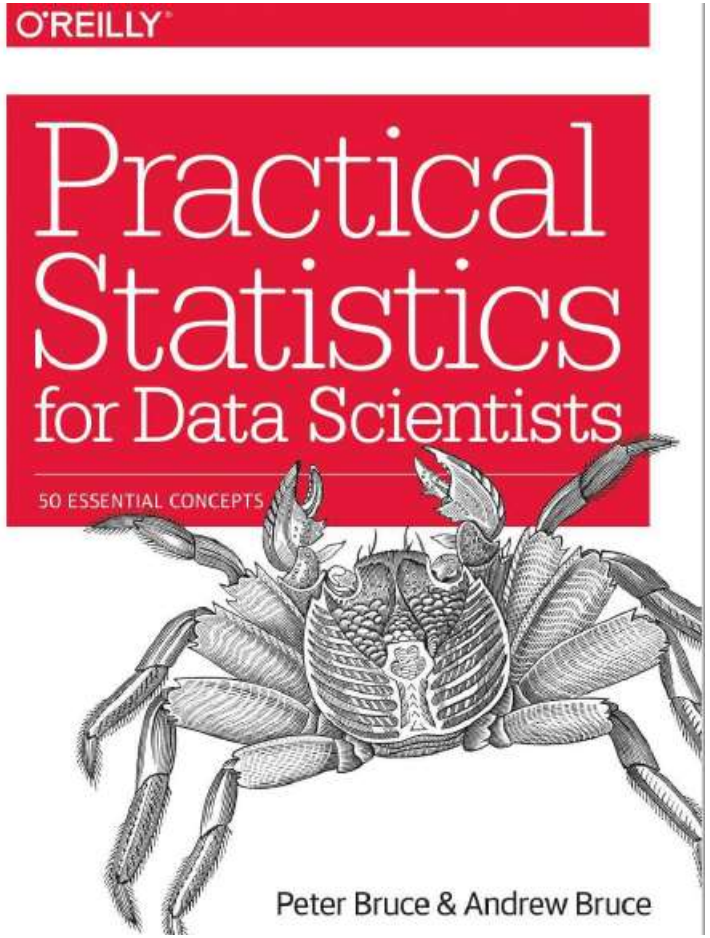
- The bootstrap does not compensate for a small sample size.
- It does not create new data, nor does it fill in holes in an existing data set. It merely informs us about how lots of additional samples would behave when drawn from a population like our original sample.
- It is also computationally intensive, and was not a feasible option before the widespread availability of computing power.

KEY IDEAS

- The bootstrap (sampling with replacement from a data set) is a powerful tool for assessing the variability of a sample statistic.
- The bootstrap can be applied in similar fashion in a wide variety of circumstances, without extensive study of mathematical approximations to sampling distributions.
- It also allows us to estimate sampling distributions for statistics where no mathematical approximation has been developed.
- When applied to predictive models, aggregating multiple bootstrap sample predictions (bagging) outperforms the use of a single model.

Resampling versus Bootstrapping

Sometimes the term *resampling* is used synonymously with the term *bootstrapping*, as just outlined. More often, the term *resampling* also includes permutation procedures.



Read Ch 3 to understand the difference.

Central Limit Theorem

- It says that the means drawn from multiple samples will resemble the familiar bell-shaped normal curve even if the source population is not normally distributed, **provided** that the **sample size is large enough** and the departure of the data from normality is not too great.

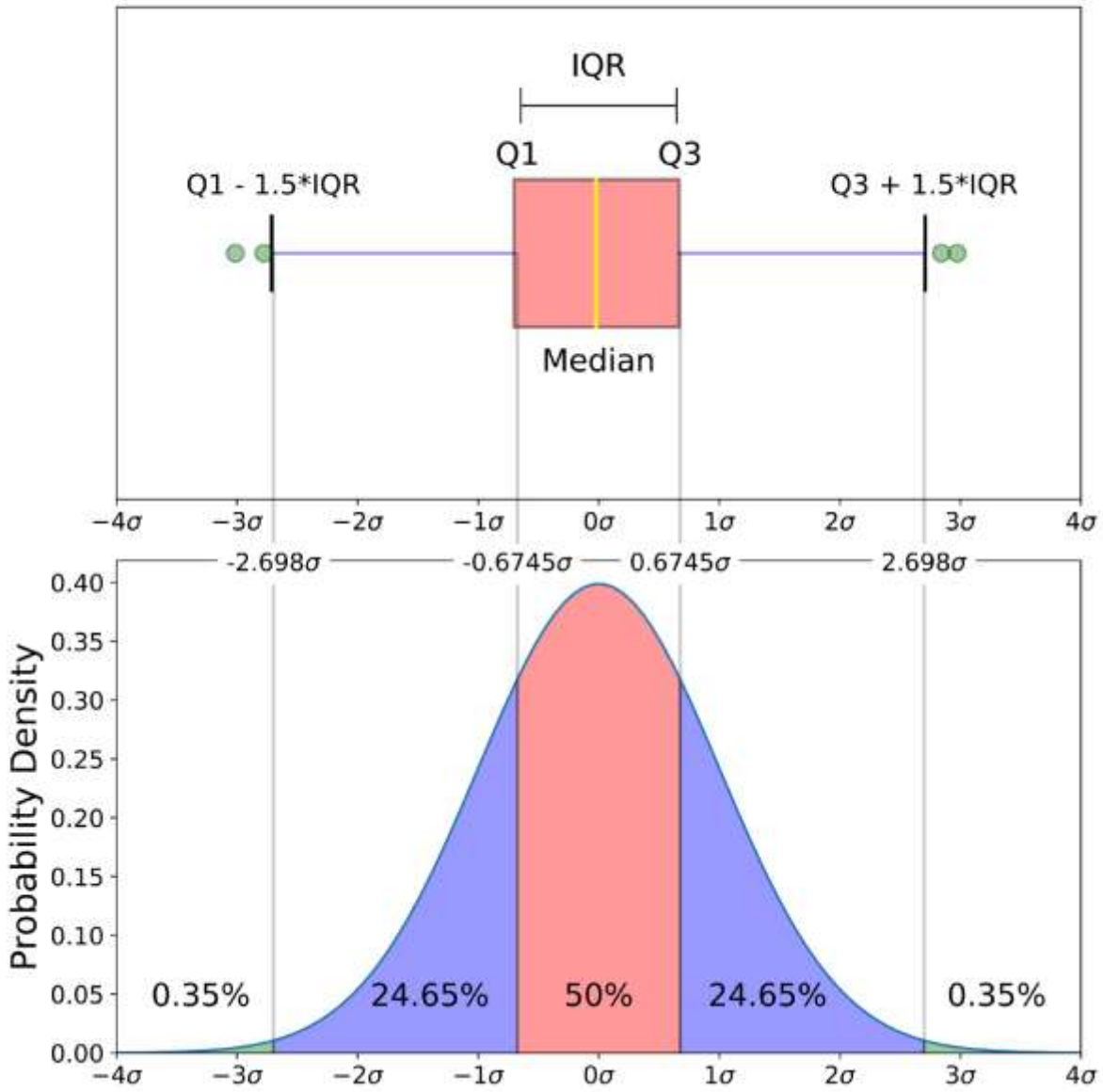
- This received alot of attention initially as it underlies the machinery of hypothesis tests and confidence intervals.
- but, since formal hypothesis tests and confidence intervals play a small role in data science, and the bootstrap is available in any case, the central limit theorem is not so central in the practice of data science.

Confidence Intervals

- The percentage associated with the confidence interval is termed the level of confidence. The higher the level of confidence, the wider the interval.
- Also, the smaller the sample, the wider the interval (i.e., the more uncertainty).
- For a data scientist, a confidence interval is a tool to get an idea of how variable a sample result might be. Data scientists would use this information not to publish a scholarly paper or submit a result to a regulatory agency (as a researcher might), but most likely to communicate the potential error in an estimate, and, perhaps, learn whether a larger sample is needed.

- Confidence intervals are the typical way to present estimates as an interval range.
- The more data you have, the less variable a sample estimate will be.
- The lower the level of confidence you can tolerate, the narrower the confidence interval will be.
- The bootstrap is an effective way to construct confidence intervals.

Statistical Distribution



Interquartile Range

$\text{range} = \text{max} - \text{min}$ $\quad$ $\text{IQR} = Q_3 - Q_1$

16, 24, 26, 26, 26, 27, 28 $\quad$ 23, 25, 28, 28, 32, 33, 35

$\text{range} = 28 - 16$ $\quad$ $\text{IQR} = 33 - 25$

The couples were given a score in each round.

The scores in the first round were

16 27 12 18 26 21 27 22 18 17

(a) Calculate the median and semi-interquartile range of these scores. 3

12, 16 17 18 18 21 22 26 27 27

$\frac{n+1}{2} = \frac{10+1}{2} = 5.5$

$Q_1 = 17$

median

$\frac{21+18}{2}$

$Q_2 = 19.5$

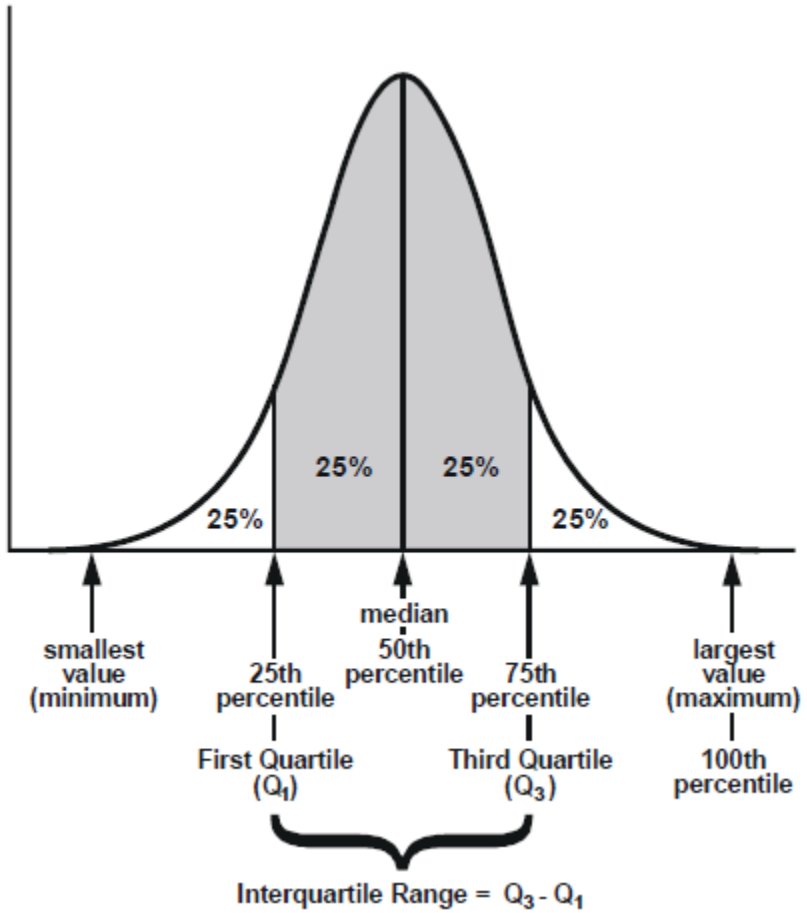
Q_2

$Q_3 = 26$

$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{26-17}{2}$

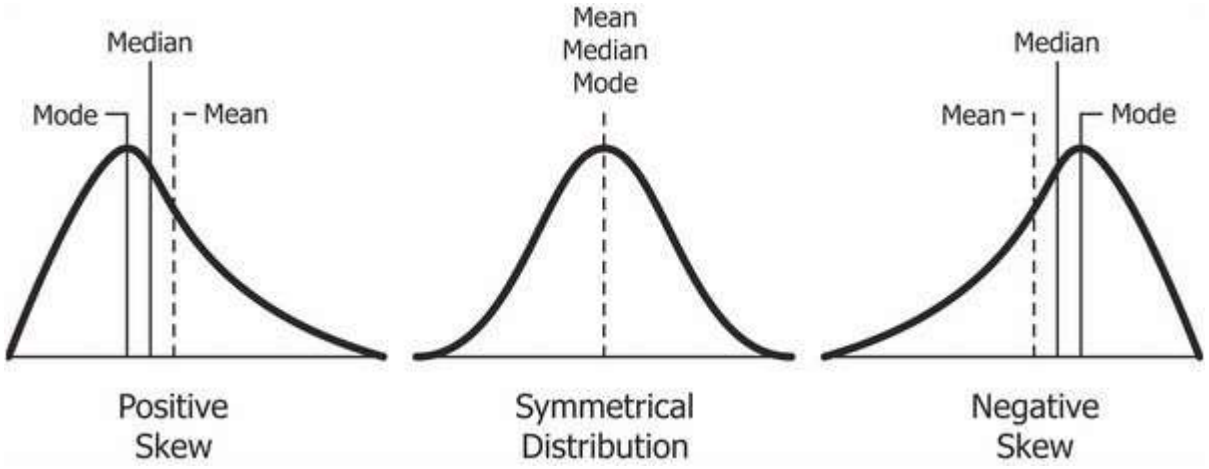
$= \frac{9}{2} = 4.5$

(b) In the second round, the median was 26 and the semi-interquartile range was 2·5.



Skewness

- Skewness means lack of symmetry
- i.e. if mean median mode don't coincide distribution is considered skewed
- Variance tells us about the amount of variability while **skewness gives the direction of variability**



Positive Skew

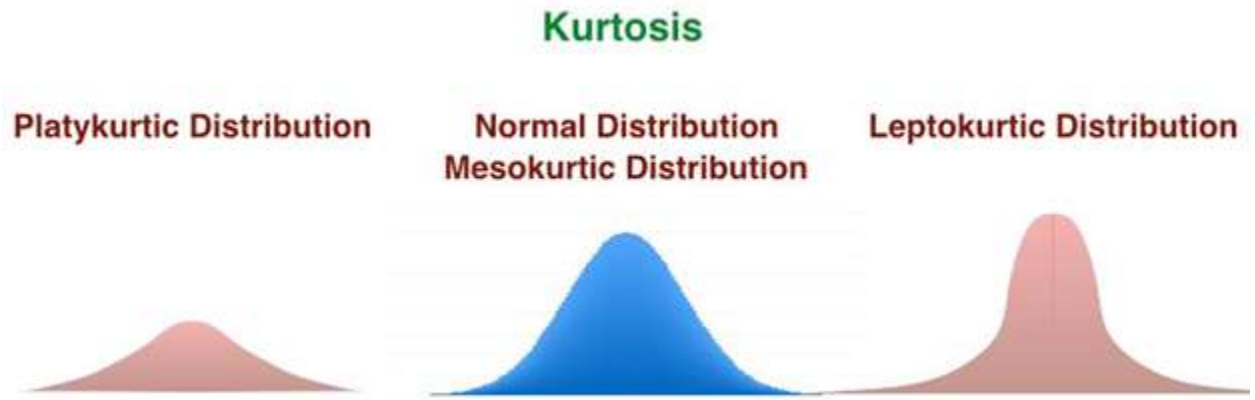
- Right tail is longer
- Mass of distribution is concentrated on left
- **Mode< Median< Mean**

Negative Skew

- Left tail is longer
- Mass of distribution is concentrated on Right
- **Mean< Median< Mode**

Kurtosis

- Kurtosis is a statistical measure that is used to describe distributions.
- Kurtosis is defined as a measure of '**peakedness**' and is generally measured **relative to normal distributions**.



Platykurtic(kurtosis < 3)

- A **lower tail** and ***stretched around center*** tails means **most of the data points are present in high proximity with mean.**

Mesokurtic (kurtosis = 3)

- Mesokurtic is the **same as the normal distribution,** which means kurtosis is near to 0..

Leptokurtic (kurtosis > 3)

- A very long and skinny tails, which means there **are more chances of outliers**.
- Indicates **more** of the numbers **are located in the tails of the distribution instead of around the mean**

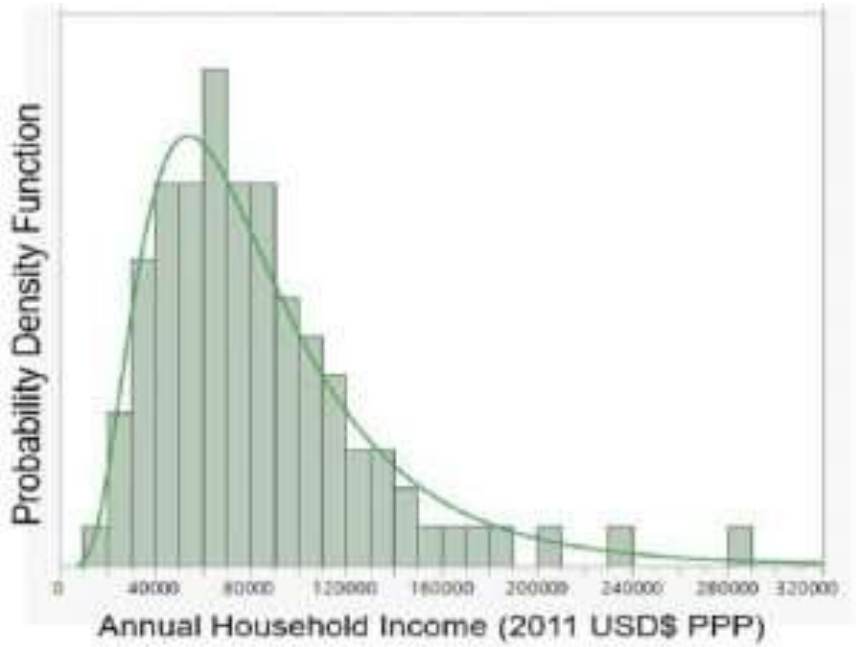
Calculating Skewness:

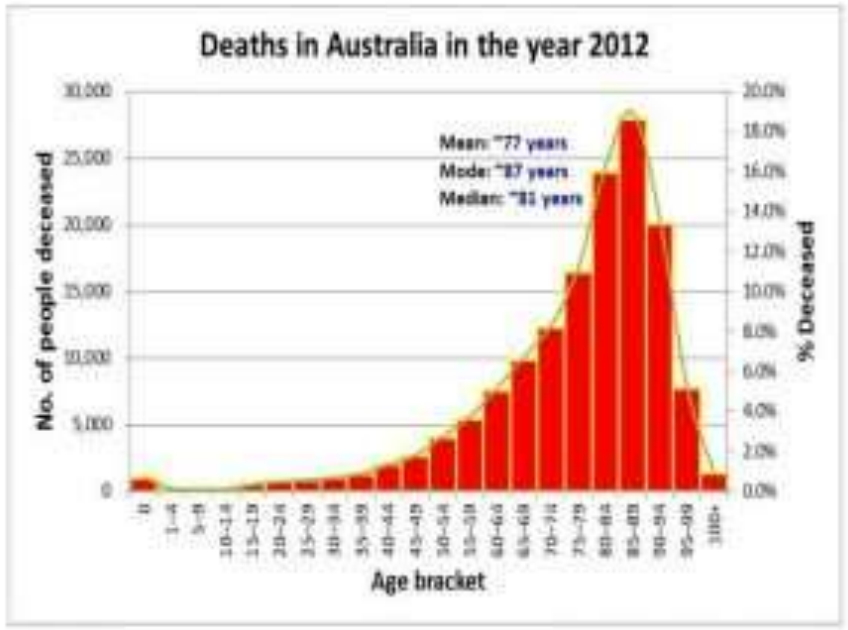
- Pearson's first Coefficient = (Mean – Mode)/ Standard Deviation
- Pearson's 2nd Coefficient = 3(Mean – Median)/Standard Deviation

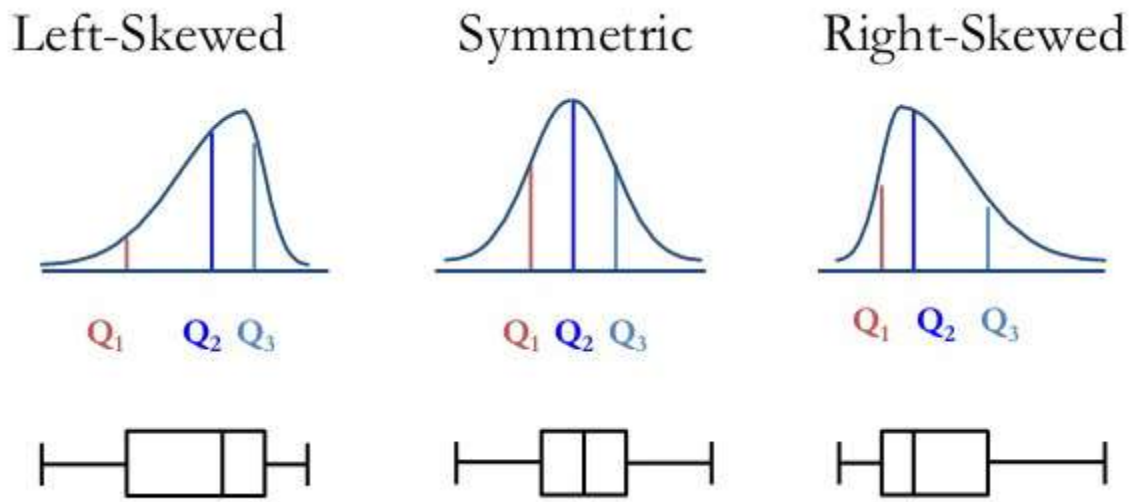
Mean – Mode ≈ 3(Mean – Median)

Calculating Kurtosis

Excess Kurtosis = Kurt - 3






Left-Skewed
Symmetric
Right-Skewed
Q1 Q2 Q3
Q1 Q2 Q3
Q1 Q2 Q3

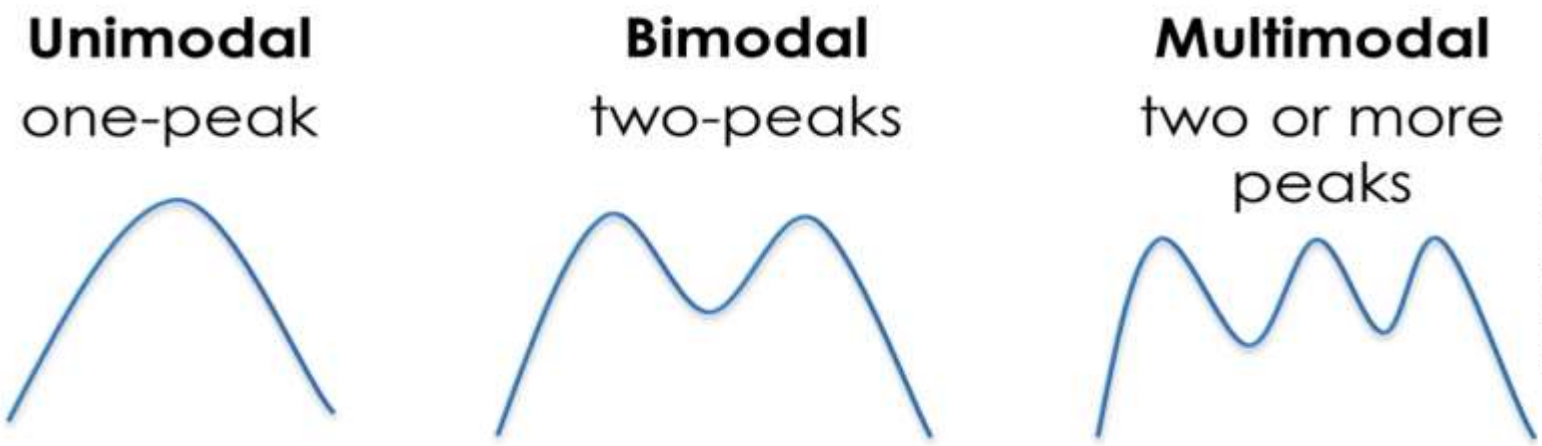
Normal Distribution

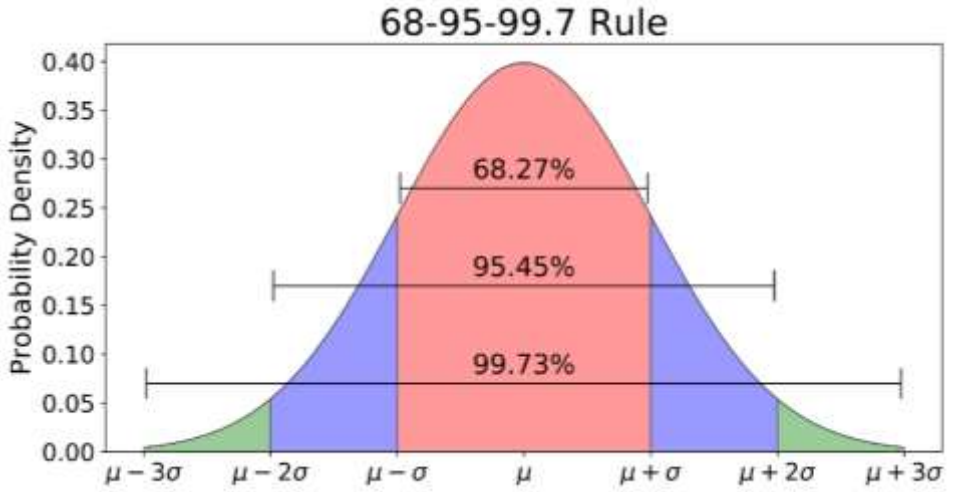
A.K.A. *Gaussian D*istribution

- Unimodal (i.e only 1 peak in middle)
- Symmetrical

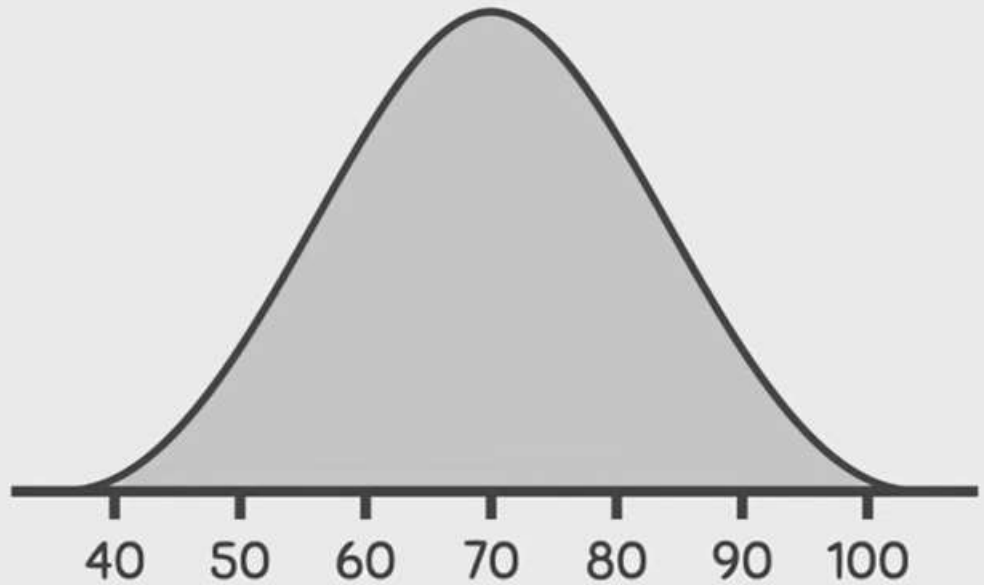
  Note: Symmetry and Modality are independent events.
- A Normal Distributed curve Never touches the x-axis it goes on for **infinity.**
- **68-95-99.7 Rule** Covered by 3 std ***is great way of approximating the of a normal distribution****. This will work for any normal distribution no matter it's shape and size.* However we can go beyond +-3 std but the area contained under this curve will be very small.

Modality

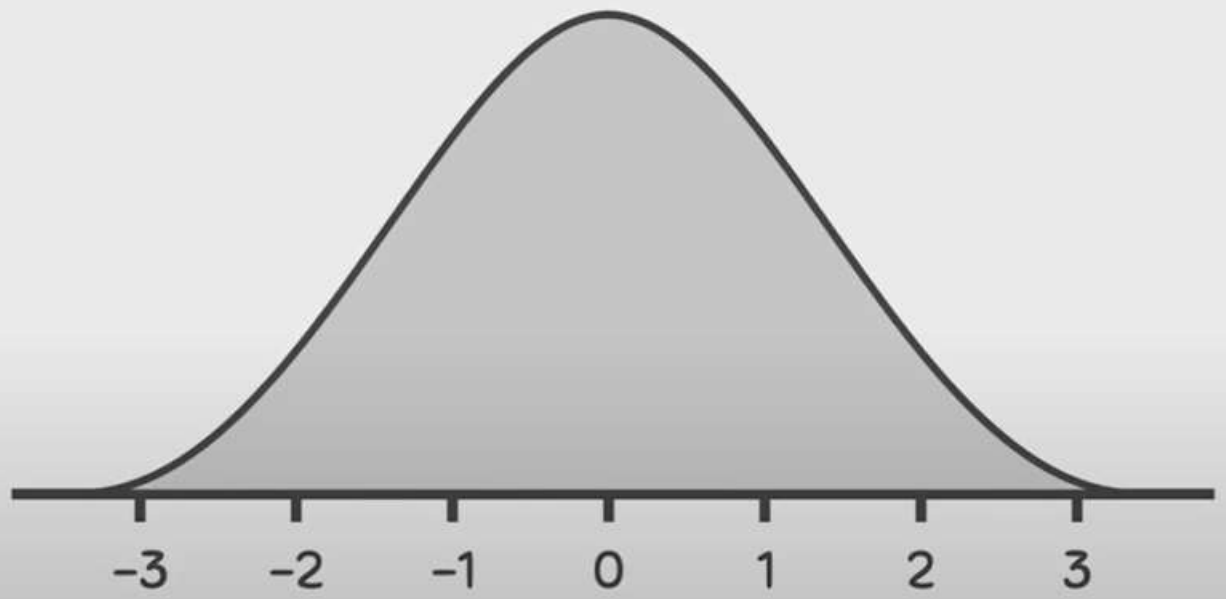




❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

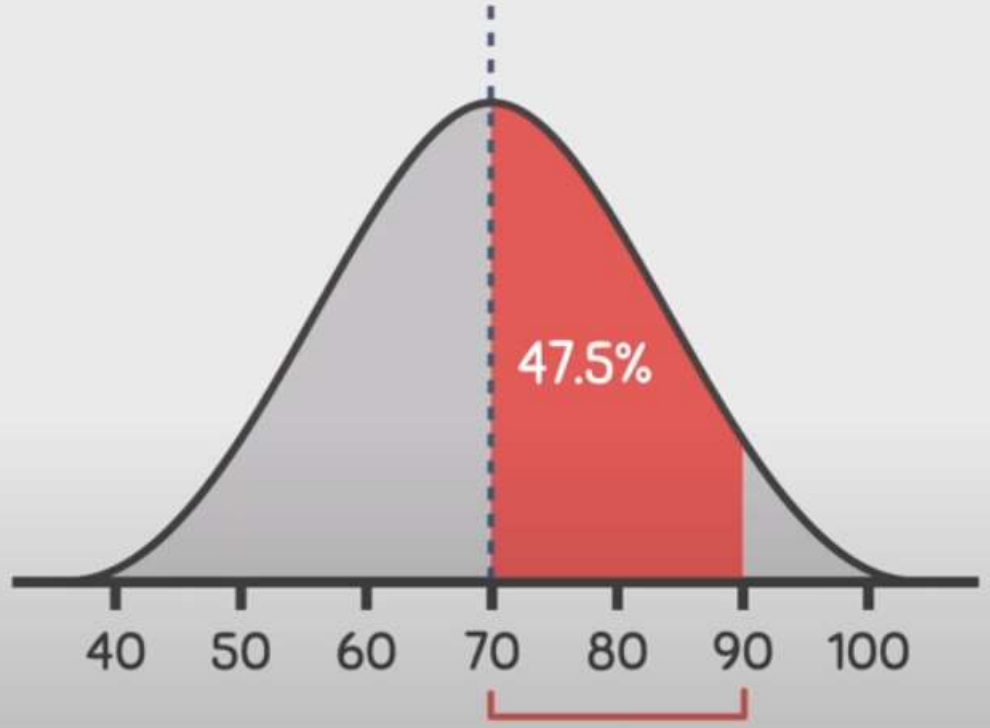


❷ For the normal distribution below, approximately what area is contained between −2 and 1?



$\mu = 70$

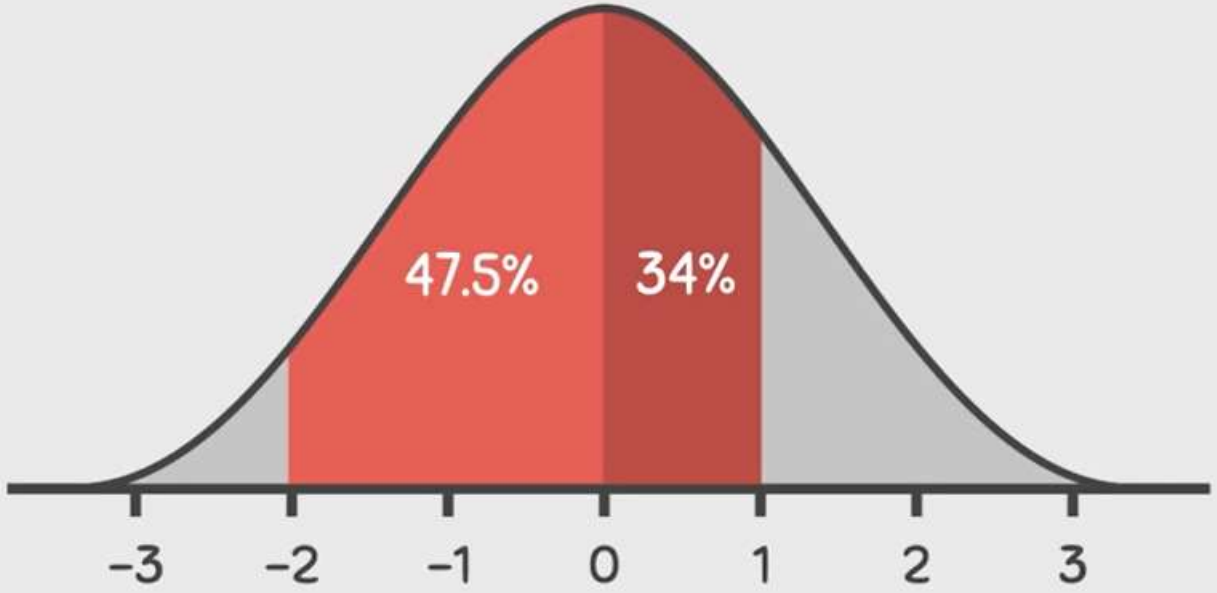
$\sigma = 10$



Answer 1

$\mu = 0$

$\sigma = 1$



Answer 2

- A *standard normal* distribution is one in which the units on the x-axis are expressed in terms of standard deviations away from the mean.
- To compare data to a standard normal distribution, you subtract the mean then divide by the standard deviation; this is also called *normalization* or *standardization.*

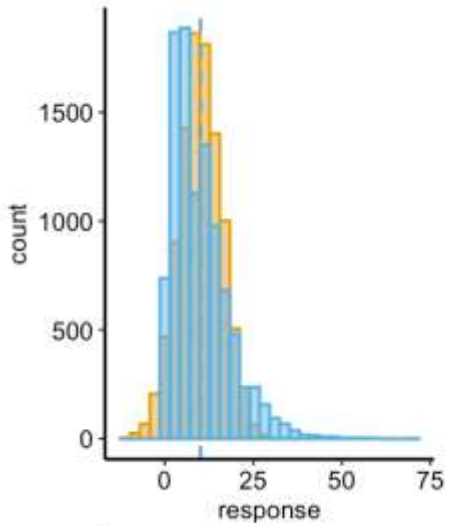
Note that "standardization" in this sense is unrelated to database record standardization (conversion to a common format). The transformed value is termed a *z-score*, and the normal distribution is sometimes called the *z-distribution*.

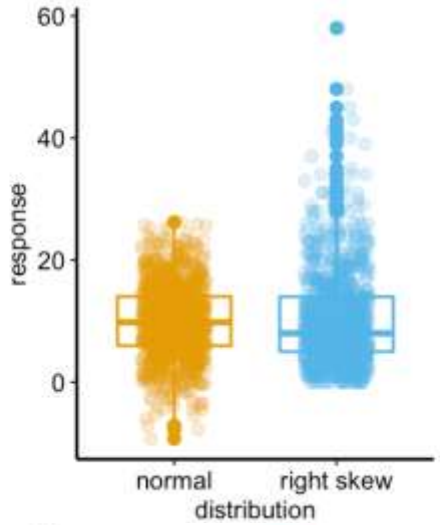
To convert data to *z*-scores, you subtract the mean of the data and divide by the standard deviation; you can then compare the data to a normal distribution.

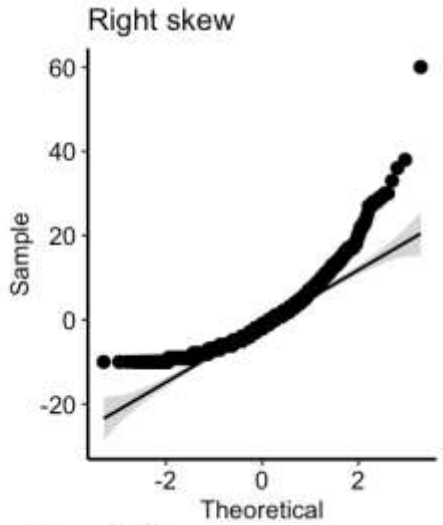
Converting data to *z*-scores (i.e., standardizing or normalizing the data) does *not* make the data normally distributed. It just puts the data on the same scale as the standard normal distribution, often for comparison purposes.

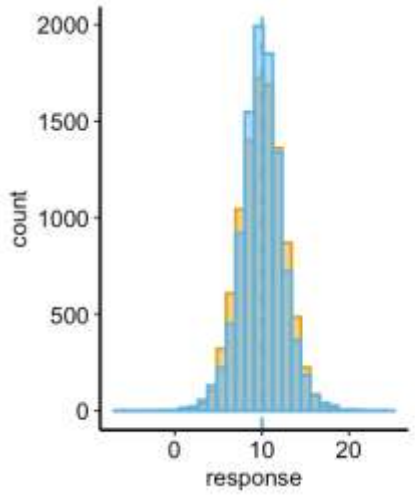
QQ Plot

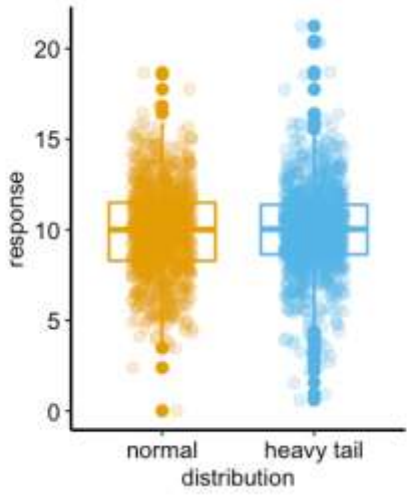
Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

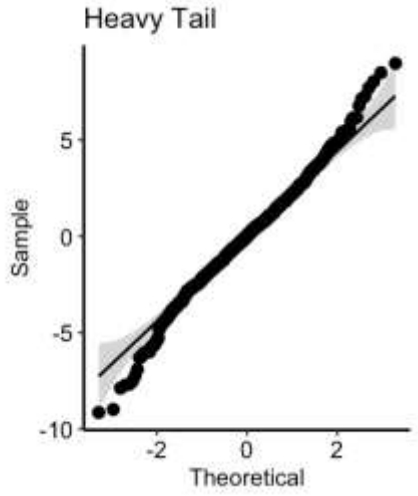


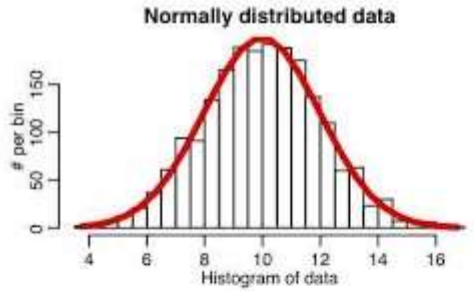


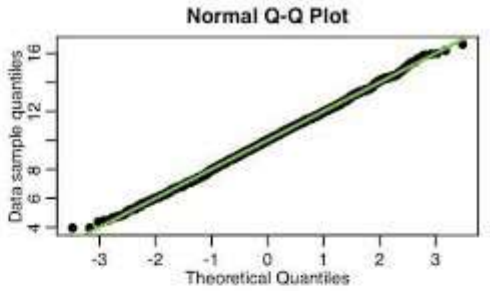


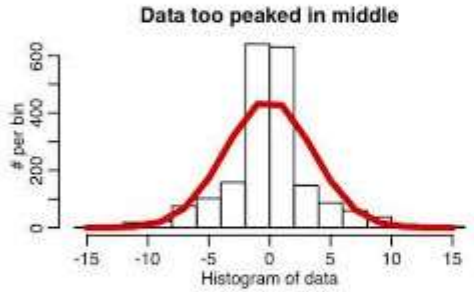


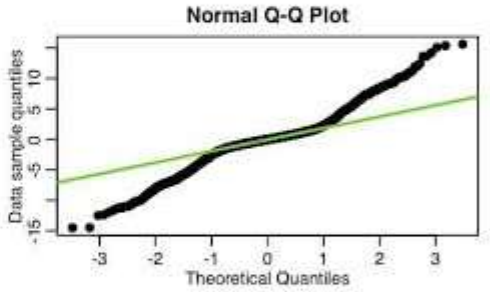


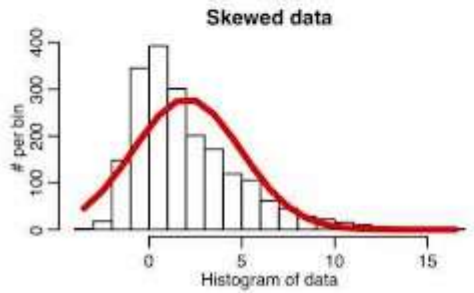


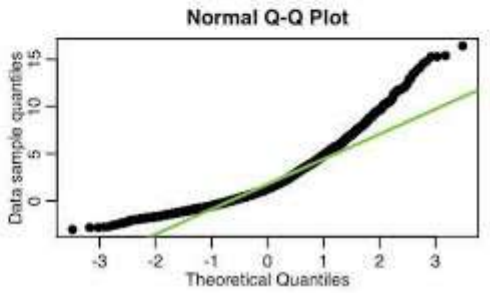




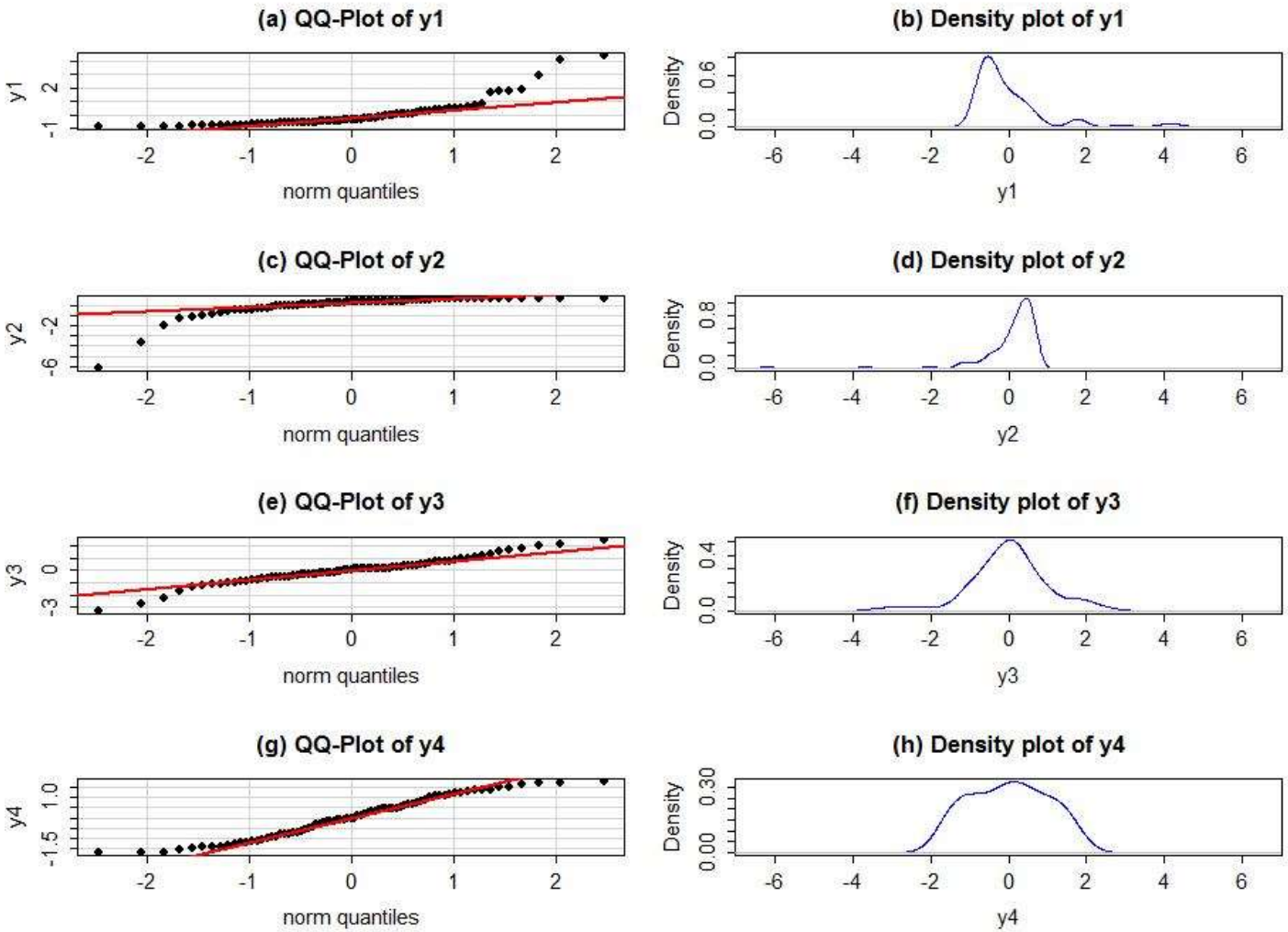


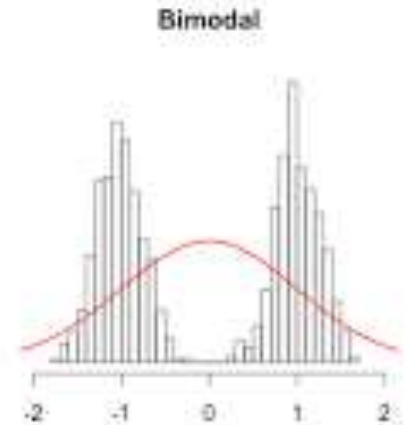


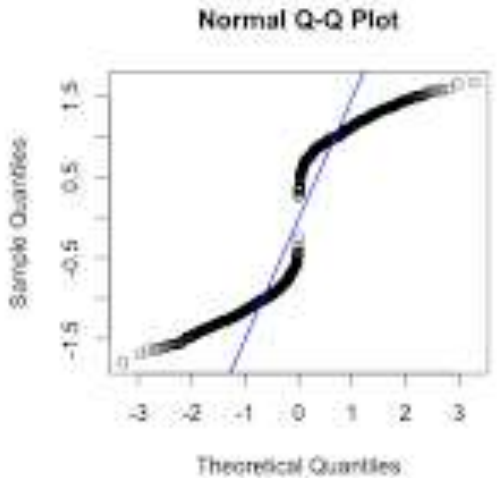


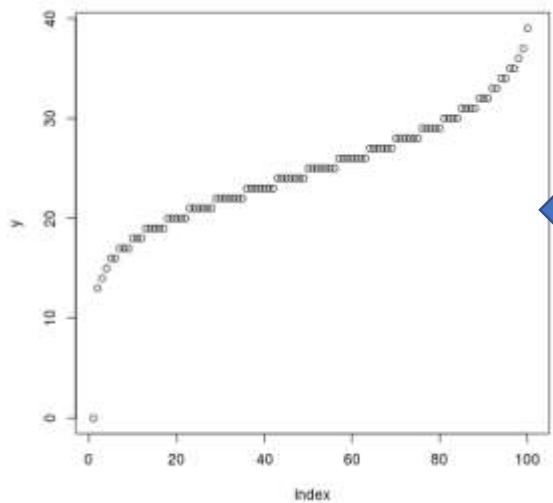


Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.









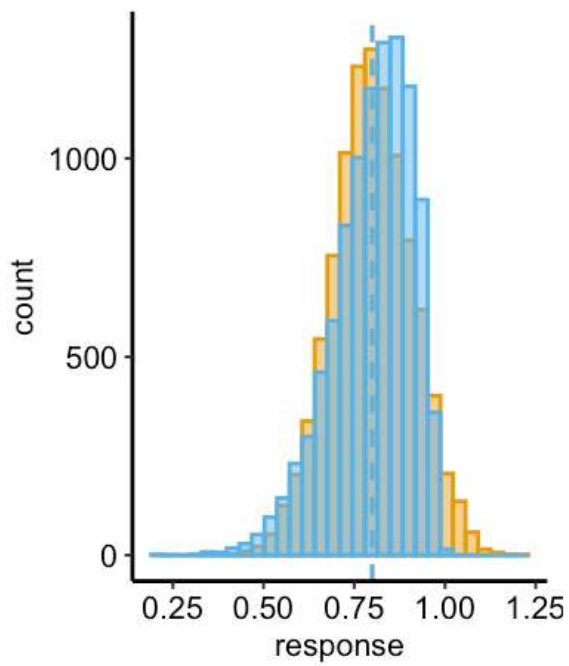
Negative Binomial Quantile Function

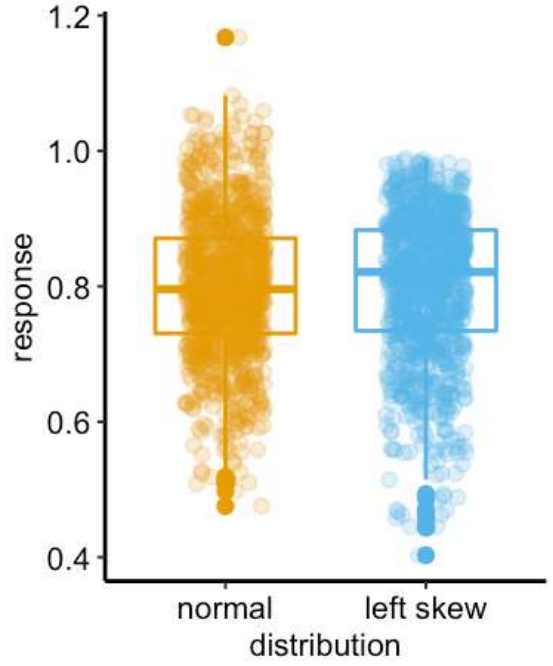
How does it work?

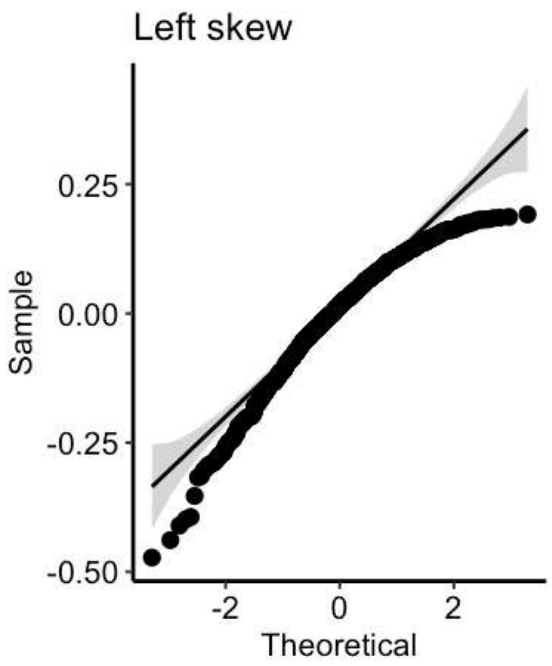
We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.

Now we have to focus on the ends of the straight line. If the points at the ends of the curve formed from the points are not falling on a straight line but indeed are scattered significantly from the positions then we cannot conclude a relationship between the x and y axes which clearly signifies that our ordered values which we wanted to calculate are not Normally distributed.

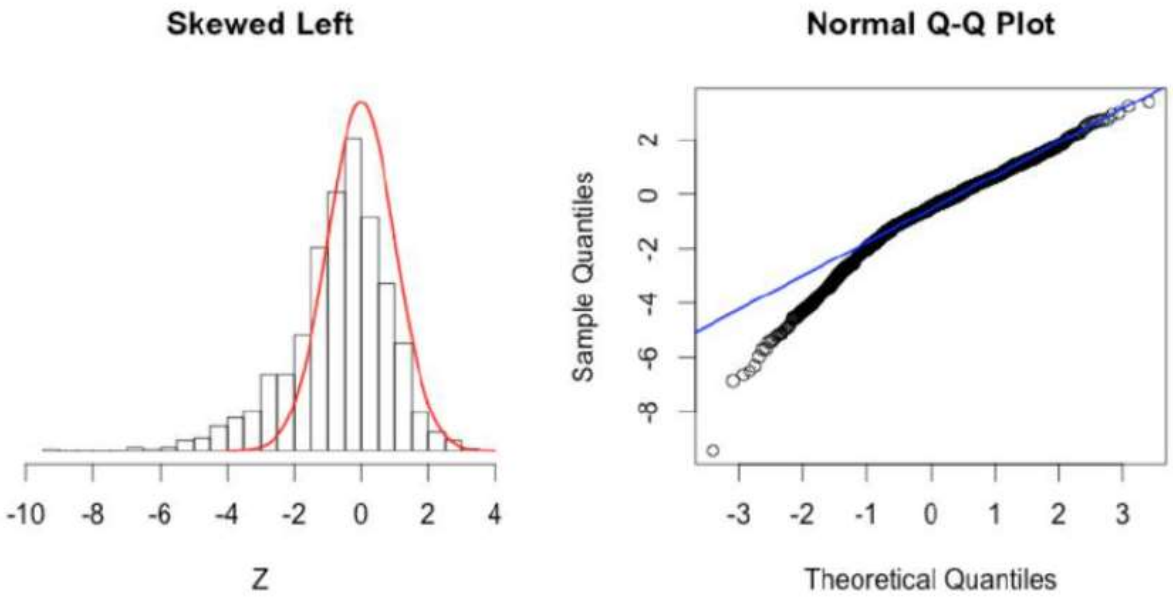
If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.



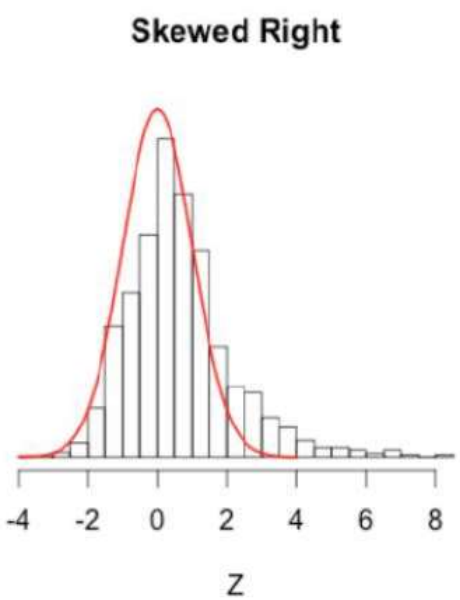


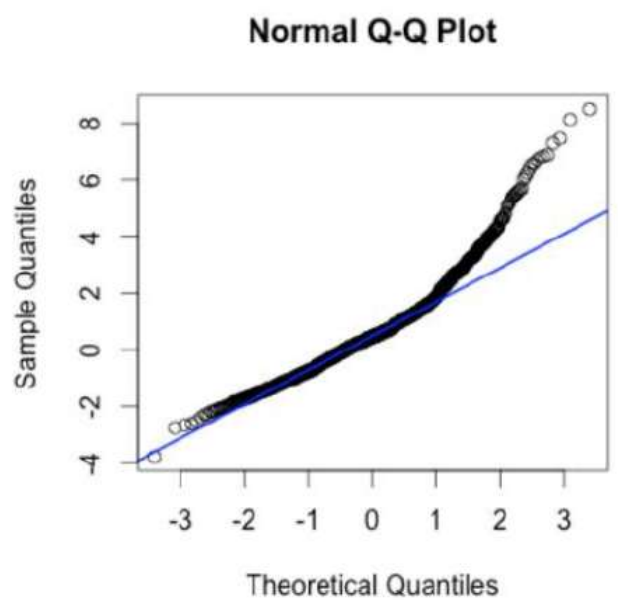


Skewed Q-Q plots



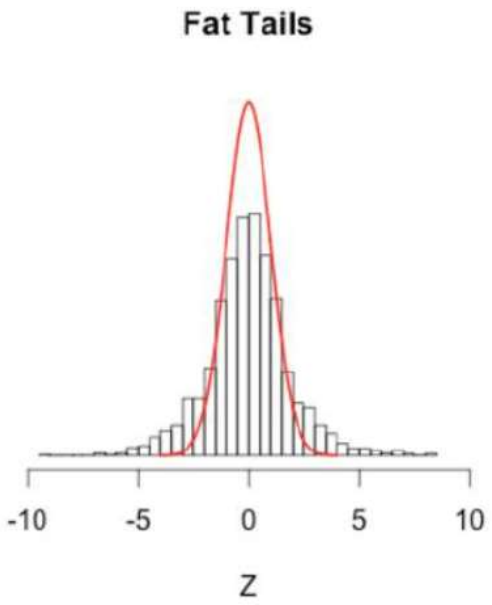
If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is **left-skewed** (or ***negatively skewed***)

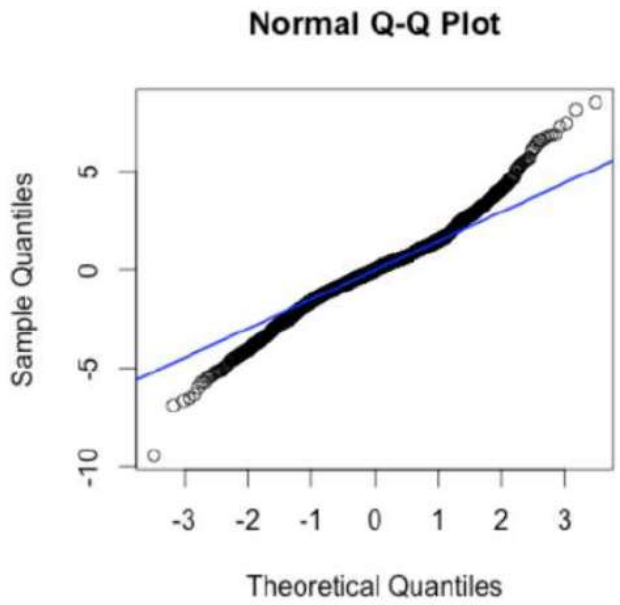




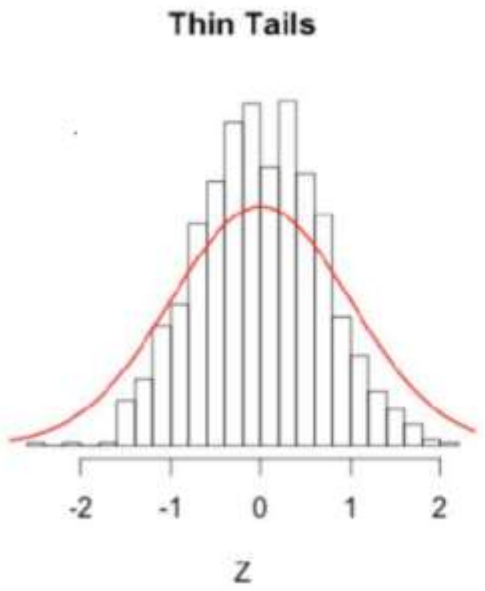
When we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is **right-skewed** (or ***positively skewed***).

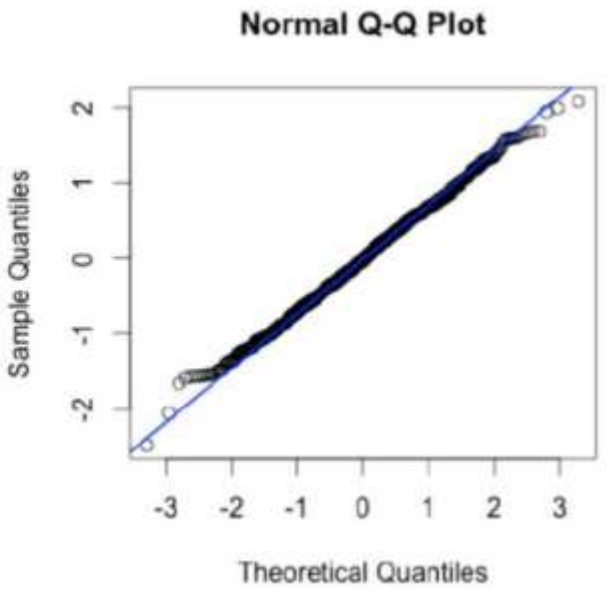
Kurtosis (a measure of *Tailedness*) QQ-Plot





The distribution with a fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line

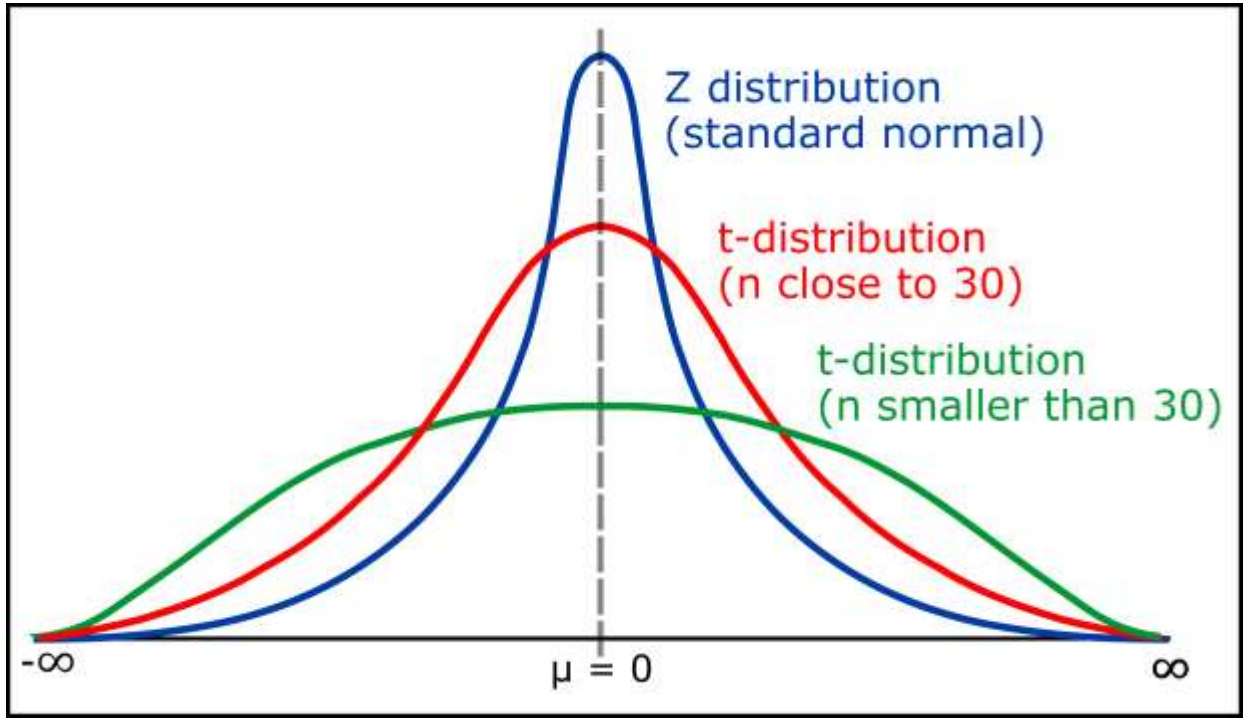




A thin-tailed distribution will form a Q-Q plot with a very less or negligible deviation at the ends thus making it a perfect fit for the Normal Distribution.

T-Distribution

- The t-distribution is often called *Student's t*
- The *t-distribution* is a normally shaped distribution, but a bit thicker and longer on the tails.
- It is used extensively in depicting distributions of sample statistics.
- Distributions of sample means are typically shaped like a t-distribution, and there is a family of t-distributions that differ depending on how large the sample is.
- The larger the sample, the more normally shaped the t-distribution becomes



The T distribution is similar to the normal distribution, just with fatter tails. Both assume a normally distributed population. T distributions have higher kurtosis than normal distributions. The probability of getting values very far from the mean is larger with a T distribution than a normal distribution.

CONs

The T distribution can skew exactness relative to the normal distribution. Its shortcoming only arises when there's a need for perfect normality. The T-distribution should only be used when population standard deviation is not known. If the population standard deviation is known and the sample size is large enough, the normal distribution should be used for better results.

Bernoulli Distribution

- The Bernoulli distribution is a special case of the bin`omial distribution where a single trial is conducted (so *n* would be 1 for such a binomial distribution). It is also a special case of the **two-point distribution**, for which the possible outcomes need not be 0 and 1.

- It has only one parameter, which is the probability of success.

How is Binomial Distribution different from Bernoulli Distribution?

- The Binomial distribution is like a bigger brother of the Bernoulli distribution.
- It models the number of successes in a situation of repeated Bernoulli experiments. So rather than focusing on the probability of success, we focus on a success count.
- The two parameters for the Binomial distribution are the number of experiments and the probability of success.

Binomial Distribution

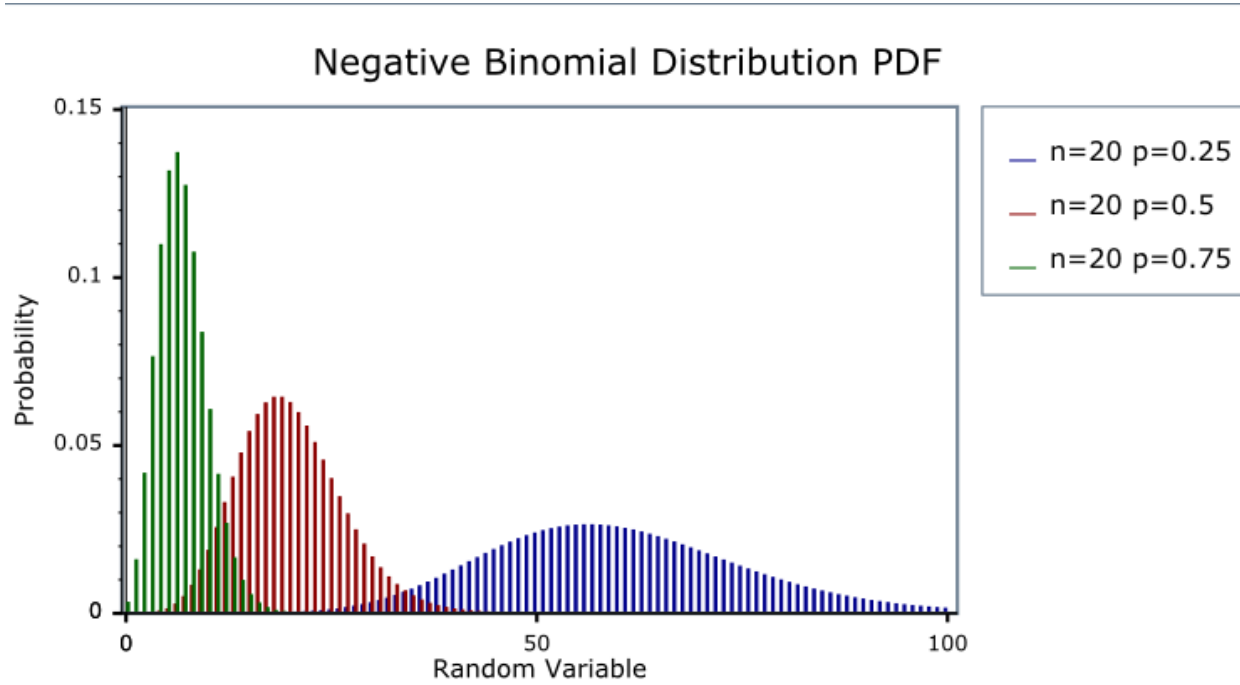
- The binomial distribution is a probability distribution that summarizes the likelihood that a value will take one of two independent values under a given set of parameters or assumptions.
- **Assumptions** of the binomial distribution are that there is only one outcome for each trial, that each trial has the same probability of success, and that each trial is mutually_exclusive, or independent of one another.
- The binomial distribution is a common **discrete distribution** used in statistics, as opposed to a continuous distribution, such as the normal distribution.
- ***Degrees of freedom:*** A parameter that allows the t-distribution to adjust to different sample sizes, statistics, and number of groups.

Notations for Binomial Distribution and the Mass Formula:

$$P(X) = {}_nC_x\, p^x q^{n-x}$$

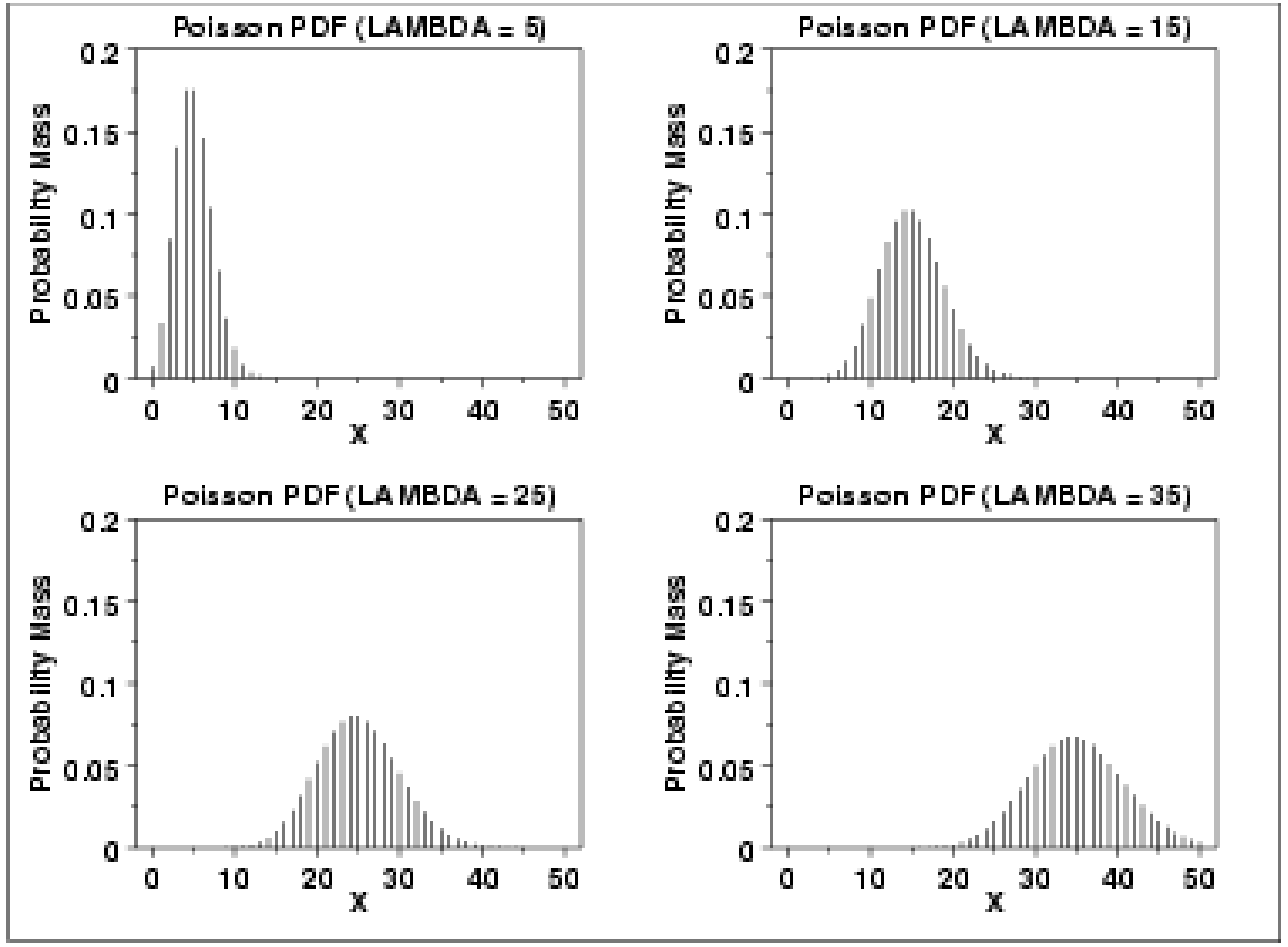
Where:

- **P** is the probability of success on any trail.
- **q = 1- P** – the probability of failure
- **n** – the number of trails/experiments
- **x** – the number of successes, it can take the values 0, 1, 2, 3, . . . n.
- **nCx = n!/x!(n-x)** and denotes the number of combinations of **n** elements taken **x** at a time.



Poisson Distribution

- Poisson distributions are used when the variable of interest is a **discrete** count variable.
- Poisson distribution describes a number of events in a fixed time frame.
- Many economic and financial data appear as count variables, such as how many times a person becomes unemployed in a given year, thus lending themselves to analysis with a Poisson distribution.

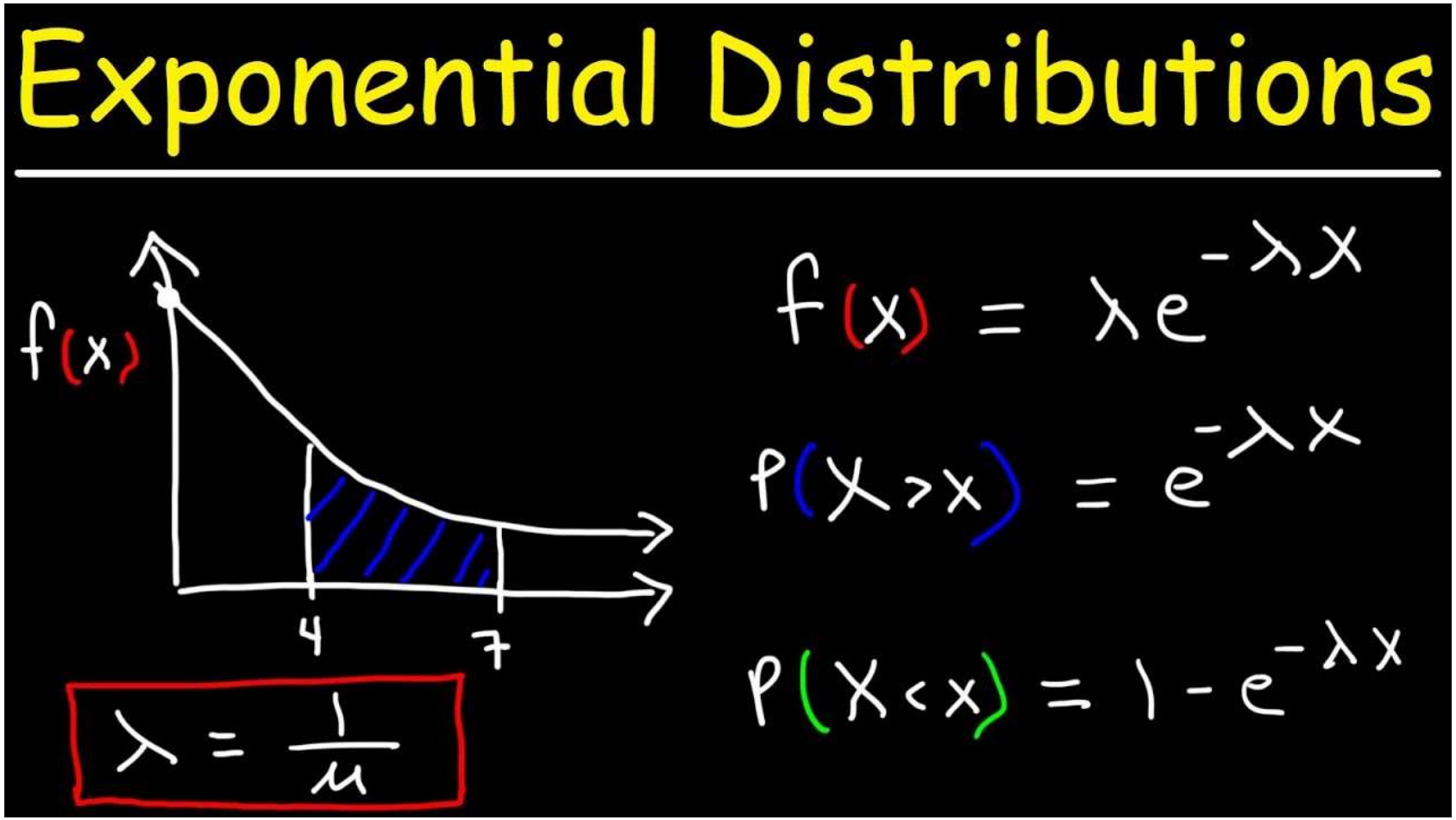


The type of event you could think about is the number of customers entering a store every 15 minutes. In this case, we keep the 15 minutes as a fixed value (unit time) so that we can ignore it in the rest of the calculations.
In this scenario, there would be an average number of customers entering each unit time, which is called the rate. This rate is called Lambda and it is the only parameter needed for the Poisson distribution

Exponential Distribution

The frequency distribution of the time or distance from one event to the next event.

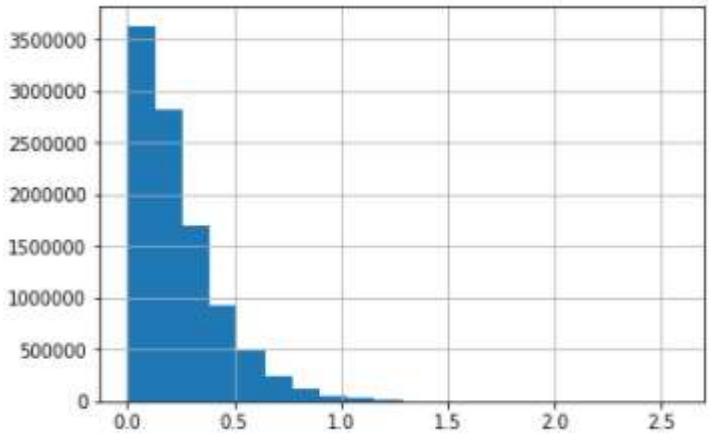
- The Exponential distribution is related to the Poisson distribution.
- Where the Poisson distribution describes the number of events per unit time, the exponential distribution describes the waiting time between events.
- It takes the same parameter as the Poisson distribution: the event rate.
- In some cases, however, (amongst others in Python's Scipy) people prefer to use the parameter 1 / event rate.



Weibull Distribution

- The Weibull distribution is another distribution that is a variation of the waiting time problem.
- It describes a waiting time for one event, if that event becomes more or less likely with time.
- The Weibull distribution takes two parameters.
  - ❖ Firstly, the rate parameter as in the Poisson and exponential distribution.
  - ❖ Secondly a c parameter. A c of 1 means that there is a constant event rate (so that is actually an exponential distribution). A c higher than one means that the event rate increases with time. A c below 1 means that the event rate decreases with time.

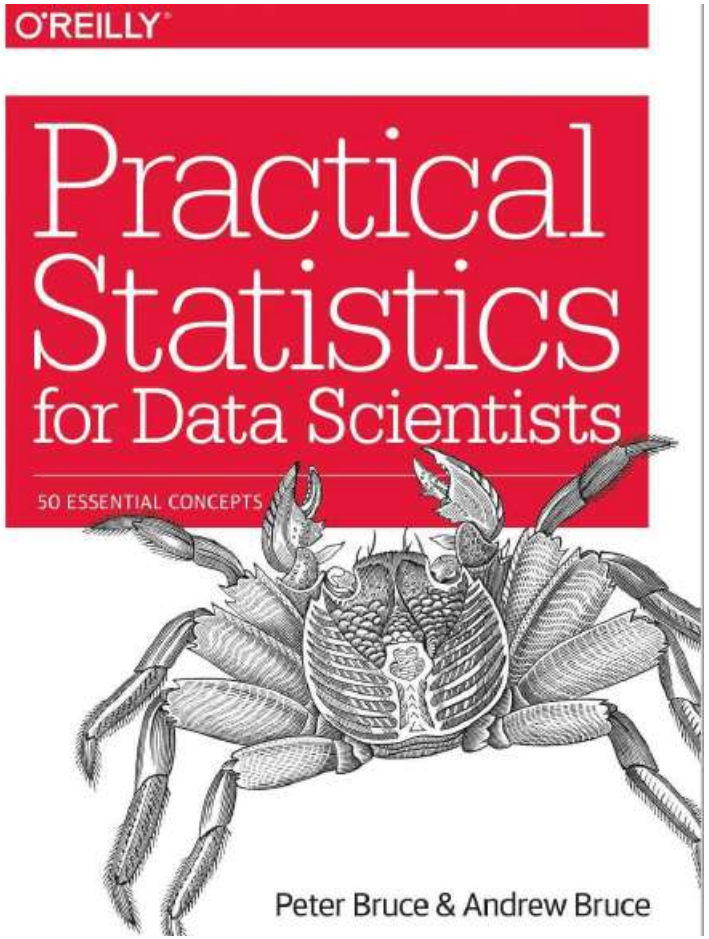
  If we take the same value of lambda as in the Poisson example (lambda = 4), and we add a value for c of 1.1 (so an increasing rate with time), we get the following result:

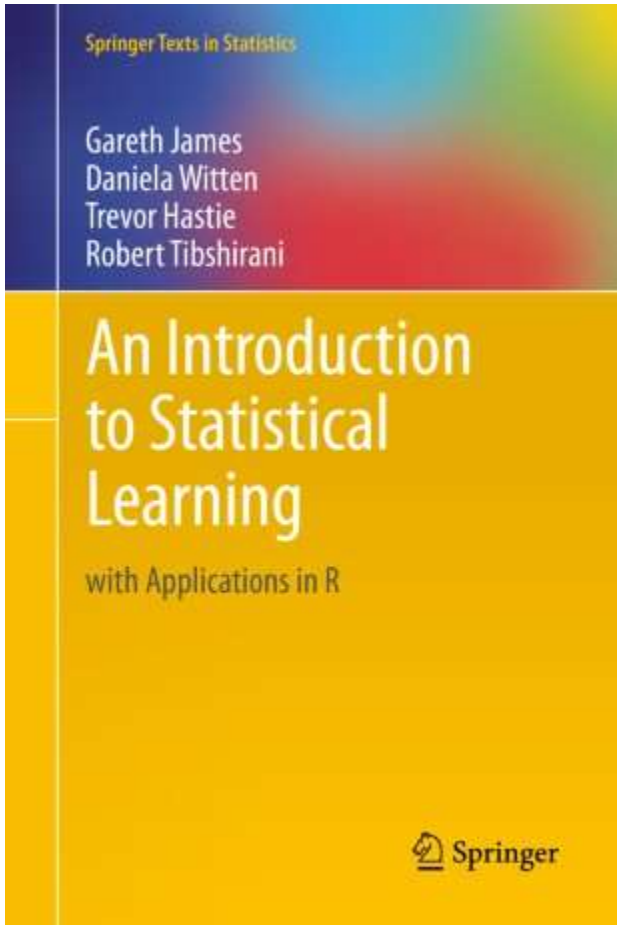


- ✓ For events that occur at a constant rate, the number of events per unit of time or space can be modeled as a Poisson distribution.
- ✓ In this scenario, you can also model the time or distance between one event and the next as an exponential distribution.
- ✓ A changing event rate over time (e.g., an increasing probability of device failure) can be modeled with the Weibull distribution.
- ✓ Random selection of data can reduce bias and yield a higher quality data set than would result from just using the conveniently available data.
- ✓ Knowledge of various sampling and data generating distributions allows us to quantify potential errors in an estimate that might be due to random variation.
- ✓ At the same time, the bootstrap (sampling with replacement from an observed data set) is an attractive "one size fits all" method to determine possible error in sample estimates.
- ✓ The t-distribution is actually a family of distributions resembling the normal distribution, but with thicker tails.
- ✓ It is widely used as a reference basis for the distribution of sample means, differences between two sample means, regression parameters, and more.

What do data scientists need to know about the t-distribution and the central limit theorem?

Not a whole lot. These distributions are used in classical statistical inference, but are not as central to the purposes of data science. Understanding and quantifying uncertainty and variation are important to data scientists, but empirical bootstrap sampling can answer most questions about sampling error. However, data scientists will routinely encounter t-statistics in output from statistical software and statistical procedures in R, for example in A-B tests and regressions, so familiarity with its purpose is helpful.





INVESTOPEDIA