A Quick Review of Basic Probability and Statistics for Data Scientists

Some universal notations used

b	binomial
μ	expected value [parameter]
n	number of trials [parameter]
N	normal
p	probability of success [parameter]
pdf	probability density function
pmf	probability mass function
RV	random variable
σ	standard deviation [parameter]
x	value for random variable X (e.g., observed number of successes for a binomial random
X	random variable X

Contents

1. Probability: The Basics

What is probability?

Probability is the mathematical language to understand uncertainty. We need to make decisions in the presence of uncertainty which is ever present.

Example 1.1 : The Earth is warming- a phenomenon that is known as Global Warming (GW). Is modern human activity the cause of GW? Many physical, economic, social phenomena are characterized by uncertainty.

What is a sample space and event?

Sample space: This is a set of all possible outcome values of an experiment. Notation Ω. So if we consider a coin flip then sample space would be {head, tail}. If one unbiased die is thrown then sample space would be Ω = {1, 2, 3, 4, 5, 6} and the sample outcome ω ϵ Ω.

Event: It is a subset of sample space $E \subseteq \Omega$. For a given event "Getting even numbers after throwing unbiased die" the subset is E = {2, 4, 6}. So every time we run experiment either the event will occur or it won't.

Why we need it: Both sample space and event helps us to determine the probability of event. Probability is nothing but ratio of number of elements in event space to number of elements of sample space.

What is the probability of an event?

The **probability of an event** is its relative frequency (expected proportion) in the long run. If an event occurs x times out of n, then its probability will *converge* on $x \div n$ as n becomes infinitely large. For example, if we flip a coin many, many times, we expect to see half the flips turn up heads, but only *in the long run*.

When n is small, the observed relative frequency (proportion) of an event is not be a reliable reflection of its probability. However, as the number of observations n increases, the observed frequency becomes a more reliable reflection of the probability.

Example 1.2: If a coin is flipped 10 times, there is no guarantee that it will turn up heads 50% of the time. (In fact, most of the time it will not show "5 of 10" heads.) However, if the coin is flipped 100,000 times, chances are pretty good that the proportion of "heads" will be pretty close to 50%.

What are the rules for working with probabilities?

Let S be the space of all possible elementary outcomes of an experiment. Then the probability P is function:

P: S → [0,1]

that satisfy the following properties (axioms):

1. $0 \le P(E) \le 1$ for each event $E \in S$
2. $P(S) = 1$
3. For each sequence $E_1, E_2, E_3, ...$ of mutually disjoint events we have

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

What are the Consequences of Probability Rules (Axioms)?

1. $P(\varnothing) = 0$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. $P(A \cup B) = P(A) + P(B)$ if $P(A \cap B) = 0$
4. If $A \subset B$ then $P(A) \le P(B)$
5. $P(A^c)$ or $P(\bar{A}) = 1 - P(A)$

where A^c or $\bar{A}$ is the complement of A.

Example 1.3: Two coin tosses. Let H1 be the event that a heads occurs on toss 1 and H2 a heads on toss 2. All events are equally likely. Sample space = {HH, HT, TH, TT}

– H1 = {HH, HT}
– H2 = {HH, TH}
– P(H1 U H2) = ½ + ½ - ¼ = ¾
–

What are independent events?

Two events A and B are **independent** if *P*(A ∩ B) = *P*(A)*P*(B). *P*(A∩B) is also written as *P*(AB) and *P*(A,B).

If A and B are disjoint events and *P*(A) > 0 and *P*(B) > 0 then A and B cannot be independent i.e. *P*(A ∩ B) = 0. Yet P(A)P(B) > 0

–

What is the conditional probability?

Very Important Concept

P(A|B) denotes the "fraction of occurrences of B in which A also occurs" and it is defined as

P(A|B) = *P*(A ∩ B)/*P*(B); *P*(B) > 0

and indicates the likelihood that A occurs given knowledge that B has occurred.

Example 1.4: Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC ? i.e., we want to find P(CD) | AC).

Show that $P(CD \mid AC) = \dfrac{P(CD \cap AC)}{P(AC)} = \dfrac{0.2}{0.7} = .2857$

Note that for a fixed B, *P*(.|B) is a probability. Therefore if A1 and A2 are disjoint then

P(A1 U A2 |B) = *P*(A1|B) + *P*(A2|B)

and *P*(A|B U C) ≠ *P*(A|B) + *P*(A|C) and *P*(A|B) ≠ *P*(B|A).

Conditional probability is fundamental to statistical modelling. In particular in modelling statistical processes, a causal connection between B and A means:

$$P(A\,|\,B) \geq P(A)$$

What is the conditional independence?

Two events A and B are independent of one another if

$$P(A\,|\,B) \geq P(A)$$

i.e. $P(A \cap B) = P(A)P(B)$. Knowledge of B's occurrence has no effect on the likelihood that A will occur.

What is the total probability?

Let A_i 's be a partition of the S (space of outcomes), i.e. $S = \bigcup_{i=1}^{n} A_i$ and $A_i \bigcap A_j = \varnothing$ for $i \neq j$

then $P(B) = \sum_{i=1}^{n} P(B\,|\,A_i)P(A_i)$.

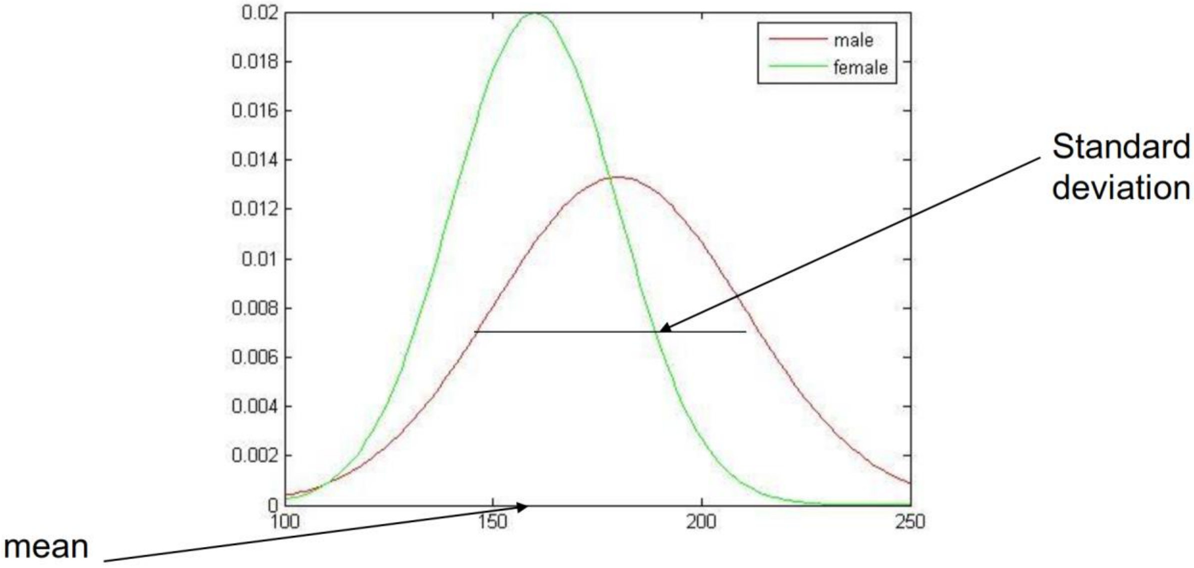
What the Bayes Theorem says?

Let A_i's be partition of S then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\Sigma_j P(B|A_j)P(A_j)}$$

2. Random Variables (RV)

We were told that the "height" of human beings is normally distributed!



Why distributions are important?

- Distribution capture the essence of data associated with a particular variable(s) (e.g., height).
- If we know height is **Normally** distributed, then a small random sample is enough to provide a very good idea about the general population.
- Can answer questions like: what is the **probability** of finding a 2-meter-tall Greek?
- **Need to understand the concept of random variable**.

What is the definition of a random variable?

Suppose that to each point of a sample space we assign a number. We then have a *function* defined on the sample space. This function is called a *random variable* (or *stochastic variable*) or more precisely a *random function* (*stochastic function*). It is usually denoted by a capital letter such as X or Y. In general, a random variable has some specified physical, geometrical, or other significance.

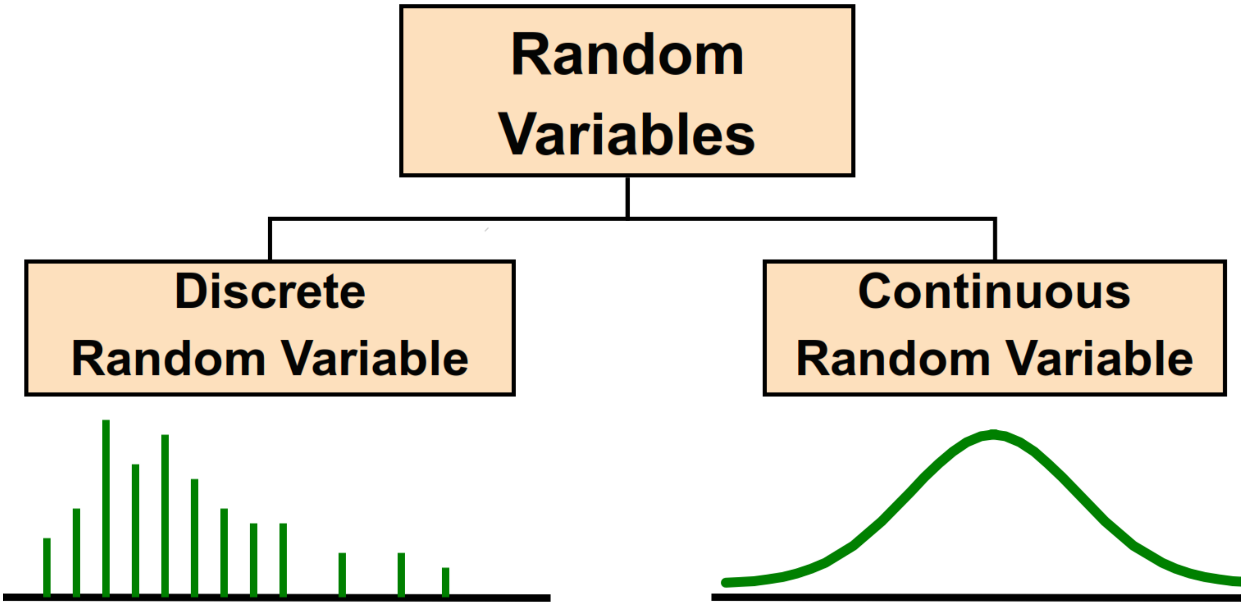
EXAMPLE 2.1 Suppose that a coin is tossed twice so that the sample space is $S = \{HH, HT, TH, TT\}$. Let X represent the number of heads that can come up. With each sample point we can associate a number for X as shown in Table 2-1. Thus, for example, in the case of HH (i.e., 2 heads), $X = 2$ while for TH (1 head), $X = 1$. It follows that X is a random variable.

Table 2-1

Sample	HH	HT	TH	TT
X	2	1	1	0

It should be noted that many other random variables could also be defined on this sample space, for example, the square of the number of heads or the number of heads minus the number of tails.

A random variable that takes on a finite or countably infinite number of values is called a *discrete random variable* while one which takes on a noncountably infinite number of values is called a *nondiscrete or continuous random variable*.

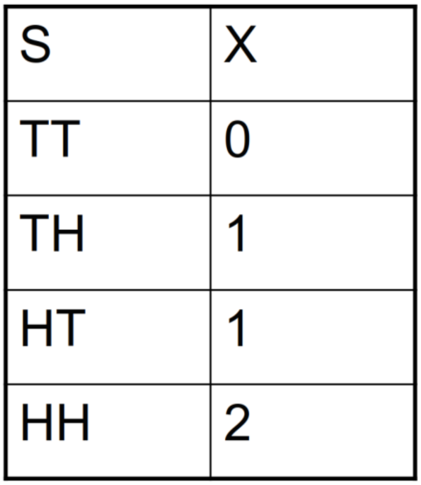


Is the random variable a function?

Let S be the sample space. A random variable X is a function

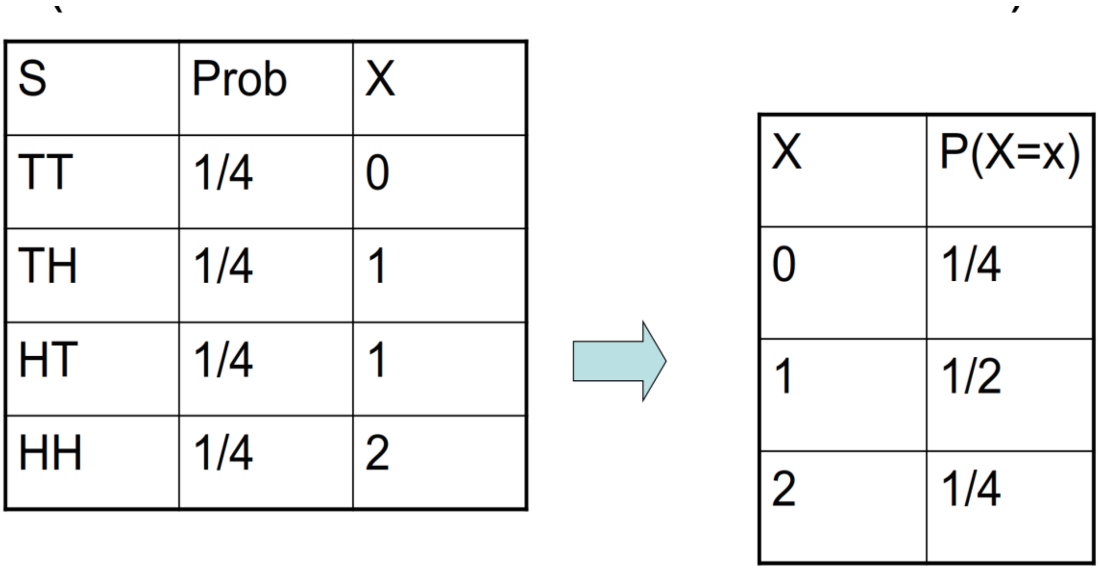
$$X: S \rightarrow \text{Real}$$

Suppose we toss a coin twice. Let X be the random variable **number of heads.**



S	X
TT	0
TH	1
HT	1
HH	2

We also associate a probability with X attaining that value.



S	Prob	X
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

X	P(X=x)
0	1/4
1	1/2
2	1/4

$$P(X = x_i) \geq 0 \qquad \textstyle\sum_{i=1}^{3} P(X = x_i) = 1$$

Random Variables follow a Distribution?

Examples

- The height of Australian soldiers is a random variable which follows a Normal distribution with mean 180 cm and standard deviation 15 cm.
- The frequency of words in a text is a random variable which follows a Zipf distribution.
- The speed of a hurricane is a random variable which follows a Cauchy distribution.

- The number of car accidents in a fixed time duration is a random variable which follows a Poisson distribution.
- The number of heads in a sequence of coin tosses is a random variable which follows a Binomial distribution.
- The number of web hits in a given time period is a r.v. which follows a Pareto distribution.
- **Many times we don't know what named distribution a r.v. follows or whether it follows any named distribution at all!**

Remember that in broad mathematical terms, there are two types of random variables: **discrete random variables** and **continuous random variables**.

What is a discrete random variable and its discrete distribution?

Discrete random variables form a countable set of outcomes. We will study binomial random variables as an example of a type of discrete random variable. Let X be a discrete r.v. then it takes countably many values {x1, x2, …}. The probability function or probability mass function or probability distribution for X is given by

$$f_X(x) = \begin{cases} P(X = x_i) & for\ i = 1, 2, ... \\ 0 & otherwise \end{cases}$$

It is convenient to introduce the *probability distribution* given

$$P(X = x) = f_X(x)$$

For $x = x_k$, this reduces to the above equation while for other values of x, $f_X(x)$=0.

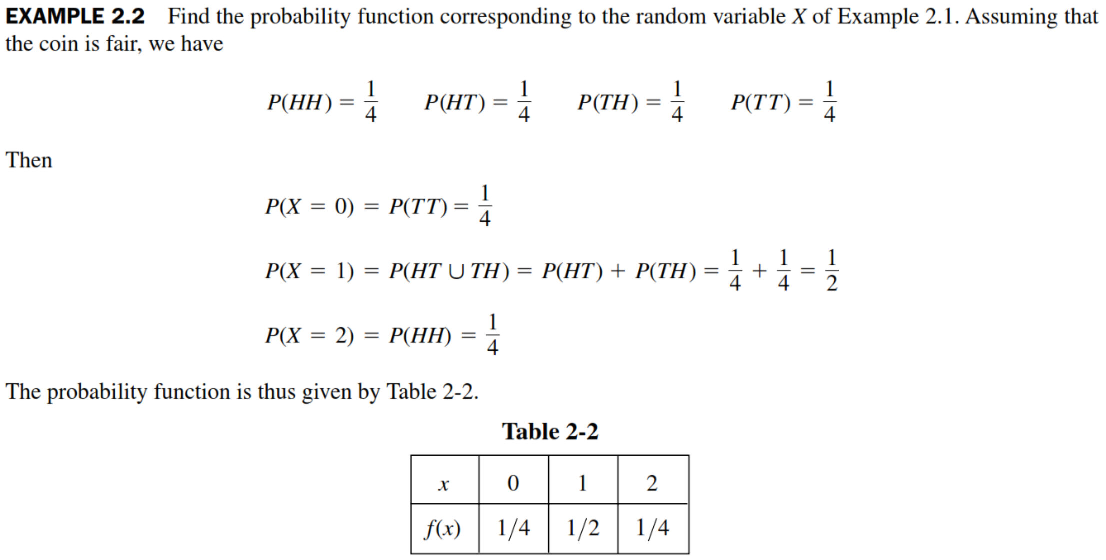
In general, $f_X(x)$ is a discrete distribution function if

1. $f_X(x) \geq 0$
2. $\sum_x f_X(x) = 1$

Where the sum in 2 is taken over all possible values of *x*.

From previous example, we derive the discrete distribution

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & otherwise \end{cases}$$



EXAMPLE 2.2 Find the probability function corresponding to the random variable X of Example 2.1. Assuming that the coin is fair, we have

$$P(HH) = \frac{1}{4} \qquad P(HT) = \frac{1}{4} \qquad P(TH) = \frac{1}{4} \qquad P(TT) = \frac{1}{4}$$

Then

$$P(X = 0) = P(TT) = \frac{1}{4}$$

$$P(X = 1) = P(HT \cup TH) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$P(X = 2) = P(HH) = \frac{1}{4}$$

The probability function is thus given by Table 2-2.

Table 2-2

x	0	1	2
$f(x)$	1/4	1/2	1/4

Distribution Functions for Random Variables

The *cumulative distribution function*, or briefly the *distribution function*, for a random variable X is defined by

$$F(x) = P(X \le x)$$

where x is any real number, i.e., $-\infty < x < \infty$.

The distribution function $F(x)$ has the following properties:

1. $F(x)$ is nondecreasing [i.e., $F(x) \le F(y)$ if $x \le y$]
2. $lim_{x->-\infty} F(x) = 0, lim_{x->\infty} F(x) = 1$
3. F(x) is continuous from the right [i.e. $lim_{h->+0} F(x+h) = F(x)$ for all x]

Distribution Functions for Discrete Random Variables

The distribution function for a discrete random variable X can be obtained from its probability function by noting that, for all x in $(-\infty, \infty)$

$$F(x) = P(X \le x) \quad = \quad \sum_{u \le x} f(u)$$

where the sum is taken over all values u taken on by X for which $u \le x$.

If X takes on only a finite number of values $x_1, x_2, \ldots, x_n$, then the distribution function is given by

$$F(x) = \begin{cases} 0 & -\infty < x < x_1 \\ f(x_1) & x_1 \leq x < x_2 \\ f(x_1) + f(x_2) & x_2 \leq x < x_3 \\ \vdots & \vdots \\ f(x_1) + \cdots + f(x_n) & x_n \leq x < \infty \end{cases}$$

EXAMPLE 2.3 (a) Find the distribution function for the random variable X of Example 2.2. (b) Obtain its graph.

(a) The distribution function is

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & 2 \leq x < \infty \end{cases}$$

(b) The graph of $F(x)$ is shown in Fig. 2-1.

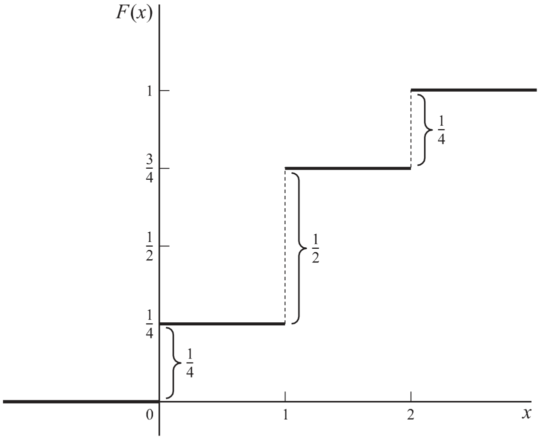


Fig. 2-1

The following things about the above distribution function, which are true in general, should be noted.

1. The magnitudes of the jumps at 0, 1, 2 are which are precisely the probabilities in Table 2-2. This fact enables one to obtain the probability function from the distribution function.

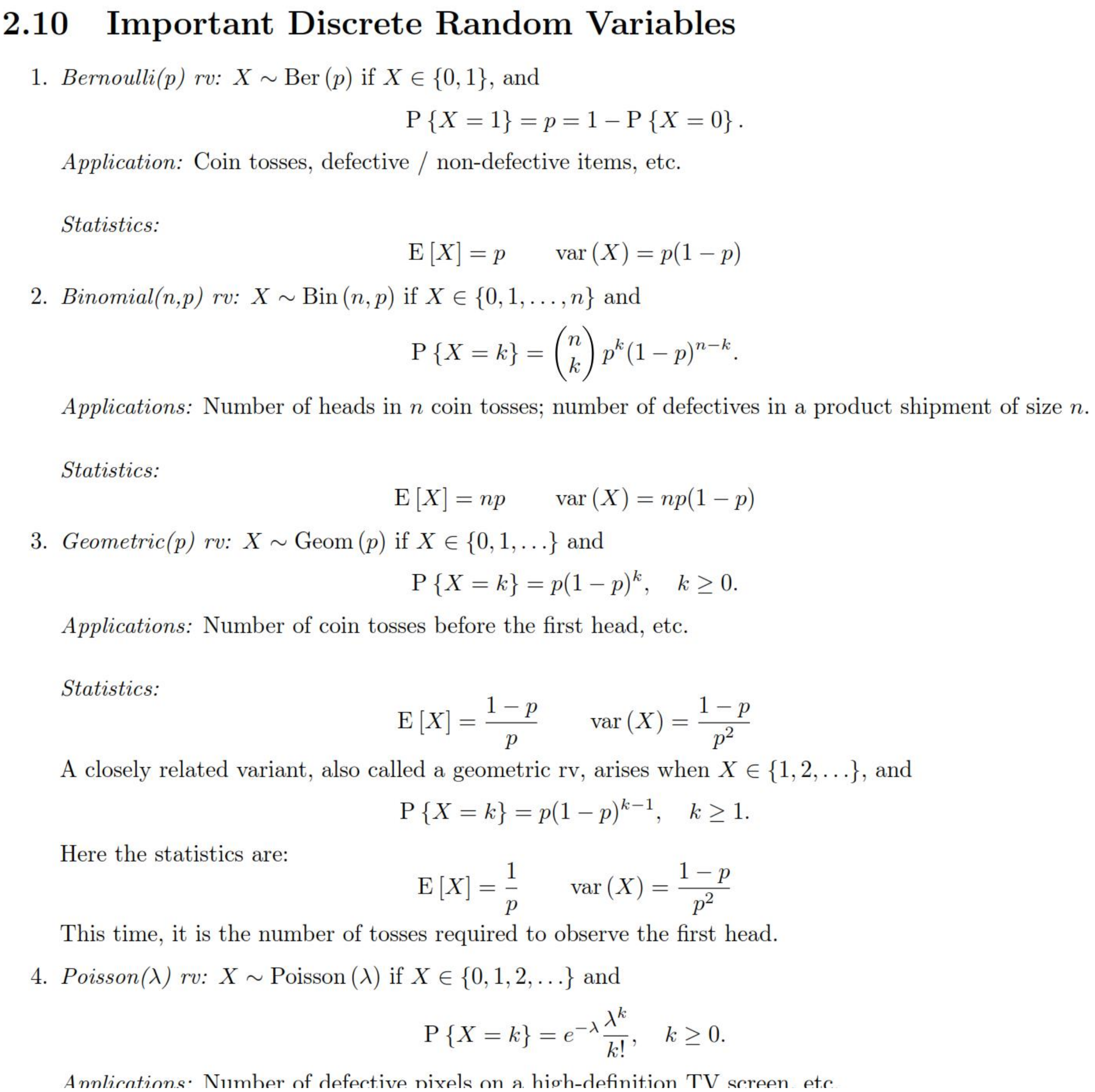
2. Because of the appearance of the graph of Fig. 2-1, it is often called a *staircase function* or *step function*. The value of the function at an integer is obtained from the higher step; thus the value at 1 is ¾ and not 1/4 . This is expressed mathematically by stating that the distribution function is *continuous from the right* at 0, 1, 2.

3. As we proceed from left to right (i.e. going *upstairs*), the distribution function either remains the same or increases, taking on values from 0 to 1. Because of this, it is said to be a *monotonically increasing function*.

It is clear from the above remarks and the properties of distribution functions that the probability function of a discrete random variable can be obtained from the distribution function by noting that

$$f(x) = F(x) - \lim_{u \to x^-} F(u)$$

Important Discrete Random Variables



2.10 Important Discrete Random Variables

1. *Bernoulli(p) rv:* $X \sim \mathrm{Ber}(p)$ if $X \in \{0,1\}$, and
$$\mathrm{P}\{X=1\} = p = 1 - \mathrm{P}\{X=0\}.$$
Application: Coin tosses, defective / non-defective items, etc.

   Statistics:
$$\mathrm{E}[X] = p \qquad \mathrm{var}(X) = p(1-p)$$

2. *Binomial(n,p) rv:* $X \sim \mathrm{Bin}(n,p)$ if $X \in \{0,1,\ldots,n\}$ and
$$\mathrm{P}\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}.$$
Applications: Number of heads in n coin tosses; number of defectives in a product shipment of size n.

   Statistics:
$$\mathrm{E}[X] = np \qquad \mathrm{var}(X) = np(1-p)$$

3. *Geometric(p) rv:* $X \sim \mathrm{Geom}(p)$ if $X \in \{0,1,\ldots\}$ and
$$\mathrm{P}\{X=k\} = p(1-p)^k, \quad k \geq 0.$$
Applications: Number of coin tosses before the first head, etc.

   Statistics:
$$\mathrm{E}[X] = \frac{1-p}{p} \qquad \mathrm{var}(X) = \frac{1-p}{p^2}$$
A closely related variant, also called a geometric rv, arises when $X \in \{1,2,\ldots\}$, and
$$\mathrm{P}\{X=k\} = p(1-p)^{k-1}, \quad k \geq 1.$$
Here the statistics are:
$$\mathrm{E}[X] = \frac{1}{p} \qquad \mathrm{var}(X) = \frac{1-p}{p^2}$$
This time, it is the number of tosses required to observe the first head.

4. *Poisson(λ) rv:* $X \sim \mathrm{Poisson}(\lambda)$ if $X \in \{0,1,2,\ldots\}$ and
$$\mathrm{P}\{X=k\} = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \geq 0.$$
Applications: Number of defective pixels on a high-definition TV screen, etc.

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & otherwise \end{cases}$$

Let X denotes a **continuous random variables and** form a continuum of possible outcomes taking values in R. The **probability density function** or probability mass function or

X is a function $f_X(x)$ such that $P(X \le x) = \int\limits_{-\infty}^{x} f_X(t)dt$ and its **distribution function** is defined as $F(x) = P(X \le x)$ Furthermore, the density function satisfies the following properties

1. $f_X > 0$
2. $\int\limits_{-\infty}^{\infty} f_X(t)dt = 1$
3. $P(a \le x \le b) = \int\limits_{a}^{b} f_X(t)dt$
4. $f_X(x) = \dfrac{dF(x)}{dx}$ (density function)

We interpret $f_X(x)$ as the "likelihood" that X takes on the value x. However we need to note that $P(X = x) = \int\limits_{x}^{x} f_X(t)dt = 0$

Example of continuous density function is

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0\, otherwise \end{cases}$$

which is leads to the Uniform (0,1) distribution.

EXAMPLE 2.4 If an individual is selected at random from a large group of adult males, the probability that his height X is precisely 68 inches (i.e., 68.000 . . . inches) would be zero. However, there is a probability greater than zero than X is between 67.000 . . . inches and 68.500 . . . inches, for example.

A function $f(x)$ that satisfies the above requirements is called a *probability function* or *probability distribution* for a continuous random variable, but it is more often called a *probability density function* or simply *den- sity function*. Any function $f(x)$ satisfying Properties 1 and 2 above will automatically be a density function, and required probabilities can then be obtained from (8).

EXAMPLE 2.5 Find the constant c such that the function

$$f(x) = \begin{cases} cx^2 & 0 < x < 3 \\ 0 & otherwise \end{cases}$$

is a density function, and (b) compute $P(1 < X < 2)$.

Solution

(a) Since $f(x)$ satisfies Property 1 if $c \geq 0$, it must satisfy Property 2 in order to be a density function. Now

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^3 cx^2\,dx = \left.\frac{cx^3}{3}\right|_0^3 = 9c$$

and since this must equal 1, we have $c = 1/9$.

(b) $$P(1 < X < 2) = \int_1^2 \frac{1}{9}x^2\,dx = \left.\frac{x^3}{27}\right|_1^2 = \frac{8}{27} - \frac{1}{27} = \frac{7}{27}$$

In case $f(x)$ is continuous, which we shall assume unless otherwise stated, the probability that X is equal to any particular value is zero. In such case we can replace either or both of the signs $<$ in (8) by $\leq$. Thus, in Example 2.5,

$$P(1 \leq X \leq 2) = P(1 \leq X < 2) = P(1 < X \leq 2) = P(1 < X < 2) = \frac{7}{27}$$

EXAMPLE 2.6 (a) Find the distribution function for the random variable of Example 2.5. (b) Use the result of (a) to find $P(1 < x \leq 2)$.

(a) We have

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,du$$

If $x < 0$, then $F(x) = 0$. If $0 \leq x < 3$, then

$$F(x) = \int_0^x f(u)\,du = \int_0^x \frac{1}{9}u^2\,du = \frac{x^3}{27}$$

If $x \geq 3$, then

$$F(x) = \int_0^3 f(u)\,du + \int_3^x f(u)\,du = \int_0^3 \frac{1}{9}u^2\,du + \int_3^x 0\,du = 1$$

Thus the required distribution function is

$$F(x) = \begin{cases} 0 & x < 0 \\ x^3/27 & 0 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Note that $F(x)$ increases monotonically from 0 to 1 as is required for a distribution function. It should also be noted that $F(x)$ in this case is continuous.

(b) We have

(b) We have

(b) We have

$$\begin{aligned} P(1 < X \leq 2) &= P(X \leq 2) - P(X \leq 1) \\ &= F(2) - F(1) \\ &= \frac{2^3}{27} - \frac{1^3}{27} = \frac{7}{27} \end{aligned}$$

as in Example 2.5.

The probability that X is between x and $x + \Delta x$ is given by

$$P(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f(u)\,du \tag{9}$$

so that if Δx is small, we have approximately

$$P(x \leq X \leq x + \Delta x) \;=\; f(x)\Delta x \tag{10}$$

We also see from (7) on differentiating both sides that

$$\frac{dF(x)}{dx} = f(x) \tag{11}$$

at all points where $f(x)$ is continuous; i.e., the derivative of the distribution function is the density function.

It should be pointed out that random variables exist that are neither discrete nor continuous. It can be shown that the random variable X with the following distribution function is an example.

$$F(x) = \begin{cases} 0 & x < 1 \\ \dfrac{x}{2} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

In order to obtain (11), we used the basic property

$$\frac{d}{dx}\int_a^x f(u)\,du = f(x) \tag{12}$$

which is one version of the Fundamental Theorem of Calculus.

Graphical Interpretations

If f (x) is the density function for a random variable X, then we can represent y = f (x) graphically by a curve as in Fig. 2-2. Since f (x) ≥ 0, the curve cannot fall below the x axis. The entire area bounded by the curve and the x axis must be 1 because of Property 2. Geometrically the probability that X is between a and b, i.e., P(a ≤ X ≤ b), is then represented by the area shown shaded, in Fig. 2-2. The distribution function F(x) = P(X ≤ x) is a monotonically increasing function which increases from 0 to 1 and is represented by a curve as in Fig. 2-3.

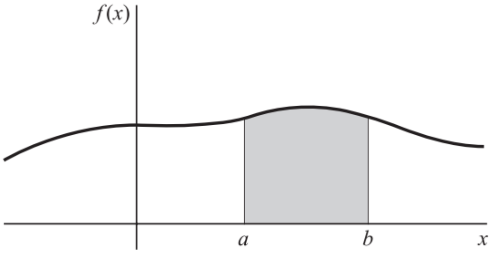


Fig. 2-2

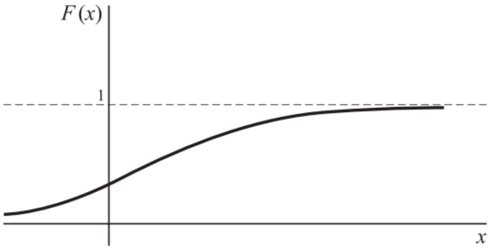


Fig. 2-3

Problem. Find a such that

Problem. Find a such that

$$f(x) = \frac{a}{1+x^2}$$

is a density function (this is the density of *Cauchy* distribution). Find the distribution function.

Solution. Check #1: $f(x) > 0$ for all x. To find a we have to calculate

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx \\ = & \left(Arc\tan(x)|_{-\infty}^{\infty} \right) \\ = & \left(\frac{\pi}{2} - (-\frac{\pi}{2}) \right) \\ = & \left(\frac{\pi}{2} + \frac{\pi}{2} \right) \\ = & \ \pi . \end{aligned}$$

That means

$$a = \frac{1}{\pi}.$$

The distribution function is

$$\begin{aligned} F(x) &= \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+t^2} dt \\ &= \frac{1}{\pi} \left(Arc\tan(x) + \frac{\pi}{2} \right) \\ &= \frac{1}{\pi} Arc\tan(x) + \frac{1}{2}. \end{aligned}$$

Mode is where the density attains its maximum. We call density $f(x)$ *unimodal* if it has one mode, i.e. $f(x)$ is increasing at left and decreasing at right of the mode. Otherwise, we call the density *multimodal.* Let m denote the mode, we call X *symmetric* if

$$\Pr(X - m \leq -a) = \Pr(X - m \geq a).$$

For a symmetric distribution: $f(m-x) = f(x+m)$ and $F(m-x) = 1 - F(x+m)$. Illustrate geometrically.

Prove, that the Cauchy distribution is symmetric, and therefore unimodal with mode=0.

Exponential distribution is not symmetric.

Problem. Find a such that

$$f(x) = \frac{a}{1+x^2}$$

is a density function (this is the density of *Cauchy* distribution). Find the distribution function.

Solution. Check #1: $f(x) > 0$ for all x. To find a we have to calculate

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{1+x^2} dx \\
= &\left(Arc\tan(x)|_{-\infty}^{\infty} \right) \\
= &\left(\frac{\pi}{2} - (-\frac{\pi}{2}) \right) \\
= &\left(\frac{\pi}{2} + \frac{\pi}{2} \right) \\
= &\pi.
\end{aligned}$$

That means

$$a = \frac{1}{\pi}.$$

The distribution function is

$$\begin{aligned}
F(x) &= \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+t^2} dt \\
&= \frac{1}{\pi} \left(Arc\tan(x) + \frac{\pi}{2} \right) \\
&= \frac{1}{\pi} Arc\tan(x) + \frac{1}{2}.
\end{aligned}$$

Mode is where the density attains its maximum. We call density $f(x)$ *unimodal* if it has one mode, i.e. $f(x)$ is increasing at left and decreasing at right of the mode. Otherwise, we call the density *multimodal*. Let m denote the mode, we call X *symmetric* if

$$\Pr(X - m \leq -a) = \Pr(X - m \geq a).$$

For a symmetric distribution: $f(m-x) = f(x+m)$ and $F(m-x) = 1 - F(x+m)$. Illustrate geometrically.

Prove, that the Cauchy distribution is symmetric, and therefore unimodal with mode=0.

Exponential distribution is not symmetric.

Problem. Find a such that

$$f(x) = \frac{a}{1+x^2}$$

is a density function (this is the density of *Cauchy* distribution). Find the distribution function.

Solution. Check #1: $f(x) > 0$ for all x. To find a we have to calculate

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx \\ = & \left(Arc\tan(x)|_{-\infty}^{\infty} \right) \\ = & \left(\frac{\pi}{2} - (-\frac{\pi}{2}) \right) \\ = & \left(\frac{\pi}{2} + \frac{\pi}{2} \right) \\ = & \pi. \end{aligned}$$

That means

$$a = \frac{1}{\pi}.$$

The distribution function is

$$\begin{aligned} F(x) &= \frac{1}{\pi} \int_{-\infty}^{x} \frac{1}{1+t^2} dt \\ &= \frac{1}{\pi} \left(Arc\tan(x) + \frac{\pi}{2} \right) \\ &= \frac{1}{\pi} Arc\tan(x) + \frac{1}{2}. \end{aligned}$$

Mode is where the density attains its maximum. We call density $f(x)$ *unimodal* if it has one mode, i.e. $f(x)$ is increasing at left and decreasing at right of the mode. Otherwise, we call the density *multimodal.* Let m denote the mode, we call X *symmetric* if

$$\Pr(X - m \leq -a) = \Pr(X - m \geq a).$$

For a symmetric distribution: $f(m-x) = f(x+m)$ and $F(m-x) = 1 - F(x+m)$. Illustrate geometrically.

Prove, that the Cauchy distribution is symmetric, and therefore unimodal with mode=0.

Exponential distribution is not symmetric.

Important Continuous Variables

2.11 Important Continuous Random Variables

1. *Uniform(a,b) rv:* $X \sim \mathrm{Unif}\,(a,b)$, $a < b$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{o.w.} \end{cases}$$

Applications: Arises in random number generation, etc.

Statistics:

$$\mathrm{E}\,[X] = \frac{a+b}{2} \qquad \mathrm{var}\,(X) = \frac{(b-a)^2}{12}$$

2. *Beta(α, β) rv:* $X \sim \mathrm{Beta}\,(\alpha, \beta)$, $\alpha, \beta > 0$, if

$$f_X(x) = \begin{cases} \frac{x^\alpha(1-x)^\beta}{\mathrm{B}(\alpha,\beta)} & 0 \le x \le 1 \\ 0 & \text{o.w.} \end{cases}$$

where $\mathrm{B}(\alpha, \beta)$ is the "normalization factor" chosen to ensure that $f_X(\cdot)$ integrates to one, i.e.

$$\mathrm{B}(\alpha, \beta) = \int_0^1 y^\alpha (1-y)^\beta dy.$$

Applications: The Beta distribution is a commonly used "prior" on the Bernoulli parameter p.

Exercise 2.1: Compute the mean and variance of a $\mathrm{Beta}\,(\alpha, \beta)$ rv in terms of the function $\mathrm{B}(\alpha, \beta)$.

3. *Exponential(λ) rv:* $X \sim \mathrm{Exp}\,(\lambda)$, $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{o.w.} \end{cases}$$

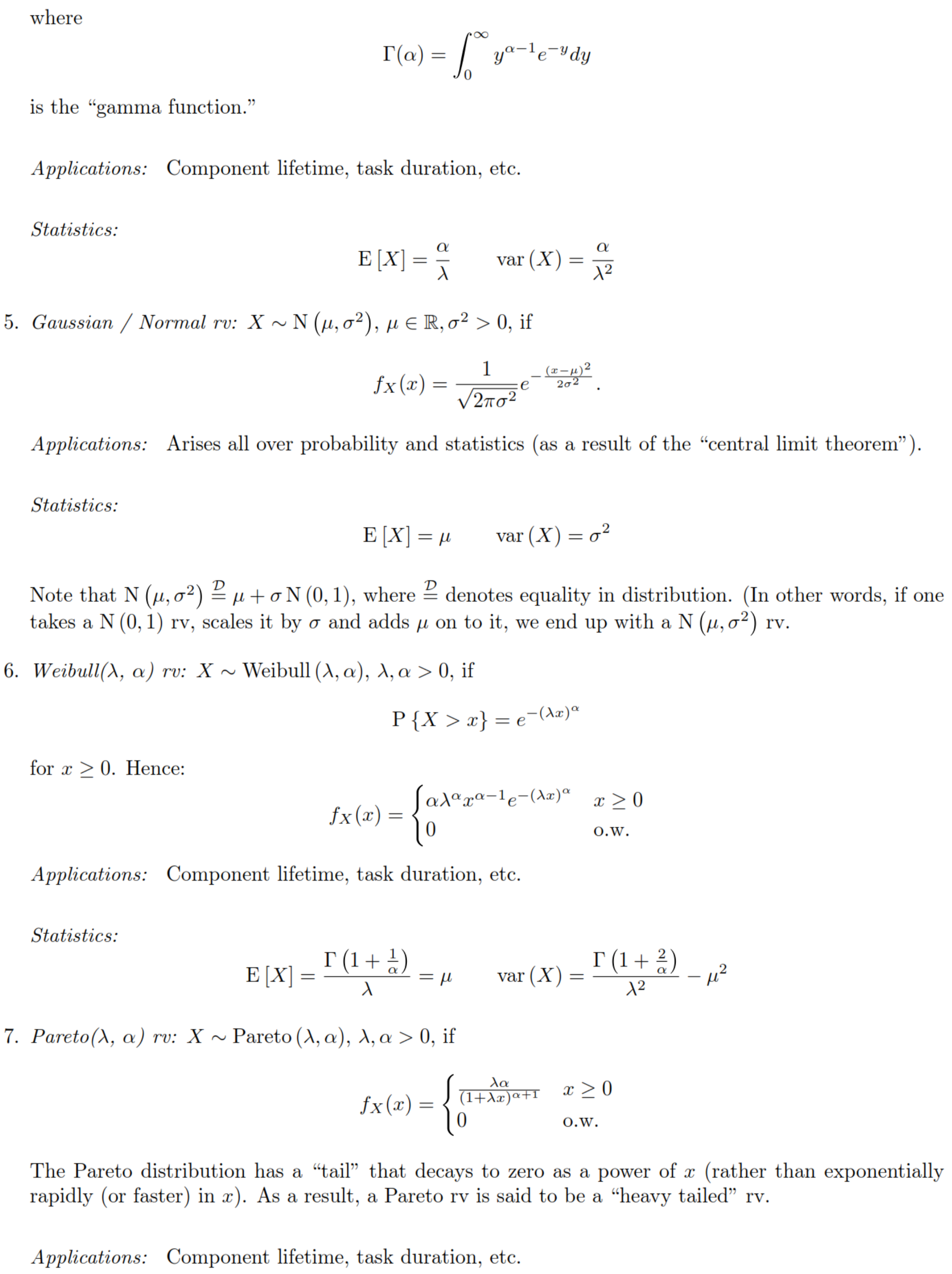
Applications: Component lifetime, task duration, etc.

Statistics:

$$\mathrm{E}\,[X] = \frac{1}{\lambda} \qquad \mathrm{var}\,(X) = \frac{1}{\lambda^2}$$

4. *Gamma(λ, α) rv:* $X \sim \mathrm{Gamma}\,(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$f_X(x) = \begin{cases} \frac{\lambda(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x \ge 0 \\ 0 & \text{o.w.} \end{cases}$$



where

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

is the "gamma function."

Applications: Component lifetime, task duration, etc.

Statistics:

$$\mathrm{E}[X] = \frac{\alpha}{\lambda} \qquad \mathrm{var}(X) = \frac{\alpha}{\lambda^2}$$

5. *Gaussian / Normal rv:* $X \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, if

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

Applications: Arises all over probability and statistics (as a result of the "central limit theorem").

Statistics:

$$\mathrm{E}[X] = \mu \qquad \mathrm{var}(X) = \sigma^2$$

Note that $\mathrm{N}(\mu, \sigma^2) \overset{\mathcal{D}}{=} \mu + \sigma \, \mathrm{N}(0,1)$, where $\overset{\mathcal{D}}{=}$ denotes equality in distribution. (In other words, if one takes a $\mathrm{N}(0,1)$ rv, scales it by σ and adds μ on to it, we end up with a $\mathrm{N}(\mu, \sigma^2)$ rv.

6. *Weibull(λ, α) rv:* $X \sim \mathrm{Weibull}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$\mathrm{P}\{X > x\} = e^{-(\lambda x)^\alpha}$$

for $x \geq 0$. Hence:

$$f_X(x) = \begin{cases} \alpha\lambda^\alpha x^{\alpha-1} e^{-(\lambda x)^\alpha} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Applications: Component lifetime, task duration, etc.

Statistics:

$$\mathrm{E}[X] = \frac{\Gamma\left(1 + \frac{1}{\alpha}\right)}{\lambda} = \mu \qquad \mathrm{var}(X) = \frac{\Gamma\left(1 + \frac{2}{\alpha}\right)}{\lambda^2} - \mu^2$$

7. *Pareto(λ, α) rv:* $X \sim \mathrm{Pareto}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$f_X(x) = \begin{cases} \frac{\lambda\alpha}{(1+\lambda x)^{\alpha+1}} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

The Pareto distribution has a "tail" that decays to zero as a power of x (rather than exponentially rapidly (or faster) in x). As a result, a Pareto rv is said to be a "heavy tailed" rv.

Applications: Component lifetime, task duration, etc.

How we compute the expected value of a random variable?

If X is a discrete real-valued rv, its expectation is defined as

$$E(X) = \sum_x x P_X(x)$$

If X is continuous rv, its expectation is just

$$E(X)=\int_{-\infty}^{\infty} f_X(x)$$

3. Commonly Used Summary Statistics

Given a rv X, the following are the most commonly used “summary statistics.”

1. ***Mean of X***: The mean of X is just its expectation E[X]. We will see later, in our discussion of the law of large numbers, why E[X] is a key characteristic of X’s distribution.
2. ***Variance of X***:

$$\text{var}(X)=E[(X-E[X])^2]$$

This is a measure of X’s variability.

3. ***Standard Deviation of X***:

$$\sigma(X)=\sqrt{\text{var}(X)}$$

This is a measure of variability that scales appropriately under a change in the units used to measure X (e.g. if X is a length, changing units from feet to inches multiplies the variance by 144, but the standard deviation by 12).

4. ***Squared Coefficient of Variation***:

$$c^2(X)=\frac{\text{var}(X)}{E[X]^2}$$

This is a dimensionless measure of variability that is widely used when characterizing the variation that is present in a non-negative rv X (e.g. task durations, component lifetimes, etc).

5. ***Median of X***: this is the value *n* having the property that

$$P(X\le m)=\frac{1}{2}=P(X\ge m)$$

(and is uniquely defined when P(X ≤ ·} is continuous and strictly increasing). It is a measure of the “central tendency” of X that complements the mean. Its advantage, relative to the mean, is that it is less sensitive to “outliers” (i.e. observations that are in the “tails” of X that have a big influence on the mean, but very little influence on the median).

6. ***p^{th} quantile of X (0<p<1)***: If X is real number random variable with distribution $F(x)=P(X\le x)$, $x\in\mathrm{R}$ then the pth quantile is the number $x_p \in$ R so that $F(x_p)=p$.
   - The pth quantile is

     $$Q(p)=F^{-1}(p),\quad p\in(0,1)$$

Note: The $\frac{1}{4}$ th quantile is called the lower quartile, the $\frac{3}{4}$ th quantile is called the upper quartile. Median is the $\frac{1}{2}$ th quantile, i.e. *F*(median) = $\frac{1}{2}$

Percentile is a quantile expressed in per cents (e.g. 75 percentile, 25 percentile, etc.). Quantiles and percentiles are used to characterize the range of the distribution.

Problem. Find the pth quantile of the exponential distribution defined by the distribution function

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-x} \text{ if } x \geq 0 \end{cases}.$$

Also, find median, the lower and the upper quartile. Find a and b such that $\Pr(a < X < b) = 0.5$

Solution. We need to solve equation

$$1 - e^{-x} = p$$

which yields the pth quantile

$$x_p = -\ln(1-p).$$

The median is

$$x_{.5} = -\ln(1 - .5) = 0.69315.$$

The lower quartile is

$$x_{.25} = -\ln(1 - .25) = .28768.$$

The upper quartile is

$$x_{.75} = -\ln(1 - .75) = 1.3863.$$

Since $\Pr(x_{.25} < X) = .25$ and $\Pr(x_{.75} > X) = .25$ we have

$$\Pr(x_{.25} < X < x_{.75}) = .5$$

that means $a = x_{.25}$ and $b = x_{.75}$.

Definition of a Quantile

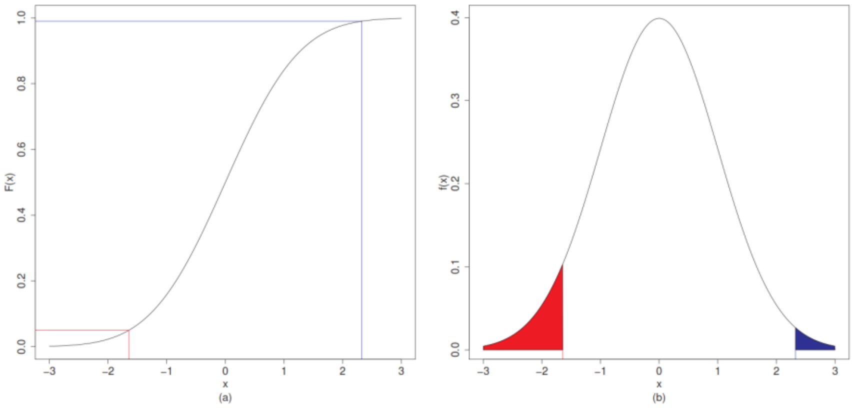


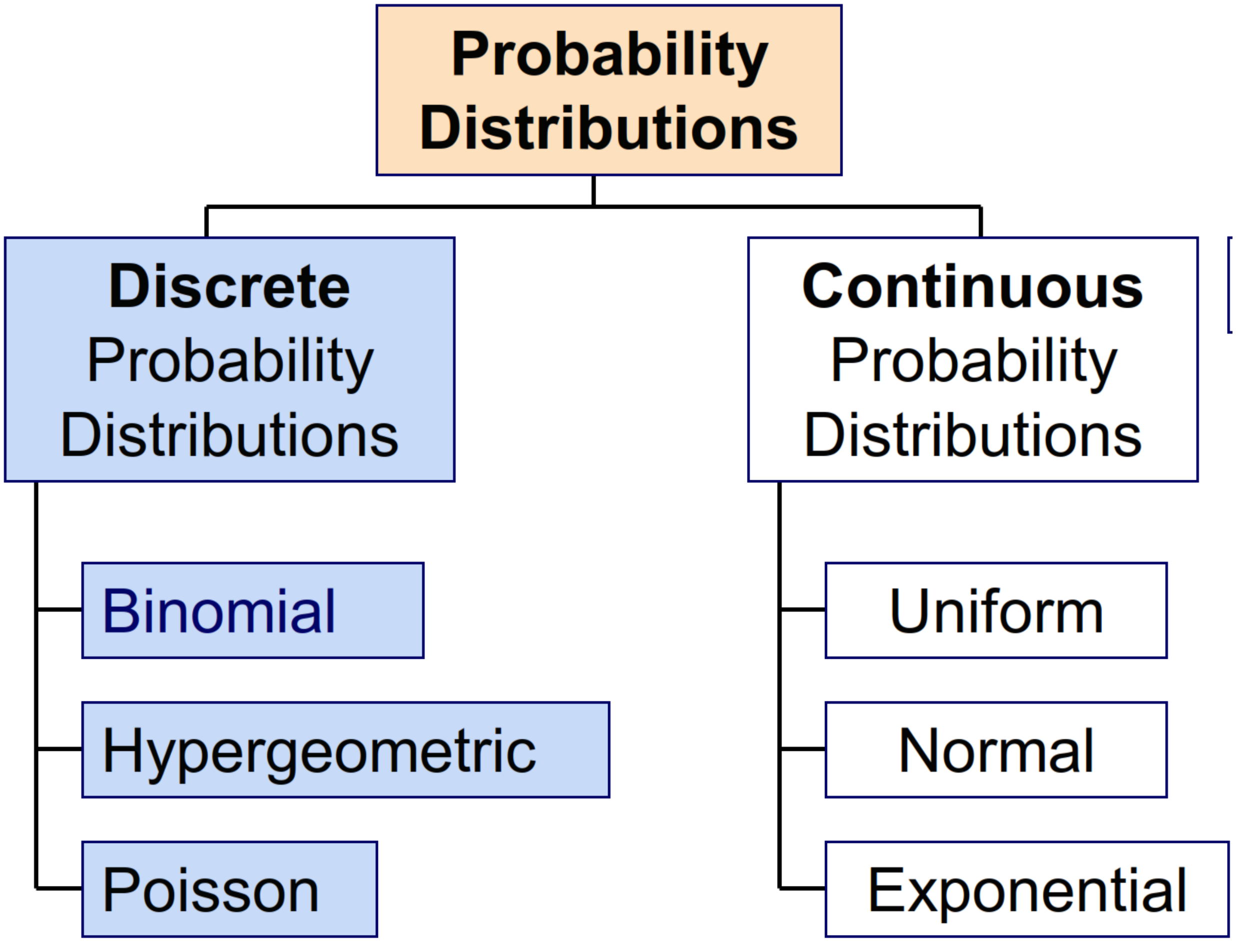
Figure 2: (a) Distribution function. (b) Density function. The 5% quantile with red. The 99% quantile with blue.

7. ***Inter-quartile range***: This is the quantity

$$F_X^{-1}(\frac{3}{4}) - F_X^{-1}(\frac{1}{4})$$

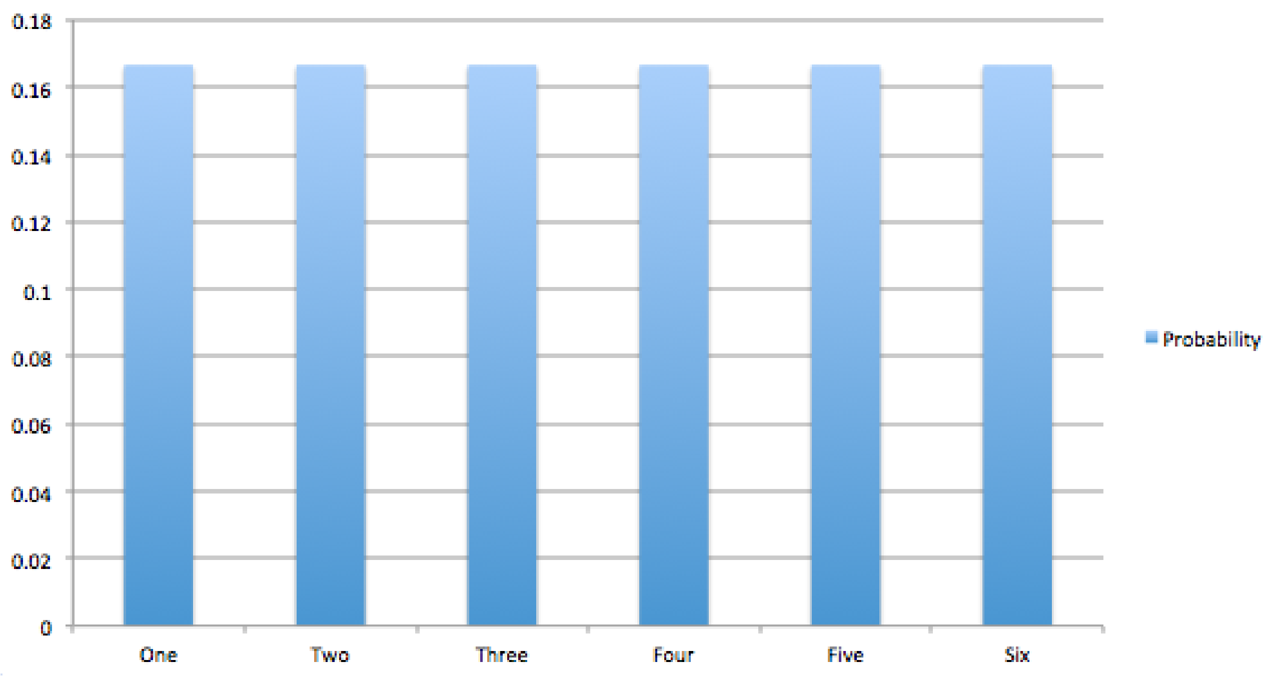
it is a measure of variability that, like the median, is (much) less sensitive to outliers than is the standard deviation.

4. Examples of Probability Distributions and Important Theorems


Probability Distributions
Discrete Probability Distributions
Continuous Probability Distributions
Binomial
Hypergeometric
Poisson
Uniform
Normal
Exponential

Probability distributions: Let's remember that it is a probability of every possible outcome in sample space.

So for a unbiased dice, probability of every outcome is equal 1/6 which look like,



When a probability distribution looks like this (equal probability of all outcomes) it is called **Uniform probability distribution**.

Important thing to consider here is sum of all probabilities is exactly equals to one for probability distribution.

Why we need it: Most of the statistical modelling methods make certain assumption about underlying probability distribution. So based on what kind of distribution data follows, we can choose appropriate methods. Sometimes we will transform (log transform, inverse transform) the data if the distribution observed is not what we would have expected or required by certain statistical methods.

We have categorized probability distribution in to three classes, **discrete probability distribution**, **continuous probability distribution** and **Cumulative Distribution Function**.

- Discrete: Sample space is collection of discrete values. e.g. Coin flip, die throw etc
- Continuous: Sample space is collection of infinite continuous values. e.g. Height of all people in US, distance travelled to reach workplace
- Cumulative: we add the probabilities for all values qualifying as "less than or equal" to the specified value

Probability Distribution Definition Again!

For a discrete random variable, its **probability distribution** (also called the **probability distribution function**) is any table, graph, or formula that gives each possible value and the probability of that value.
Remember: The total of all probabilities across the distribution must be 1, and each individual probability must be between 0 and 1, inclusive.
Example 4.1

What if we flipped a fair coin four times? What are the possible outcomes and what is the probability of each?

Figure 1 below is a probability distribution for the number of heads in 4 flips of a coin. Given that P(Heads)=.50, the probability of not flipping heads at all is 1/16, or .0625. In 6.25% of all trials, we can expect that there will be no heads. This may be written as P(X=0)=.0625. Similarly, the probability of flipping heads once in four trials is 4/16, or .25. In 25% of all trials, we can expect that heads will be flipped exactly once. This may be written as P(X=1)=.25.

This probability distribution could be constructed by listing all 16 possible sequences of heads and tails for four flips (i.e., HHHH, HTHH, HTTH, HTTT, etc.), and then counting how many sequences there are for each possible number of heads. Or, in section 5.4 you will see how these could be computed using binomial random variable techniques.

Figure 1. Probability Distribution for Number of Heads in 4 Flips of a Coin

Heads	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16

Definition: Cumulative Probabilities

Cumulative probability: Likelihood of an outcome less than or equal to a given value occurring. To find a **cumulative probability** we add the probabilities for all values qualifying as "less than or equal" to the specified value.

Example 4.2

Suppose we want to know the probability that the number of heads in four flips is less than two. If we let X represent number of heads we get on four flips of a coin, then:

Because this is a discrete distribution, the probability of flipping less than two heads is equal to flipping one or zero heads:

$P(X<2)=P(X=0\cup 1)$

The probability of flipping 1 head and the probability of flipping 0 heads are mutually exclusive events. Thus,

$P(0\cup 1)=P(X=0)+P(X=1)$

We can use the values from Figure 1 above to solve this equation.

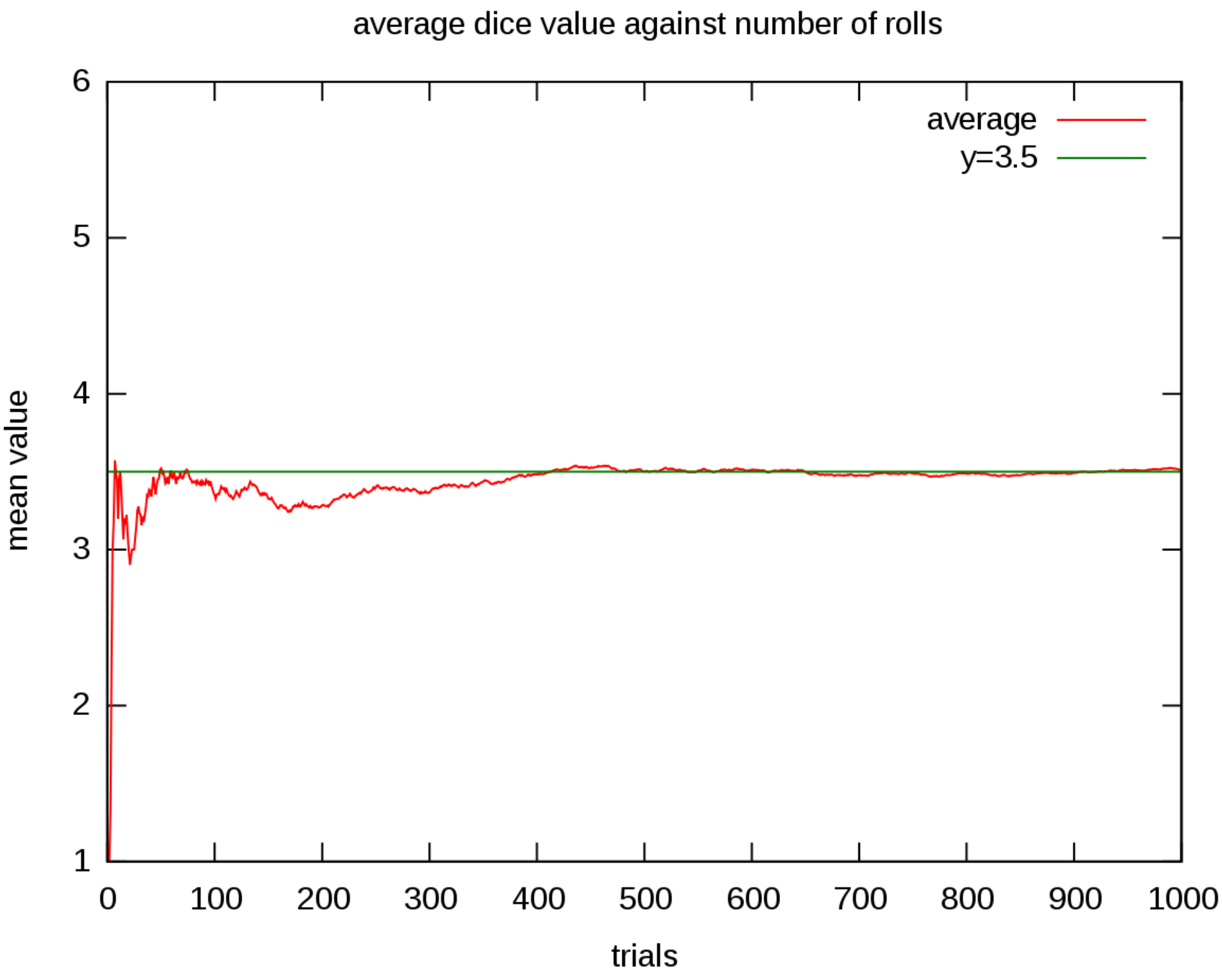
$P(X=0)+P(X=1)=(1/16)+(4/16)=5/16$

Theorem Law of Large Numbers

Law of Large numbers: The law of large numbers implies larger the sample size, closer is our sample mean to the true (population) mean.

Why we need it: Have you ever wondered, if probability of any outcome (head or tail) for a fair coin is exactly half but for 10 trials you might actually get different results (e.g. 6 heads and 4 tails). Well, Law of Large numbers provides answer to it. It says as we will increase number of trials, mean of all trials will come closer to expected value.

Another simple example is, for an unbiased die probability of every outcome {1,2,3,4,5,6} is exactly same (1/6) so the mean should be 3.5.



As we can see in the image above, only after large number of trials the mean approaches to 3.5.

Central Limit Theorem

Central Limit theorem: Regardless of the underlying distribution, if we draw large enough samples and plot each sample mean then it approximates to normal distribution.

Why we need it: If we know given data is normally distributed then it provides more understanding about data as compared to unknown distribution. And the Central Limit Theorem enables us to actually use the real world data (near-normal or non-normal) with statistical methods making assumption about normality of the data.

An article on about.com summarizes the practical use of CLT as follows,

"The assumption that data is from a normal distribution simplifies matters, but seems a little unrealistic. Just a little work with some real-world data shows that outliers, skewness, multiple peaks and asymmetry show up quite routinely. We can get around the problem of data from a population that is not normal. The use of an appropriate sample size and the central limit theorem help us to get around the problem of data from populations that are not normal.

Thus, even though we might not know the shape of the distribution where our data comes from, the central limit theorem says that we can treat the sampling distribution as if it were normal."

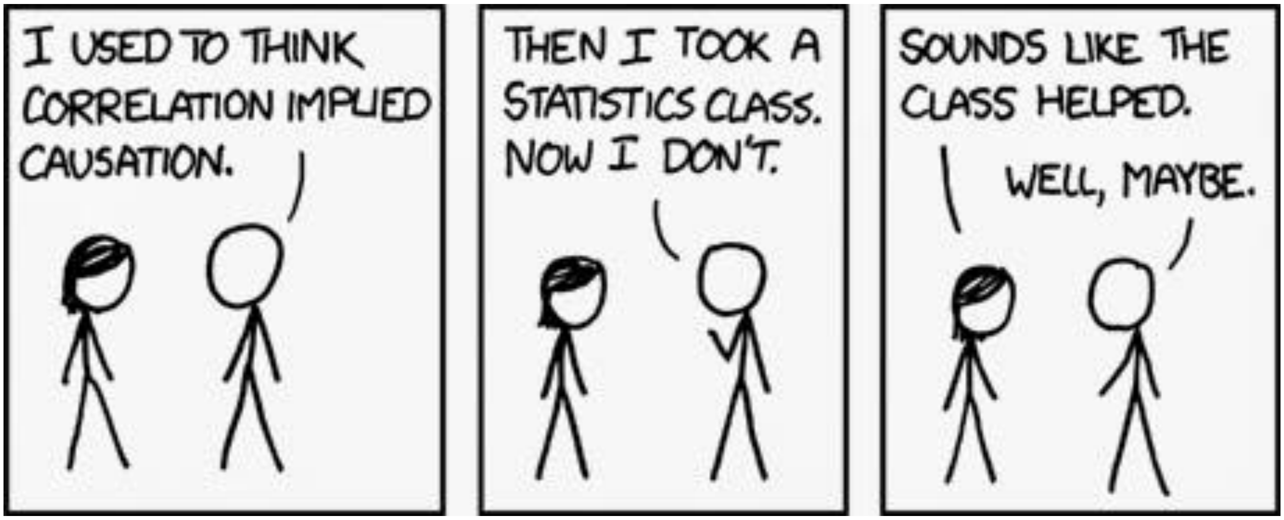
Correlation

Correlation: A number representing strength of association between two variables. A high value of correlation coefficient implies both variables are strongly associated.

One way to measure it is Person's correlation coefficient. It most widely used method which can measure only linear relationship between variables. The coefficient value varies from -1 to 1.

The correlation coefficient value of zero means, there is no relationship between two variables. A negative value means as one variable increases the other decreases.

The most important thing to remember here is, correlation does not necessarily mean there is a causation. It represents how two variables are associated with each other.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Peter Flom, a statistical consultant explains the difference in simple words as following:
"Correlation means two things go together. Causation means one thing causes another."

Once we find correlation, controlled experiments can be conducted to check if any causation exists. There are few statistical methods which help us to check non-linear relationship between two variables like, maximal correlation.

Why we need it: A correlation coefficient tells us how strongly two variables are associated and direction of the association.

Binomial random variables

Definition

There are many types of discrete random variables. Here, we introduce the binomial family.

Binomial random variables are discrete RVs of "counts" that describe the number of "successes" (X) in n

independent Bernoulli trials,[a] where each Bernoulli trial has probability of success designated as p.

Binomial random variables have two **parameters**, n and p.

$n \equiv$ the number of independent "Bernoulli" trials
$p \equiv$ the probability of success for each trial (which does not change from trial to trial)

EXAMPLE. Consider the number of successful treatments (random variable X) in 3 patients (n = 3) where the probability of success in each instance (p) is 0.25. X can take on the discrete values of 0, 1, 2, or 3.

Notation. Let "b" represent "binomial distribution" and "~" represent "distributed as." Thus, X~b(n, p) is read as "random variable X is distributed as a binomial random variable with parameters n and p."

EXAMPLE. X~b(3, .25) is read "X is distributed as a binomial random variable with parameters n=3 and p=.25."

More notation.

- Let P($X = x$) represent "the probability that random variable X takes on a value of x."
- Let P($X \leq x$) represent "the probability random variable X takes on a value less than or equal to x." This is the **cumulative probability** of the event.

DEFINITION. The **probability mass function (pmf)** assigns probabilities for all possible outcomes of a discrete random variable.

EXAMPLE. The pmf for X~b(3, .25) is shown in Table 1. Probabilities for each potential outcome are shown in the second column. Cumulative probabilities are shown in the third column.

TABLE 1. The pmf for X~b(3, .25).

X **Number of successes**	**Pr(*X* = *x*)** **Probability**	**Pr(*X* ≤ *x*)** **Cumulative Probability**
0 (event A)	0.4219	0.4219
1 (event B)	0.4219	0.8438
2 (event C)	0.1406	0.9844
3 (event D)	0.0156	1.0000

INTERPRETATION. How we calculated these probabilities is not currently the issue. Instead, let us focus on meaning. The above pmf states that for *X*~b(3, .25) we expect to see 0 successes 0.4219 of the time, 1 success

0.4219 of the time, 2 successes 0.1406 of the time, and 3 successes 0.0156 of the time.
Calculations.

[a] A **Bernoulli trial** is a random event that can take on one of two possible outcomes. One possible outcome is arbitrarily designated as a "success." The other outcome is designated a "failure." Outcomes are also designated as either 0 ("failure") or 1 ("success").

Rules for working with probabilities revisited

Notation:

- A ≡ event A
- B ≡ event B
- P(A) ≡ the probability of event A
- Ā ≡ the *complement* of event A ≡ not A (i.e., anything other than A)
- ⋃ ≡ union of events. For example, A ⋃ B means that either A *or* B occur.
- ⋂ ≡ intersection of events. For example, A ⋂ B means that both A *and* B occur.

Rule 1: Probabilities can be no less than 0% and no more than 100%. An event with probability 0 can never occur. An event with probability 1 is certain or always occurs.

$$0 \le P(A) \le 1$$

Note that an all the events in Table 1 obey this rule.

Rule 2: All possible outcomes taken together have probability exactly equal to 1.

P(all possible outcomes) = 1

Note that in Table 1, P(all possible outcomes) = 0.4129 + 0.4129 + .1406 + 0.0156 = 1.

Rule 3: When two events are disjoint (cannot occur together), the probability of their union is the sum of their individual probabilities.

P(A ∪ B) = P(A) + P(B), if A and B are disjoint

In Table 1 let A ≡ 0 successes and A ≡ 1 success. P(A ∪ B) = 0.4219 + 0.4219 = 0.8438.

Rule 4: The probability of a complement is equal to 1 minus the probability of the event.

P(Ā) = 1 – P(A)

In Table 1, Ā ≡ (1, 2, or 3 successes) and P(Ā) = 1 – 0.4219 = 0.578.

The area under the curve (AUC)

Probability mass functions (pmfs) can be drawn as pmf **histograms**. The area under the bars of pmf histograms correspond to probabilities. For example, the pmf histogram for the random variable in Table 1 is as follows:

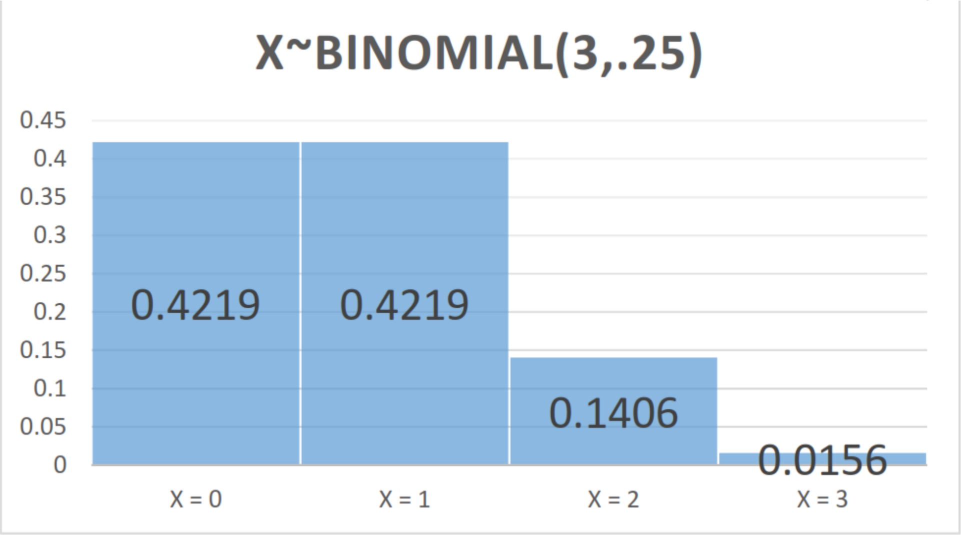


Figure 1. X~b(3, .25).

Area of the first bar: P(X = 0). The height of the bar = 0.4219. On the horizontal axis, the first bar stretches from 0 to 1. Therefore, this rectangle has base = 1. The **area** of this bar = height × base = 0.4219× 1 = 0.4219. This is also the probability that zero events occur. Therefore, P(*X* = 0) = area of the bar =0.4219.

The area under the bars of a pmf histogram corresponds to its probability.

Area of the second bar: P(X = 1). The second bar has height = 0.4219, base = 1 (from 1 to 2), and area (i.e., probability) = h × b = 0.4219 × 1 = 0.4219.

Area of the first two bars, i.e., P(X = 0) ∪P(X = 1). The combined area of the first two bars = 0.4219 + 0.4219 = 0.8438, corresponding to the probability of 0 or 1 successes.

The area under the pmf histogram **("area under the curve")** between any two points is equal to the probability of the corresponding outcomes.

The “Rule of Complements”

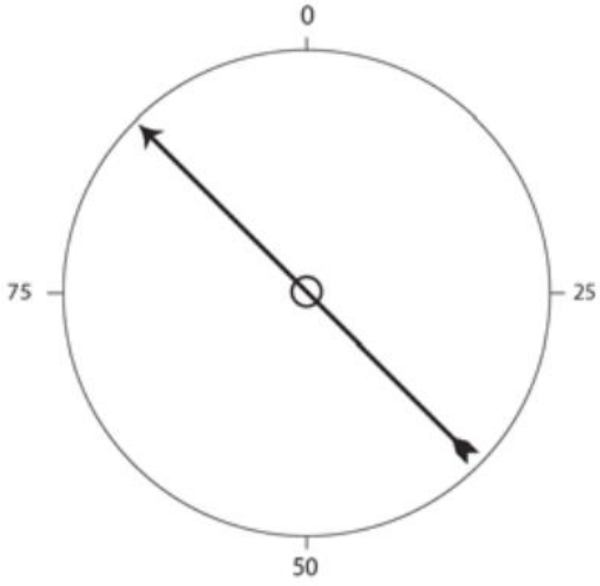
Recall that Ā ≡ the *complement* of event A, i.e., “not A,” i.e., anything other than A. Rule 4 (prior page) says P(Ā) = 1 – Pr(A).

EXAMPLE. Consider the pmf in Table 1 (X~b(3, .25). Let A ≡ 0 successes. Therefore Ā ≡ 1, 2, or 3 successes. This corresponds to the AUC in the “right tail” of the pmf historgram. By the rule of complements, Pr(Ā) = 1 – 0.4219 = 0.5781.

Normal Distributions

Recall that **continuous random variables** form a continuum of possible outcomes. There are many different types of continuous random variables. These random variable types occur in families (e.g., uniform random variables, normal random variables, chi-squared random variables, etc.).

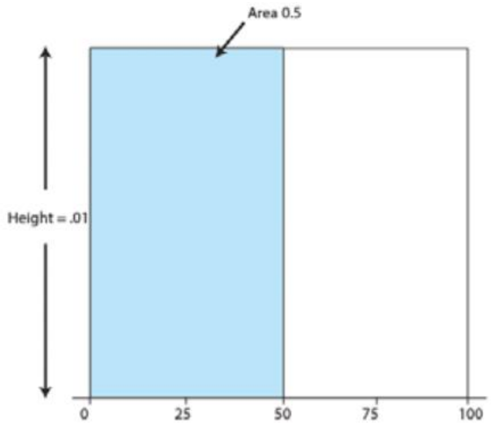
Consider the spinner below. This spinner will generate a continuous uniform random variable with values between 0 and 100. This is a continuous random variable with a range of 0 to 100. The spinner can land on any value between any two points. For example, between 27.5 and 28, it can land on 27.5, 27.75, 27.875, 27.27.9375, etc.



Continuous random number generator: Uniform(0, 100).

To understand continuous random variables, you must accept the thought experiment that the probability of landing on any specific number is 0 (or at least not determinable). For example, P(X = 50) = 0. However, the probability of landing between any two values is determinable. For example, the probability of the above random spinner landing on a value between 0 and 50 is .5, i.e., $P(0 \leq X \leq 50) = .50$.

Probability density functions (pdf) assign probabilities for all possible outcomes for continuous random variables. pdfs cannot be shown in tabular form. They can, however, be represented with integral functions (calculus). They can also be drawn. For example, the pdf for the above random number spinner looks like this:



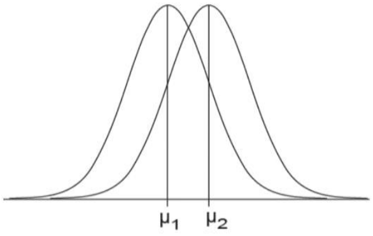
pdf for continuous random number between 0 and 100 showing Pr(0 <= X <= 50)

Importantly, that the **area under the curve (AUC) concept** introduced on the prior page also applies to pdf graphs. For example, the AUC between 0 and 50 (shaded above) = height × base = .01 × 50 = .50, or 50%. Therefore, $P(0 \leq X \leq 50) = .50$ for this particular continuous random variable.

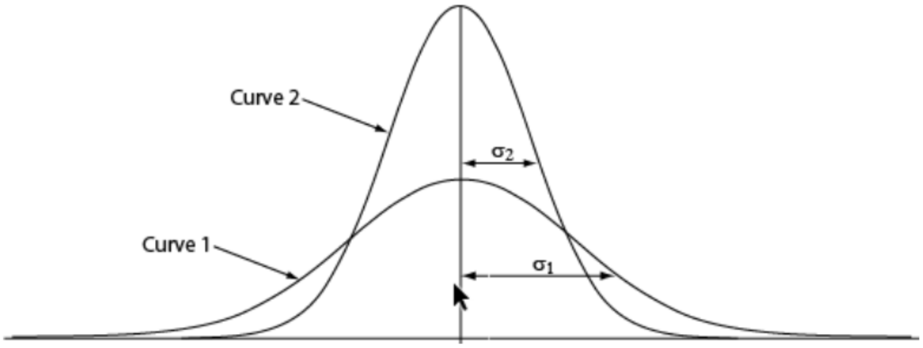
The Normal Distribution

Normal random variables are a *family* of continuous random variables. Each family member is characterized by two **parameters**, μ (“mu”) and σ (“sigma”).

- μ ≡ the pdf’s mean or expected value (indicating central location)
- σ ≡ the pdf’s standard deviation (indicating spread) When μ changes, the location of the pdf changes.



When σ changes, the spread of the pdf changes.



The parameters μ and σ are the analogues (but not the same as) the statistics $\overline{x}$ and *s*. However, you cannot calculate μ and σ . μ and σ are not from any data source.

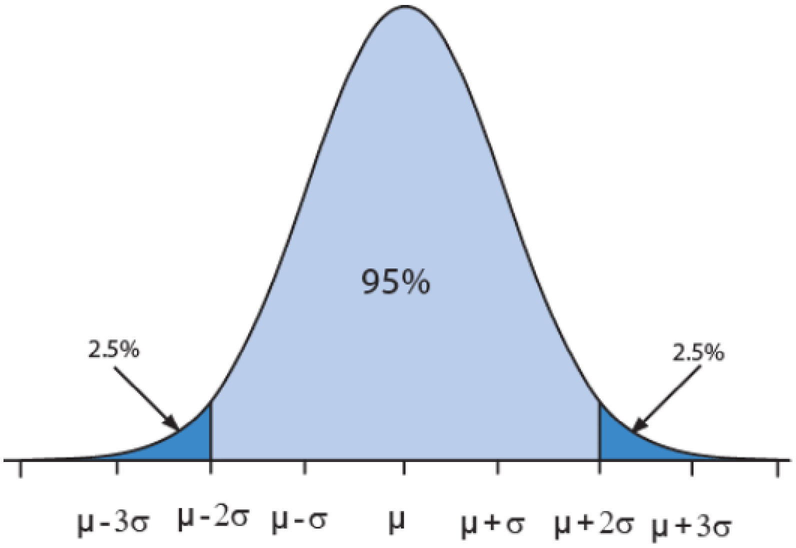
You can visualize the size of σ on a normal pdf plot by identifying the curve's **points of inflection**. This is where the curve begins to change slope. Trace the slope of the normal curve with your finger. As you "ski" down the slope, the point of inflection is where the slope *begins* to flatten. The left inflection point marks the location $\mu - \sigma$. This is one σ-unit below the mean.
The right point of inflection marks the location of $\mu - \sigma$. This is one σ-unit below the mean.

Normal probabilities

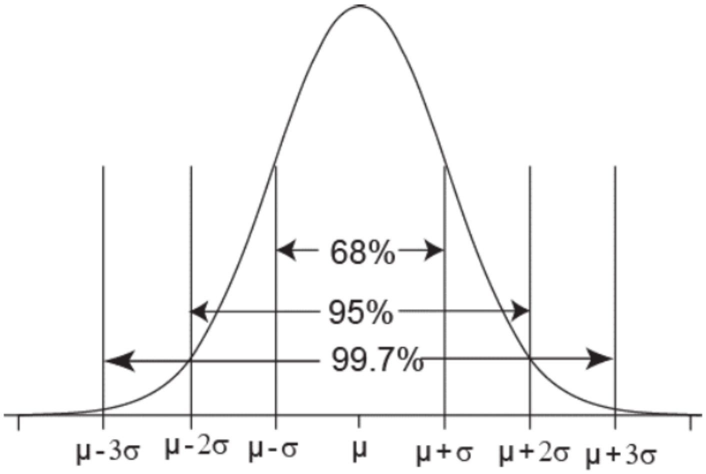
The 68-95-99.7 rule helps get a grip on normal probabilities.[b]

- 68% of the AUC for normal RVs lies in the region $\mu \pm \sigma$
- 95% of the AUC for normal RVs lies in the region $\mu \pm 2\sigma$
- 99.7% of the AUC for normal RVs lies in the region $\mu \pm 3\sigma$

These rules apply only to normal random variables. Visually, the "95" part of the rule looks like this:



Think in terms of these landmarks:



Although μ and σ vary from one normal random variable to the next, you can apply the 68-95-99.7 rule to any normal random variable if you keep these facts in mind: (1) probability = AUC; (2) The total AUC

for the pdf = 1; (3) Values for the random variable lie on the horizontal axis

EXAMPLE. The Wechsler Intelligence Scale is calibrated to produce a normal distribution with μ = 100 and σ = 15 within each age group.

Notation. Let X~N(μ, σ) represent a normal random variable with mean μ and standard deviation σ. Using this notation, Wechsler Intelligence scale scores in a population *X*~N(100, 15). This is stated as "*X* is distributed as a normal random variable with mean 100 and standard deviation 15."

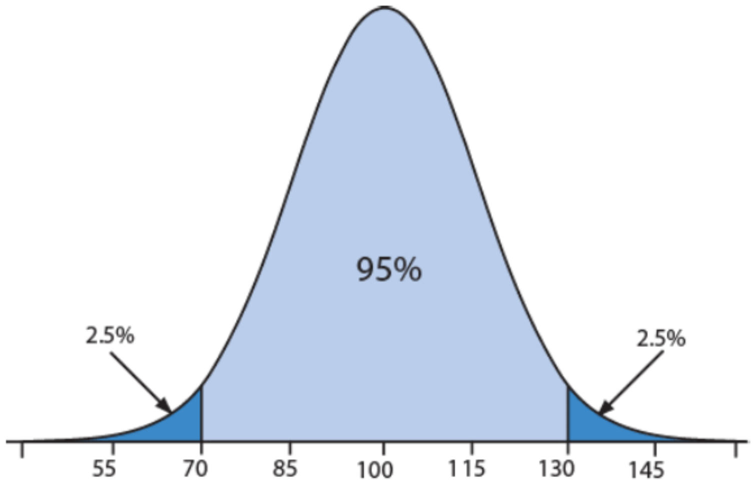
The 68–95–99.7 rule states that for *X*~N(100, 15):

[b] You must accept the fact that the area under the curve (AUC) represents probabilities.

- 68% of the AUC lies in the range μ ± σ = 100 ± 15 = 85 to 115
- 95% of the AUC lies in the range μ ± 2σ = 100 ± (2)(15) = 70 to 130

- 99.7% of the AUC lies in the range μ ± 3σ = 100 ± (3)(15) = 55 to 145

This next figure shows the AUC for X~N(100, 15). Notice the center of the curve is on μ. Also notice landmarks at ±1σ, ±2σ, ±3σ on the horizontal axis.



Finding AUCs with for normal random variable app

"In the old days, we found normal probabilities with a a tedious process that relied on tables. We can now use a app for the purpose. Either way, the key concept is the AUC between any two points corresponds to probability.

We can use this app to calculate AUCs between any two points for any X~N(μ, σ): [http://onlinestatbook.com/2/calculators/normal_dist.html.](http://onlinestatbook.com/2/calculators/normal_dist.html) There is a link to this app on www.sjsu.edu/faculty/gerstman/StatPrimer .

Example. Plug in values for X~(100,15). The AUC between 130 and ∞ corresponds to the right tail of the pdf. Note that this AUC (probability) is 0.0228 (roughly 2.5%).

37