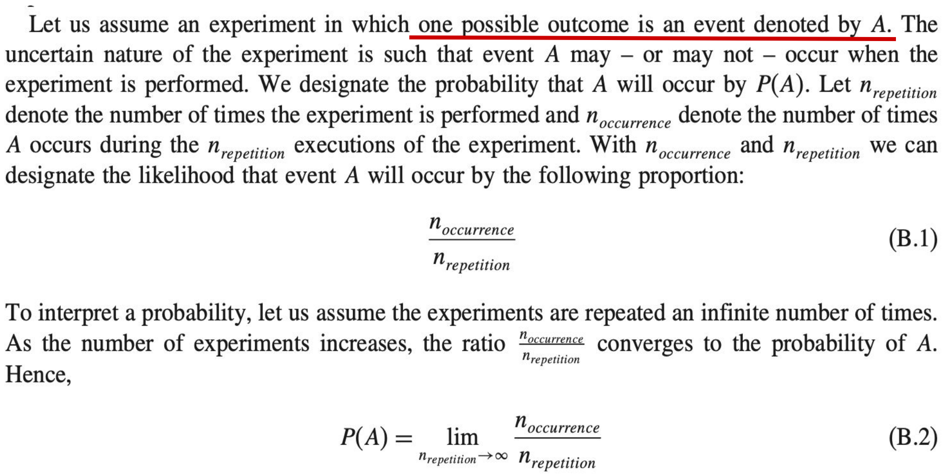
CIS 678 - Machine Learning

Basics of Probability

Basic Probability - Data Science



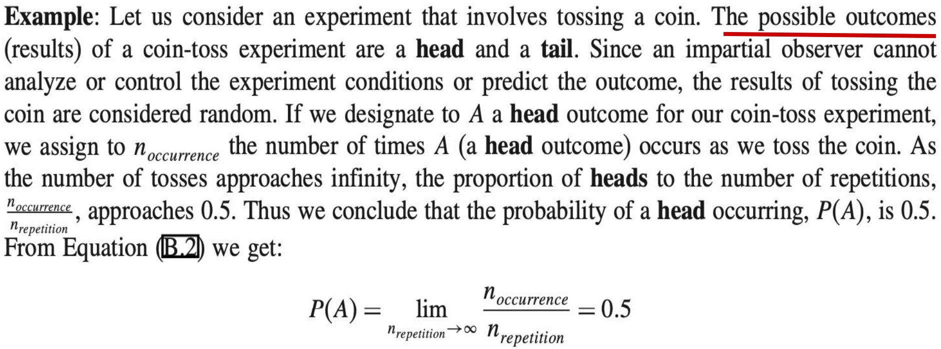
Let us assume an experiment in which one possible outcome is an event denoted by A. The uncertain nature of the experiment is such that event A may – or may not – occur when the experiment is performed. We designate the probability that A will occur by $P(A)$. Let $n_{repetition}$ denote the number of times the experiment is performed and $n_{occurrence}$ denote the number of times A occurs during the $n_{repetition}$ executions of the experiment. With $n_{occurrence}$ and $n_{repetition}$ we can designate the likelihood that event A will occur by the following proportion:

$$\frac{n_{occurrence}}{n_{repetition}} \tag{B.1}$$

To interpret a probability, let us assume the experiments are repeated an infinite number of times. As the number of experiments increases, the ratio $\frac{n_{occurrence}}{n_{repetition}}$ converges to the probability of A. Hence,

$$P(A) = \lim_{n_{repetition}\to\infty} \frac{n_{occurrence}}{n_{repetition}} \tag{B.2}$$

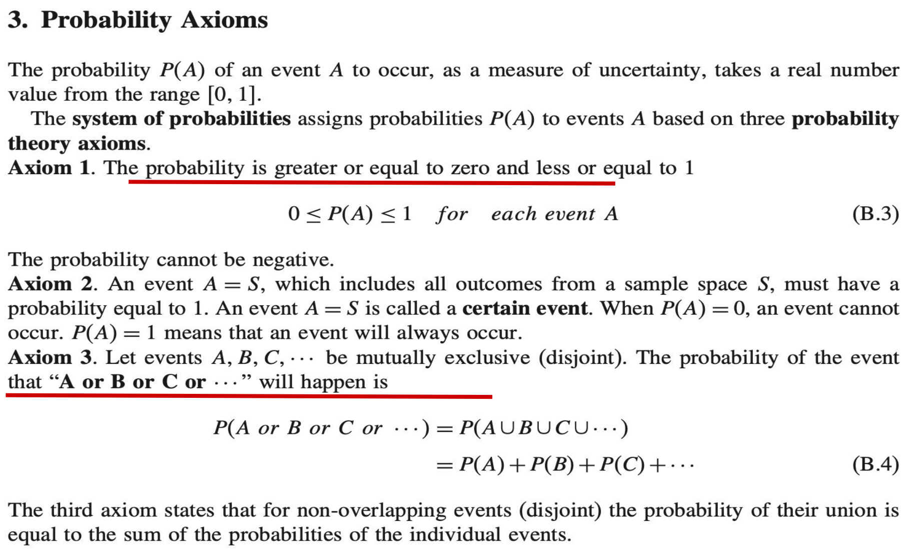
Basic Probability - Data Science



Example: Let us consider an experiment that involves tossing a coin. The possible outcomes (results) of a coin-toss experiment are a **head** and a **tail**. Since an impartial observer cannot analyze or control the experiment conditions or predict the outcome, the results of tossing the coin are considered random. If we designate to A a **head** outcome for our coin-toss experiment, we assign to $n_{occurrence}$ the number of times A (a **head** outcome) occurs as we toss the coin. As the number of tosses approaches infinity, the proportion of **heads** to the number of repetitions, $\frac{n_{occurrence}}{n_{repetition}}$, approaches 0.5. Thus we conclude that the probability of a **head** occurring, $P(A)$, is 0.5. From Equation (B.2) we get:

$$P(A) = \lim_{n_{repetition} \to \infty} \frac{n_{occurrence}}{n_{repetition}} = 0.5$$

Basic Probability - Data Science



3. Probability Axioms

The probability $P(A)$ of an event A to occur, as a measure of uncertainty, takes a real number value from the range $[0, 1]$.

The **system of probabilities** assigns probabilities $P(A)$ to events A based on three **probability theory axioms**.

Axiom 1. The probability is greater or equal to zero and less or equal to 1

$$0 \leq P(A) \leq 1 \quad for \quad each\ event\ A \tag{B.3}$$

The probability cannot be negative.

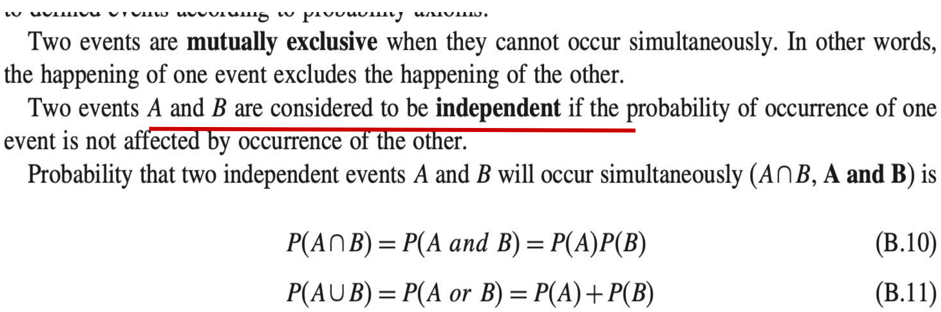
Axiom 2. An event $A = S$, which includes all outcomes from a sample space S, must have a probability equal to 1. An event $A = S$ is called a **certain event**. When $P(A) = 0$, an event cannot occur. $P(A) = 1$ means that an event will always occur.

Axiom 3. Let events $A, B, C, \cdots$ be mutually exclusive (disjoint). The probability of the event that "**A or B or C or** $\cdots$" will happen is

$$\begin{aligned} P(A\ or\ B\ or\ C\ or\ \cdots) &= P(A \cup B \cup C \cup \cdots) \\ &= P(A) + P(B) + P(C) + \cdots \end{aligned} \tag{B.4}$$

The third axiom states that for non-overlapping events (disjoint) the probability of their union is equal to the sum of the probabilities of the individual events.

Basic Probability - Data Science



to defined events according to probability axioms.

Two events are **mutually exclusive** when they cannot occur simultaneously. In other words, the happening of one event excludes the happening of the other.

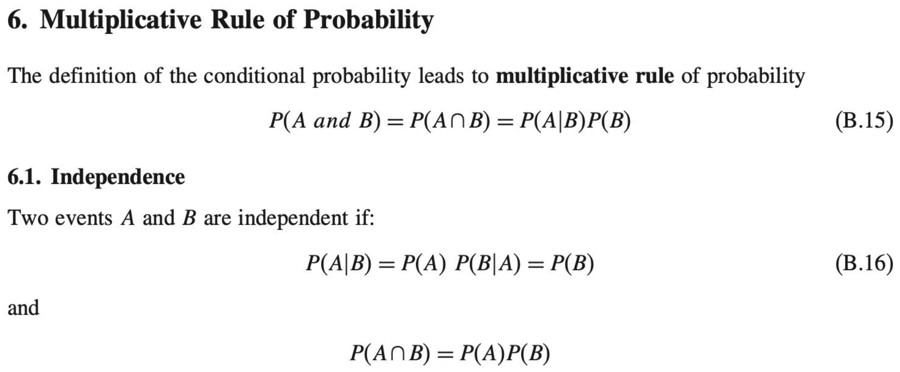
Two events A and B are considered to be **independent** if the probability of occurrence of one event is not affected by occurrence of the other.

Probability that two independent events A and B will occur simultaneously ($A \cap B$, **A and B**) is

$$P(A \cap B) = P(A \ and \ B) = P(A)P(B) \tag{B.10}$$

$$P(A \cup B) = P(A \ or \ B) = P(A) + P(B) \tag{B.11}$$

Basic Probability - Data Science



6. Multiplicative Rule of Probability

The definition of the conditional probability leads to **multiplicative rule** of probability

$$P(A\ and\ B) = P(A \cap B) = P(A|B)P(B) \tag{B.15}$$

6.1. Independence

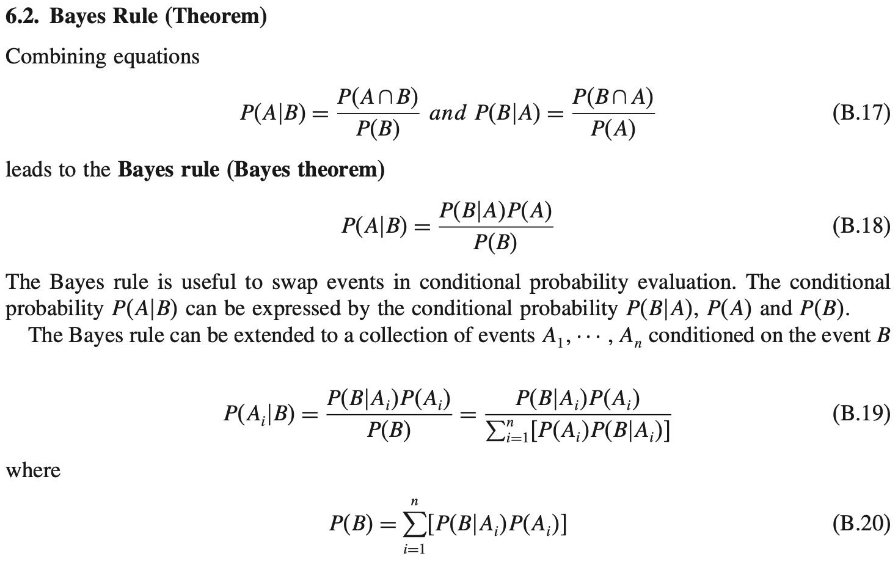
Two events A and B are independent if:

$$P(A|B) = P(A) \;\; P(B|A) = P(B) \tag{B.16}$$

and

$$P(A \cap B) = P(A)P(B)$$

Basic Probability - Data Science



6.2. Bayes Rule (Theorem)

Combining equations

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \; and \; P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{B.17}$$

leads to the **Bayes rule (Bayes theorem)**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{B.18}$$

The Bayes rule is useful to swap events in conditional probability evaluation. The conditional probability $P(A|B)$ can be expressed by the conditional probability $P(B|A)$, $P(A)$ and $P(B)$.

The Bayes rule can be extended to a collection of events $A_1, \cdots, A_n$ conditioned on the event B

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n}[P(A_i)P(B|A_i)]} \tag{B.19}$$

where

$$P(B) = \sum_{i=1}^{n}[P(B|A_i)P(A_i)] \tag{B.20}$$

Probability distributions

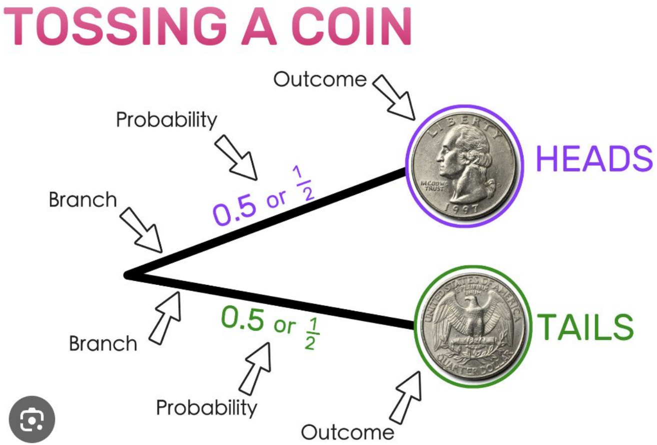
Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data (**variable**) and how often they occur.

A very important Math/Stat/DS concept

Probability distributions

Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data and how often they occur.

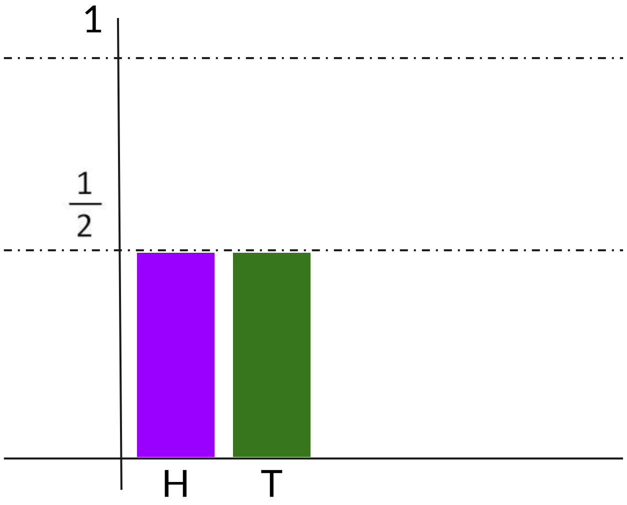
A very important Math/Stat/DS concept



Probability distributions

Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data and how often they occur.

A very important Math/Stat/DS concept



Probability distributions

Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data and how often they occur.

- Uniform distribution (discrete and continuous)
- Binomial distribution (discrete, binary)
- Multinomial distribution (discrete, general)
- Normal/Gaussian distribution (continuous)

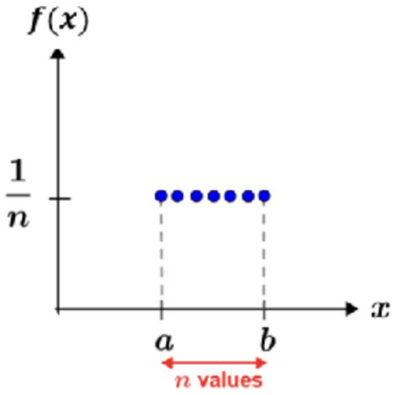
A very important Math/Stat/DS concept

Discrete Uniform Distribution

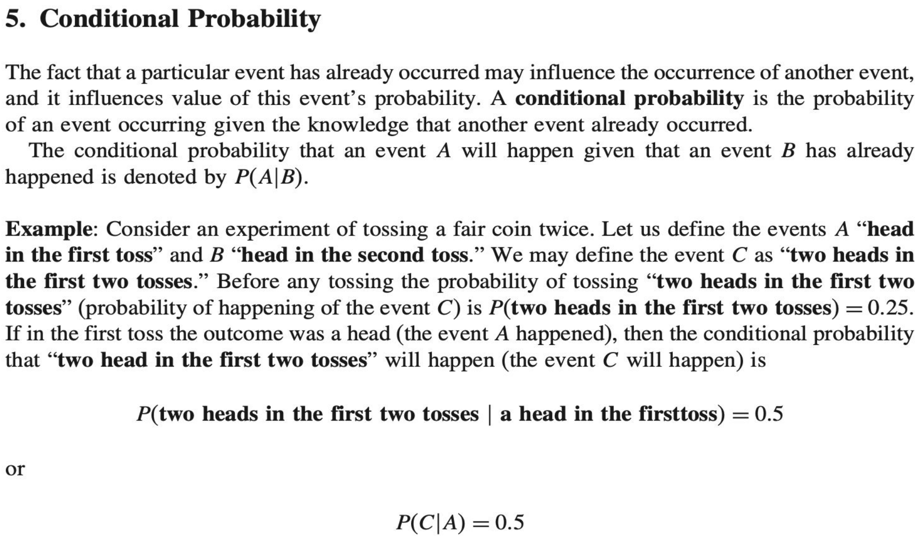
- Uniform/Equal probability
- ***X*** is a random variable
- ***n*** is the number of different choices ***X*** has

$$f(x) = \frac{1}{n}\ ,\ \mathrm{x} = 1, 2, 3, ..., \mathrm{n}$$

Tossing a (**fair**) coin, throwing a (**fair**) dice



Basic Probability - Data Science



5. Conditional Probability

The fact that a particular event has already occurred may influence the occurrence of another event, and it influences value of this event's probability. A **conditional probability** is the probability of an event occurring given the knowledge that another event already occurred.

The conditional probability that an event A will happen given that an event B has already happened is denoted by $P(A|B)$.

Example: Consider an experiment of tossing a fair coin twice. Let us define the events A "**head in the first toss**" and B "**head in the second toss**." We may define the event C as "**two heads in the first two tosses**." Before any tossing the probability of tossing "**two heads in the first two tosses**" (probability of happening of the event C) is $P(\textbf{two heads in the first two tosses}) = 0.25$. If in the first toss the outcome was a head (the event A happened), then the conditional probability that "**two head in the first two tosses**" will happen (the event C will happen) is

$$P(\textbf{two heads in the first two tosses} \mid \textbf{a head in the firsttoss}) = 0.5$$

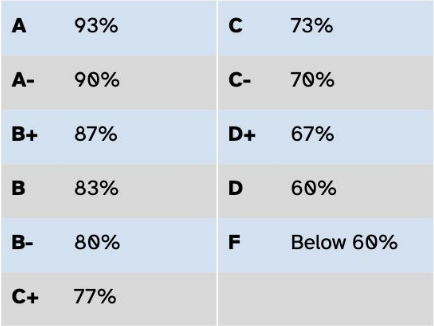
or

$$P(C|A) = 0.5$$

Continuous Non-Uniform Distribution

- Non-Uniform/Un-equal probability

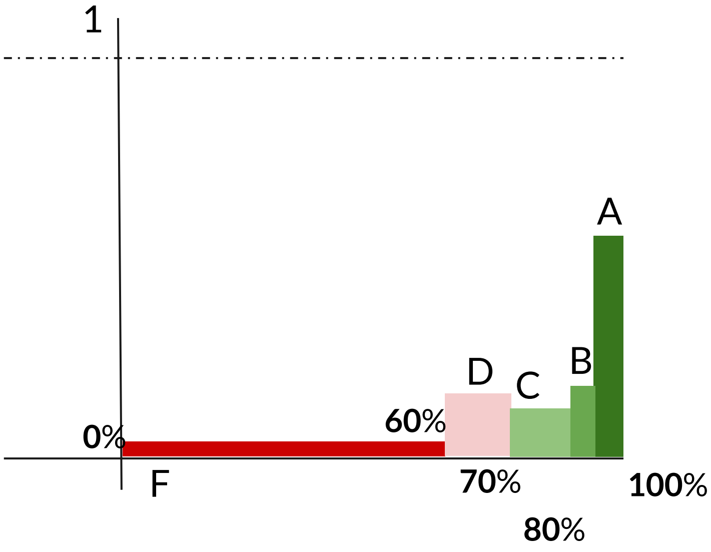
Student grades, blood pressure levels, etc



A	93%	**C**	73%
A-	90%	**C-**	70%
B+	87%	**D+**	67%
B	83%	**D**	60%
B-	80%	**F**	Below 60%
C+	77%		

Probability distributions

Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data and how often they occur.



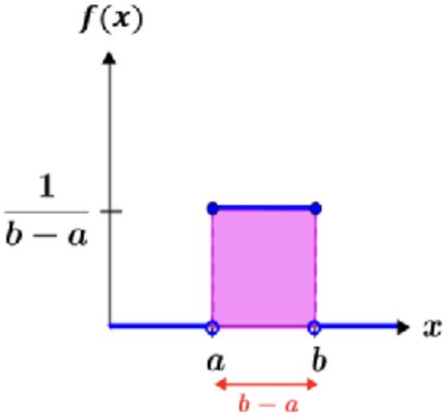
A very important Math/Stat/DS concept

Continuous Uniform Distribution

- Uniform/Equal probability

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases}$$

Throwing a (**fair**) dirt

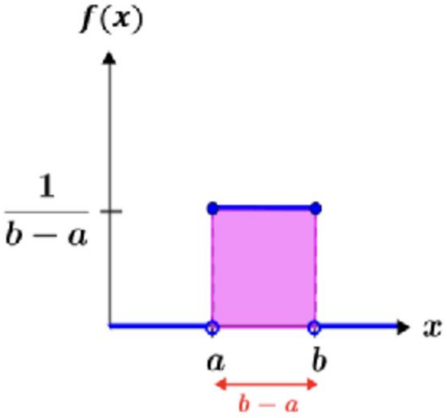


Continuous Uniform Distribution

- Uniform/Equal probability

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & otherwise \end{cases}$$

Throwing a (**fair**) dirt



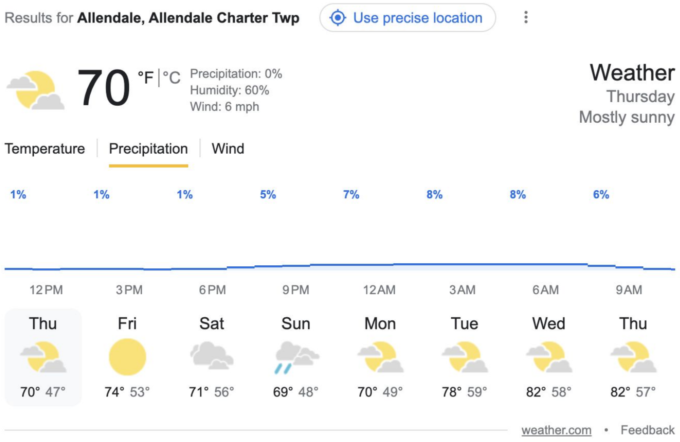
Probability distributions

Distribution: Generally, **a function** showing **all possible values (or intervals)** of the data (**variable**) and how often they occur.

A very important Math/Stat/DS concept

Binomial Distribution

- Binary variables (two outcomes) such as tossing a fair coin follows a **uniform distribution**; however,
- Not all variables follow the same: such as answering a yes/no questions
  - **Is it going to rain tomorrow?**
  - Am I going to get a ‘A’ grade?



Results for **Allendale, Allendale Charter Twp**

70 °F | °C

Precipitation: 0%
Humidity: 60%
Wind: 6 mph

Weather
Thursday
Mostly sunny

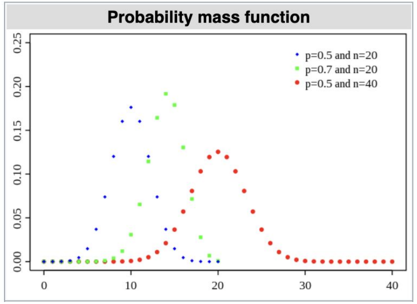
12 PM	3 PM	6 PM	9 PM	12 AM	3 AM	6 AM	9 AM
1%	1%	1%	5%	7%	8%	8%	6%

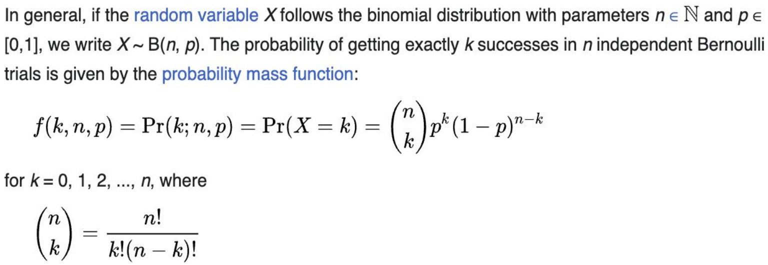
Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
70° 47°	74° 53°	71° 56°	69° 48°	70° 49°	78° 59°	82° 58°	82° 57°

Binomial Distribution

Doing a repeated independent trials/experiments

- If you throw a coin n time, what will be the probability of getting 1 - n (denoted as *k*) heads



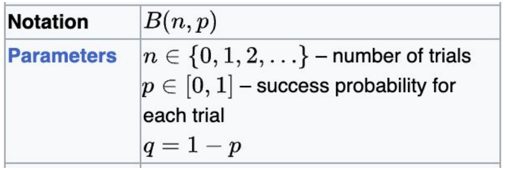


In general, if the random variable X follows the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, we write $X \sim B(n, p)$. The probability of getting exactly k successes in n independent Bernoulli trials is given by the probability mass function:

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, ..., n$, where

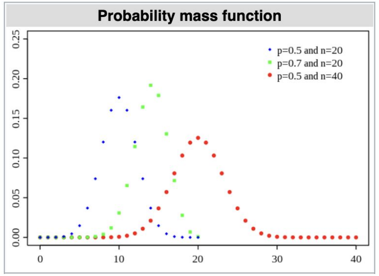
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$



Notation	$B(n,p)$
Parameters	$n \in \{0,1,2,\ldots\}$ – number of trials $p \in [0,1]$ – success probability for each trial $q = 1 - p$

Binomial Distribution

- Binary variable (two outcomes): tossing a coin, answering a yes/no question, etc.
- Doing a repeated independent trials/experiments
- If you throw a coin n time, what will be the probability of getting 1 - n (denoted as *k*) heads



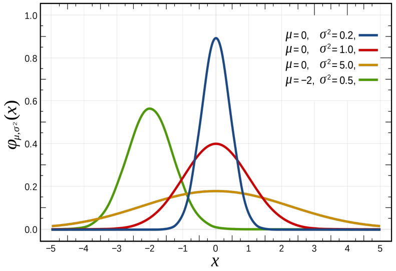
Probability mass function: The probabilities associated with all (hypothetical) values must be non-negative and sum up to 1,

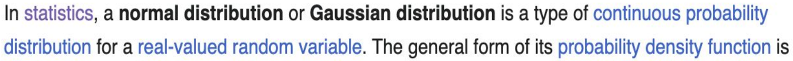
$$\boxed{\sum_x p_X(x) = 1}$$

$$\boxed{p_X(x) \geq 0.}$$

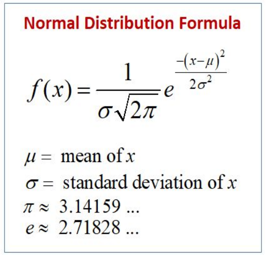
Normal (Gaussian) Distribution

- Continuous (real/measurement) variable (infinite outcomes): height or weight of students, salary of a group of professional with similar background etc.
-





In statistics, a **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is



Normal Distribution Formula

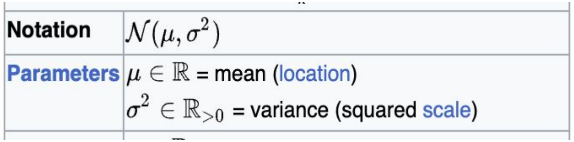
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x

$\sigma =$ standard deviation of x

$\pi \approx$ 3.14159 ...

$e \approx$ 2.71828 ...



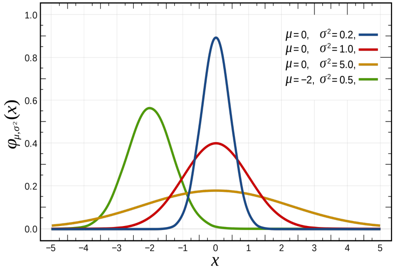
Notation	$\mathcal{N}(\mu,\sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 \in \mathbb{R}_{>0}$ = variance (squared scale)

Normal (Gaussian) Distribution

- Continuous (real/measurement) variable (infinite outcomes): your blood pressure, expected salary.
-

Carl Friedrich Gauss

Portrait by Christian Albrecht Jensen, 1840 (copy from Gottlieb Biermann, 1887)[1]



Probability density function: The probabilities associated with all (hypothetical) values must be non-negative and sum up to 1,

$$\boxed{\int_{-\infty}^{\infty} f(x)\ dx = 1}$$

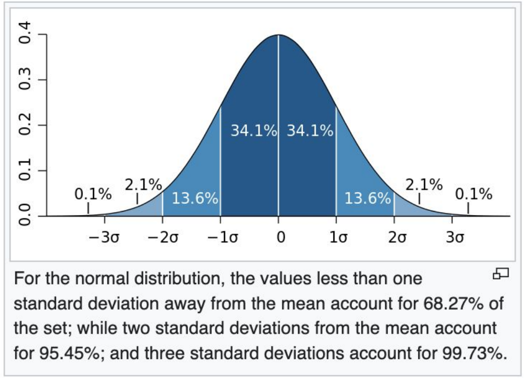
$$\boxed{P[a \le X \le b] = \int_a^b f(x)\ dx}$$

Basic Statistics - Data Science

Distribution: Generally, a function showing all possible values (or intervals) of the data and how often they occur.

(Normal distribution as an example): Normal distribution, also known as the Gaussian distribution, is a probability distribution that is **symmetric about the mean**, showing that **data near the mean are more frequent** in occurrence than data far from the mean. In graphical form, the normal distribution appears as a **"bell curve"**.

Probability mass



For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%.

ref link

Normal distribution

Distribution and normal distribution

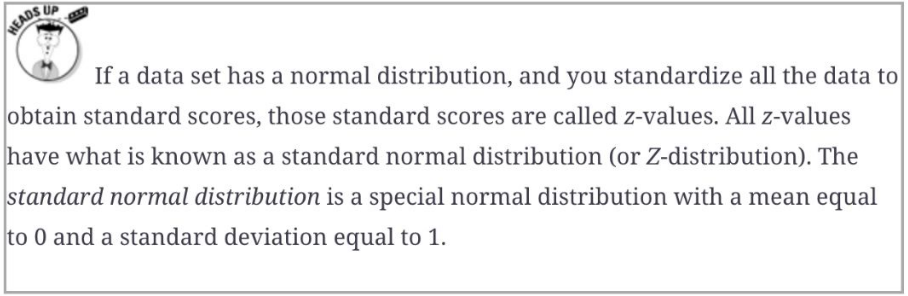
The *distribution* of a data set (or a population) is a listing or function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individuals in each group. When a distribution of numerical data is organized, they're often ordered from smallest to largest, broken into reasonably sized groups (if appropriate), and then put into graphs and charts to examine the shape, center, and amount of variability in the data.

The world of statistics includes dozens of different distributions for categorical and numerical data; the most common ones have their own names. One of the most well-known distributions is called the *normal distribution,* also known as the *bell-shaped curve.* The normal distribution is based on numerical data that is continuous; its possible values lie on the entire real number line. Its overall shape, when the data are organized in graph form, is a symmetric bell-shape. In other words, most (around 68%) of the data are centered around the mean (giving you the middle part of the bell), and as you move farther out on either side of the mean, you find fewer and fewer values (representing the downward sloping sides on either side of the bell).

Basic Statistics - Data Science

Standard-score/Z-score/Z-distribution: Z-score is a statistical 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.

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



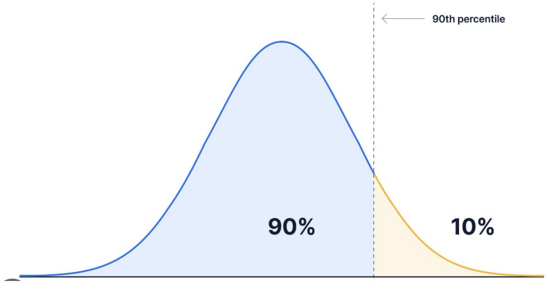
If a data set has a normal distribution, and you standardize all the data to obtain standard scores, those standard scores are called *z*-values. All *z*-values have what is known as a standard normal distribution (or *Z*-distribution). The *standard normal distribution* is a special normal distribution with a mean equal to 0 and a standard deviation equal to 1.

ref link

Basic Statistics - Data Science

Percentile: The **percentile** reported for a given score is the percentage of values in the data set that fall below that certain score. For example,

- If **your score** was reported to be at the **90th percentile**, that means 90% of the other people who took the test scored lower than you did.


90th percentile