Basics of Probability

Emre Özer

April 2019

Part of the “Statistics of Measurement” course at Imperial College. The statistics part isn’t in these notes.

Contents

1 Probability

1.1 Definitions and properties

We are concerned with probability in the context of *experiments*.

Definition (Trial)**.** A *trial* is a single performance of an experiment.

Definition (Outcome)**.** Each possible result of a trial is called an *outcome*.

Definition (Sample space)**.** In a given experiment, the set of all possible outcomes of an individual trial is called the *sample space*, denoted S.

Example. For a coin flip, the sample space is $S = \{\text{Heads}, \text{Tails}\}$. For a die, we have $S = \{1, 2, 3, 4, 5, 6\}$.

Definition (Event)**.** An *event* is a subset of the sample space. We may denote $A \subseteq S$.

Definition (Mutually exclusive events)**.** Two events $A, B \subseteq S$ are said to be *mutually exclusive* if and only if $A \cap B = \emptyset$.

Definition (Complement of event)**.** Given an event $A \subseteq S$, its *complement* is defined as the set of all outcomes in the sample space not in A, given by $\overline{A} := \{x \in S | x \notin A\}$.

Definition (Frequentist probability)**.** The *probability* $\Pr(A)$ of some event A is the expected relative frequency of the event in a large number of trials. If there is a total of n_S outcomes, and n_A of those correspond to the event A, then the probability $\Pr(A)$ is given by

$$\Pr(A) = \frac{n_A}{n_S}. \tag{1.1}$$

Proposition (Properties of probabilities)**.** From (1.1) we deduce the following set of axioms:

1. Given a sample space S, we have

   $$0 \leq \Pr(A) \leq 1, \quad \forall\, A \subseteq S. \tag{1.2}$$

2. We are certain to obtain one of the outcomes:

   $$\Pr(S) = \frac{n_S}{n_S} \equiv 1. \tag{1.3}$$

3. Given two events $A, B \in S$, we have

   $$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{1.4}$$

   This follows from (1.1), noting that $n_{A\cup B} = n_A + n_B + n_{A\cap B}$. If A and B are mutually exclusive, we obtain the special case

   $$\Pr(A \cup B) = \Pr(A) + \Pr(B). \tag{1.5}$$

4. Complement events are mutually exclusive, so consider some $A \subseteq S$ and its complement $\overline{A}$. We have

   $$1 = \Pr(S) = \Pr(A \cup \overline{A}) = \Pr(A) + \Pr(\overline{A}),$$

   from which we obtain the *complement law*:

   $$\Pr(\overline{A}) = 1 - \Pr(A). \tag{1.6}$$

These properties can be extended to multiple events, although unions become more complicated.

1.2 Conditional probability

Definition (Conditional probability)**.** The *conditional probability* is the probability that a particular event occurs *given* the occurrence of another, possibly related, event.

Suppose we are interested in the probability that an event $B \subseteq S$ will occur, given that $A \subseteq S$ has happened. We denote this by $\Pr(B|A)$. Now, notice that the probability of A and B will happen is given by

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A) = \Pr(B)\Pr(A|B).$$

So, we have

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}, \tag{1.7}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{1.8}$$

Note (Reduced sample space)**.** We may think of $\Pr(B|A)$ as the probability of B in the *reduced sample space* defined by A. Hence, if A and B are mutually exclusive, we have

$$\Pr(A \cap B) = \Pr(B \cap B) = \Pr(\emptyset) \equiv 0.$$

Definition (Statistically independent events)**.** Two events $A, B \subseteq S$ are said to be *statistically independent* if $\Pr(A|B) = \Pr(A) \Leftrightarrow \Pr(B|A) = \Pr(B)$. In other words, the probability of one is independent of the occurrence of the other.

If A and B are statistically independent, it follows from (1.7) and (1.8) that

$$\Pr(A \cap B) = \Pr(A)\Pr(B). \tag{1.9}$$

Equation (1.9) may be taken as the definition of statistical independence. The concept can be extended to a set of events $\{A_i\}$, which are said to be *mutually independent* if

$$\Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = \prod_{i=1}^{n} \Pr(A_i).$$

Proposition (Addition law for conditional probabilities)**.** Suppose $A \subseteq S$ is a union of n *mutually exclusive* events A_i. Given another event $B \subseteq S$, we have

$$\Pr(A|B) = \sum_i \Pr(A_i|B) \Longleftrightarrow \Pr(A)\Pr(B|A) = \sum_i \Pr(A_i)\Pr(B|A_i) \tag{1.10}$$

Proof. Consider the probability $\Pr(A \cap B)$, we have

$$\Pr(A \cap B) = \Pr((A_1 \cap B) \cup \ldots \cup (A_n \cap B)) = \sum_i \Pr(A_i \cap B),$$

where we used equation (1.5) since each A_i is mutually exclusive. Substituting using (1.7) and (1.8) yields the two equivalent relations in (1.10) respectively. □

Proposition (Total probability law)**.** In the special case where the events A_i *exhaust* the sample space S such that $A = S$, we have $A \cap B = S \cap B = B$, and (1.10) implies

$$\Pr(B) = \sum_i \Pr(A_i)\Pr(B|A_i). \tag{1.11}$$

Proof. From (1.10), we have

$$\sum_i \Pr(A_i)\Pr(B|A_i) = \Pr(A)\Pr(B) = \Pr(S)\Pr(B) = \Pr(B). \quad \square$$

1.2.1 Bayes' theorem

This result follows directly from (1.7) and (1.8).

Theorem (Bayes' theorem)**.** Given two events $A, B \subseteq S$, we have

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{1.12}$$

In most cases, $\Pr(B)$ will not be known, so let's look at different ways to express it. Firstly, note that the event A, together with its complement $\overline{A}$ form a mutually exclusive set which exhausts S. Hence, by (1.11) we have

$$\Pr(B) = \Pr(A)\Pr(B|A) + \Pr(\overline{A})\Pr(B|\overline{A}),$$

so that Bayes' theorem becomes

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(A)\Pr(B|A) + \Pr(\overline{A})\Pr(B|\overline{A})}. \tag{1.13}$$

We may go further and note that *each outcome* is mutually exclusive. Therefore, considering each outcome as an event, we may write

$$\Pr(B) = \sum_i \Pr(x_i)\Pr(B|x_i), \quad \text{where } x_i \in S.$$

This yields

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\sum_i \Pr(x_i)\Pr(B|x_i)}. \tag{1.14}$$

Finally, we may look at *relative probabilities*. Let $A, B, C \subseteq S$ and consider the relative probabilities of A and C given the occurrence of B. Then, we have

$$\frac{\Pr(A|B)}{\Pr(C|B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|C)\Pr(C)}. \tag{1.15}$$

1.3 Permutations and combinations

When calculating probabilities, it is necessary to count the number of occurrences of events. The way we count depends on whether the occurrences are distinguishable, and if the order matters.

1.3.1 Permutations

Consider a set of n distinguishable objects. How many ways can we arrange them (how many *permutations* exist)? We have n options for the first position, $(n-1)$ for the second and so on. So, it is easy to see that n objects may be arranged in

$$n \times (n-1) \times (n-2) \times \ldots \times (1) = n!$$

different ways. We may generalise by considering choosing $k < n$ objects from n. The number of possible permutations is

$$\underbrace{n \times (n-1) \times \ldots \times (n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!} := P(n,k).$$

So far we have considered objects sampled *without replacement*. If, on the other hand we sample k objects from a set of n *with replacement*, it is clear that the number of permutations is n^k.

Finally, we consider the case of *indistinguishable objects.* Assume that n_1 objects are of type 1, n_2 objects are of type 2 and so on. The number of distinguishable permutations of a group of n such objects is

$$\frac{n!}{n_1! \times n_2! \times \ldots \times n_m!}.$$

1.3.2 Combinations

Now, consider the number of *combinations* of objects when the order is not important. Since there are $P(n,k)$ permutations, and k objects may be arranged in $k!$ ways, we have

$$C(n,k) := \frac{n!}{(n-k)!k!}.$$

Note that these are also the *binomial coefficients.*

Another case is consider dividing n objects into m piles, with n_i objects in the i^{th} pile. The number of ways to do so is

$$\frac{n!}{n_1! \times n_2! \times \ldots \times n_m!}.$$

These are the *multinomial coefficients.* Note that this is identical to distinguishable permutations of groups of indistinguishable objects.

1.4 Random variables

Definition (Random variable)**.** Given a sample space S, a *random variable* assigns a real number to *each* possible outcome. In a sense, we can treat a random variable as a map $X : S \to \mathbb{R}$.

Furthermore, assuming that a probability can be assigned to all possible outcomes in a sample space S, it is possible to assign a probability distribution to any random variable.

1.4.1 Discrete random variables

Definition (Discrete random variable)**.** A random variable X is a *discrete random variable* if it takes values from a discrete set, such that $X \in \{x_1, x_2, \ldots, x_n\}$, with probabilities $\{p_1, p_2, \ldots, p_n\}$.

Definition (Probability function)**.** Given a discrete random variable, we can define a *probability function* $f(x)$ which assigns probabilities to the domain of X,

$$f(x) = \Pr(X = x) = \begin{cases} p_i & \text{if } x = x_i, \\ 0 & \text{otherwise.} \end{cases} \tag{1.16}$$

We require

$$\sum_i f(x_i) = 1.$$

Definition (Cumulative probability function)**.** A *cumulative probability function* $F(x)$ of some random variable X is the probability that $X \leq x$, so that

$$F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} f(x_i). \tag{1.17}$$

Proposition. The probability that the random variable X lies between two values $a < b \in \mathbb{R}$ is given by

$$\Pr(a < X \leq b) = F(b) - F(a),$$

where $F(x)$ is the cumulative probability function.

Proof. Given a probability function $f(x)$, we have

$$\Pr(a < X \leq b) = \sum_{a<x_i<b} f(x_i) = \sum_{x_i \leq b} f(x_i) - \sum_{x_i \leq a} f(x_i) \equiv F(b) - F(a) \quad \square$$

1.4.2 Continuous random variables

Definition (Continuous random variables)**.** A random variables X is said to be *continuous* if it is defined over a continuous domain.

Definition (Probability density function)**.** A *probability density function* (PDF) $f(x)$ of a continuous random variable X is defined such that

$$\Pr(x < X \leq x + \mathrm{d}x\,) = f(x)\,\mathrm{d}x\,. \tag{1.18}$$

It follows from this definition that $f(x) \geq 0$ for all $x \in D$ where D is the domain over which the random variable is defined.

Similar to a probability function, we require the probabilities to sum to one. So, we require

$$\int_{x \in D} f(x)\,\mathrm{d}x\, = 1.$$

The probability that X lies in the interval $[a, b]$ for some $a < b \in D$ is then given by

$$\Pr(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x\,.$$

Definition (Cumulative probability density function)**.** We define the *cumulative density* $F(x)$ by

$$F(x) = \Pr(X < x) = \int_{x_l}^{x} f(x')\,\mathrm{d}x'$$

where x_l is the infimum of the domain D.

Proposition. The probability density and the cumulative density are related so that

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

Proof. Consider the probability that X lies in some interval $[a, b]$.

$$\Pr(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x\, = \int_{x_l}^b f(x)\,\mathrm{d}x\, - \int_{x_l}^a f(x)\,\mathrm{d}x\, = F(b) - F(a). \quad \square$$

Proposition. A *discrete random variable* can be treated as continuous, with a probability density of the form

$$f(x) = \sum_i p_i \delta(x - x_i),$$

where $\delta(x - x_i)$ is the Dirac delta function.

Proof. Simply calculate the probability $\Pr(a < X \leq b)$:

$$\Pr(a < X \leq b) = \sum_i \int_a^b p_i \delta(x - x_i)\,\mathrm{d}x\, = \sum_i p_i,$$

where the final sum only includes values of x_i that lie between a and b.

1.4.3 Multiple random variables

We can consider multiple random variables. Generally, the variables may depend on each other and are described by a *joint probability density function.* Let X and Y be two random variables. If they are discrete, we have

$$\Pr(X = x_i, Y = y_i) = f(x_i, y_i),$$

and if they are continuous,

$$\Pr(x < X \leq x + \,\mathrm{d}x\,, y < Y \leq y + \,\mathrm{d}y\,) = f(x,y)\,\mathrm{d}x\,\,\mathrm{d}y\,.$$

If the two variables are independent, we may use equation (1.9), which implies

$$f(x,y) = g(x)h(y),$$

where $g(x)$ and $h(y)$ are density functions for the two variables.

1.5 Functions of random variables

Suppose x is some random variable for which the probability density function $f(x)$ is known. In many cases, we are more interested in a related random variable $y = y(x)$. What is the probability density function $g(y)$ for the new variable y?

1.5.1 Discrete random variables

If x is discrete, then we have $x \in \{x_1, x_2, \ldots, x_n\}$. In this case, y must also be discrete, and is given by $y_i = y(x_i)$.

If the function $y(x)$ is single-valued, then there exists an inverse $x(y)$ and the probability function becomes very simple:

$$g(y) = \begin{cases} f(x(y_i)) & \text{if } y = y_i, \\ 0 & \text{otherwise.} \end{cases}$$

The complications arise when the function $y(x)$ is not necessarily single-valued. In this case, we need to sum over all x_j such that $y_i = y(x_j)$. The probability function is

$$g(y) = \begin{cases} \sum_j f(x_j) & \text{if } y = y_i, \\ 0 & \text{otherwise.} \end{cases}$$

The sum over j is performed over all j such that $y_i = y(x_j)$.

1.5.2 Continuous random variables

The probability that y lies in the range $[y, y + \,\mathrm{d}y\,]$ is given by

$$g(y)\,\mathrm{d}y\, = \int_{\mathrm{d}S} f(x)\,\mathrm{d}x\,, \tag{1.19}$$

where $\mathrm{d}S\, = \{x \in D : y \leq y(x) \leq y + \,\mathrm{d}y\,\}$ corresponds to all values of x for which $y(x)$ lies in the range $[y, y + \,\mathrm{d}y\,]$. This is completely general.

Let's start with the single-valued case again. Then, we simply have

$$g(y)\,\mathrm{d}y\, = \int_{x(y)}^{x(y+\,\mathrm{d}y\,)} f(x)\,\mathrm{d}x\, = f(x(y))(x(y + \,\mathrm{d}y\,) - x(y))$$

$$\implies g(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| f(x(y)).$$

Generally, let $\{x_i\}$ be the set of x values which satisfy $y = y(x_i)$. Then, we have

$$\begin{aligned} g(y)\,\mathrm{d}y &= \sum_{x_i} \left| \int_{x_i(y)}^{x_i(y+\mathrm{d}y)} f(x)\,\mathrm{d}x \right| \\ \implies g(y) &= \sum_{x_i} \left|\frac{\mathrm{d}x_i}{\mathrm{d}y}\right| f(x_i(y)). \end{aligned}$$

1.6 Properties of distributions

We define some useful measures such as *expectation values* and *variances*. These give us useful information about the distributions.

Definition (Expectation value)**.** The *expectation value* $E[g(X)]$ of any function $g(X)$ of the random variable X is defined as

$$E[g(x)] = \int_{x\in D} g(x) f(x)\,\mathrm{d}x\,. \tag{1.20}$$

For discrete distributions, the integral is replaced by a sum so that

$$E[g(X)] = \sum_i f(x_i) g(x_i) = \sum_{x_i} p_i g(x_i). \tag{1.21}$$

The expectation value has the following properties:

1. if $a \in \mathbb{R}$, then $E[a] = a$,
2. for any $a \in \mathbb{R}$, $E[ag(x)] = aE[g(x)]$,
3. if $g(x) = s(x) + t(x)$, then $E[g(x)] = E[s(x)] + E[t(x)]$.

Definition (Mean)**.** The *mean* is the expectation value of the random variable and is given by

$$\mu := E[x] = \int_{x\in D} x f(x)\,\mathrm{d}x\,. \tag{1.22}$$

Definition (Mode)**.** The *mode* of a distribution is the value of the random variable x at which the distribution has its maximum value. There may be multiple modes.

Definition (Median)**.** The *median* of a distribution is the value of the random variable at which the cumulative distribution has value 1/2.

Definition (Upper and lower quartiles)**.** Given a cumulative distribution $F(x)$, the *upper and lower quartiles* q_u and q_l are defined such that

$$F(q_u) = \frac{3}{4} \quad F(q_l) = \frac{1}{4}.$$

Definition (Percentile)**.** The n^{th} percentile, P_n, of a distribution $f(x)$ is given in terms of the cumulative distribution $F(x)$ as follows:

$$F(P_n) = \frac{n}{100}.$$

Definition (Variance)**.** The *variance*, $V[x]$, of a distribution $f(x)$ is defined as

$$V[x] := E[(x-\mu)^2] = \int_{x\in D} (x-\mu)^2 f(x)\,\mathrm{d}x\,. \tag{1.23}$$

For a discrete distribution, we have

$$V[x] = \sum_i p_i (x_i - \mu)^2. \tag{1.24}$$

Note that the variance is a strictly positive quantity, since the integrand is always positive. It equals zero if and only if the probability of the mean is one.

Definition (Standard deviation)**.** The *standard deviation*, σ is defined as the positive square root of the variance,

$$\sigma := \sqrt{V[x]}. \tag{1.25}$$

The standard deviation gives a measure of how spread out the distribution is around the mean.

Proposition (Chebyshev's inequality)**.** The upper limit on the probability that random variable x takes values outside a given range centred on the mean is given by

$$\Pr(|x-\mu| \geq c) \leq \frac{\sigma^2}{c^2}, \tag{1.26}$$

for any $c > 0$.

Proof. First, we note that the probability is given by the integral

$$\Pr(|x-\mu| \geq c) = \int_{|x-\mu|\geq c} f(x)\,\mathrm{d}x\,.$$

Now, since the integrand in equation (1.23) is strictly positive, we have

$$\sigma^2 \geq \int_{|x-\mu|\geq c} (x-\mu)^2 f(x)\,\mathrm{d}x \;\geq c^2 \int_{|x-\mu|\geq c} f(x)\,\mathrm{d}x \;= c^2 \Pr(|x-\mu| \geq c).$$

The result follows. □

Definition (Moments)**.** The k^{th} *moment* of a distribution is defined as

$$\mu_k := E[x^k] = \int_{x\in D} x^k f(x)\,\mathrm{d}x\,. \tag{1.27}$$

Note that the mean is also the first moment.

Proposition. The variance, mean and the second moment of a distribution is related by

$$V[x] = E[x^2] - E[x]^2 = \mu_2 - \mu_1^2.$$

Proof. The variance is given by

$$V[x] = E[(x-\mu)^2] = \int f(x)(x^2 + \mu_2 - 2\mu x)\,\mathrm{d}x \;= \mu_2 - \mu_1^2 \quad \square.$$

Definition (Central moment)**.** The k^{th} *central moment* of a distribution is defined by

$$\nu_k := E[(x-\mu)^k] = \int_{x\in D} (x-\mu)^2 f(x)\,\mathrm{d}x\,. \tag{1.28}$$

We see that $\nu_1 = 0$ and $\nu_2 = \sigma^2$.

Definition (Normalized central moments)**.** The *normalized* or *standardized* moment of degree k is defined as

$$\gamma_k := \frac{\nu_k}{\sigma^k}. \tag{1.29}$$

Definition (Skewness)**.** The *skewness* of a distribution is equal to γ_3. It equals zero if the distribution is symmetric about its mean, it is negative if the distribution is skewed to values of x less than the mean, and positive otherwise.

Definition (Curtosis)**.** The *curtosis* of a distribution is given by γ_4. The curtosis of a Gaussian distribution is 3. So, we define *excess curtosis* as $\gamma_4 - 3$. A positive value of the excess kurtosis implies a relatively narrower peak and wider wings than the Gaussian distribution with the same mean and variance. A negative excess kurtosis implies a wider peak and shorter wings.

1.7 Important discrete distributions

1.7.1 The binomial distribution

The binomial distribution describes processes with two possible outcomes, A and $B = \overline{A}$. We call these *success* and *failure* respectively.

Given $\Pr(A) = p$, we deduce $\Pr(B) = 1-p$. If we perform n *trials*, then the discrete random variable

$$X = \text{the number of times A occurs,}$$

is described by the binomial distribution. So let's derive it.

The probability of obtaining k successes and $n-k$ failures, in that order is simply

$$p^k(1-p)^{n-k}.$$

This is a single permutation. Since we do not care about the ordering, the number of different ways we can obtain k successes is given by the combination $C(n,k)$. Hence, the probability of obtaining k successes from n trials, with success rate p is given by

$$\Pr(X = k) = \frac{n!}{k!(n-k)!}\, p^k(1-p)^{n-k} = B(k; p, n). \tag{1.30}$$

This is the ***binomial distribution.*** Now, let's look at some of its properties.

Notice that the distribution is normalized, since

$$\sum_{k=0}^{n} f(k) = \sum_{k=0}^{n} C(n,k)p^k(1-p)^{n-k} = (p + (1-p))^n = 1.$$

The first moment of the binomial distribution is given by

$$\mu = E[k] = \sum_{k=0}^{n} kC(n,k)p^k(1-p)^{n-k} = np.$$

The variance is

$$\sigma^2 = V[k] = np(1-p) \implies \sigma = \sqrt{np(1-p)}$$

1.7.2 Multinomial distributions

Instead of two outcomes, we can consider multiple. In that case, we can simply apply the binomial distribution multiple times - *sort of.*

Consider the case of 3 outcomes, the generalization to n outcomes is trivial. Let the three outcomes have probabilities p_1, p_2, p_3 respectively. What is the probability that out of n trials, we get k_1 of outcome 1 and k_2 of outcome 2?

Firstly, note that we must have $k_3 \equiv n - k_1 - k_2$ of outcome 3. Then, the probability for a single permutation is simply

$$p_1^{k_1} \times p_2^{k_2} \times p_3^{k_3}.$$

The question is, how many combinations are there? We can start by choosing k_1 from n and then k_2 from $n - k_1$. Equivalently, we can start by choosing k_3 out of n and k_2 from $n - k_3$, the order does not matter. The expression is

$$C(n, k_1) \times C(n - k_1, k_2) = \frac{n!}{k_1!(n - k_1)!} \times \frac{(n - k_1)!}{(k_2)!(k_3)!} = \frac{n!}{(k_1)!(k_2)!(k_3)!}.$$

Hence, the multinomial distribution for n outcomes is

$$M(k_i; p_i, n) = n! \prod_i \frac{p_i^{k_i}}{k_i!}. \tag{1.31}$$

1.7.3 Geometric and negative binomial distributions

We obtain the *geometric distribution* if we consider, as the random variable,

$$X = \text{number of trials required to obtain the first success.}$$

The probability that k trials are required to obtain the first success is

$$\Pr(X = k) = (1 - p)^{k-1} p,$$

and this is the geometric distribution.

Another random variable to consider is

$$X = \text{number of failures before the } r^{\text{th}} \text{ success.}$$

This yields the *negative binomial distribution.* What is the probability that $X = k$? We must have $r - 1$ successes and k failures, the number of ways to have this is simply $C(k + r - 1, k)$. The probability for a single permutation is $p^r(1 - p)^k$. Hence, the distribution is

$$\Pr(X = k) = C(k + r - 1, k) p^r (1 - p)^k = \frac{(k + r - 1)!}{k!(r - 1)!} p^r (1 - p)^k.$$

1.7.4 Hypergeometric distribution

So far, we have considered *independent trials.* What happens if we sample without replacement?

Consider drawing a random set of balls from a bag containing M red and $N - M$ white balls. What is the probability of drawing k red balls out of n draws? Now, notice that the problem is different since we are reducing the number of balls in the bag by drawing them. The random variable is

$$X = \text{number of red balls drawn.}$$

How many ways are there for drawing k red balls from a set of M? This is given by $C(M, k)$. Now, how many ways can we draw $n - k$ white balls from a set of $N - M$? Again, it is $C(N - M, n - k)$. The total number of ways to draw n balls is given by $C(N, n)$. Hence, the probability distribution is

$$\Pr(X = k) = \frac{C(M, k) C(N - M, n - k)}{C(N, n)} = \frac{(Np)!(Nq)!n!(N - n)!}{x!(Np - x)!(n - x)!(Nq - n + x)!N!}$$

where $p \equiv M/N$ and $q \equiv 1 - p$. This is the *hypergeometric distribution.*

1.7.5 Poisson distribution

The *Poisson distribution* describes the probability that exactly k events will occur in a given interval given a mean occurrence μ.

We treat the occurrence of an event as *success,* in which case it becomes obvious that we can use the binomial distribution. However, we cannot quantify the trial number n, since now we have a fixed average μ. We can take the limit as $n \to \infty$, with $np \to \mu$.

We start by letting $p = \mu/n$, and take the limit as $n \to \infty$.

$$\lim_{n\to\infty} B(k; \mu/n, n) = \lim_{n\to\infty} \frac{(\mu/n)^k (1-\mu/n)^{n-k} n!}{k!(n-k)!}.$$

In the limit, we have

$$\frac{n!}{(n-k)!} = n \times (n-1) \times \ldots \times (n-k+1) \to n^k.$$

Also,

$$\begin{aligned}\lim_{n\to\infty}(1-\mu/n)^{n-k} = \lim_{n\to\infty}(1-\mu/n)^n &= \lim_{n\to\infty}\sum_{i=0}^{n}(1)^{n-i}\left(-\frac{\mu}{n}\right)^i \frac{n!}{(n-i)!(i)!} \\ &= \lim_{n\to\infty}\sum_{i=0}^{n}\left(-\frac{\mu}{n}\right)^i \frac{n^i}{(i)!} \\ &= \sum_{i=0}^{\infty}\frac{(-\mu)^i}{(i)!} = \mathrm{e}^{-\mu}.\end{aligned}$$

Hence, we can evaluate the limit:

$$\lim_{n\to\infty} \frac{(\mu/N)^k (1-\mu/n)^{n-k} n!}{k!(n-k)!} = \frac{\mu^k \mathrm{e}^{-\mu}}{k!} \equiv P(k;\mu). \tag{1.32}$$

This is the ***Poisson distribution.***

We may check for normalization as follows:

$$1 \equiv \mathrm{e}^{-\mu}\mathrm{e}^{\mu} = \sum_{k=0}^{\infty}\frac{\mu^k}{k!}\mathrm{e}^{-\mu} = \sum_{k=0}^{\infty} P(k;\mu).$$

The mean is given by

$$\begin{aligned}E[x] = \sum_{x=0}^{\infty} xP(x;\mu) &= \sum_{x=0}^{\infty}\frac{x\mu^x \mathrm{e}^{-\mu}}{x!} \\ &= \mathrm{e}^{-\mu}\sum_{x=1}^{\infty}\frac{\mu_{x-1}\mu}{(x-1)!} \\ &= \mu\mathrm{e}^{-\mu}\mathrm{e}^{\mu} = \mu.\end{aligned}$$

For the variance, it is easiest to calculate the second moment.

$$\begin{aligned}E[x^2] = \sum_{x=0}^{\infty}\frac{x^2\mu^x\mathrm{e}^{-\mu}}{x!} &= \mu\mathrm{e}^{-\mu}\sum_{x=1}^{\infty}\frac{\mu^{x-1}}{(x-1)!}(x-1+1) \\ &= \mu\mathrm{e}^{-\mu}\left(\sum_{x=2}^{\infty}\frac{\mu^{x-2}\mu}{(x-2)!} + \sum_{x=1}^{\infty}\frac{\mu^{x-1}}{(x-1)!}\right) \\ &= \mu\mathrm{e}^{-\mu}\left(\mu\mathrm{e}^{\mu} + \mathrm{e}^{\mu}\right) = \mu^2 + \mu.\end{aligned}$$

The variance is then given by

$$V[x] = E[x^2] - E[x]^2 = \mu.$$

1.8 Important continuous distributions

1.8.1 Uniform distribution

The *uniform distribution* describes a continuous random variable that has a constant PDF over its interval. If $x \in [a, b]$, then the uniform distribution is given by

$$f(x) = \begin{cases} 1/(b-a), & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases} . \tag{1.33}$$

It is obviously normalized. The mean is given by

$$\mu = E[x] = \int_a^b \frac{x \,\mathrm{d}x}{(b-a)} = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2},$$

as we would expect. The second moment is given by

$$E[x^2] = \frac{1}{b-a} \int_a^b x^2 \,\mathrm{d}x = \frac{a^2 + b^2 + ab}{3}.$$

Hence, the variance is

$$V[x] = E[x^2] - E[x]^2 = \frac{(a-b)^2}{12}.$$

The standard deviation is

$$\sigma = \frac{b-a}{2\sqrt{3}}.$$

1.8.2 Exponential distribution

The *exponential distribution* describes the length of intervals between Poisson events, or equivalently, the distribution of the interval before the first event. So, let's derive it!

What is the probability that the first event occurs in the interval $[x, x + \mathrm{d}x\,]$? This is given by the probability that no events occur until x, times the probability that one event occurs in the interval. The probability that no events occur until x is

$$P(0; \lambda x) = \frac{(\lambda x)^0 \mathrm{e}^{-\lambda x}}{0!} = \mathrm{e}^{-\lambda x}$$

where λ is the *rate* of occurrence. The probability that one event will occur in interval $[x, x + \mathrm{d}x\,]$ is

$$P(1; \lambda \,\mathrm{d}x\,) = \frac{(\lambda \,\mathrm{d}x\,)^1 \mathrm{e}^{-\lambda \,\mathrm{d}x}}{1!} = \lambda \,\mathrm{d}x\,.$$

Hence, the probability that the first event occurs in the interval $[x, x + \mathrm{d}x\,]$ is

$$\lambda \mathrm{e}^{-\lambda x} \,\mathrm{d}x \equiv f(x) \,\mathrm{d}x\,,$$

where $f(x)$ is the ***exponential distribution.*** It is normalized:

$$\int_{x \in D} f(x) \,\mathrm{d}x = \int_0^\infty \lambda \mathrm{e}^{-\lambda x} \,\mathrm{d}x = 1,$$

where we note that the domain of x is $[0, \infty)$. The mean is given by

$$\mu = E[x] = \int_0^\infty x \lambda \mathrm{e}^{-\lambda x} \,\mathrm{d}x = \frac{1}{\lambda}.$$

The second moment is

$$E[x^2] = \int_0^\infty x^2 \lambda \mathrm{e}^{-\lambda x} \,\mathrm{d}x = \frac{2}{\lambda^2}.$$

Therefore, the variance and the standard deviation are

$$V[x] = \frac{1}{\lambda^2}, \quad \sigma = \frac{1}{\lambda}.$$

1.8.3 Gamma distribution

We may generalize the exponential distribution to consider the interval between every r^{th} Poisson events, or the interval until the r^{th} Poisson event.

The probability that $r-1$ events occur until x, given a rate λ, is

$$P(r-1;\lambda x) = \frac{(\lambda x)^{r-1}\mathrm{e}^{-\lambda x}}{(r-1)!}.$$

The probability that an event will occur in the interval $[x, x+\,\mathrm{d}x\,]$ is $\lambda\,\mathrm{d}x$, so the probability that the r^{th} event will occur at x is

$$\lambda\,\mathrm{d}x\,P(r-1;\lambda x) = \lambda\,\mathrm{d}x\,\frac{(\lambda x)^{r-1}\mathrm{e}^{-\lambda x}}{(r-1)!} = f(x)\,\mathrm{d}x\,,$$

where $f(x)$ is the ***gamma distribution*** of order r with parameter λ.

The mean and the variance are

$$E[x] = \frac{r}{\lambda}, \quad V[x] = \frac{r}{\lambda^2}.$$

1.8.4 Gaussian distribution

This is the most important distribution, due to the *central limit theorem.* The Gaussian probability density for the random variable x, with mean μ and standard deviation σ is

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \tag{1.34}$$

where $x \in \mathbb{R}$. The factors in front ensure it is normalised

$$\int_{-\infty}^{\infty} G(x;\mu,\sigma)\,\mathrm{d}x\, = 1, \quad \forall\ \mu,\sigma.$$

Notice that changing μ simply shifts the curve along the x-axis, and changing σ broadens or narrows the curve. So, it may be more convenient to consider the standard form by defining a random variable $Z = (x-\mu)/\sigma$ - called the *standard score.* In that case, $\mathrm{d}x\, = \sigma\,\mathrm{d}Z\,$, therefore

$$G(Z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-Z^2/2},$$

which is the *standard Gaussian distribution.*

We define the cumulative probability density for a Gaussian distribution as

$$F(u) = \Pr(x<u) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{u}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x\,. \tag{1.35}$$

This integral cannot be evaluated analytically. We use tables of values of the cumulative probability function for the standard Gaussian distribution:

$$\Phi(z) = \Pr(Z<z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\mathrm{e}^{-Z^2/2}\,\mathrm{d}Z\,. \tag{1.36}$$

We may also use the *error function*, defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_{0}^{x}\mathrm{e}^{-u^2}\,\mathrm{d}u\,. \tag{1.37}$$

1.9 The central limit theorem

The *central limit theorem* states that the average of N independent random variables will tend to have a Gaussian PDF as N gets large. Note that the probability densities of the individual random variables does not matter!

Theorem (Central limit theorem)**.** Let $\{x_1, x_2, \ldots, x_N\}$ be N *independent* random variables with probability densities $\rho_i(x)$. Define a new random variable X such that

$$X := \frac{\sum_{i=1}^{N} x_i}{N}$$

is the mean of the x_i. The *central limit theorem* states that the random variable X has the following properties:

(i) its expectation value is given by $E[X] = (\sum_i \mu_i)/N$,

(ii) its variance is given by $V[X] = \sum_i \sigma_i^2/N^2$,

(iii) as $n \to \infty$, the probability density of X tends to a Gaussian with corresponding mean and variance.

The first two statements are straightforward to prove, the second one requiring x_i to be independent. The last statement is the most important, and can be proven by considering the moment generating functions.