



**MET CS688 C1**

# ***WEB ANALYTICS AND MINING***

**ZLATKO VASILKOSKI**

STATISTICS & PROBABILITY

# Data Samples and Precision - Motivation

Online polls vs standard polls

Which is better? Large number of people or a smaller subset?

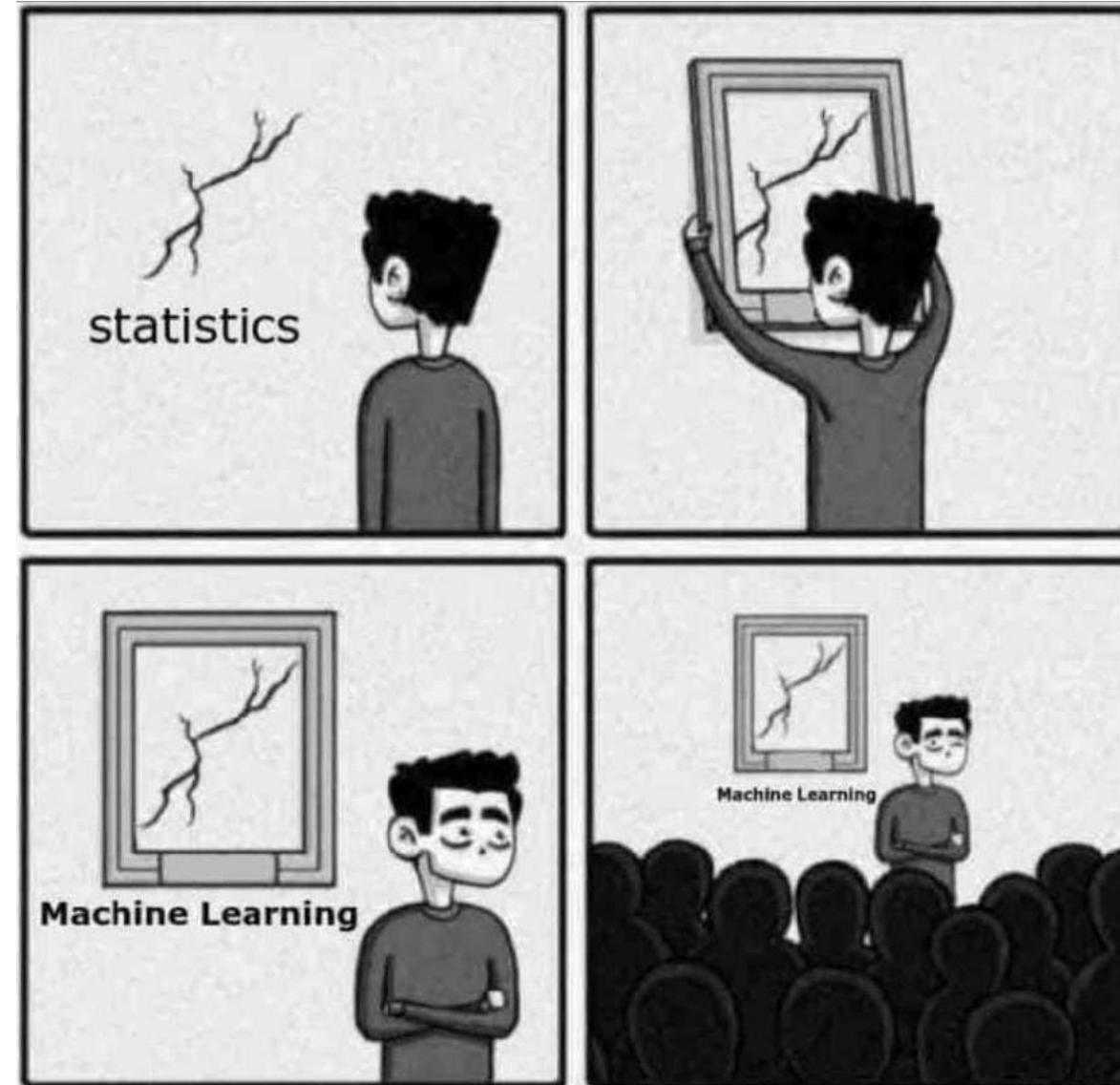
- In 1936, Literary Digest sent ballots to 10 million ( $\frac{1}{4}$ ) of the electorate regarding the U.S. presidency. 2.4 million people replied. Their conclusion was that Alfred Landon will win against Franklin Roosevelt with a large margin of 55% vs 41%. Roosevelt won by 67% vs 31%.
- George Gallup, founder of Gallup, Inc. (1935), on a much smaller sample correctly predicted that Franklin Roosevelt will win, making Gallup, Inc. a leader in American polling.
- Neglecting sampling bias (names from a phone book and car registration) in favor of size. Are they representative sample? Gallup carefully selected a representative sample.
- Very, very relevant to online sampling, larger data not always better.
- 1998 US university poll - the most influential person in the last millennium – Jamie Pollock - played for Manchester City, a team which then fell into what was then Division Two. Pollock scored a bizarre own goal over his own goalkeeper. The own goal condemned Manchester City to relegation to the third tier for the first time, whilst keeping the other team in the division. As a result, a group of QPR fans thanked him by voting him the "most influential man of the past 2,000 years" in an internet poll, where "Jesus and Marx came second, and third."

Sources: <http://www.bbc.co.uk/programmes/p049k2dj>

# Statistics and Probability

One view of interpretation of Statistics and Probability as  
Machine Learning

But lately Machine Learning seems to offer more.



# Statistics and Probability - Basics

## **Descriptive Statistics**

- Methods for organizing and summarizing the data through graphs, charts, tables, averages, measures of variation, and percentiles.

## **Inferential Statistics**

- Methods for drawing conclusions and their reliability about a population based on information drawn from a sample of the population.

# Descriptive Statistics - Basics

Describing data in graphs, charts, tables, averages, measures of variation..., important aspect to consider is how the data is distributed with respect to some central position.

## Measures of Central Tendency (center of the data set)

- Relates to the most typical value of the data set (measures of central tendency).
  - **Mean** (*or average*) - sum of the values divided by the number of the values in the data set.
  - **Median** - the number that splits the data between the bottom  $\frac{1}{2}$  and the top  $\frac{1}{2}$ .
  - **Mode** - the most frequently occurring value in the data set.

# Measures of Variation - Basics

It is also important to consider how data varies with respect to its central position.

The measures of center (*mean*, *median*, and *mode*) allow us to compare two data sets and draw conclusions. Two data sets can have similar measures of center but differ significantly.

The most used measures of data variation are the

- **Range** - the difference between the largest (maximum) and the smallest (minimum) dataset values.
- **Standard deviation S** (or square root of the **variance**) - measures of the deviation from the mean.
- **Interquartile range (IQR)** - is the difference between the third and first **quartiles**.

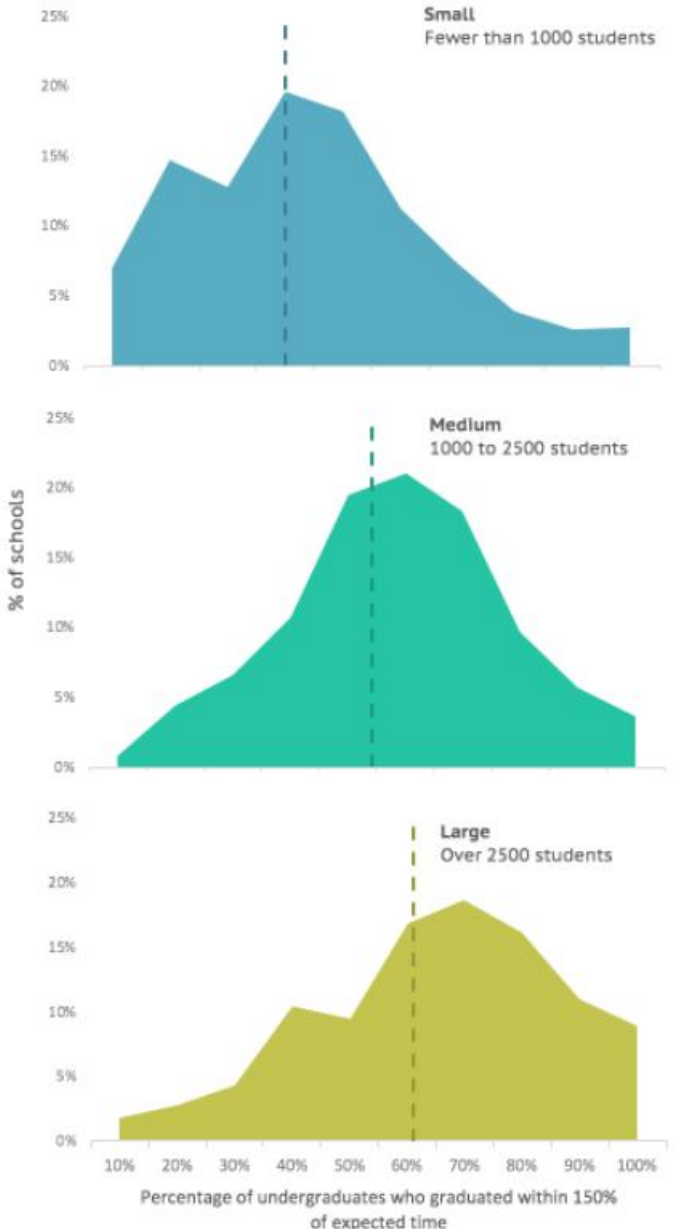
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$S^2$  = sample variance  
 $x_i$  = the value of the one observation  
 $\bar{x}$  = the mean value of all observations  
 $n$  = the number of observations

Source: <https://www.storytellingwithdata.com/blog/2019/2/21/various-views-of-variability>

## Graduation Rates from Private US Colleges

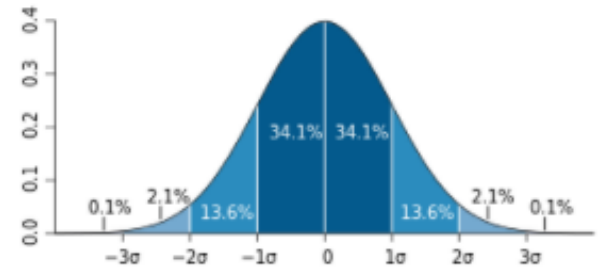
Graduation rates are correlated to school size, with higher graduation rates at larger schools. There is a large variation between schools within each size category.



# Measures of Variation in Data - Basics

**Standard deviation** (or square root of the **variance**) - measures of the deviation from the mean.

- The more variation there is in the data set, the larger the standard deviation.
- For data that has a bell-shaped distribution, the empirical rule states that
  - ~99.7% of the observations lie within **three** standard deviations of either side of the mean,
  - ~95% of the observations lie within **two** standard deviations of either side of the mean, and
  - ~58% of the observations lie within **one** standard deviation of either side of the mean.



**Interquartile range (IQR)** - is the difference between the third and first **quartiles**.

- Since the mean and the standard deviation is very sensitive to outliers (extreme values), therefore measures based on **percentiles & quartiles** are used.
- **Percentiles** divide the data set into 100 equal parts
  - The first percentile divides the bottom 1% from the top 99%
  - The second percentile divides the bottom 2% from the top 98%, etc.
- **Quartiles**
  - $Q_1$  - first quartile divides bottom 25% data from the top 75%.
  - $Q_2$  - second quartile (median) divides bottom 50% data from the top 50%.
  - $Q_3$  - third quartile divides bottom 75% data from the top 25%.
- $IQR = Q_3 - Q_1$

# Outliers - Basics

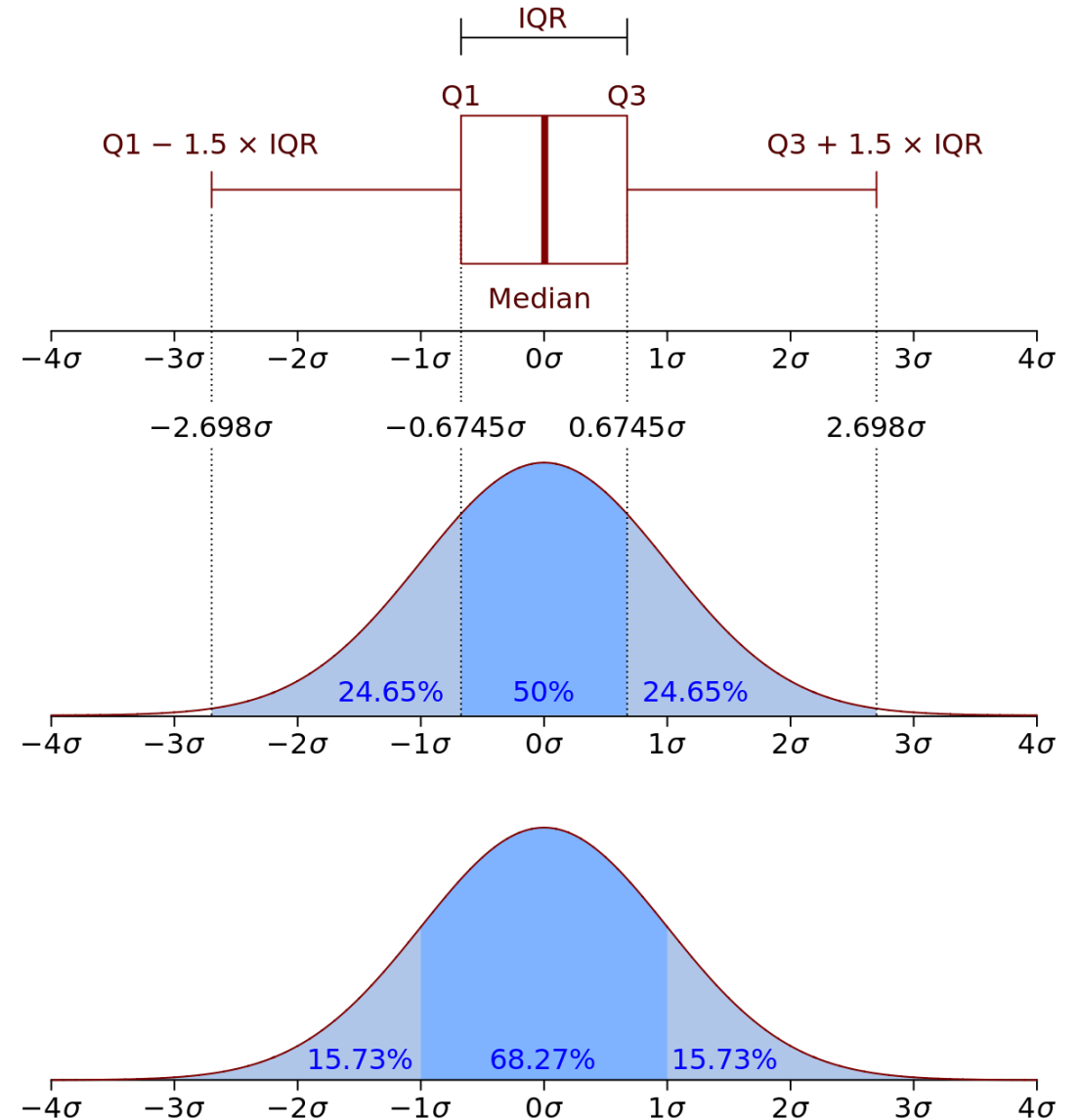
The lower limit of a data set is

$$Q_1 - 1.5 * IQR$$

And the upper limit of a data set is

$$Q_3 + 1.5 * IQR$$

The values that lie below the lower limit or above the upper limit are considered as **outliers**.





# Descriptive Measures—Population versus Sample

- The **population mean** is the mean of all observations for the entire population (for population size  $N$ ).
- The **sample mean** is the mean of a sample of size  $n$ .

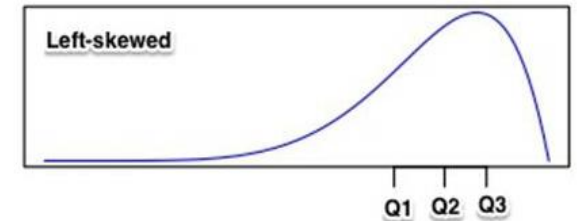
There is only **one** population mean whereas there are **many** sample means

# Shape of Data

The shape of data is the distribution of data values throughout the range of the data.

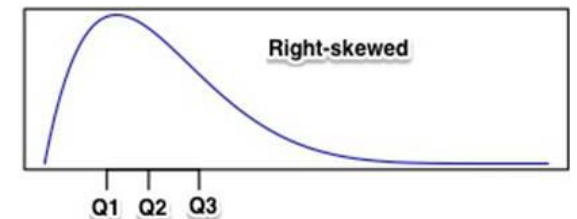
In a **left-skewed distribution**, the following properties hold:

- The distance (range) from the **minimum** value to the **median** is greater than the distance from the **median** to the **maximum** value.
- The distance (range) from the minimum value to the first quartile ( $Q_1$ ) is greater than the distance from the third quartile ( $Q_3$ ) to the maximum value.
- The distance (range) from the first quartile ( $Q_1$ ) to the median is greater than the distance from the median to the third quartile ( $Q_3$ ).



In a **right-skewed distribution**, the following properties hold:

- The distance (range) from the **minimum** value to the **median** is less than the distance from the **median** to the **maximum** value.
- The distance (range) from the minimum value to the first quartile ( $Q_1$ ) is less than the distance from the third quartile ( $Q_3$ ) to the maximum value.
- The distance (range) from the first quartile ( $Q_1$ ) to the median is less than the distance from the median to the third quartile ( $Q_3$ ).



In a **symmetric distribution**, all the above distances are the same.

# A Definition of Statistics

- **Descriptive:** It is used to describe the characteristics of the data, identifying narratives in the dataset. In particular, it enables us to draw some conclusion about the dataset.
- **Inferential:** It is used to make inferences from data and find something from the data, which is the same reason we are using machine learning.
- **Variable vs Value**
  - Value: **Continuous** vs **Discrete**
  - Variables: **Categorical (nominal), numerical**
  - Probability

# Types of Categorical Data

- **Ordinal:**

{high, medium, low} - {heavy, light} - {Saturday, Sunday,...}

- **Interval:**

- ordinal but with range

- {"very weak-to-weak", "weak-to-ok", "ok-to-good", "good-to-fantastic"}.
    - {"<30%" , "between 30% and 60%" and "more than 60%"}

- **Ratio:**

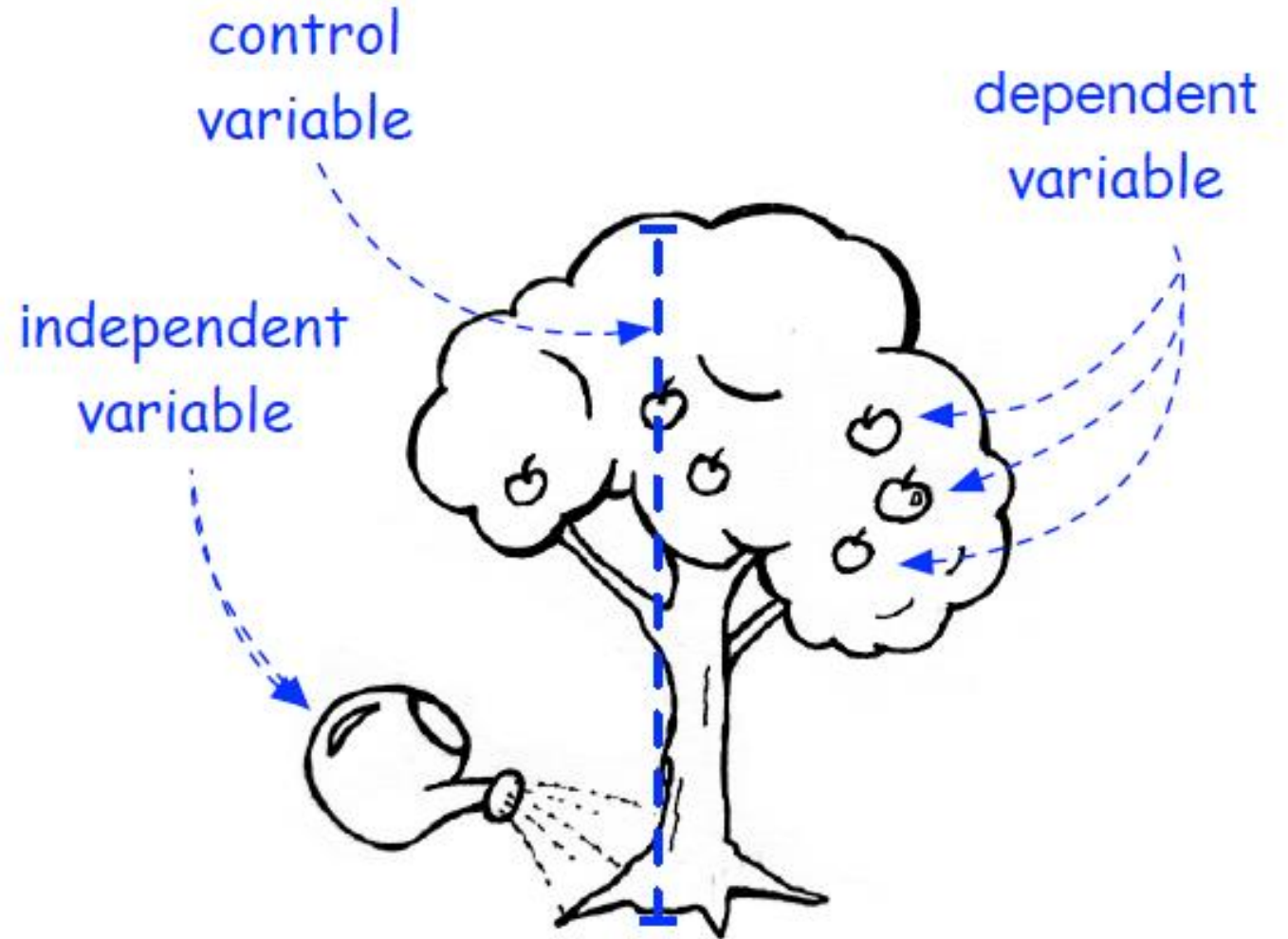
- ratio data has all properties of an interval data, and also has a clear definition of 0, which means no object for this variable exists, i.e. NULL.
    - the amount of air we breath is an interval data (e.g. very high, high, medium, low, very low,...), but not ratio because there is no zero air.

# Data - Basics

Types of variables

- **Control (constant)** variable
- **Dependent (output)** variable
- **Independent (input)** variable

**Independent** variable is a thing that causes changes to the **dependent** variable.



# First insight on the data

- **Arithmetic Mean**

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

- **Harmonic Mean**

Good for dealing with outliers. e.g. calculating the speed of moving object, profit of a company in consecutive years, ...

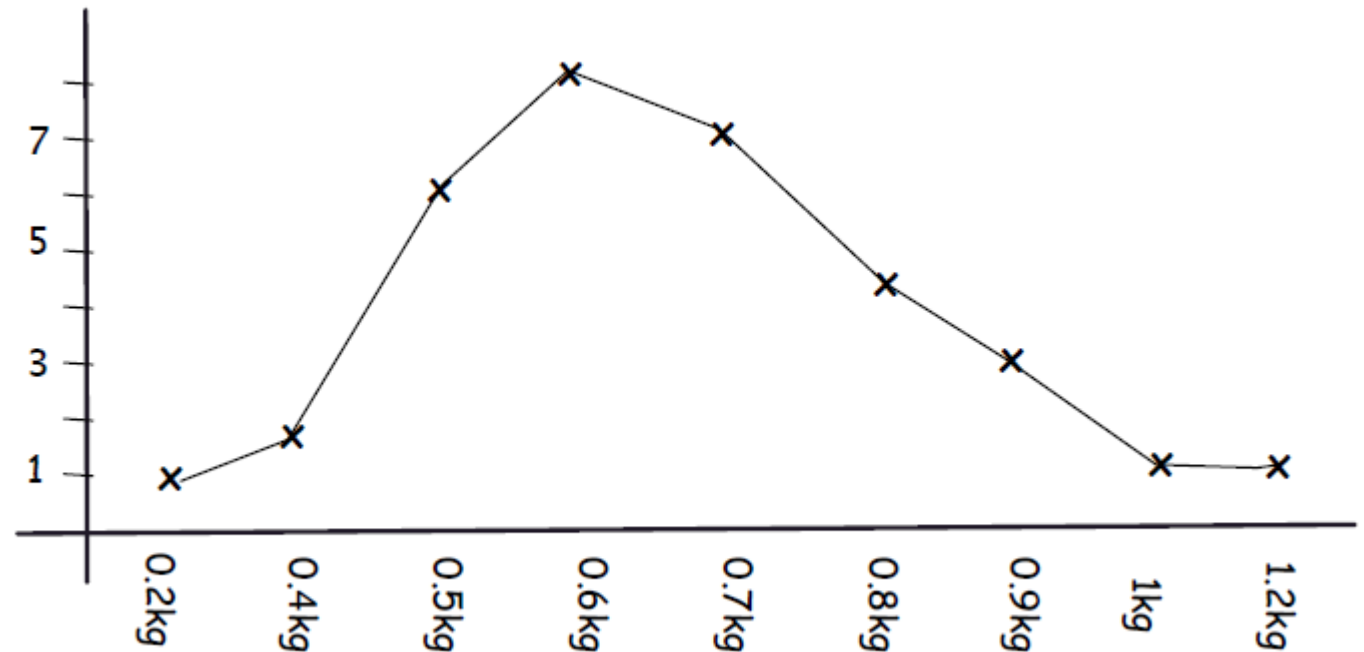
$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

- **Geometrical Mean**

Very sensitive to outliers and zero e.g. compare rooms with their capacity (width, height and length)

$$GM = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \dots x_n}$$

Weight	0.2kg	0.4kg	0.5kg	0.6kg	0.7kg	0.8kg	0.9kg	1kg	1.2kg
Number of chickens	1	2	6	8	7	4	3	1	1



# Median and Mode (for discrete data)

## Discrete Data

- **Multimodal** means a dataset that has more than one peak in its distribution, which is different than multivariate.
- **Multivariate** is a dataset that it has more than one information sources.

**Median** – middle value separating the higher half from the lower half of a data sample

- Even number of data points  
 $median(1, 2, 2, 3, 4, 9, 13, 100) = 3.5$
- Odd number of data points  
 $median(1, 2, 2, 3, 9, 13, 100) = 3$

{1, 1, 1, 2, 4, 4, 6, 7, 8}

↑  
median

{1, 2, 3, 4, 5, 6, 7, 8, 85, 88}

↑ ↑  
median

**Mode** is the data that has the highest frequency in the sample

{1, 1, 1, 2, 4, 4, 6, 7, 8} mode = 1

# Variance, Standard Deviation and Covariance

The **standard deviation**  $\sigma$  is the square root of the **variance**

$$\text{var}(X, X) = \sigma^2$$

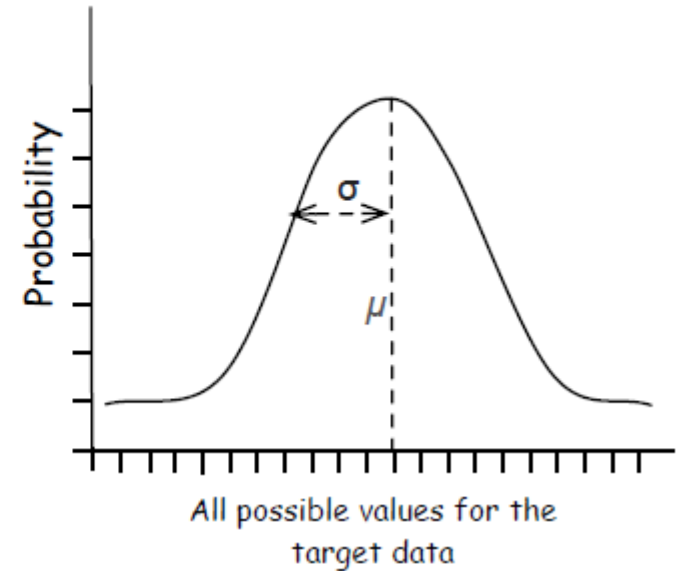
Variance

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

A single data point

Mean

Number of... data points



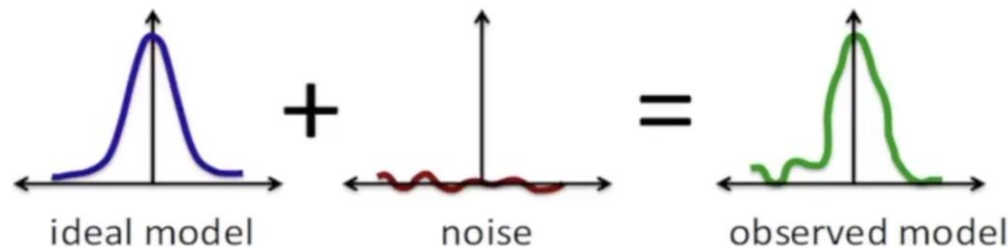
$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{n - 1}$$



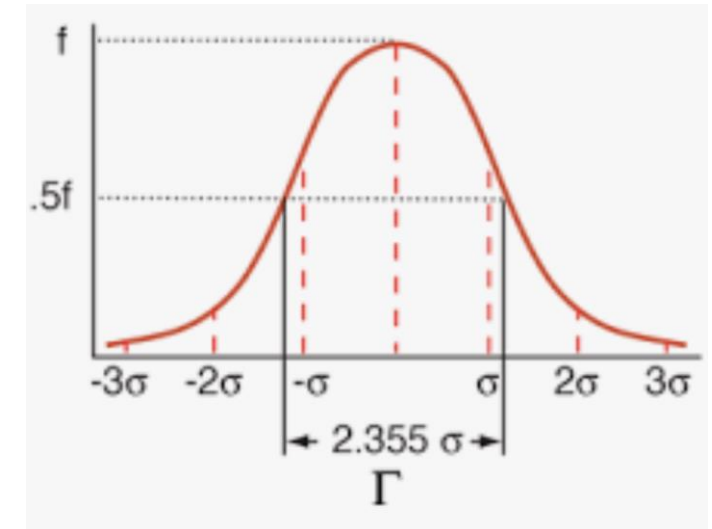
# General About Statistics

**Parameter Estimation** – General type of problem (more than 100 years old)

- Given samples that belong to some unknown but nicely defined distribution.
- Prototypical example - Samples given from a Gaussian  $f(x, \mu, \sigma)$ 
  - Determine the parameters  $(\mu, \sigma)$  that uniquely determine the distribution
  - For a Gaussian, this is a simple task. Just take a bunch of samples and calculate:
    - The **Mean**  $\mu = \frac{1}{N} \sum_{i=1}^n X_i$
    - The **Variance**  $\sigma^2 = \frac{1}{N} \sum_{i=1}^n (X_i - \mu)^2$
- The determination of these parameters for a given sample is an example of a more general paradigm that started with the work of R. A. Fisher between 1910 and 1920, called **Maximum Likelihood Estimator (MLE)**.
  - MLE fastest convergence toward the parameters as the number of samples increases to infinity. Many times, it is very hard (computationally) to achieve this.
  - Even more, in reality we have models with a noise, addressed by J. W. Tukey in 1960.



Gaussian (Normal Distribution)



$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

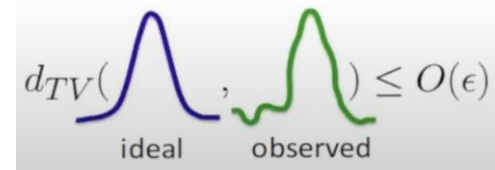
$\frac{1}{\sqrt{2\pi}\sigma}$  – **Mode**, height of the peak  
 $\mu$  – **Mean**, position of the center of the peak  
 $\sigma$  – **Standard deviation**  
 $\sigma^2$  – **Variance**

# General About Statistics

How to perform parameter estimation for model with a noise? Can they be de-noised?

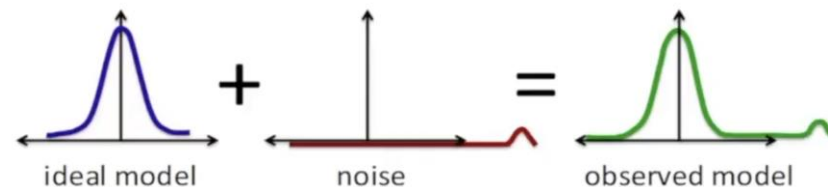
- Consider the  $L_1$  norm of the noise to be at most  $O(\epsilon)$ .
- One way to think about this is to consider that each sample contains  $\epsilon$  fraction of corrupt data.
  - Called Huber's contamination model
  - Outliers – data that was corrupt
  - Inliers – data from the intact distribution
- What do we mean by MLE (Maximum Likelihood Estimator) in a presence of noise?
  - Total variation distance which is  $\frac{1}{2}$  of the  $L_1$  distance between the two distributions (Gaussian and noise).

$$d_{TV}(f(x), g(x)) = \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| dx$$



The diagram shows two probability density functions. On the left is a smooth, symmetric blue curve labeled 'ideal'. On the right is a green curve labeled 'observed' which is mostly symmetric but has a small, irregular bump on its right tail. Between the two curves is the text  $d_{TV}(\text{ideal}, \text{observed}) \leq O(\epsilon)$ .

- Think about the goal in few different ways:
  - Given something that is  $\epsilon$  close to being a Gaussian, find MLE that is  $\epsilon$  close to the original non corrupted Gaussian
  - The data is not a Gaussian at all, but can we find a nice distribution such a Gaussian that will best represent the data.
- Do MLE estimates of the mean  $\mu$  and the standard deviation  $\sigma$  work in presence of noise?
  - No, they don't!



# General About Statistics

How do we make **MLE** work on a noisy data?

- Instead of empirical **mean**  $\mu$ , use the **median**
- Instead of empirical **variance**  $\sigma^2$ , use **MAD** - Median Absolute Deviation (median of medians)  
$$MAD = \text{median}(|X_i - \text{median}(X_1, X_2, \dots, X_n)|)$$

How to use Median of medians (MAD)

- Calculate the median for all of the samples
- Look at the absolute value of the difference between the sample and the median of the samples.
- Take median of this absolute value difference

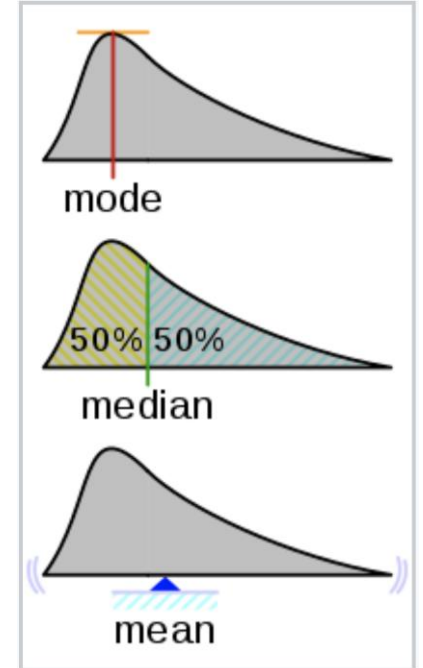
All this works nicely in 1 dimension, called robust estimates of location and scale.

$$\hat{\mu} = \text{median}(X)$$
$$\hat{\sigma} = \frac{MAD}{\Phi^{-1}(3/4)}$$

Here the correction term  $\Phi$  to MAD is just to make sure that MAD is unbiased estimator.

Thus  $\hat{\mu}$  and  $\hat{\sigma}$  are within  $\varepsilon$  of the original Gaussian.

- Computationally robust and very efficient (median calculations very efficient)



# Gaussian Mixture Models (GMM)

Clustering - number of clusters enclosed in a dataset, but we do not know where these clusters are as well as how they are shaped.

Standard approach KNN

- The KNN model in the overlapping areas of 2 clusters is not so accurate.
- KNN clusters are **circular** shaped whilst the data can be of **ellipsoid** or any other shape.
- Another weak point of KNN in its original form is that each point is allocated to one cluster, that is, **each point** either belongs to **cluster 1** or to **cluster 2**. This leads to inclusion of any points between the clusters.

Better algorithm is Gaussian Mixture Models (GMM)

- Assume that the clusters are not defined by simple circles but by more complex, ellipsoid shapes.
- Accomplishes this by trying to fit a mixture of gaussians to the dataset.
- Allocate to each point a likelihood to belong to each of the gaussians.

KNN vs GMM, [https://www.python-course.eu/expectation\\_maximization\\_and\\_gaussian\\_mixture\\_models.php](https://www.python-course.eu/expectation_maximization_and_gaussian_mixture_models.php)

# Basic Probability Concepts

$$P(\text{desired event}) = \frac{\text{Number of desired events}}{\text{Total number of events}}$$

- Probability of an **opposite** event  $A'$

$$P(A') = 1 - P(A)$$

- Probability of two **independent** events, for example, if two coins are flipped, then the chance of both being heads is

$$P(A \text{ and } B) = P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

- **Joint** Probability of either  $A$  or  $B$  occurring ( $A \cup B$ ,  $A$  or  $B$ )

- For two **mutually exclusive** events such as rolling a 3 or 5 on a six-sided die is

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) = P(3 \text{ or } 5) = P(3) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

- For two **NON mutually exclusive** events such as drawing a card from a deck of cards and getting a diamond ( $\frac{13}{52}$ ) or a face card (J,Q,K) ( $\frac{12}{52}$ ) or both ( $\frac{3}{52}$ ) is

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$$

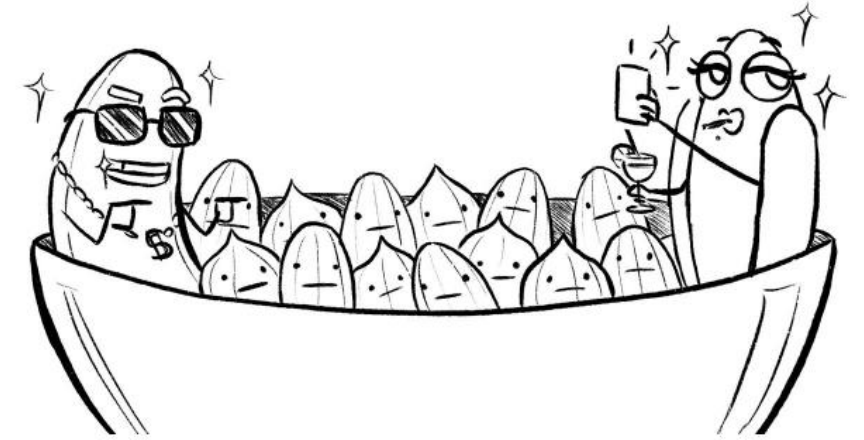
- **Conditional** Probability - "the probability of A, given B" is

$$P(A|B) = P(A \text{ given } B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3} \text{ or } \frac{2}{3}$$

- For example, in a bag of 2 white balls and 2 black balls (4 balls in total), the probability of drawing a white (or black) ball is  $\frac{1}{2}$ , however, when taking a second ball, the probability of it being either a white ball or a black ball depends on the ball previously taken and it can be  $\frac{1}{3}$  or  $\frac{2}{3}$  depending if a white or black ball respectively was taken out previously.

# Exercise

10% pistachio  
20% cashew  
40% hazelnut  
30% almonds



A mixture of nuts and the use of logical “**OR**” and “**AND**” operations. Consider events of picking a nut **independent** and **mutually exclusive**

- What is the probability of drawing a pistachio **or** a cashew?

Logical **OR** for **independent and mutually exclusive** events reduces to addition of the probabilities

$$A \text{ or } B = A \cup B = P(A) + P(B)$$

$$A = P(\text{Pistachio}) = 0.1$$

$$B = P(\text{Cashew}) = 0.2$$

$$P(\text{Pistachio or Cashew}) = P(A) + P(B) = 0.1 + 0.2 = 0.3$$

- What is the probability of drawing a pistachio **and** a cashew?

Logical **AND** for **independent and mutually exclusive** events reduces to multiplication of the probabilities

$$A \cap B = A \text{ and } B = P(A) \cdot P(B)$$

$$P(\text{Pistachio and Cashew}) = P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02$$

# Expected Value (Mean)

**Expected Value** is used when we want to calculate the mean of probability events.

Mean is used when we want to calculate the average value of a given sample.

- So **expected value** is a **mean** but with **weight** or **probability of each values**.
- **Example 1:**
  - In rolling a dice, the probability of getting 6 is  $1/6$ . Therefore, we can say the **expected value** of rolling a dice twice and getting 6 both times is

$$E = 1/6 * 1/6 = 1/36$$

- **Example 2:**
  1. The class “Web Mining for beginners” costs \$3000
  2. The class “Web Mining for intermediates” costs \$7,000
  3. The class “Web Mining for expert” costs \$10,000

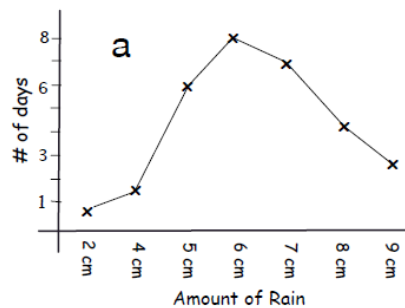
100 people subscribe for a course. 2% have subscribed for the expert level, 8% for intermediate level and 90% for beginner level.

  - What is the **expected value**?

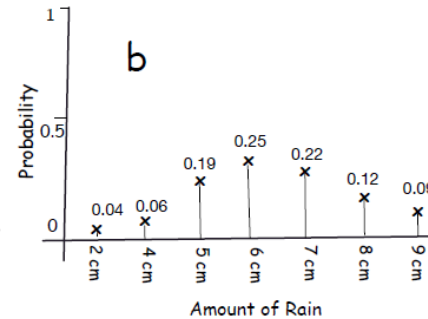
$$E = 0.02 * 10000 + 0.08 * 7000 + 0.90 * 3000 = 20 + 56 + 270 = 346$$

# Few other Probability Quantities

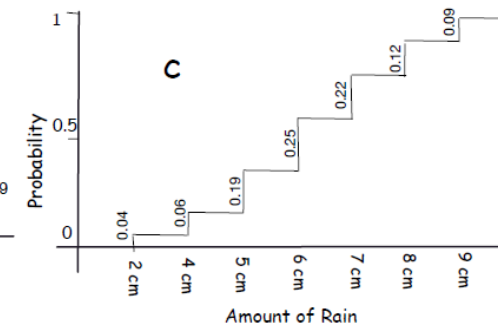
- **Probability Mass Function (PMF)** shows how **dense** is the probability at **each data point** in a **continuous vector**.
  - Probability Mass Function (PMF) shows how **dense** is the probability at **each data point** in a **continuous vector**.
- **Probability Density Function (PDF)** shows how **dense** is the probability at **each data point** in a **discrete vector**.
  - Probability Density Function (PDF) shows how **dense** is the probability at **each data point** in a **discrete vector**.
- **Cumulative Distribution Function (CDF)** is a function that describes a **distribution** of a variable (either discrete or continuous variable).
  - Cumulative Distribution Function CDF is a function that describes a **distribution** of a variable (either discrete or continuous variable).
- CDF is the probability of being  $x$  ( $x$  is a value), CDF is the cumulative PDF or PMF.



↑  
distribution of rains  
dnorm



↑  
PDF  
pnorm

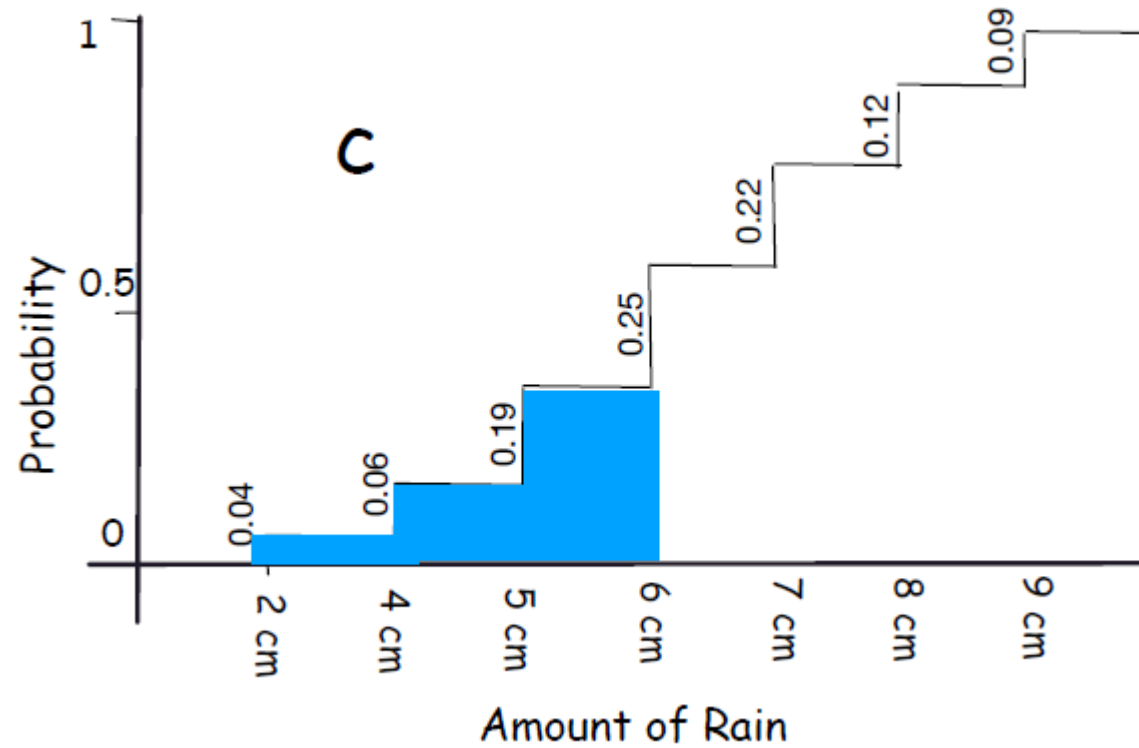


↑  
CDF  
qnorm



# CDF Example

- What is the probability that raining is going to be less than 6cm or  $P(x < 6\text{cm})$ ?

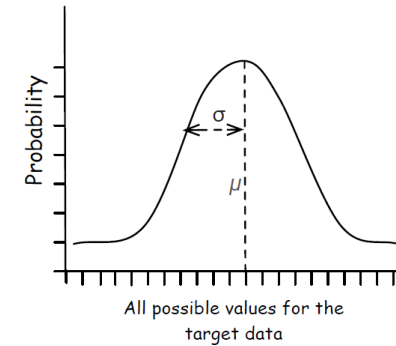


# Distributions

- Usually, we use distributions to plot a variable in a two-dimensional space
- A distribution presents characteristics of a dataset (descriptive statistics)
- Remember:
  - *Any data object must include repetitive values in a dataset, because any scientific phenomena should be reproducible.*
- The distribution presents all possible values of a single variable or **how often they occur** (probability or frequency of their occurrences).

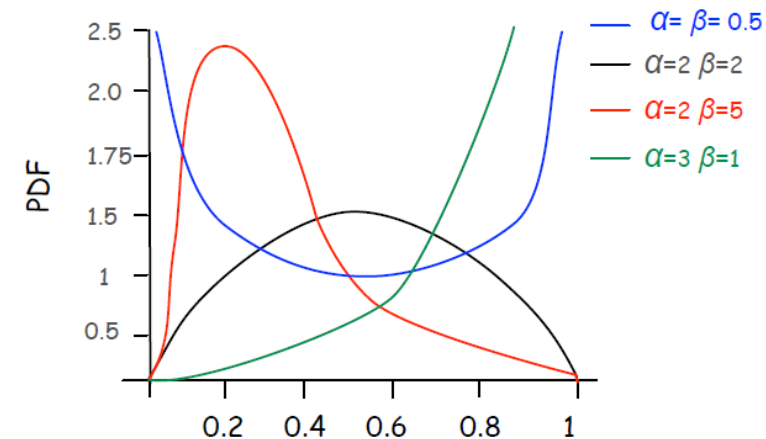
# Distributions - examples

- **Normal (Gaussian) Distribution** (described before)



- **Beta Distribution**

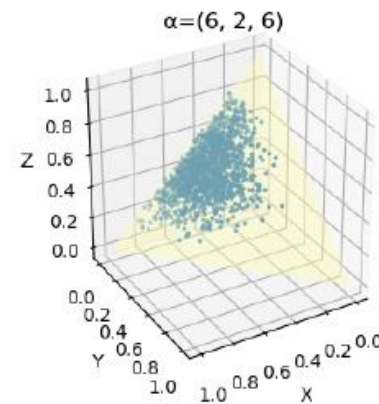
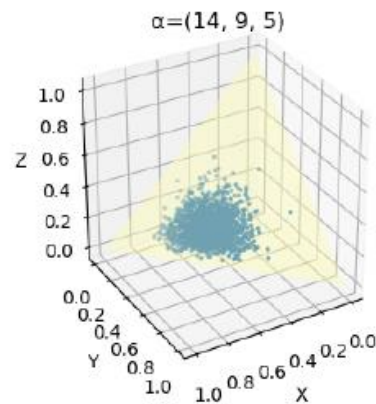
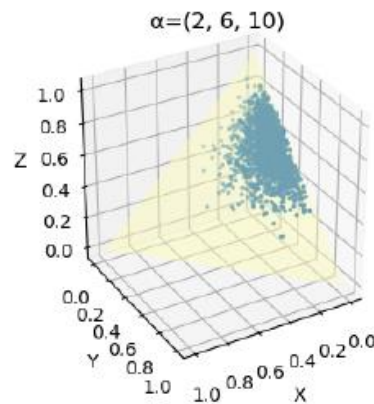
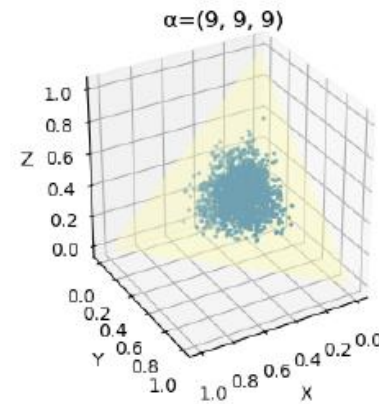
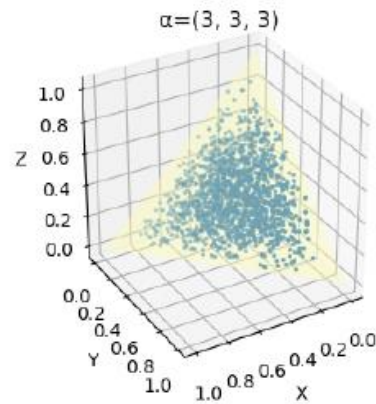
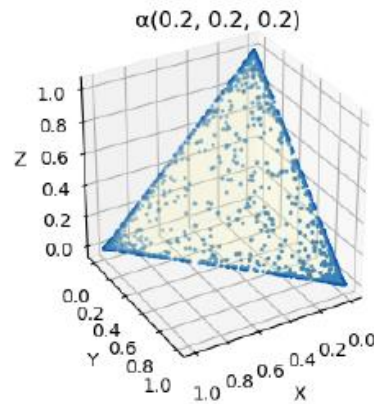
- The Beta distribution is suitable for the **random behavior of a binary trail** (yes/no, success/failure, coin head/tail,...).
- In other words, when there is an uncertainty in a binary probability result, we can use Beta distribution to understand the **conditional distribution** of a success rate (success is a binary state such as true/ok/yes...).
- It is useful when we **do not have any information** about the probability.



# Dirichlet Distribution

- **Dirichlet Distribution**

- Similar to Beta distribution but for two variable instead of one variable.
- $\alpha$  is a vector of non-zero numbers.
- $k$  is the dimension number of the vector



# Binomial Distribution

**Binomial Distribution**  $P(k) = \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$

$p$  – probability of a **successful** event

$k$  – number of **successful** events

$n$  – total number of **independent** trials

- There are cases that the variable is **discrete** and **binary** (it has only two discrete states), such as flipping a coin (head or tail), the door state (open or close) or whether you find this course helpful (yes, no). For these cases, the **binomial distribution** is used.
- **Independent trail:** each trial (coin flipping event) does “not” have any connection to the previous trial, e.g. flipping coin, dice,...
- **Dependent trail:** the outcome of a trail depends on its previous trails, e.g. taking anti depressant pills. (1st is good, 2nd is great, 3rd sleepy, 4th more depressed)
- Binomial distribution is appropriate if these three conditions are all true.
  1. We have a series of **independent trails**.
  2. Trail **output is binary**, and denoted as success or failure, but it could be other information as well such as yes/no, true/false, etc.
  3. **There are finite number of trails**.

**Example:** Suppose a biased coin comes up heads with probability  $p = 0.3$  (probability of a **successful** event) when tossed. The probability of seeing exactly  $k = 4$  heads (number of **successful** events) in  $n = 6$  tosses (total number of **independent** trials) is

$$P(k) = \binom{6}{4} 0.3^4 (1 - 0.3)^{6-4} = 0.0595$$

# Geometric Distribution

**Geometric Distribution**  $P(X = k) = (1 - p)^{k-1}p$

$p$  – probability for number of events for one **successful** event  
 $(1 - p)$  – probability for number of **failures** before the first success  
 $k = 1, 2, \dots$  – the number of trials

Assumptions:

- (i) there is a series of **independent** trails.
- (ii) trail's **output is binary**, e.i. success/failure, yes/no, true/false, etc.
- (iii) As the desired binary state is acquired the **trial stops immediately**.

**Example:** A flying insect can collide with car's window in  $k$  passing cars (trials) making it dirty. Define:

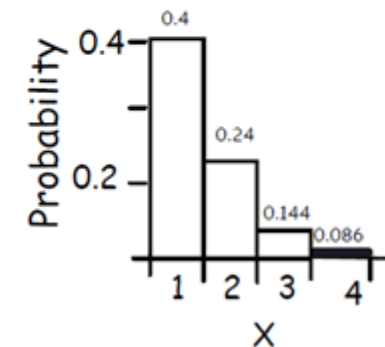
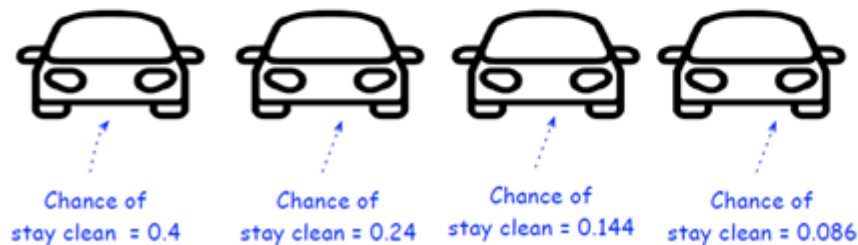
- $p = 0.4$  - Probability of a **successful** event - insect does not collide with car's clean window.
- $(1 - p) = 1 - 0.4 = 0.6$  - Probability of a **failure** event - insect collides with car's clean window.

$$P(X=1) = P(\text{success in the 1st trial}) = 0.4$$

$$P(X=2) = P(\text{failure in the 1st trial}) \times P(\text{success in the 2nd trial}) = 0.6 \times 0.4 = 0.24$$

$$P(X=3) = P(\text{failure in the 1st trial}) \times P(\text{failure in the 2nd trial}) \times P(\text{success in the 3rd trial}) \\ = 0.6 \times 0.6 \times 0.4 = 0.144$$

$$P(X=4) = P(\text{failure in the 1st trial}) \times P(\text{failure in the 2nd trial}) \times P(\text{failure in the 3rd trial}) \times \\ P(\text{success in the 4th trial}) \\ = 0.6 \times 0.6 \times 0.6 \times 0.4 = 0.086$$



# Poisson Distribution

- There are **rare** events happening in a system, such as malfunctions of a machine. We know the average occurrences of these rare events (in time) and their frequency is not changing.
- Poisson distribution is being used to model the **intervals** of **rare events**.
- The **value of the mean ( $\lambda$ ) and variance are equal**. Important not to forget is that the mean  **$\lambda$  is not changing**.
- The number of rare events is  **$r$** .

The diagram shows the Poisson distribution formula  $P(X = r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!}$ . Blue dashed arrows point from text annotations to parts of the formula: one from "'e' is a mathematical constant similar to  $\pi$  and it is 2.718" to the  $e^{-\lambda}$  term; another from "Mean" to the  $\lambda^r$  term; and a third from "The number of rare event occurrences" to the  $r!$  term in the denominator.

$$P(X = r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!}$$

## Example:

An earthquake occur twice ( $\lambda = 2$ ) every 100 years on average. Assuming the Poisson model is appropriate, what is the probability of  $r = 3$  earthquakes occurring in 100 years?

$$P(r = 3 \text{ earthquakes in 100 years}) = \frac{e^{-\lambda} \lambda^r}{r!} = \frac{e^{-2} 2^3}{3!} = \frac{e^{-2} 2^3}{3 \cdot 2 \cdot 1} = 0.18$$

# Poisson Distribution - Examples

- We have a rare event (a failure) with a mean occurrences  $\lambda = 2$  per month. This means that in 4 months we can expect 4 failures ( $\lambda = 2 \cdot 4 = 8$ ). In other words, the mean of four months is 8. What is the probability of  $r = 3$  of these rare event occurring in 4 months?

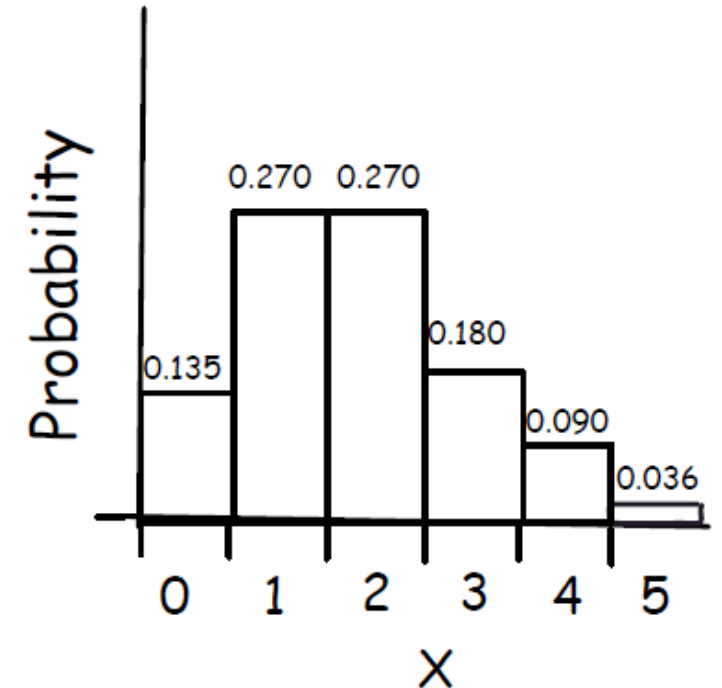
$$P(r = 3) = \frac{e^{-\lambda} \cdot \lambda^r}{r!} = \frac{e^{-8} \cdot 8}{3!} = \frac{e^{-8} \cdot 8}{3 \cdot 2 \cdot 1} = 0.286$$

- What is the probability that we get zero failure ( $r = 3$ ) in a month (per month  $\lambda =$

$$P(r = 0) = \frac{e^{-2} \cdot 2^0}{0!} = \frac{e^{-2} \cdot 1}{1} = 0.135$$

- For a single failure we would have

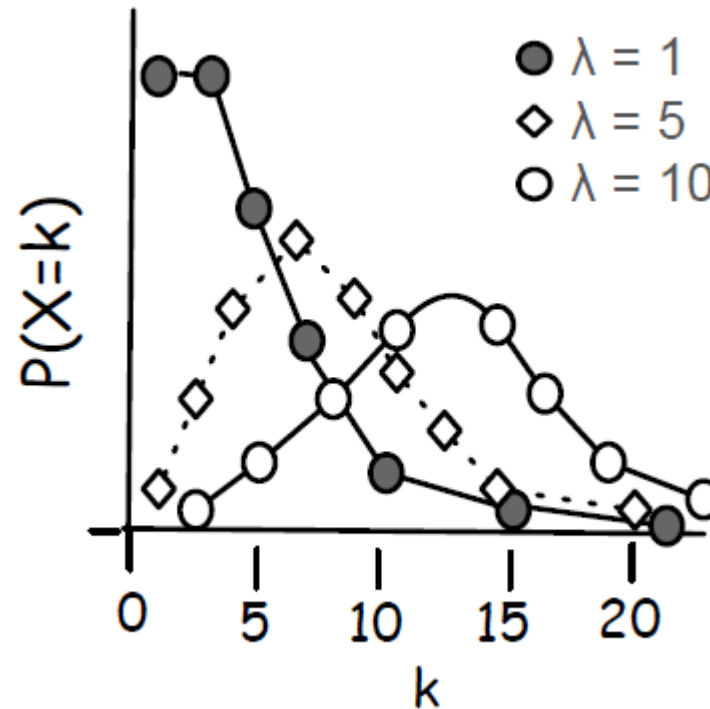
$$P(r = 1) = \frac{e^{-2} \cdot 2^1}{1!} = 0.270$$





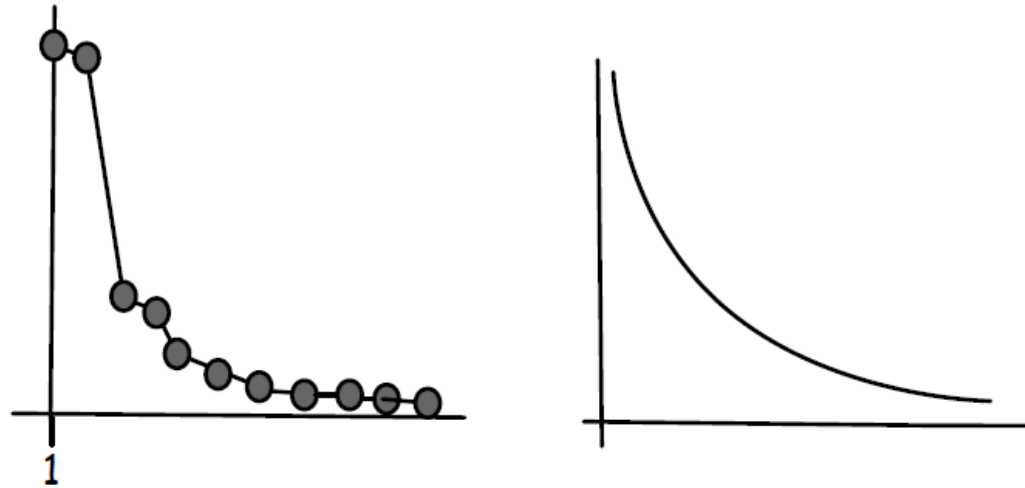
# Poisson Distribution - Notes

- If the rare events occur at a constant rate, Poisson distribution is appropriate.
- Nevertheless, if they occur at random rate and time, and we cannot identify its rate, we can use **Weibull distribution**, which we will not explain it here.



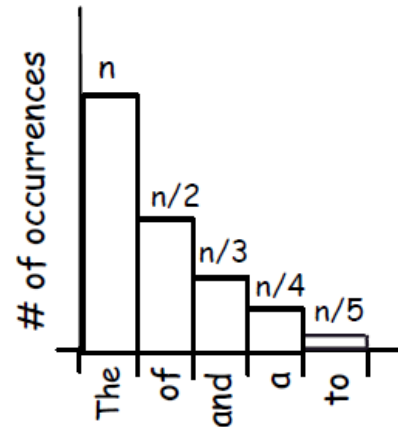
# Power law Distribution

- Examples of **Power Law Distribution** were mentioned before



$Y=X^{\alpha}$ . ' $\alpha$ ' is called power law exponent and causes this exponential changes, it is constant.

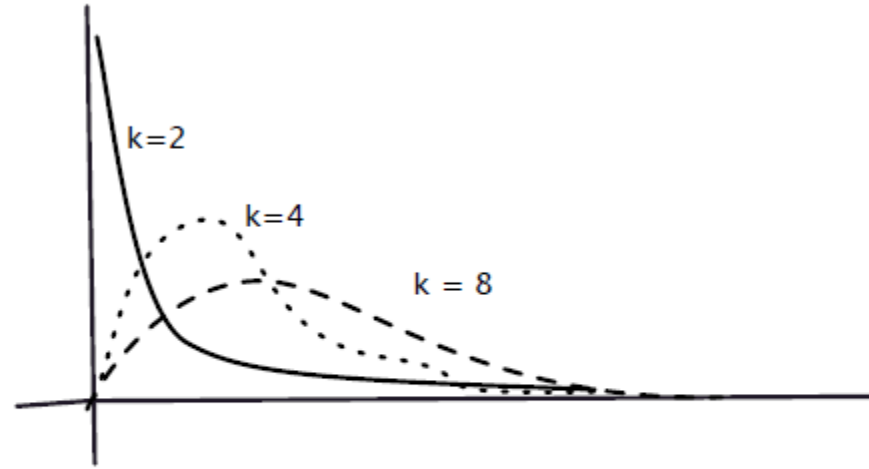
**Zipflaw**



# Chi-Square Distribution

Relates to  $k$  - Degree of Freedom

- The degree of freedom  $k$  is the number of independent variables in a study that can vary freely.



Chi-square distribution with three different degree of freedom.

# Summary

- We use **binomial distribution**, when we like to know the “probability of getting the certain number of success”.
- We use **geometric distribution**, when we like to know “how many trials do we need before the first success”.
- For both **binomial** and **geometric distribution**, the probability of success in each trial should be equal. Otherwise, none of those distributions could be used.
- **Geometric distribution** and binomial distribution are very similar. However, geometric distribution “stops” as the first failure or success or any other target Boolean variable it may encounter.
- In **Poisson distribution**,  $\lambda$  (lambda) is being used to present mean and not  $\mu$ , because in Poisson distribution, variance is equal to the mean. Therefore, using  $\mu$  or  $\sigma^2$  might be confusing.
- Use **Poisson distribution** if the events are independent. For instance, malfunction events occur in a given interval, and we know the value of  $\lambda$  in that interval.
- **Binomial, geometric and Poisson distributions** are for discrete data and for discrete data we use histogram. Nevertheless, since the number of data points are usually large, a line chart is being used to demonstrate distribution.
- When the number of sample is too large, it is better to use **Poisson** distribution rather than **binomial** distribution. Because when  $n$  is large, and the system must calculate  $n!$  will eat lots of computer memory.
- **Normal, Chi-square and Power-Law** distributions are also used for continuous variables.

# R versions for These Distributions

- For the beta distribution see [dbeta](#).
- For the binomial (including Bernoulli) distribution see [dbinom](#).
- For the Cauchy distribution (type of normal distribution) see [dcauchy](#).
- For the chi-squared distribution see [dchisq](#).
- For the exponential distribution see [dexp](#).
- For the F distribution see [df](#).
- For the gamma distribution see [dgamma](#).
- For the geometric distribution see [dgeom](#). (This is also a special case of the negative binomial.)
- For the hypergeometric distribution see [dhyper](#).
- For the log-normal distribution see [dlnorm](#).
- For the multinomial distribution see [dmultinom](#).
- For the negative binomial distribution see [dnbinom](#).
- For the normal distribution see [dnorm](#).
- For the Poisson distribution see [dpois](#).
- For the Student's t distribution see [dt](#).
- For the uniform distribution see [dunif](#).
- For the Weibull distribution see [dweibull](#).

A

- A