17. Independence and conditioning of events

Definition 112. Let A, B be two events in the same probability space.

(1) If $\mathbf{P}(B) \neq 0$, we define the *conditional probability of A given B* as

$$\mathbf{P}\left(A \,\middle|\, B\right) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

(2) We say that A and B are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. If $\mathbf{P}(B) \neq 0$, then A and B are independent if and only if $\mathbf{P}(A \,|\, B) = \mathbf{P}(A)$ (and similarly with the roles of A and B reversed). If $\mathbf{P}(B) = 0$, then A and B are necessarily independent since $\mathbf{P}(A \cap B)$ must also be 0.

What do these notions mean intuitively? In real life, we keep updating probabilities based on information that we get. For example, when playing cards, the chance that a randomly chosen card is an ace is $1/13$, but having drawn a card, the probability for the next card may not be the same - if the first card was seen to be an ace, then the chance of the second being an ace falls to $3/51$. This updated probability is called a conditional probability. Independence of two events A and B means that knowing whether or not A occured does not change the chance of occurrence of B. In other words, the conditional probability of A given B is the same as the unconditional (original) probability of A.

Example 113. Let $\Omega = \{(i,j) : 1 \leq i, j \leq 6\}$ with $p_{(i,j)} = \frac{1}{36}$. This is the probability space corresponding to a throw of two fair dice. Let $A = \{(i,j) : i \text{ is odd}\}$ and $B = \{(i,j) : j \text{ is 1 or 6}\}$ and $C = \{(i,j) : i + j = 4\}$. Then $A \cap B = \{(i,j) : i = 1, 3, \text{ or } 5, \text{ and } j = 1 \text{ or } 6\}$. Then, it is easy to see that

$$\mathbf{P}(A \cap B) = \frac{6}{36} = \frac{1}{6}, \;\; \mathbf{P}(A) = \frac{18}{36} = \frac{1}{2}, \;\; \mathbf{P}(B) = \frac{12}{36} = \frac{1}{3}.$$

In this case, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ and hence A and B are independent. On the other hand,

$$\mathbf{P}(A \cap C) = \mathbf{P}\{(1,3),(2,2)\} = \frac{1}{18}, \;\; \mathbf{P}(C) = \mathbf{P}\{(1,3),(2,2),(3,1)\} = \frac{1}{12}.$$

Thus, $\mathbf{P}(A \cap C) \neq \mathbf{P}(A)\mathbf{P}(C)$ and hence A and C are not independent.

This agrees with the intuitive understanding of independence, since A is an event that depends only on the first toss and B is an event that depends only on the second toss. Therefore, A and B ought to be independent. However, C depends on both tosses, and hence cannot be expected to be independent of A. Indeed, it is easy to see that $\mathbf{P}(C \,|\, A) = \frac{1}{9}$.

Example 114. Let $\Omega = S_{52}$ with $p_\pi = \frac{1}{52!}$. Define the events

$$A = \{\pi : \pi_1 \in \{10,20,30,40\}\}, \qquad A = \{\pi : \pi_2 \in \{10,20,30,40\}\}.$$

Then both $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{13}$. However, $\mathbf{P}(B \,|\, A) = \frac{3}{51}$. One can also see that $\mathbf{P}(B \,|\, A^c) = \frac{4}{51}$.

In words, A (respectively B) could be the event that the first (respectively second) card is an ace. Then $\mathbf{P}(B) = 4/52$ to start with. When we see the first card, we update the probability. If the first card was not an ace, we update it to $\mathbf{P}(B \,|\, A^c)$ and if the first card was an ace, we update it to $\mathbf{P}(B \,|\, A)$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$

equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

Exercise 115. If A and B are independent, show that the following pairs of events are also independent.

(1) A and B^c.
(2) A^c and B.
(3) A^c and B^c.

Total probability rule and Bayes' rule: Let $A_1, \ldots, A_n$ be pairwise disjoint and mutually exhaustive events in a probability space. Assume $\mathbf{P}(A_i) > 0$ for all i. This means that $A_i \cap A_j = \emptyset$ for any $i \neq j$ and $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$. We also refer to such a collection of events as a partition of the sample space.

Let B be any other event.

(1) (Total probability rule). $\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)$.
(2) (Bayes' rule). Assume that $\mathbf{P}(B) > 0$. Then, for each $k = 1, 2, \ldots, n$, we have

$$\mathbf{P}(A_k \mid B) = \frac{\mathbf{P}(A_k)\mathbf{P}(B \mid A_k)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}.$$

PROOF. The proof is merely by following the definition.

(1) The right hand side is equal to

$$\mathbf{P}(A_1)\frac{\mathbf{P}(B \cap A_1)}{\mathbf{P}(A_1)} + \ldots + \mathbf{P}(A_n)\frac{\mathbf{P}(B \cap A_n)}{\mathbf{P}(A_n)} = \mathbf{P}(B \cap A_1) + \ldots + \mathbf{P}(B \cap A_n)$$

which is equal to $\mathbf{P}(B)$ since A_i are pairwise disjoint and exhaustive.
(2) Without loss of generality take $k = 1$. Note that $\mathbf{P}(A_1 \cap B) = \mathbf{P}(A_1)\mathbf{P}(B \cap A_1)$. Hence

$$\begin{aligned}\mathbf{P}(A_1 \mid B) &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \ldots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n)}\end{aligned}$$

where we used the total probability rule to get the denominator. ■

Exercise 116. Suppose A_i are events such that $\mathbf{P}(A_1 \cap \ldots \cap A_n) > 0$. Then show that

$$\mathbf{P}(A_1 \cap \ldots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2)\ldots\mathbf{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

Example 117. Consider a rare disease X that affects one in a million people. A medical test is used to test for the presence of the disease. The test is 99% accurate in the sense that if a person has no disease, the chance that the test shows positive is 1% and if the person has disease, the chance that the test shows negative is also 1%.

Suppose a person is tested for the disease and the test result is positive. What is the chance that the person has the disease X?

Let A be the event that the person has the disease X. Let B be the event that the test shows positive. The given data may be summarized as follows.

(1) $\mathbf{P}(A) = 10^{-6}$. Of course $\mathbf{P}(A^c) = 1 - 10^{-6}$.
(2) $\mathbf{P}(B \mid A) = 0.99$ and $\mathbf{P}(B \mid A^c) = 0.01$.

What we want to find is $\mathbf{P}(A\,|\,B)$. By Bayes' rule (the relevant partition is $A_1 = A$ and $A_2 = A^c$),

$$\mathbf{P}(A\,|\,B) = \frac{\mathbf{P}(B\,|A)\mathbf{P}(A)}{\mathbf{P}(B\,|A)\mathbf{P}(A)+\mathbf{P}(B\,|A^c)\mathbf{P}(A^c)} = \frac{0.99\times 10^{-6}}{0.99\times 10^{-6}+0.01\times(1-10^{-6})} = 0.000099.$$

The test is quite an accurate one, but the person tested positive has a really low chance of actually having the disease! Of course, one should observe that the chance of having disease is now approximately 10^{-4} which is considerably higher than 10^{-6}.

Independence of three or more events:

Definition 118. Events $A_1,\ldots,A_n$ in a common probability space are said to be independent if $\mathbf{P}(A_{i_1}\cap A_{i_2}\cap\ldots\cap A_{i_m}) = \mathbf{P}(A_{i_1})\mathbf{P}(A_{i_2})\ldots\mathbf{P}(A_{i_m})$ for every choice of $m \le n$ and every choice of $1 \le i_1 < i_2 < \ldots < i_m \le n$.

The independence of n events requires us to check 2^n equations (that many choices of $i_1, i_2, \ldots$). Should it not suffice to check that each pair of A_i and A_j are independent? The following example shows that this is not the case!

Example 119. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega} \in \Omega$. Define the events $A = \{\underline{\omega} : \omega_1 = 0\}$, $A = \{\underline{\omega} : \omega_2 = 0\}$ and $C = \{\underline{\omega} : \omega_1 + \omega_2 = 0 \text{ or } 2\}$. In words, we toss a fair coin n times and A denotes the event that the first toss is a tail, B denotes the event that the second toss is a tail and C denotes the event that out of the first two tosses are both heads or both tails. Then $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{4}$. Further,

$$\mathbf{P}(A\cap B) = \frac{1}{4},\ \mathbf{P}(B\cap C) = \frac{1}{4},\ P(A\cap C) = \frac{1}{4},\ \mathbf{P}(A\cap B\cap C) = \frac{1}{4}.$$

Thus, A, B, C are independent *pairwise*, but not independent by our definition because $\mathbf{P}(A\cap B\cap C) \neq \frac{1}{8} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$.

Intuitively this is right. Knowing A does not given any information about C (similarly with A and B or B and C), but knowing A and B tells us completely whether or not C occurred! Thus is is right that the definition should not declare them to be independent.

Exercise 120. Let $A_1,\ldots,A_n$ be events in a common probability space. Then, $A_1, A_2, \ldots, A_n$ are independent if and only if the following equalities hold.

For each i, define B_i as A_i and A_i^c. Then

$$\mathbf{P}(B_1\cap B_2\cap\ldots\cap B_n) = \mathbf{P}(B_1)\mathbf{P}(B_2)\ldots\mathbf{P}(B_n).$$

Note: This should hold for any possible choice of B_is. In other words, the system of 2^n equalities in the definition of independence may be replaced by this new set of 2^n equalities. The latter system has the advantage that it immediately tells us that if $A_1,\ldots,A_n$ are independent, then $A_1, A_2^c, A_3, \ldots$ (for each i choose A_i or its complement) are independent.

18. Independence and conditioning of random variables

Definition 121. Let $\mathbf{X} = (X_1,\ldots,X_m)$ be a random vector (this means that X_i are random variables on a common probability space). We say that X_i are *independent* if $F_{\mathbf{X}}(t_1,\ldots,t_m) = F_1(t_1)\ldots F_m(t_m)$ for all $t_1,\ldots,t_m$.

Remark 122. Recalling the definition of independence of events, the equality $F_{\mathbf{X}}(t_1,\dots,t_m) = F_1(t_1)\dots F_m(t_m)$ is just saying that the events $\{X_1 \le t_1\},\dots\{X_m \le t_m\}$ are independent. More generally, it is true that $X_1,\dots,X_m$ are independent if and only if $\{X_1 \in A_1\},\dots,\{X_m \in A_m\}$ are independent events for any $A_1,\dots,A_m \subseteq \mathbb{R}$.

Remark 123. In case $X_1,\dots,X_m$ have a joint pmf or a joint pdf (which we denote by $f(t_1,\dots,t_m)$), the condition for independence is equivalent to

$$f(t_1,\dots,t_m) = f_1(t_1)f_2(t_2)\dots f_m(t_m)$$

where f_i is the marginal density (or pmf) of X_i. This fact can be derived from the definition easily. For example, in the case of densities, observe that

$$\begin{aligned} f(t_1,\dots,t_m) &= \frac{\partial^m}{\partial t_1\dots\partial t_m} F(t_1,\dots,t_m) \quad \text{(true for any joint density)} \\ &= \frac{\partial^m}{\partial t_1\dots\partial t_m} F_1(t_1)\dots F_m(t_m) \quad \text{(by independence)} \\ &= F_1'(t_1)\dots F_m'(t_m) \\ &= f_1(t_1)\dots f_m(t_m). \end{aligned}$$

When we turn it around, this gives us a quicker way to check independence.

Fact: Let $X_1,\dots,X_m$ be random variables with joint pdf $f(t_1,\dots,t_m)$. Suppose we can write this pdf as $f(t_1,\dots,t_m) = cg_1(t_1)g_2(t_2)\dots g_m(t_m)$ where c is a constant and g_i are some functions of one-variable. Then, $X_1,\dots,X_m$ are independent. Further, the marginal density of X_k is $c_k g_k(t)$ where $c_k = \frac{1}{\int_{-\infty}^{+\infty} g_k(s)ds}$. An analogous statement holds when $X_1,\dots,X_m$ have a joint pmf instead of pdf.

Example 124. Let $\Omega = \{0,1\}^n$ with $p_{\underline{\omega}} = p^{\sum \omega_k} q^{n-\sum \omega_k}$. Define $X_k : \Omega \to \mathbb{R}$ by $X_k(\underline{\omega}) = \omega_k$. In words, we are considering the probability space corresponding to n tosses of a fair coin and X_k is the result of the kth toss. We claim that $X_1,\dots,X_n$ are independent. Indeed, the joint pmf of $X_1,\dots,X_n$ is

$$f(t_1,\dots,t_n) = p^{\sum t_k} q^{n-\sum t_k} \quad \text{where } t_i = 0 \text{ or } 1 \text{ for each } i \le n.$$

Clearly $f(t_1,\dots,t_m) = g(t_1)g(t_2)\dots g(t_n)$ where $g(s) = p^s q^{1-s}$ for $s = 0$ or 1 (this is just a terse way of saying that $g(s) = p$ if $s = 1$ and $g(s) = q$ if $s = 0$). Hence $X_1,\dots,X_n$ are independent and X_k has pmf g (i.e., $X_k \sim \text{Ber}(p)$).

Example 125. Let (X,Y) have the bivariate normal density

$$f(x,y) = \frac{\sqrt{ab-c^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}(a(x-\mu_1)^2 + b(y-\mu_2)^2 + 2c(x-\mu_1)(y-\mu_2))}.$$

If $c = 0$, we observe that

$$f(x,y) = C_0 e^{-\frac{a(x-\mu_1)^2}{2}} e^{-\frac{b(y-\mu_2)^2}{2}} \qquad (C_0 \text{ is a constant, exact value unimportant})$$

from which we deduce that X and Y are independent and $X \sim N(\mu_1, \frac{1}{a})$ while $Y \sim N(\mu_2, \frac{1}{b})$.

Can you argue that if $c \neq 0$, then X and Y are not independent?

Example 126. Let (X,Y) be a random vector with density $f(x,y) = \frac{1}{\pi}\mathbf{1}_{x^2+y^2\le 1}$ (i.e., it equals 1 if $x^2+y^2 \le 1$ and equals 0 otherwise). This corresponds to picking a point at random from the disk of radius 1 centered at $(0,0)$. We claim that X and Y are not independent. A quick way to see this is that if $I = [0.8,1]$, then $\mathbf{P}\{(X,Y) \in [0.8,1]\times[0.8,1]\} = 0$ whereas

$\mathbf{P}\{(X,Y) \in [0.8,1]\}\mathbf{P}\{(X,Y) \in [0.8,1]\} \neq 0$ (If X,Y were independent, we must have had $\mathbf{P}\{(X,Y) \in [a,b] \times [c,d]\} = \mathbf{P}\{X \in [a,b]\}\mathbf{P}\{Y \in [c,d]\}$ for any $a < b$ and $c < d$).

A very useful (and intuitively acceptable!) fact about independence is as follows.

Fact: Suppose $X_1, \ldots, X_n$ are independent random variables. Let $k_1 < k_2 < \ldots < k_m = n$. Let $Y_1 = h_1(X_1, \ldots, X_{k_1})$, $Y_2 = h_2(X_{k_1+1}, \ldots, X_{k_2}), \ldots Y_m = h_m(X_{k_{m-1}}, \ldots, X_{k_m})$. Then, $Y_1, \ldots, Y_m$ are also independent.

Remark 127. In the previous section we defined independence of events and now we have defined independence of random variables. How are they related? We leave it to you to check that events $A_1, \ldots, A_n$ are independent (according the definition of the previous section) if and only if the random variables $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_m}$ are independent (according the definition of this section)

The next part, about conditioning on random variables and conditional densities was not covered in class and is not included in syllabus.

Conditioning on random variables: Let $X_1, \ldots, X_{k+\ell}$ be random variables on a common probability space. Let $f(t_1, \ldots, t_{k+\ell})$ be the pmf of $(X_1, \ldots, X_{k+\ell})$ and let $g(t_1, \ldots, t_\ell)$ be the pmf of $(X_{k+1}, \ldots, X_{k+\ell})$ (of course we can compute g from f by summing over the first k indices). Then, for any $s_1, \ldots, s_\ell$ such that $\mathbf{P}\{X_{k+1} = s_1, \ldots X_m = s_\ell\} > 0$, we can define

(2)
$$h_{s_1,\ldots,s_\ell}(t_1, \ldots, t_k) = \mathbf{P}\{X_1 = t_1, \ldots, X_k = t_k \,\big|\, X_{k+1} = s_1, \ldots X_m = s_\ell\} = \frac{f(t_1, \ldots, t_k, s_1, \ldots, s_\ell)}{g(s_1, \ldots, s_\ell)}.$$

It is easy to see that $h_{s_1,\ldots,s_\ell}(\cdot)$ is a pmf on $\mathbb{R}^k$. It is called the conditional pmf of $(X_1, \ldots, X_k)$ given that $X_{k+1} = s_1, \ldots X_m = s_\ell$.

Its interpretation is as follows. Originally we had random observables $X_1, \ldots, X_k$ which had a certain joint pmf. Then we observe the values of the random variables $X_{k+1}, \ldots, X_{k+\ell}$, say they turn out to be $s_1, \ldots, s_\ell$, respectively. Then we update the distribution (or pmf) of $X_1, \ldots, X_k$ according to the above recipe. The conditional pmf is the new function $h_{s_1,\ldots,s_\ell}(\cdot)$.

Exercise 128. Let $(X_1, \ldots, X_{n-1})$ be a random vector with multinomial distribution with parameters $r, n, p_1, \ldots, p_n$. Let $k < n-1$. Given that $X_{k+1} = s_1, \ldots, X_{n-1} = s_{n-k+1}$, show that the conditional distribution of $(X_1, \ldots, X_k)$ is multinomial with parameters $r', n', q_1, \ldots, q_{k+1}$ where $r' = r - (s_1 + \ldots + s_{n-k+1})$, $n' = k+1$, $q_j = p_j/(p_1 + \ldots + p_k + p_n)$ for $j \leq k$ and $q_{k+1} = p_n/(p_1 + \ldots + p_k + p_n)$.

This looks complicated, but is utterly obvious if you think in terms of assigning r balls into n urns by putting each ball into the urns with probabilities $p_1, \ldots, p_n$ and letting X_j denote the number of balls that end up in the j^{th} urn.

Conditional densities Now suppose $X_1, \ldots, X_{k+\ell}$ have joint density $f(t_1, \ldots, t_{k+\ell})$ and let $g(s_1, \ldots, s_\ell)$ by the density of $(X_{k+1}, \ldots, X_{k+\ell})$. Then, we define the conditional density of $(X_1, \ldots, X_k)$ given $X_{k+1} = s_1, \ldots, X_{k+\ell} = s_\ell$ as

$$h_{s_1,\ldots,s_\ell}(t_1, \ldots, t_k) = \frac{f(t_1, \ldots, t_k, s_1, \ldots, s_\ell)}{g(s_1, \ldots, s_\ell)}. \tag{3}$$

This is well-defined whenever $g(s_1, \ldots, s_\ell) > 0$.

Remark 129. Note the difference between (2) and (3). In the latter we have left out the middle term because $\mathbf{P}\{X_{k+1} = s_1, \ldots, X_{k+\ell} = s_\ell\} = 0$. In (2) the definition of pmf comes from the definition of conditional probability of events but in (3) this is not so. We simply define the conditional density by analogy with the case of conditional pmf. This is similar to the difference between interpretation of pmf ($f(t)$ is actually the probability of an event) and pdf ($f(t)$ is not the probability of an event but the density of probability near t).

Example 130. Let (X,Y) have bivariate normal density $f(x,y) = \frac{\sqrt{ab-c^2}}{2\pi} e^{-\frac{1}{2}(ax^2+by^2+2cxy)}$ (so we assume $a > 0, b > 0, ab - c^2 > 0$). In the mid-term you showed that the marginal distribution of Y is $N(0, \frac{a}{ab-c^2})$, that is it has density $g(y) = \frac{\sqrt{ab-c^2}}{\sqrt{2\pi a}} e^{-\frac{ab-c^2}{2a}y^2}$. Hence, the conditional density of X given $Y = y$ is

$$h_y(x) = \frac{f(x,y)}{g(y)} = \frac{\sqrt{a}}{\sqrt{2\pi}} e^{-\frac{a}{2}(x+\frac{c}{a}y)^2}.$$

Thus the conditional distribution of X given $Y = y$ is $N(-\frac{cy}{a}, \frac{1}{a})$. Compare this with marginal (unconditional) distribution of X which is $N(0, \frac{b}{ab-c^2})$.

In the special case when $c = 0$, we see that for any value of y, the conditional distribution of X given $Y = y$ is the same as the unconditional distribution of X. What does this mean? It is just another way of saying that X and Y are independent! Indeed, when $c = 0$, the joint density $f(x,y)$ splits into a product of two functions, one of x alone and one of y alone.

Exercise 131. Let (X,Y) have joint density $f(x,y)$. Let the marginal densities of X and Y be $g(x)$ and $h(y)$ respectively. Let $h_x(y)$ be the conditional density of Y given $X = x$.

(1) If X and Y are independent, show that for any x, we have $h_x(y) = h(y)$ for all y.
(2) If $h_x(y) = h(y)$ for all y and for all x, show that X and Y are independent.

Analogous statements hold for the case of pmf.

19. Mean and Variance

Let X be a random variable with distribution F. We shall assume that it has pmf or pdf denoted by f.

Definition 132. The *expected value* (also called *mean*) of X is defined as the quantity $\mathbf{E}[X] = \sum_t t f(t)$ if f is a pmf and $\mathbf{E}[X] = \int_{-\infty}^{+\infty} t f(t) dt$ if f is a pdf (provided the sum or the integral converges absolutely).

Note that this agrees with the definition we gave earlier for random variables with pmf. it is possible to define expect value for distributions without pmf or pdf, but we shall not do it here.

Properties of expectation: Let X, Y be random variables both having pmf f, g or pdf f, g, respectively.

(1) Then, $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$ for any $a, b \in \mathbb{R}$. In particular, for a constant random variable (i.e., $X = a$ with probability 1 for some a, $\mathbf{E}[X] = a$). This is called *linearity* of expectation.
(2) If $X \geq Y$ (meaning, $X(\omega) \geq Y(\omega)$ for all ω), then $\mathbf{E}[X] \geq \mathbf{E}[Y]$
(3) If $\varphi : \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}[\varphi(X)] = \begin{cases} \sum_t \varphi(t) f(t) & \text{if } f \text{ is a pmf.} \\ \int_{-\infty}^{+\infty} \varphi(t) f(t) dt & \text{if } f \text{ is a pdf.} \end{cases}$$

(4) More generally, if $(X_1, \ldots, X_n)$ has joint pdf $f(t_1, \ldots, t_n)$ and $V = T(X_1, \ldots, X_n)$ (here $T : \mathbb{R}^n \to \mathbb{R}$), then $\mathbf{E}[V] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} T(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \ldots dx_n$.

For random variables on a discrete probability space (then they have pmf), we have essentially proved all these properties (or you can easily do so). For random variables with pmf, a proper proof require a bit of work. So we shall just take these for granted. We state one more property of expectations, its relationship to independence.

Lemma 133. *Let X, Y be random variables on a common probability space. If X and Y are independent, then $\mathbf{E}[H_1(X)H_2(Y)] = \mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$ for any functions $H_1, H_2 : \mathbb{R} \to \mathbb{R}$ (for which the expectations make sense). In particular, $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$.*

PROOF. Independence means that the joint density (analogous statements for pmf omitted) of (X, Y) is f the form $f(t, s) = g(t)h(s)$ where $g(t)$ is the density of X and $h(s)$ is the density of Y. Hence,

$$\mathbf{E}[H_1(X)H_2(Y)] = \iint H_1(t)H_2(s)f(t,s)dtds = \left(\int_{-\infty}^{\infty} H_1(t)g(t)dt \right) \left(\int_{-\infty}^{\infty} H_2(s)g(s)ds \right)$$

which is precisely $\mathbf{E}[H_1(X)]\mathbf{E}[H_2(Y)]$. ∎

Expectation is a very important quantity. Using it, we can define several other quantities of interest.

Discussion: For simplicity let us take random variables to have densities in this discussion. You may adapt the remarks to the case of pmf easily. The density has all the information we need about a random variable. However, it is a function, which means that we have to know $f(t)$ for every t. In real life often we have random variables whose pdf is unknown

or impossible to determine. It would be better to summarize the main features of the distribution (i.e., the density) in a few numbers. That is what the quantities defined below try to do.

Mean: Mean is another term for expected value.

Quantiles: Let us assume that the CDF F of X is strictly increasing and continuous. Then $F^{-1}(t)$ is well defined for every $t \in (0,1)$. For each $t \in (0,1)$, the number $Q_t = F^{-1}(t)$ is called the t-quantile. For example, the $1/2$-quantile, also called *median* is the number x such that $F(x) = \frac{1}{2}$ (unique when the CDF is strictly increasing and continuous). Similarly one defines $1/4$-quantile and $3/4$-quantile and these are sometimes called quartiles.[12]

Moments: The quantity $\mathbf{E}[X^k]$ (if it exists) is called the k^{th} *moment* of X.

Variance: Let $\mu = \mathbf{E}[X]$ and define $\sigma^2 := \mathbf{E}\left[(X-\mu)^2\right]$. This is called the *variance* of X, also denoted by $\text{Var}(X)$. It can be written in other forms. For example,

$$\begin{aligned}\sigma^2 &= \mathbf{E}[X^2 + \mu^2 - 2\mu X] \qquad \text{(by expanding the square)}\\ &= \mathbf{E}[X^2] + \mu^2 - 2\mu\mathbf{E}[X] \qquad \text{(by property (1) above)}\\ &= \mathbf{E}[X^2] - \mu^2.\end{aligned}$$

That is $\text{Var}(X) = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$.

Standard deviation: The standard deviation of X is defined as $\text{s.d.}(X) := \sqrt{\text{Var}(X)}$.

Mean absolute deviation: The mean absolute deviation of X is defined as the $\mathbf{E}[|X - \text{med}(X)|]$.

Coefficient of variation: The coefficient of variation of X is defined as $\text{c.v.}(X) = \frac{\text{s.d.}(X)}{|\mathbf{E}[X]|}$.

Covariance: Let X, Y be random variables on a common probability space. The *covariance* of X and Y is defined as $\text{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. It can also be written as $\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$.

Correlation: Let X, Y be random variables on a common probability space. Their *correlation* is defined as $\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$.

[12] Another familiar quantity is the percentile, frequently used in reporting performance in competitive exams. For each x, the x-percentile is nothing but $F(x)$. For exam scores, it tells the proportion of exam-takers who scored less than or equal to x.

Entropy: The entropy of a random variable X is defined as

$$\text{Ent}(X) = \begin{cases} -\sum_i f(t)\log(f(t_i)) & \text{if } X \text{ has pmf } f. \\ -\int f(t)\log(f(t)) & \text{if } X \text{ has pdf} f. \end{cases}$$

If $\mathbf{X} = (X_1, \ldots, X_n)$ is a random vector, we can define its entropy exactly by the same expressions, except that we use the joint pmf or pdf of $\mathbf{X}$ and the sum or integral is over points in $\mathbb{R}^n$.

Discussion: What do these quantities mean?

Measures of central tendency Mean and median try to summarize the distribution of X by a single number. Of course one number cannot capture the whole distribution, so there are many densities and mass functions that have the same mean or median. Which is better - mean or median? This question has no unambiguous answer. Mean has excellent mathematical properties (mainly linearity) which the median lacks ($\text{med}(X+Y)$ bears no general relationship to $\text{med}(X)+\text{med}(Y)$). In contrast, mean is sensitive to outliers, while the median is far less so. For example, if the average income in a village of 50 people is 1000 Rs. per month, the immigration of multi-millionaire to the village will change the mean drastically but the median remains about the same. This is good, if by giving one number we are hoping to express the state of a typical individual in the population.

Measures of dispersion: Suppose the average height of people in a city is 160 cm. This could be because everyone is 160 cm exactly or because half the people are 100 cm. while the other half are 220 cm., or alternately the heights could be uniformly spread over 150-170 cm., etc. How widely the distribution is spread is measured by standard deviation and mean absolute deviation. Since we want deviation from mean, $\mathbf{E}[X - \mathbf{E}[X]]$ looks natural, but this is zero because of cancellation of positive and negative deviations. To prevent cancellation, we may put absolute values (getting to the m.a.d, but that is usually taken around the median) or we may square the deviations before taking expectation (giving the variance, and then the standard deviation). Variance and standard deviation have much better mathematical properties (as we shall see) and hence are usually preferred.

The standard deviation has the same units as the quantity. Fo example, if mean height is 160cm measured in centimeters with a standard deviation of 10cm, and the mean weight is 55kg with a standard deviation of 5kg, then we cannot say which of the two is less variable. To make such a comparison we need a dimension free quantity (a pure number). Coefficient of variation is such a quantity, as it measure the standard deviation per mean. For the height and weight data just described, the coefficients of variation are 1/16 and 1/11, respectively. Hence we may say that height is less variable than weight in this example.

Measures of association: The marginal distributions do not determine the joint distribution. For example, if (X,Y) is a point chosen at random from the unit square (with vertices $(0,0),(1,0),(0,1),(1,1)$) then X,Y both have marginal distribution that is uniform on $[0,1]$. If (U,V) is a point picked at random from the diagonal line (the line segment from $(0,0)$ to $(1,1)$, then again U and V have marginals that are uniform on $[0,1]$. But the two joint distributions are completely different. In particular, giving the means and standard deviations of X and Y does not tell anything about possible relationships between the two.

Covariance is the quantity that is used to measure the "association" of Y and X. Correlation is a dimension free quantity that measures the same. For example, we shall see that if $Y = X$, then $\text{Corr}(X,Y) = +1$, if $Y = -X$ then $\text{Corr}(X,Y) = -1$. Further, if X and Y are independent, then $\text{Corr}(X,Y) = 0$. In general, if an increase in X is likely to mean an increase in Y, then the correlation is positive and if an increase in X is likely to mean a decrease in Y then the correlation is negative.

Example 134. Let $X \sim N(\mu,\sigma^2)$. Recall that its density is $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We can compute

$$\mathbf{E}[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \mu.$$

On the other hand

$$\begin{aligned}
\text{Var}(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \;=\; \sigma^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}}\, du \qquad (\text{substitute } x = \mu + \sigma u) \\
&= \sigma^2 \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} u^2 e^{-\frac{u^2}{2}}\, du \;=\; \sigma^2 \frac{2\sqrt{2}}{\sqrt{2\pi}} \int_{0}^{+\infty} \sqrt{t} e^{-t}\, dt \qquad (\text{substitute } t = u^2/2) \\
&= \sigma^2 \frac{2\sqrt{2}}{\sqrt{2\pi}} \Gamma(3/2) \;=\; \sigma^2.
\end{aligned}$$

To get the last line, observe that $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2)$ and $\Gamma(1/2) = \sqrt{\pi}$. Thus we now have a meaning for the parameters μ and σ^2 - they are the mean and variance of the $N(\mu,\sigma^2)$ distribution. Again note that the mean is the same for all $N(0,\sigma^2)$ distributions but the variances are different, capturing the spread of the distribution.

Exercise 135. Let $X \sim N(0,1)$. Show that $\mathbf{E}[X^n] = 0$ if n is odd and if n is even then $\mathbf{E}[X^n] = (n-1)(n-3)\ldots(3)(1)$ (product of all odd numbers up to and including $n-1$). What happens if $X \sim N(0,\sigma^2)$?

Exercise 136. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Geo}(p)$. $\mathbf{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{q}{p^2}$.
(2) $X \sim \text{Bin}(n,p)$. $\mathbf{E}[X] = np$ and $\text{Var}(X) = npq$.
(3) $X \sim \text{Pois}(\lambda)$. $\mathbf{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.
(4) $X \sim \text{Hypergeo}(N_1, N_2, m)$. $\mathbf{E}[X] = \frac{mN_1}{N_1+N_2}$ and $\text{Var}(X) =$??.

Exercise 137. Calculate the mean and variance for the following distributions.

(1) $X \sim \text{Exp}(\lambda)$. $\mathbf{E}[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.
(2) $X \sim \text{Gamma}(\nu,\lambda)$. $\mathbf{E}[X] = \frac{\nu}{\lambda}$ and $\text{Var}(X) = \frac{\nu}{\lambda^2}$.
(3) $X \sim \text{Unif}[0,1]$. $\mathbf{E}[X] = \frac{1}{2}$ and $\text{Var}(X) = \frac{1}{12}$.
(4) $X \sim \text{Beta}(p,q)$. $\mathbf{E}[X] = \frac{p}{p+q}$ and $\text{Var}(X) = \frac{pq}{(p+q)^2(p+q+1)}$.

Properties of covariance and variance: Let X, Y, X_i, Y_i be random variables on a common probability space. Small letters a, b, c etc will denote scalars.

(1) (Bilinearity): $\text{Cov}(aX_1 + bX_2, Y) = a\text{Cov}(X_1,Y) + b\text{Cov}(X_2,Y)$ and $\text{Cov}(Y, aX_1 + bX_2) = a\text{Cov}(Y,X_1) + b\text{Cov}(Y,X_2)$
(2) (Symmetry): $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.

(3) (Positivity): $\mathrm{Cov}(X,X) \geq 0$ with equality if and only if X is a constant random variable. Indeed, $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Exercise 138. Show that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$ (hence $\mathrm{sd}(cX) = |c|\mathrm{sd}(X)$). Further, if X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Note that the properties of ovariance are very much like properties of inner-products in vector spaces. In particular, we have the following analogue of the well-known inequality for vectors $(\mathbf{u}\cdot\mathbf{v})^2 \leq (\mathbf{u}\cdot\mathbf{u})(\mathbf{v}\cdot\mathbf{v})$.

Cauchy-Schwarz inequality: If X and Y are random variables with finite variances, then $(\mathrm{Cov}(X,Y))^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y)$ with equality if and only if $Y = aX + b$ for some scalars a, b.

If not convinced, follow the proof of Cauchy-Schwarz inequality that you have seen for vectors (basically, note that $\mathrm{Var}(X+tY) \geq 0$ for any scalar t and choose an appropriate t to get the Cauchy-Schwarz's inequality.

20. Makov's and Chebyshev's inequalities

Let X be a non-negative integer valued random variable with pmf $f(k)$, $k = 0, 1, 2, \ldots$. Fix any number m, say $m = 10$. Then

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} kf(k) \geq \sum_{k=10}^{\infty} 10f(k) = 10\mathbf{P}\{X \geq 10\}.$$

More generally $m\mathbf{P}\{X \geq m\} \leq \mathbf{E}[X]$. This shows that if the expected value is finite This idea is captured in general by the following inequality.

Markov's inequality: Let X be a non-negative random variable with finite expectation. Then, for any $t > 0$, we have $\mathbf{P}\{X \geq t\} \leq \frac{1}{t}\mathbf{E}[X]$.

PROOF. Fix $t > 0$ and let $Y = X\mathbf{1}_{X<t}$ and $Z = X\mathbf{1}_{X\geq t}$ so that $X = Y + Z$. Both Y and Z are non-negative random variable and hence $\mathbf{E}[X] = \mathbf{E}[Y] + \mathbf{E}[Z] \geq \mathbf{E}[Z]$. On the other hand, $Z \geq t\mathbf{1}_{X\geq t}$ (why?). Therefore $\mathbf{E}[Z] \geq t\mathbf{E}[\mathbf{1}_{X\geq t}] = t\mathbf{P}\{X \geq t\}$. Putting these together we get $\mathbf{E}[X] \geq t\mathbf{P}\{X \geq t\}$ as desired to show. ■

Markov's inequality is simple but surprisingly useful. Firstly, one can apply it to functions of our random variable and get many inequalities. Here are some.

Variants of Markov's inequality:

(1) If X is a non-negative random variable with finite p^{th} moment, then $\mathbf{P}\{X \geq t\} \leq t^{-p}\mathbf{E}[X^p]$ for any $t > 0$.
(2) If X is a random variable with finite second moment, then $\mathbf{E}[|X - \mu| \geq t] \leq \frac{1}{t^2}\mathrm{Var}(X)$. [*Chebyshev's inequality*]
(3) IF X is a random variable with finite exponential moments, then $\mathbf{P}(X > t) \leq e^{-\lambda t}\mathbf{E}[e^{\lambda X}]$ for any $\lambda > 0$.

Thus, if we only know that X has finite mean, the tail probability $\mathbf{P}(X > t)$ must decay at least as fast as $1/t$. But if we knew that the second moment was finite we could assert that the decay must be at least as fast as $1/t^2$, which is better. If $\mathbf{E}[e^{\lambda X}] < \infty$, then we get much faster decay of the tail, like $e^{-\lambda t}$.

Chebyshev's inequality captures again the intuitive notion that variance measures the spread of the distribution about the mean. The smaller the variance, lesser the spread. An alternate way to write Chebyshev's inequality is

$$\mathbf{P}(|X-\mu| > r\sigma) \leq \frac{1}{r^2}$$

where $\sigma = \text{s.d.}(X)$. This measures the deviations in multiples of the standard deviation. This is a very general inequality. In specific cases we can get better bounds than $1/r^2$ (just like Markov inequality can be improved using higher moments, when they exist).

One more useful inequality we have already seen is the Cauchy-Schwarz inequality: $(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]$ or $(\text{Cov}(X,Y))^2 \leq \text{Var}(X)\text{Var}(Y)$.

21. Weak law of large numbers

Let $X_1, X_2, \ldots$ be i.i.d random variables (independent random variables each having the same marginal distribution). Assume that the second moment of X_1 is finite. Then, $\mu = \mathbf{E}[X_1]$ and $\sigma^2 = \text{Var}(X_1)$ are well-defined.

Let $S_n = X_1 + \ldots + X_n$ (partial sums) and $\overline{X}_n = \frac{S_n}{n} = \frac{X_1+\ldots+X_n}{n}$ (*sample mean*). Then, by the properties of expectation and variance, we have

$$\mathbf{E}[S_n] = n\mu, \;\; \text{Var}(S_n) = n\sigma_1^2, \qquad \mathbf{E}[\overline{X}_n] = \mu, \;\; \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

In particular, $\text{s.d.}(\overline{X}_n) = \sigma/\sqrt{n}$ decreases with n. If we apply Chebyshev's inequality to $\overline{X}_n$, we get for any $\delta > 0$ that

$$\mathbf{P}\{|\overline{X}_n - \mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2 n}.$$

This goes to zero as $n \to \infty$ (with $\delta > 0$ being fixed). This means that for large n the sample mean is unlikely to be far from μ (sometimes called "population mean"). This is consistent with our intuitive idea that if we toss a p-coin many times, we can get a better guess of what the value of p is.

Weak law of large numbers (Jacob Bernoulli): With the above notations, for any $\delta > 0$, we have

$$\mathbf{P}\{|\overline{X}_n - \mu| \geq \delta\} \leq \frac{\sigma^2}{\delta^2 n} \to 0 \quad \text{as } n \to \infty.$$

This is very general, in that we only assume the existence of variance. If X_k are assumed to have more moments, one can get better bounds. For example, when X_k are i.i.d. $\text{Ber}(p)$, we have the following theorem.

Hoeffding's inequality: Let $X_1, \ldots, X_n$ be i.i.d. $\text{Ber}(p)$. Then

$$\mathbf{P}\{|\overline{X}_n - p| \geq \delta\} \leq 2e^{-n\delta^2/2}.$$

22. Monte-Carlo integration

In this section we give a simple application of WLLN. Let $\varphi : [0,1] \to \mathbb{R}$ be a continuous function. We would like to compute $I = \int_0^1 \varphi(x)dx$. Most often we cannot compute the integral explicitly and for an approximate value we resort to numerical methods. Here is an idea to use random numbers.

Let $U_1, U_2, \ldots, U_n$ be i.i.d. Unif$[0,1]$ random variables and let $X_1 = \varphi(U_1), \ldots, X_n = \varphi(U_n)$. Then, X_k are i.i.d. random variables with common mean and variance

$$\mu = \int_0^1 \varphi(x)dx = I, \qquad \sigma^2 := \mathrm{Var}(X_1) = \int_0^1 (\varphi(x) - I)^2 dx.$$

This gives the following method of finding I. Fix a large number N appropriately and pick N uniform random numbers U_k, $1 \le k \le N$. Then define $\hat{I}_N := \frac{1}{N}\sum_{k=1}^N \varphi(U_k)$. Present $\hat{I}_N$ as an approximate value of I.

In what sense is this an approximation of I and why? Indeed, by WLLN $\mathbf{P}\{|\hat{I}_n - I| \ge \delta\} \to 0$ and hence we expect $\hat{I}_n$ to be close to I. How large should n be? For this, we fix two numbers $\varepsilon = 0.01$ and $\delta = 0.001$ (you may change the numbers). By Chebyshev's inequality, observe that $\mathbf{P}\{|\hat{I}_n - I| \ge \delta\} \to \sigma^2/N\delta^2$.

First find N so that $\sigma^2/N\delta^2 < \varepsilon$, i.e., $N = \lceil \frac{\sigma^2}{\delta^2} \rceil$. Then, the random variable $\hat{I}_N$ is within δ of I with probability greater than $1-\varepsilon$. This is a probabilistic method, hence there is a possibility of large error, but with a small probability. Observe that N grows proportional to *square* of $1/\delta$. To increase the accuracy by 10, you must increase the number of samples by a factor of 100.

One last point. To find N we need σ^2 which involves computing another integral involving φ which we do not know how to compute! Here we do not need the exact value of the integral. For example, if our functions satisfies $-M \le \varphi(x) \le M$ for all $x \in [0,1]$, then also $-M \le I \le M$ and hence $(\varphi(x) - \mathbf{1}_)^2 \le 4M^2$. This means that $\sigma^2 \le 4M^2$. Therefore, if we take $N = \lceil \frac{4M^2}{\delta^2} \rceil$ then the value of N is larger than required for the desired accuracy. We can work with this N. Note that the dependence of of N on δ does not change.

Exercise 139. We know that $\int_0^1 \frac{1}{1+x^2}dx = \frac{\pi}{4}$. Based on this, devise a method to find an approximate value of π. Use any software you like to implement your method and see how many sample you need to get an approximation to 1, 2 and 3 decimal places consistently (consistently means with a large enough probability, say 0.9).

Exercise 140. Devise a method to approximate e and π (there are many possible integrals).

This method can be used to evaluate integrals over any interval. For instance, how would you find $\int_a^b \varphi(t)dt$ or $\int_0^\infty \varphi(t)e^{-t}dt$ or $\int_{-\infty}^\infty \varphi(t)e^{-t^2}dt$ where φ is a function on the appropriate interval? It can also be used to evaluate multiple integrals (and consequently to find the areas and volumes of sets). The only condition is that it should be possible to evaluate the given function φ at a point x on the computer. To illustrate, consider the problem of finding the area of a region $\{(x,y) : 0 \le x,y, \le 1,\ 2x^3y^2 \ge 1,\ x^2 + 2y^2 \le 2.3\}$. It is complicated to work with such regions analytically, but given a point (x,y), it is easy to check on a computer whether all the constraints given are satisfied.

As a last remark, how do Monte-Carlo methods compare with the usual numerical methods? In the latter, usually a number N and a set of points $x_1, \ldots, x_N$ are fixed along with some weights $w_1, \ldots, w_N$ that sum to 1. Then one presents $\tilde{I} := \sum_{k=1}^N w_k \varphi(x_k)$ as the approximate value of I. Lagrange's method, Gauss quadrature etc are of this type. Under certain assumptions on φ, the accuracy of these integrals can be like $1/N$ as opposed to $1/\sqrt{N}$ in Monte-Carlo. But when those assumptions are not satisfied, $\tilde{I}$ can be way off I. One may regard this as a game of strategy as follows.

I present a function φ (say bounded between -1 and 1) and you are expected to give an approximation to φ. Quadrature methods do a good job generically, but if I knew the

procedure you use, then I can give a function for which your result is entirely wrong (for example, I pick a function φ which vanishes at each of the quadrature points!). However, with Monte-Carlo methods, even if I know the procedure, there is no way to prevent you from getting an approximation of accuracy $1/\sqrt{N}$. This is because neither of us know where the points U_k will fall!

23. Central limit theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We saw that $\overline{X}_n$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

This roughly means that $\overline{X}_n$ is close to μ, within a few multiples of $\sigma/\sqrt{n}$ (as shown by Chebyshev's inequality). Now we look at $\overline{X}_n$ with a finer microscope. In other words, we ask for the probability that $\overline{X}_n$ is within the tiny interval $[\mu + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{n}}]$ for any $a < b$. The answer turns out to be surprising and remarkable!

Central limit theorem: Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We assume that $0 < \sigma^2 < \infty$. Then, for any $a < b$, we have

$$\mathbf{P}\left\{\mu + a\frac{\sigma}{\sqrt{n}} \le \overline{X}_n \le \mu + b\frac{\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

What is remarkable about this? The end result does not depend on the distribution of X_is at all! Only the mean and variance of the distribution were used! As this is one of the most important theorems in all of probability theory, we restate it in several forms, all equivalent to the above.

Restatements of central limit theorem: Let X_k be as above. Let $S_n = X_1 + \ldots + X_n$. Let Z be a $N(0,1)$ random variable. Then of course $\mathbf{P}\{a < Z < b\} = \Phi(b) - \Phi(a)$.

(1) $\mathbf{P}\{a < \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \le b\} \to \Phi(b) - \Phi(a) = \mathbf{P}\{a < Z < b\}$. Put another way, this says that for large n, the random variable $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ has $N(0,1)$ distribution, approximately. Equivalently, $\sqrt{n}(\overline{X}_n - \mu)$ has $N(0, \sigma^2)$ distribution, approximately.

(2) Yet another way to say the same is that S_n has approximately normal distribution with mean $n\mu$ and variance $n\sigma^2$. That is,

$$\mathbf{P}\{a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\} \to \mathbf{P}\{a < Z < b\}.$$

The central limit theorem so deep and surprising and useful. The following example gives a hint as to why.

Example 141. Let $U_1, \ldots, U_n$ be i.i.d. Uniform($[-1,1]$) random variables. Let $S_n = U_1 + \ldots + U_n$, let $\overline{U}_n = S_n/n$ (sample mean) and let $Y_n = S_n/\sqrt{n}$. Consider the problem of finding the distribution of any of these. Since they are got from each other by scaling, finding the distribution of one is the same as finding that of any other. For uniform $[-1,1]$, we know that $\mu = 0$ and $\sigma^2 = 1/3$. Hence, CLT tells us that

$$\mathbf{P}\{\frac{a}{\sqrt{3}} < Y_n < \frac{b}{\sqrt{3}}\} \to \Phi(b) - \Phi(a).$$

or equivalently, $\mathbf{P}\{a < Y_n < b\} \to \Phi(b\sqrt{3}) - \Phi(a\sqrt{3})$. For large n (practically, $n = 50$ is large enough) we may use this limit as a good aproximation to the probability we want.

Why is this surprising? The way to find the distribution of Y_n would be this. Using the convolution formula n times successively, one can find the density of $S_n = U_1 + \ldots + U_n$ (in principle! the actual integration may be intractable!). Then we can find the density of Y_n by another change of variable (in one dimension). Having got the density of Y_n, we integrate it from a to b to get $\mathbf{P}\{a < Y_n < b\}$. This is clearly a daunting task (if you don't feel so, just try it for $n = 5$).

The CLT cuts short all this and directly gives an approximate answer! And what is even more surprising is that the original distribution does not matter - we only need to know the mean and variance of the original distribution!

We shall not prove the central limit theorem in general. But we indicate how it is done when X_k come from $\text{Exp}(\lambda)$ distribution. This is optional and may be skipped.

CLT FOR EXPONETIALS. Let X_k be i.i.d. $\text{Exp}(1)$ random variables. They have mean $\mu = 1$ and variance $\sigma^2 = 1$. We know that (this was an exercise), $S_n = X_1 + \ldots + X_n$ has $\text{Gamma}(n,1)$ distribution. Its density is given by $f_n(t) = e^{-t}t^{n-1}/(n-1)!$ for $t > 0$.

Now let $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n}{\sqrt{n}}$. By a change of variable (in one-dimension) we see that the density of Y_n is given by $g_n(t) = \sqrt{n} f_n(n + t\sqrt{n})$. Let us analyse this.

$$\begin{aligned}
g_n(t) &= \sqrt{n}\, \frac{1}{(n-1)!} e^{-(n+t\sqrt{n})} (n + t\sqrt{n})^{n-1} \\
&= \sqrt{n}\, \frac{n^{n-1}}{(n-1)!} e^{-n-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1} \\
&\approx \sqrt{n}\, \frac{n^{n-1}}{\sqrt{2\pi}(n-1)^{n-\frac{1}{2}} e^{-n+1}} e^{-n-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1} \quad \text{(by Stirling's formula)} \\
&= \frac{1}{\sqrt{2\pi}(1-\frac{1}{n})^{n-\frac{1}{2}} e^{1}} e^{-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1}.
\end{aligned}$$

To find the limit of this, first observe that $(1 - \frac{1}{n})^{n-\frac{1}{2}} \to e^{-1}$. It remains to find the limit of $w_n := e^{-t\sqrt{n}} \left(1 + \frac{t}{\sqrt{n}}\right)^{n-1}$. Easiest to do this by taking logarithms. Recall that $\log(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \ldots$. Hence

$$\begin{aligned}
\log w_n &= -t\sqrt{n} + (n-1)\log\left(1 + \frac{t}{\sqrt{n}}\right) \\
&= -t\sqrt{n} + (n-1)\left[\frac{t}{\sqrt{n}} - \frac{t^2}{2n} + \frac{t^3}{3n^{3/2}} - \ldots\right] \\
&= -\frac{t^2}{2} + [\ldots]
\end{aligned}$$

where in $[\ldots]$ we have put all terms which go to zero as $n \to \infty$. Since there are infinitely many, we should argue that even after adding all of them, the total goes to zero as $n \to \infty$. Let us skip this step and simply conclude that $\log w_n \to -t^2/2$. Therefore, $g_n(t) \to \varphi(t) := \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ which is the standard normal density.

What we wanted was $\mathbf{P}\{a < Y_n < b\} = \int_a^b g_n(t)dt$. Since $g_n(t) \to \varphi(t)$ for each t, it is believable that $\int_a^b g_n(t)dt \to \int_a^b \varphi(t)dt$. This too needs justification but we skip it. Thus,

$$\mathbf{P}\{a < Y_n < b\} \to \int_a^b \varphi(t)dt = \Phi(b) - \Phi(a).$$

This proves CLT for the case of exponential random variables. ■