# Probability and Applications

by Patrick Boily; inspired by Rafal Kulik

Data analysis is sometimes presented in a "point-and-click manner", with tutorials often bypassing foundations in probability and statistics to focus on software use and specific datasets. While modern analysts do not always need to fully understand the theory underpinning the methods that they use, understanding some of the basic concepts can only lead to long-term benefits.

In this chapter, we introduce some of the crucial probabilistic notions that will help analysts get the most out of their data.

## 6.1 Basic Notions

**Probability theory** is the mathematical discipline relating to the numerical description of the likelihood of an event.

## 6.1.1 Sample Spaces and Events

Throughout, we will deal with **random experiments** (e.g., measurements of speed/ weight, number and duration of phone calls, etc.).

For any "experiment," the **sample space** is defined as the set of all its **possible outcomes**, often denoted by the symbol *S*. A sample space can be **discrete** or **continuous**.

An **event** is a collection of outcomes from the sample space &. Events will be denoted by *A*, *B*, *E*<sub>1</sub>, *E*<sub>2</sub>, etc.

#### Examples

- Toss a fair coin the corresponding (discrete) sample space is
   \$ = {Head, Tail}.
- Roll a die the corresponding (discrete) sample space is S = {1,2,3,4,5,6}, with various events represented by
  - rolling an even number: {2, 4, 6};
  - rolling a prime number:  $\{2, 3, 5\}$ .
- Suppose we measure the weight (in grams) of a chemical sample the (continuous) sample space can be represented by *S* = (0,∞), the positive half line, and various events by subsets of *S*, such as
  - sample is less than 1.5 grams: (0, 1.5);
  - sample exceeds 5 grams:  $(5, \infty)$ .

6.1	Basic Notions	253
	Sample Spaces and Events .	253
	Counting Techniques	254
	Ordered Samples	255
	Unordered Samples	257
	Probability of an Event	257
	Conditionality Probability	260
	Bayes' Theorem	266
6.2	Discrete Distributions	272
	Random Variables	272
	Expectation	275
	Binomial R.V	277
	Geometric R.V	282
	Negative Binomial R.V	282
	<b>Poisson R.V.</b>	283
	Other Discrete R.V	288
6.3	Continuous Distributions .	288
	Continuous R.V.	288
	Expectation	294
	Normal R.V	296
	Exponential R.V.	301
	Gamma R.V.	304
	Binomial Approximations .	305
	Other Continuous R.V	307
6.4	Joint Distributions	307
6.5	<b>CLT/Sampling Distributions</b>	313
	Sampling Distributions	313
	Central Limit Theorem	316
	Sampling Distributions II .	323
6.6	Exercises	327
	Chapter References	336

For any events  $A, B \subseteq \mathcal{S}$ :

- the **union** *A* ∪ *B* of *A* and *B* are all outcomes in *S* contained in either *A* or *B*;
- the intersection A ∩ B of A and B are all outcomes in S contained in both A and B;
- the complement A<sup>c</sup> of A (sometimes denoted A or -A) is the set of all outcomes in S that are not in A.

If *A* and *B* have no outcomes in common, they are **mutually exclusive**; which is denoted by  $A \cap B = \emptyset$  (the empty set). In particular, *A* and  $A^c$  are always mutually exclusive.<sup>1</sup>

## Examples

- Roll a die and let  $A = \{2, 3, 5\}$  (a prime number) and  $B = \{3, 6\}$  (multiples of 3). Then  $A \cup B = \{2, 3, 5, 6\}$ ,  $A \cap B = \{3\}$  and  $A^c = \{1, 4, 6\}$ .
- 100 plastic samples are analyzed for scratch and shock resistance.

		shock resistance			
		high low			
scratch	high	70	4		
resistance low		1	25		

If *A* is the event that a sample has high shock resistance and *B* is the event that a sample has high scratch residence, then  $A \cap B$  consists of 70 samples.

## 6.1.2 Counting Techniques

A **two-stage procedure** can be modeled as having k bags, with  $m_1$  items in the first bag, ...,  $m_k$  items in k-th bag.

The **first stage** consists of picking a bag, and the **second stage** consists of drawing an item out of that bag. This is equivalent to picking one of the  $m_1 + \cdots + m_k$  total items.

If all the bags have the same number of items,  $m_1 = \cdots = m_k = n$ , then there are kn items in total, and this is the **total number of ways** the two-stage procedure can occur.

#### Examples

• How many ways are there to first roll a die and then draw a card from a (shuffled) 52–card pack?

**Answer:** there are 6 ways the first step can turn out, and for each of these (the stages are independent, in fact) there are 52 ways to draw the card. Thus there are  $6 \times 52 = 312$  ways this can turn out.

1: Events can be represented graphically using Venn diagrams – mutually exclusive events are those which do not have a common intersection. • How many ways are there to draw two tickets numbered 1 to 100 from a bag, the first with the right hand and the second with the left hand?

**Answer:** There are 100 ways to pick the first number; for *each of these* there are 99 ways to pick the second number. Thus, the task has  $100 \times 99 = 9900$  possible outputs.

## **Multi-Stage Procedures**

A *k*-stage process is a process for which:

- there are *n*<sub>1</sub> possibilities at stage 1;
- regardless of the 1st outcome there are n<sub>2</sub> possibilities at stage 2,
- . . .
- regardless of the previous outcomes, there are *n<sub>k</sub>* choices at stage *k*.

There are thus  $n_1 \times n_2 \cdots \times n_k$  total ways the process can turn out.

## 6.1.3 Ordered Samples

Suppose we have a bag of *n* billiard balls numbered 1, ..., n. We can draw an **ordered sample** of size *r* by picking balls from the bag:

- with replacement, or
- without replacement.

With how many different collection of *r* balls can we end up in each of those cases (each is an *r*-stage procedure)?

**Key Notion:** all the object (balls) can be differentiated (using numbers, colours, etc.)

#### Sampling With Replacement (Order Important)

If we replace each ball into the bag after it is picked, then every draw is the same (there are *n* ways it can turn out). According to our earlier result, there are

$$\underbrace{n \times n \times \dots \times n}_{r \text{ stages}} = n'$$

ways to select an ordered sample of size r with replacement from a set with n objects  $\{1, 2, ..., n\}$ .

#### Sampling Without Replacement (Order Important)

If we **do not** replace each ball into the bag after it is drawn, then the choices for the second draw depend on the result of the first draw, and there are only n - 1 possible outcomes.

Whatever the first two draws were, there are n - 2 ways to draw the third ball, and so on.

Thus there are

$$\underbrace{n \times (n-1) \times \cdots \times (n-r+1)}_{r \text{ stages}} = {}_{n}P_{r} \quad \text{(common symbol)}$$

ways to select an ordered sample of size  $r \le n$  without replacement from a set of *n* objects  $\{1, 2, ..., n\}$ .

### **Factorial Notation**

For a positive integer *n*, write

$$n! = n(n-1)(n-2)\cdots 1.$$

There are two possibilities:

- when r = n, nPr = n!, and the ordered selection (without replacement) is called a permutation;
- when r < n, we can write

$${}_{n}P_{r} = \frac{n(n-1)\cdots(n-r+1)}{(n-r)\cdots1}\frac{(n-r)\cdots1}{(n-r)\cdots1}$$
$$= \frac{n!}{(n-r)!} = n \times \cdots \times (n-r+1).$$

By convention, we set 0! = 1, so that

$$_{n}P_{r} = \frac{n!}{(n-r)!}, \quad \text{for all } r \le n.$$

#### **Examples:**

• In how many different ways can 6 balls be drawn *in order* without replacement from a bag of balls numbered 1 to 49?

Answer: We compute

$${}_{49}P_6 = 49 \times 48 \times 47 \times 46 \times 45 \times 44 = 10,068,347,520.$$

This is the number of ways the actual drawing of the balls can occur for Lotto 6/49 in real-time (balls drawn one by one).

• How many 6-digits PIN codes can you create from the set of digits {0,1,...,9}?

Answer: If the digits may be repeated, we see that

 $10 \times 10 \times 10 \times 10 \times 10 \times 10 = 10^{6} = 1,000,000.$ 

If the digits may not be repeated, we have instead

$${}_{10}P_6 = 10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151,200.$$

## 6.1.4 Unordered Samples

Suppose that we **cannot** distinguish between different ordered samples; when we look up the Lotto 6/49 results in the newspaper, for instance, we have no way of knowing the order in which the balls were drawn:

$$1 - 2 - 3 - 4 - 5 - 6$$

could mean that the first drawn ball was ball # 1, the second drawn ball was ball # 2, etc., but it could also mean that the first ball drawn was ball # 4, the second one, ball # 3, etc., or **any combination** of the first 6 balls.

Denote the (as yet unknown) number of unordered samples of size r from a set of size n by  ${}_{n}C_{r}$ . We can derive the expression for  ${}_{n}C_{r}$  by noting that the following two processes are equivalent:

- take an ordered sample of size r (there are <sub>n</sub>P<sub>r</sub> ways to do this);
- take an unordered sample of size r (there are <sub>n</sub>C<sub>r</sub> ways to do this) and then rearrange (permute) the objects in the sample (there are r! ways to do this).

Thus

$$_{n}P_{r} = {}_{n}C_{r} \times r! \implies {}_{n}C_{r} = \frac{{}_{n}P_{r}}{r!} = \frac{n!}{(n-r)! r!} = \binom{n}{r};$$

these are known as **binomial coefficients**, read as "*n*-choose-*r*".

**Example** In how many ways can the "Lotto 6/49 draw" be reported in the newspaper (if they are always reported in increasing order)?

This number is the same as the number of *unordered samples* of size 6 (different re-orderings of same 6 numbers are indistinguishable), so

$${}_{49}C_6 = \binom{49}{6} = \frac{49 \times 48 \times 47 \times 46 \times 45 \times 44}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$$
$$= \frac{10,068,347,520}{720} = 13,983,816.$$

There is a variety of binomial coefficient identities, such as

$$\binom{n}{k} = \binom{n}{n-k}, \quad \text{for all } 0 \le k \le n,$$
$$\sum_{k=0}^{n} \binom{n}{k} = 2^{n}, \quad \text{for all } 0 \le n,$$
$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}, \quad \text{for all } 0 \le k \le n-1$$
$$\sum_{j=k}^{n} \binom{j}{k} = \binom{n+1}{k+1}, \quad \text{for all } 0 \le n, \text{etc.}$$

## 6.1.5 Probability of an Event

For situations where we have a random experiment which has exactly *N* possible **mutually exclusive**, **equally likely** outcomes, we can assign

a probability to an event A by counting the number of outcomes that correspond to A – its **relative frequency**. If that count is a, then

$$P(A) = \frac{a}{N}$$

The probability of each individual outcome is thus 1/N.

#### Examples

- Toss a fair coin the sample space is S = {Head, Tail}, i.e., N = 2. The probability of observing a Head on a toss is thus <sup>1</sup>/<sub>2</sub>.
- Throw a fair six sided die. There are N = 6 possible outcomes. The sample space is

$$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}.$$

If *A* corresponds to observing a multiple of 3, then  $A = \{3, 6\}$  and a = 2, so that

Prob(number is a multiple of 3) =  $P(A) = \frac{2}{6} = \frac{1}{3}$ .

• The probabilities of seeing an even/odd number are:

Prob{even} = 
$$P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2};$$
  
Prob{prime} =  $P(\{2, 3, 5\}) = 1 - P(\{1, 4, 6\}) = \frac{1}{2}.$ 

 In a group of 1000 people it is known that 545 have high blood pressure. 1 person is selected randomly. What is the probability that this person has high blood pressure?

**Answer:** the relative frequency of people with high blood pressure is 0.545.

This approach to probability is called the **frequentist interpretation**. It is based on the idea that the theoretical probability of an event is given by the behaviour of the empirical (observed) relative frequency of the event over long-run repeatable and independent experiments.<sup>2</sup>

This is the classical definition, and the one used in these notes, but there are competing interpretations which may be more appropriate depending on the context; chiefly, the **Bayesian interpretation** (see [2] and Chapter 25 for details) and the **propensity interpretation**.<sup>3</sup>

## **Axioms of Probability**

The modern definition of probability is **axiomatic** (according to Kolmogorov's seminal work [@KOL]).

The **probability of an event**  $A \subseteq S$  is a numerical value satisfying the following properties:

- 1. for any event  $A, 1 \ge P(A) \ge 0$ ;
- 2. for the complete sample space  $\mathcal{S}$ ,  $P(\mathcal{S}) = 1$ ;
- 3. for the empty event  $\emptyset$ ,  $P(\emptyset) = 0$ , and

2: Such as when  $N \to \infty$ .

3: Introducing causality as a mechanism.

4. for two **mutually exclusive** events *A* and *B*, the probability that *A* or *B* occurs is  $P(A \cup B) = P(A) + P(B)$ .

Since  $S = A \cup A^c$ , and A and  $A^c$  are mutually exclusive, then

$$1 \stackrel{A2}{=} P(8) = P(A \cup A^{c}) \stackrel{A4}{=} P(A) + P(A^{c})$$
$$\implies P(A^{c}) = 1 - P(A).$$

## Examples

• Throw a single six sided die and record the number that is shown. Let *A* and *B* be the events that the number is a multiple of or smaller than 3, respectively. Then  $A = \{3, 6\}, B = \{1, 2\}$  and *A* and *B* are mutually exclusive since  $A \cap B = \emptyset$ . Then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) = \frac{2}{6} + \frac{2}{6} = \frac{2}{3}.$$

An urn contains 4 white balls, 3 red balls and 1 black ball. Draw one ball, and denote the following events by W = {the ball is white}, R = {the ball is red} and B = {the ball is black}. Then

$$P(W) = 1/2$$
,  $P(R) = 3/8$ ,  $P(B) = 1/8$ 

and P(W or R) = 7/8.

## **General Addition Rule**

This useful rule is a direct consquence of the axioms of probability:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**Example** An electronic gadget consists of two components, *A* and *B*. We know from experience that P(A fails) = 0.2, P(B fails) = 0.3 and P(both A and B fail) = 0.15. Find P(at least one of A and B fails) and P(neither A nor B fails).

Write A for "A fails" and similarly for B. Then we are looking to compute

$$P(\text{at least one fails}) = P(A \cup B)$$
$$= P(A) + P(B) - P(A \cap B) = 0.35;$$
$$P(\text{neither fail}) = 1 - P(\text{at least one fails}) = 0.65.$$

If *A*, *B* are mutually exclusive,  $P(A \cap B) = P(\emptyset) = 0$  and

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B).$$

With three events, the addition rule expands as follows:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$
$$- P(A \cap B) - P(A \cap C) - P(B \cap C)$$
$$+ P(A \cap B \cap C).$$

## 6.1.6 Conditional Probability and Independent Events

Any two events *A* and *B* satisfying

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

4: This is a purely mathematical definition, but it agrees with the intuitive notion of independence in simple examples. are said to be **independent**.<sup>4</sup> When events are not independent, we say that they are **dependent** or **conditional**.

Mutual exclusivity and independence are unrelated concepts. The only way for events *A* and *B* to be mutually exclusive **and** independent is for either *A* or *B* (or both) to be a non-event (the empty event):

$$\emptyset = P(A \cap B) = P(A) \times P(B) \implies P(A) = 0 \text{ or } P(B) = 0$$
$$\implies A = \emptyset \text{ or } B = \emptyset.$$

## Examples

Flip a fair coin twice – the 4 possible outcomes are all equally likely:
 \$ = {HH, HT, TH, TT}. Let

$$A = \{HH\} \cup \{HT\}$$

denote "head on first flip",  $B = \{HH\} \cup \{TH\}$  "head on second flip". Note that  $A \cup B \neq \emptyset$  and  $A \cap B = \{HH\}$ . By the general addition rule,

$$P(A) = P(\{HH\}) + P(\{HT\}) - P(\{HH\} \cap \{HT\})$$
$$= \frac{1}{4} + \frac{1}{4} - P(\emptyset) = \frac{1}{2} - 0 = \frac{1}{2}.$$

Similarly,  $P(B) = P({HH}) + P({TH}) = \frac{1}{2}$ , and so  $P(A)P(B) = \frac{1}{4}$ . But  $P(A \cap B) = P({HH})$  is also  $\frac{1}{4}$ , so A and B are independent.

• A card is drawn from a regular well-shuffled 52-card North American deck. Let *A* be the event that it is an ace and *D* be the event that it is a diamond. These two events are independent. Indeed, there are 4 aces  $P(A) = \frac{4}{52} = \frac{1}{13}$ 

and 13 diamonds

$$P(D) = \frac{13}{52} = \frac{1}{4}$$

in such a deck, so that

$$P(A)P(D) = \frac{1}{13} \times \frac{1}{4} = \frac{1}{52}$$
,

and exactly 1 ace of diamonds in the deck, so that  $P(A \cap D)$  is also  $\frac{1}{52}$ .

A six-sided die numbered 1 – 6 is loaded in such a way that the probability of rolling each value is *proportional* to that value. Find *P*(3).

Let & = {1, 2, 3, 4, 5, 6} be the value showing after a single toss; for some proportional constant v, we have P(k) = kv, for  $k \in \&$ . By

Axiom **A2**,  $P(S) = P(1) + \dots + P(6) = 1$ , so that

$$1 = \sum_{k=1}^{6} P(k) = \sum_{k=1}^{6} kv = v \sum_{k=1}^{6} k = v \frac{(6+1)(6)}{2} = 21v.$$

Hence v = 1/21 and P(3) = 3v = 3/21 = 1/7.

• Now the die is rolled twice, the second toss *independent* of the first. Find  $P(3_1, 3_2)$ .

The experiment is such that  $P(3_1) = 1/7$  and  $P(3_2) = 1/7$ , as seen in the previous example. Since the die tosses are independent,<sup>5</sup> then

$$P(3_1 \cap 3_2) = P(3_1)P(3_2) = 1/49$$

• Is a 2-engine plane more likely to be forced down than a 3-engine plane?

This question is easier to answer if we assume that **engines fail independently** (this is no doubt convenient, but the jury is still out as to whether it is realistic). In what follows, let p be the probability that an engine fails.<sup>6</sup>

The next step is to decide what type engine failure will force a plane down:<sup>7</sup>

- A 2-engine plane will be forced down if both engines fail the probability is p<sup>2</sup>;
- A 3-engine plane will be forced down if any pair of engines fail, or if all 3 fail.
  - Pair: the probability that exactly 1 pair of engines will fail independently (i.e., two engines fail and one does not) is

$$p \times p \times (1-p).$$

The order in which the engines fail does not matter: there are  ${}_{3}C_{2} = \frac{3!}{2!1!} = 3$  ways in which a pair of engines can fail: for 3 engines A, B, C, these are AB, AC, BC.

All 3: the probability of all three engines failing independently is p<sup>3</sup>.

The probability  $\geq 2$  engines failing is thus

$$P(2 + \text{ engines fail}) = 3p^2(1-p) + p^3 = 3p^2 - 2p^3.$$

Basically it's safer to use a 2-engine plane than a 3-engine plane: the 3-engine plane will be forced down more often, assuming it needs 2 engines to fly.

This "makes sense": the 2-engine plane need 50% of its engines working, while the 3-engine plane needs 66% (see Figure 6.1 to get a sense of what the probabilities are for  $0 \le p \le 1$ ).

 (Taken from [3]) Air traffic control is a safety-related activity – each piece of equipment is designed to the highest safety standards and in many cases duplicate equipment is provided so that if one item fails another takes over. 5: Is it clear what is meant by "independent tosses' '?

6: What are some realistic values of *p*?

7: There is nothing to that effect in the problem statement, so we have to make another set of assumptions.



**Figure 6.1:** Failure probability for the 2-engine and 3-engine planes.

A new system is to be provided passing information from Heathrow Airport to Terminal Control at West Drayton. As part of the system design a decision has to be made as to whether it is necessary to provide duplication.

The new system takes data from the *Ground Movements Radar* (GMR) at Heathrow, combines this with data from the *National Airspace System NAS*, and sends the output to a display at *Terminal Control* (a conceptual model is shown in Figure 6.2).



**Figure 6.2:** Conceptual model of air traffic control security system.

For all existing systems, records of failure are kept and an experimental probability of failure is calculated annually using the previous 4 years.

The reliability of a system is defined as R = 1 - P, where P = P(failure). We assume that  $R_{\text{GMR}} = R_{\text{NAS}} = 0.9999$ ,<sup>8</sup> and that the components' failure probabilities are independent.

If a single module is used, the reliability of the **single thread design** (STD) is

$$R_{\rm STD} = R_{\rm GMR} \times R_{\rm NEW} \times R_{\rm NAS}.$$

If the module is duplicated, the reliability of this **dual thread design** (DTD) is

$$R_{\text{DTD}} = R_{\text{GMR}} \times (1 - (1 - R_{\text{NEW}})^2) \times R_{\text{NAS}}.$$

Duplicating the module causes an improvement in reliability of

$$\rho = \frac{R_{\rm DTD}}{R_{\rm STD}} = \frac{(1 - (1 - R_{\rm NEW})^2)}{R_{\rm NEW}} \times 100\% \,.$$

For the module, no historical data is available. Instead, we work out the improvement achieved by using the dual thread design for various values of  $R_{\text{NEW}}$ .

$R_{\rm NEW}$	0.1	0.2	0.5	0.75	0.99	0.999	0.9999	0.99999
ρ(%)	190	180	150	125	101	100.1	100.01	100.001

If the module is **very unreliable** (i.e.,  $R_{\text{NEW}}$  is small), then there is a significant benefit in using the dual thread design ( $\rho$  is large).<sup>9</sup> If the new module is **as reliable as** GMR and NAS, that is, if

$$R_{\rm GMR} = R_{\rm NEW} = R_{\rm NAS} = 0.9999,$$

then the single thread design has a combined reliability of 0.9997 (i.e., 3 failures in 10,000 hours), whereas the dual thread design has a combined reliability of 0.9998 (i.e., 2 failures in 10,000 hours).

If the probability of failure is independent for each component, we could conclude from this that the reliability gain from a dual thread design probably does not justify the extra cost.

In the last two examples, we had to make **additional assumptions** in order to answer the questions – this is often the case in practice.

#### **Conditional Probability**

The **conditional probability** of an event *B* given that another event *A* has occurred is defined by

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Note that this definition only makes sense when "*A* can happen" i.e., P(A) > 0. If P(A)P(B) > 0, then

$$P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$$

A and B are thus independent if  $P(B \mid A) = P(B)$  and  $P(A \mid B) = P(A)$ .

9: But why would we install a module which we know to be unreliable in the first place?

8: That is to say, 1 failure in 10,000 hours.

#### Examples

• From a group of 100 people, 1 is selected. What is the probability that this person has high blood pressure (HBP)?

If we know nothing else about the population, this is an **(unconditional) probability**, namely

 $P(\text{HBP}) = \frac{\text{\# individuals with HBP in the population}}{100}$ 

If instead we first filter out all people with low cholesterol level, and then select 1 person. What is the probability that this person has HBP?

We are looking for the conditional probability

 $P(\text{HBP} \mid \text{high cholesterol});$ 

the probability of selecting a person with HBP, given high cholesterol levels, presumably different from  $P(\text{HBP} \mid \text{low cholesterol})$ .

• A sample of 249 individuals is taken and each person is classified by blood type and tuberculosis (TB) status.

	0	Α	В	AB	Total
TB	34	37	31	11	113
no TB	55	50	24	7	136
Total	89	87	55	18	249

The (unconditional) probability that a random individual has TB is  $P(\text{TB}) = \frac{\#\text{TB}}{249} = \frac{113}{249} = 0.454$ . Among those individuals with type **B** blood, the (conditional) probability of having TB is

$$P(\text{TB} \mid \text{type } \mathbf{B}) = \frac{P(\text{TB} \cap \text{type } \mathbf{B})}{P(\text{type } \mathbf{B})} = \frac{31}{55} = \frac{31/249}{55/249} = 0.564$$

• A family has two children (not twins). What is the probability that the youngest child is a girl given that at least one of the children is a girl? Assume that boys and girls are equally likely to be born.

Let *A* and *B* be the events that the youngest child is a girl and that at least one child is a girl, respectively:

$$A = \{GG, BG\} \text{ and } B = \{GG, BG, GB\},\$$

 $A \cap B = A$ . Then  $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}$  (and not  $\frac{1}{2}$ , as might naively be believed).

Incidentally,  $P(A \cap B) = P(A) \neq P(A) \times P(B)$ , which means that *A* and *B* are **not** independent events.



Figure 6.3: Decomposition of *B* via *A*.

## Law of Total Probability

Let *A* and *B* be two events. From set theory, we have

$$B = (A \cap B) \cup (\overline{A} \cap B),$$

\_\_\_\_\_

as illustrated in Figure 6.3. Note that  $A \cap B$  and  $\overline{A} \cap B$  are mutually exclusive, so that, according to Axiom **A4**, we have

$$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$

Now, assuming that  $\emptyset \neq A \neq \emptyset$ , we have

$$P(A \cap B) = P(B \mid A)P(A)$$
 and  $P(\overline{A} \cap B) = P(B \mid \overline{A})P(\overline{A}),$ 

so that

$$P(B) = P(B \mid A)P(A) + P(B \mid A)P(A).$$

This generalizes as follows: if  $A_1, ..., A_k$  are **mutually exclusive** and **exhaustive** (i.e.,  $A_i \cap A_j = \emptyset$  for all  $i \neq j$  and  $A_1 \cup ..., \cup A_k = \emptyset$ ), then for any event *B* 

$$P(B) = \sum_{j=1}^{k} P(B \mid A_j) P(A_j) = P(B \mid A_1) P(A_1) + \dots + P(B \mid A_k) P(A_k).$$

**Example** With the Law of Total Probability (the rule above), compute P(TB) using the data from one of the previous example.

The blood types {**O**, **A**, **B**, **AB**} form a mutually exclusive partition of the population, with

$$P(\mathbf{O}) = \frac{89}{249}, \ P(\mathbf{A}) = \frac{87}{249}, \ P(\mathbf{B}) = \frac{55}{249}, \ P(\mathbf{AB}) = \frac{18}{249}.$$

It is easy to see that  $P(\mathbf{O}) + P(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{AB}) = 1$ . Furthermore,

$$P(\text{TB} \mid \mathbf{O}) = \frac{P(\text{TB}\cap\mathbf{O})}{P(\mathbf{O})} = \frac{34}{89}, \ P(\text{TB} \mid \mathbf{A}) = \frac{P(\text{TB}\cap\mathbf{A})}{P(\mathbf{A})} = \frac{37}{87},$$
$$P(\text{TB} \mid \mathbf{B}) = \frac{P(\text{TB}\cap\mathbf{B})}{P(\mathbf{B})} = \frac{31}{55}, \ P(\text{TB} \mid \mathbf{AB}) = \frac{P(\text{TB}\cap\mathbf{AB})}{P(\mathbf{AB})} = \frac{11}{18}.$$

According to the law of total probability,

$$P(TB) = P(TB | \mathbf{O})P(\mathbf{O}) + P(TB | \mathbf{A})P(\mathbf{A})$$
$$+ P(TB | \mathbf{B})P(\mathbf{B}) + P(TB | \mathbf{AB})P(\mathbf{AB}),$$

so that

$$P(\text{TB}) = \frac{34}{89} \cdot \frac{89}{249} + \frac{37}{87} \cdot \frac{87}{249} + \frac{31}{55} \cdot \frac{55}{249} + \frac{11}{18} \cdot \frac{18}{249}$$
$$= \frac{34 + 37 + 31 + 11}{249} = \frac{113}{249} = 0.454,$$

which matches the previous obtained result.

## 6.1.7 Bayes' Theorem

After an experiment generates an outcome, we are often interested in the probability that a certain condition was present given an outcome.<sup>10</sup>

We have noted before that if P(A)P(B) > 0, then

 $P(A \cap B) = P(A) \times P(B \mid A) = P(B) \times P(A \mid B) = P(B \cap A);$ 

this can be re-written as Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}.$$

Bayes' Theorem is a powerful tool in probability analysis, but it is a simple corollary of the rules of probability.

#### **Central Data Analysis Question**

Given everything that was known prior to the experiment, does the observed data support the hypothesis? The **problem** is that this is usually impossible to compute directly. Bayes' Theorem offers a **possible solution**:

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{data})}$$
$$\propto P(\text{data} \mid \text{hypothesis}) \times P(\text{hypothesis}),$$

in which the terms on the right might be easier to compute than the term on the left.

#### **Bayesian Vernacular**

In Bayes' Theorem:

- *P*(hypothesis) is the prior the probability of the hypothesis being true prior to the experiment;
- *P*(hypothesis | data) is the **posterior** the probability of the hypothesis being true once the experimental data is taken into account;
- *P*(data | hypothesis) is the likelihood the probability of the experimental data being observed assuming that the hypothesis is true.

The theorem is often presented as posterior  $\propto$  likelihood  $\times$  prior, which is to say, **beliefs should be updated in the presence of new information**.

10: Or that a particular hypothesis was valid, say.

## **Formulations**

If *A*, *B* are events for which P(A)P(B) > 0, then Bayes' Theorem can be re-written, using the law of total probability, as

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})},$$

or, in the general case where  $A_1, ..., A_k$  are **mutually exclusive** and **exhaustive** events, then for any event *B* and for each  $1 \le i \le k$ ,

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + \dots + P(B \mid A_k)P(A_k)}$$

## Examples

- In 1999, Sinnas sold three car models in North America: Sarten (S), Minima (M), and Papader (PA). Of the vehicles sold that year, 50% were S, 30% were M and 20% were PA; 12% of the S, 15% of the M, and 25% of the PA had a particular defect *D*.
  - 1. If you own a 1999 Sinnas, what is the probability that it has the defect?

In the language of conditional probability,

$$P(S) = 0.5, P(M) = 0.3, P(Pa) = 0.2,$$
  
 $P(D \mid S) = 0.12, P(D \mid M) = 0.15, P(D \mid PA) = 0.25,$ 

so that

$$P(D) = P(D \mid S) \times P(S) + P(D \mid M) \times P(M) + P(D \mid Pa) \times P(Pa)$$
  
= 0.12 \cdot 0.5 + 0.15 \cdot 0.3 + 0.25 \cdot 0.2  
= 0.155 = 15.5%.

2. If a 1999 Sinnas has defect *D*, what model is it likely to be?

In the first part we computed the total probability P(D); in this part, we compare the posterior probabilities P(M | D), P(S | D), and P(Pa | D) (and not the priors!), computed using Bayes' Theorem:

$$P(S \mid D) = \frac{P(D|S)P(S)}{P(D)} = \frac{0.12 \times 0.5}{0.155} \approx 38.7\%$$
$$P(M \mid D) = \frac{P(D|M)P(M)}{P(D)} = \frac{0.15 \times 0.3}{0.155} \approx 29.0\%$$
$$P(Pa \mid D) = \frac{P(D|Pa)P(Pa)}{P(D)} = \frac{0.25 \times 0.2}{0.155} \approx 32.3\%$$

Even though Sartens are least likely to have the defect *D*, their overall prevalence in the population carries more weight.

- Suppose that a test for a particular disease has a very high success rate. If a patient:
  - 1. has the disease, the test is 'positive' with probability 0.99;
  - 2. does not have the disease, the test reports a 'negative' with prob 0.95.

Assume that only 0.1% of the population has the disease. What is the probability that a patient who tests positive does not have the disease?

Let *D* be the event that the patient has the disease, and *A* be the event that the test is positive. The probability of a true positive is

$$P(D \mid A) = \frac{P(A \mid D)P(D)}{P(A \mid D)P(D) + P(A \mid D^{c})P(D^{c})}$$
$$= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} \approx 0.019.$$

The probability of a false positive is thus  $1 - 0.019 \approx 0.981$ . Despite the apparent high accuracy of the test, the incidence of the disease is so low (1 in a 1000) that the vast majority of patients who test positive (98 in 100) do not have the disease.

The 2 in 100 who are true positives still represent 20 times the proportion of positives found in the population (before the outcome of the test is known).<sup>11</sup>

 [Monty Hall Problem] On a game show, you are given the choice of three doors. Behind one of the doors is a prize; behind the others, dirty and smelly rubbish bins (as is skillfully rendered in Figure 6.4).

You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, behind which is a bin. She then says to you, "Do you want to switch from door No. 1 to No. 2?"

Is it to your advantage to do so?



In what follows, let and be the events that switching to another door is a successful strategy and that the prize is behind the original door, respectively.

- Let's first assume that the host opens no door. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

If the prize is behind the original door, switching would succeed 0% of the time:  $P(S | D) = 0.^{12}$  If the prize is not behind the original door, switching would succeed 50% of the time:  $P(S | D^c) = 1/2.^{13}$  Thus,

11: It is important to remember that when dealing with probabilities, **both** the likelihood and the prevalence have to be taken into account.

**Figure 6.4:** The Monty Hall set-up (personal file, ... but that was probably obvious from the artistic quality ).

12: Note that the prior is P(D) = 1/3.

13: Note that the prior is  $P(D^c) = 2/3$ .

$$P(S) = P(S \mid D)P(D) + P(S \mid D^{c})P(D^{c})$$
  
=  $0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} \approx 33\%.$ 

- Now let's assume that the host opens one of the other two doors to show a rubbish bin. What is the probability that switching to another door in this scenario would prove to be a successful strategy?

If the prize is behind the original door, switching would succeed 0% of the time:  $P(S | D) = 0.^{14}$  If the prize is not behind the original door, switching would succeed 100% of the time:  $P(S | D^c) = 1.^{15}$  Thus,

$$P(S) = P(S \mid D)P(D) + P(S \mid D^{c})P(D^{c})$$
  
= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} \approx 67\%.

If no door is opened, switching is not a winning strategy, resulting in success only 33% of the time. If a door is opened, however, switching becomes the winning strategy, resulting in success 67% of the time.

The Monty Hall problem has attracted a lot of attention over the years due to its counter-intuitive result, but there is no paradox when we understand conditional probabilities.

Perhaps it would be easier to see what happens in practice: if we could pit two players against one another (one who never switches and one who always does so) in a series of Monty Hall games, which one would come out on top in the long run?

We start by setting a number of games N (not too small, or we won't be able to observe long-run behaviour) and a replicability seed (so that we may all obtain the same results).

N=500 set.seed(1234)

Next, for each game, we will place the prize behind one of the 3 doors: *A*, *B*, or *C*.

locations = sample(c("A", "B", "C"), N, replace = TRUE)

We verify that the prize gets placed behind each door roughly 33% of the time:

table(locations)/N

locations A B C 0.302 0.344 0.354 14: Note that the prior is P(D) = 1/3.

15: Note that the prior is  $P(D^c) = 2/3$ .

Let us now obtain a player's first guess for each game – this guess is completely independent of the actual prize location:

```
player.guesses = sample(c("A","B","C"), N, replace = TRUE)
```

Finally, we create a data frame telling the analyst where the prize actually is, and what door the player has selected as their original guess.

```
games = data.frame(locations, player.guesses)
head(games)
```

	locations	player.guesses
1	В	В
2	В	В
3	А	В
4	C	C
5	А	C
6	А	Α

In this example (that is, with the data generated above), how often had the player guessed correctly, before a door was opened and they were given a chance to switch?

table(games\$locations==games\$player.guesses)

FALSE TRUE 333 167

This should not come as a surprise.

We now initialize the process to find out which door the host opens. For each game, the host opens a door which is not the one selected by the player, nor the one behind which the prize is found.

```
games$open.door <- NA
for(j in 1:N){
  games$open.door[j] <- sample(setdiff(c("A", "B", "C"),
     union(games$locations[j],games$player.guesses[j])), 1)
}
head(games)</pre>
```

locations player.guesses open.door

В	В	С
В	В	С
А	В	С
С	С	Α
А	С	В
А	А	В
	B A C A A	BBBBABCCACAA

The union() call enumerates the doors that the host cannot open; the setdiff() call finds the complement of the doors that the host cannot open (i.e.: the doors that she can open), and the sample() call picks one of those doors.

If the player never switches, they win whenever they had originally guessed the location of the prize correctly:

games\$no.switch.win <- games\$player.guess==games\$locations</pre>

We find which door the player would have selected if they always switched (the door that is neither the location of the prize nor the one they had originally selected):

```
games$switch.door <- NA
for(j in 1:N){
  games$switch.door[j] <- sample(setdiff(c("A", "B", "C"),
     union(games$open.door[j],games$player.guesses[j])), 1)
}</pre>
```

If the player always switches, they win whenever their switched guess is where the prize is located:

```
games$switch.win <- games$switch.door==games$locations
head(games)</pre>
```

	locations	player.guesses	open.door	no.switch.win	switch.door	switch.win
1	В	В	C	TRUE	A	FALSE
2	В	В	C	TRUE	A	FALSE
3	А	В	C	FALSE	A	TRUE
4	C	C	А	TRUE	В	FALSE
5	Α	C	В	FALSE	А	TRUE
6	А	А	В	TRUE	C	FALSE

The chances of winning by not switching are thus:

table(games\$no.switch.win)/N

FALSE TRUE 0.666 0.334

while the chances of winning by switching are:

table(games\$switch.win)/N

FALSE TRUE 0.334 0.666

Pretty wild, eh? Numerical simulations show, beyond the shadow of a doubt, that switching IS the better strategy.

16: Note that the principles of probability theory introduced in the previous section remain valid in all cases.

17: For the purpose of these notes, a discrete set is one in which all points are **isolated**:  $\mathbb{N}$  and finite sets are discrete, but  $\mathbb{Q}$  and  $\mathbb{R}$  are not.

# 6.2 Discrete Distributions

In the next sections, we discuss how some of the probability computations can be made easier with the use of **(theoretical) distributions**.<sup>16</sup>

## 6.2.1 Random Variables and Distributions

Recall that, for any random "experiment", the set of all possible outcomes is denoted by  $\mathscr{S}$ . A **random variable** (r.v.) is a function  $X : \mathscr{S} \to \mathbb{R}$ , which is to say, it is a rule that associates a (real) number to every outcome of the experiment;  $\mathscr{S}$  is the **domain** of the r.v. X and  $X(\mathscr{S}) \subseteq \mathbb{R}$  is its **range**.

A **probability distribution function** (p.d.f.) is a function  $f : \mathbb{R} \to \mathbb{R}$  which specifies the probabilities of the values in the range  $X(\mathcal{S})$ . When  $\mathcal{S}$  is **discrete**,<sup>17</sup> we say that X is a **discrete r.v.** and the p.d.f. is called a **probability mass function** (p.m.f.).

#### Notation

Throughout, we use the following notation:

- capital roman letters (*X*, *Y*, etc.) denote r.v., and
- corresponding lower case roman letters (x, y, etc.) denote *generic* values taken by the r.v.

A discrete r.v. can be used to **define events** – if *X* takes values  $X(S) = \{x_i\}$ , then we can define the events  $A_i = \{s \in S : X(s) = x_i\}$ :

- the p.m.f. of X is  $f(x) = P(\{s \in S : X(s) = x\}) := P(X = x);$
- its cumulative distribution function (c.d.f.) is  $F(x) = P(X \le x)$ .

### Properties

If *X* is a discrete random variable with p.m.f. f(x) and c.d.f. F(x), then

- $0 < f(x) \le 1$  for all  $x \in X(\mathcal{S})$ ;  $\sum_{s \in \mathcal{S}} f(X(s)) = \sum_{x \in X(\mathcal{S})} f(x) = 1$ ;
- for any event  $A \subseteq \mathcal{S}$ ,  $P(X \in A) = \sum_{x \in A} f(x)$ ;
- for any  $a, b \in \mathbb{R}$ ,

$$P(a < X) = 1 - P(X \le a) = 1 - F(a)$$
  

$$P(X < b) = P(X \le b) - P(X = b) = F(b) - f(b)$$

• for any  $a, b \in \mathbb{R}$ ,

$$P(a \le X) = 1 - P(X < a) = 1 - (P(X \le a) - P(X = a)) = 1 - F(a) + f(a).$$

We can use these results to compute the probability of a **discrete** r.v. *X* falling in various intervals:

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a);$$
  

$$P(a \le X \le b) = P(a < X \le b) + P(X = a) = F(b) - F(a) + f(a);$$
  

$$P(a < X < b) = P(a < X \le b) - P(X = b) = F(b) - F(a) - f(b);$$
  

$$P(a \le X < b) = P(a \le X \le b) - P(X = b) = F(b) - F(a) + f(a) - f(b).$$

#### Examples

• Flip a fair coin – the outcome space is  $\mathscr{S} = \{\text{Head}, \text{Tail}\}$ . Let  $X : S \to \mathbb{R}$  be defined by X(Head) = 1 and X(Tail) = 0. Then X is a discrete random variable.<sup>18</sup>

If the coin is fair, the p.m.f. of *X* is  $f : \mathbb{R} \to \mathbb{R}$ , where

$$f(0) = P(X = 0) = 1/2, f(1) = P(X = 1) = 1/2,$$
  
 $f(x) = 0$  for all other x.

Roll a fair die – the outcome space is S = {1,...,6}. Let X : S → R
 be defined by X(i) = i for i = 1,...,6. Then X is a discrete r.v.

If the die is fair, the p.m.f. of *X* is  $f : \mathbb{R} \to \mathbb{R}$ , where

$$f(i) = P(X = i) = 1/6$$
, for  $i = 1, ..., 6$ ,  
 $f(x) = 0$  for all other  $x$ .

• For the random variable *X* from the previous example, the c.d.f. is  $F : \mathbb{R} \to \mathbb{R}$ , where

$$F(x) = P(X \le x) = \begin{cases} 0 & \text{if } x < 1\\ i/6 & \text{if } i \le x < i+1, i = 1, \dots, 6\\ 1 & \text{if } x \ge 6 \end{cases}$$

• For the same random variable, we can compute the probability  $P(3 \le X \le 5)$  directly:

$$P(3 \le X \le 5) = P(X = 3) + P(X = 4) + P(X = 5)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

or we can use the c.d.f.:

$$P(3 \le X \le 5) = F(5) - F(3) + f(3) = \frac{5}{6} - \frac{3}{6} + \frac{1}{6} = \frac{1}{2}.$$

- The number of calls received over a specific time period, X, is a discrete random variable, with potential values 0, 1, 2, . . ..
- Consider a 5–card poker hand consisting of cards selected at random from a 52–card deck. Find the probability distribution of *X*, where *X* indicates the number of red cards (◆ and ♥) in the hand.

In all, there are  $\binom{52}{5}$  ways to select poker hands. By construction, *X* can take on values x = 0, 1, 2, 3, 4, 5.

If X = 0, then none of the 5 cards in the hands are  $\blacklozenge$  or  $\blacklozenge$ , and all of the 5 cards in the hands are  $\blacklozenge$  or  $\clubsuit$ . There are thus  $\binom{26}{0} \cdot \binom{26}{5}$  5–card hands that only contain black cards, and

$$P(X=0) = \frac{\binom{26}{0} \cdot \binom{26}{5}}{\binom{52}{5}}.$$

In general, if X = x, x = 0, 1, 2, 3, 4, 5, there are  $\binom{26}{x}$  ways of having  $x \\ \bullet$  or  $\bullet$  in the hand, and  $\binom{26}{5-x}$  ways of having  $5 - x \\ \bullet$  and  $\bullet$  in the

18: As a convenience, we write X = 1 and X = 0.

hand, so that

$$f(x) = P(X = x) = \begin{cases} \frac{\binom{26}{x} \cdot \binom{26}{5-x}}{\binom{52}{5}}, \ x = 0, 1, 2, 3, 4, 5; \\ 0 \text{ otherwise} \end{cases}$$

• Find the c.d.f. of a discrete r.v. *X* with p.m.f. f(x) = 0.1x if x = 1, 2, 3, 4 and f(x) = 0 otherwise.

f(x) is indeed a p.m.f. as  $0 < f(x) \le 1$  for all x and

$$\sum_{x=1}^{4} 0.1x = 0.1(1+2+3+4) = 0.1\frac{4(5)}{2} = 1.$$

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

$$F(x) = \begin{cases} 0 & \text{if } x < 1\\ 0.1 & \text{if } 1 \le x < 2\\ 0.3 & \text{if } 2 \le x < 3\\ 0.6 & \text{if } 3 \le x < 4\\ 1 & \text{if } x \ge 4 \end{cases}$$

The p.m.f. and the c.m.f. for this r.v. are shown in Figure 6.5.



**Figure 6.5:** P.m.f. and c.m.f. for the discrete r.v. *X* defined in the last example.

## 6.2.2 Expectation of a Discrete Random Variable

The **expectation** of a discrete random variable X is

$$\mathbf{E}[X] = \sum_{x} x \cdot P(X = x) = \sum_{x} x f(x),$$

where the sum extends over all values of *x* taken by X.<sup>19</sup> The definition can be extended to a general function of *X*:

$$\mathbb{E}[u(X)] = \sum_{x} u(x)P(X=x) = \sum_{x} u(x)f(x).$$

As an important example, note that

$$E[X^{2}] = \sum_{x} x^{2} P(X = x) = \sum_{x} x^{2} f(x).$$

#### Examples

• What is the expectation on the roll *Z* of 6–sided die?

If the die is fair, then

$$\mathbf{E}[Z] = \sum_{z=1}^{6} z \cdot P(Z=z) = \frac{1}{6} \sum_{z=1}^{6} z = \frac{1}{6} \cdot \frac{6(7)}{2} = 3.5.$$

• For each 1\$ bet in a gambling game, a player can win 3\$ with probability  $\frac{1}{3}$  and lose 1\$ with probability  $\frac{2}{3}$ . Let *X* be the net gain/loss from the game. Find the expected value of the game.

*X* takes on the value 2\$ for a win and -2\$ for a loss.<sup>20</sup> The expected value of *X* is thus

$$\mathbf{E}[X] = 2 \cdot \frac{1}{3} + (-2) \cdot \frac{2}{3} = -\frac{2}{3}.$$

• If Z is the number showing on a roll of a fair 6–sided die, find  $E[Z^2]$  and  $E[(Z - 3.5)^2]$ .

$$E[Z^{2}] = \sum_{z} z^{2} P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} z^{2} = \frac{1}{6} (1^{2} + \dots + 6^{2}) = \frac{91}{6}$$
$$E[(Z - 3.5)^{2}] = \sum_{z=1}^{6} (z - 3.5)^{2} \times P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} (z - 3.5)^{2}$$
$$= \frac{(1 - 3.5)^{2} + \dots + (6 - 3.5)^{2}}{6} = \frac{35}{12}.$$

## Mean and Variance

We can interpret the expectation as the average or the **mean** of *X*, which we often denote by  $\mu = \mu_X$ . For instance, in the example of the fair die,

$$\mu_Z = E[Z] = 3.5$$

19: The expectation of a random variable is simply the average value that it takes, over all possible values.

20: That is, win/loss = outcome – bet.

Note that in the final example, we could have written

$$E[(Z - 3.5)^2] = E[(Z - E[Z])^2]$$

This is an important quantity associated to a random variable *X*, its **variance** Var[*X*].

The variance of a discrete random variable *X* is the **expected squared difference from the mean**:

$$Var(X) = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 P(X = x)$$
  
=  $\sum_x (x^2 - 2x\mu_X + \mu_X^2) f(x)$   
=  $\sum_x x^2 f(x) - 2\mu_X \sum_x x f(x) + \mu_X^2 \sum_x f(x)$   
=  $E[X^2] - 2\mu_X \mu_X + \mu_X^2 \cdot 1$   
=  $E[X^2] - \mu_X^2$ .

This is also sometimes written as  $Var[X] = E[X^2] - E^2[X]$ .

## **Standard Deviation**

The **standard deviation** of a discrete random variable *X* is defined directly from the variance:

$$SD[X] = \sqrt{Var[X]}$$
.

The mean is a measure of **centrality** and it gives an idea as to where the **bulk** of a distribution is located; the variance and standard deviation provide information about the **spread** – distributions with higher variance/SD are **more spread out about the average**.

**Example** Let *X* and *Y* be random variables with the following p.d.f.

x	P(X = x)	у	P(Y = y)
-2	1/5	-4	1/5
-1	1/5	-2	1/5
0	1/5	0	1/5
1	1/5	2	1/5
2	1/5	4	1/5

We have E[X] = E[Y] = 0 and

$$2 = \operatorname{Var}[X] < \operatorname{Var}[Y] = 8,$$

meaning that we expect both distributions to be centered at 0, but Y should be more **spread-out** than X (because its variance is greater, see Figure 6.6).



Figure 6.6: R.v. X (left) and Y (right) for two uniform distributions, as defined in the example.

## **Properties**

Let *X*, *Y* be random variables and  $a \in \mathbb{R}$ . Then

- E[aX] = aE[X];
- E[X + a] = E[X] + a;
- E[X + Y] = E[X] + E[Y];
- in general,  $E[XY] \neq E[X]E[Y]$ ;
- $\operatorname{Var}[aX] = a^2 \operatorname{Var}[X], \operatorname{SD}[aX] = |a| \operatorname{SD}[X];$
- $\operatorname{Var}[X + a] = \operatorname{Var}[X], \operatorname{SD}[X + a] = \operatorname{SD}[X].$

## 6.2.3 Binomial Distributions

Recall that the number of unordered samples of size r from a set of size *n* is

$$_{n}C_{r} = \binom{n}{r} = \frac{n!}{(n-r)!r!}.$$

## Examples

- $2! \times 4! = (1 \times 2) \times (1 \times 2 \times 3 \times 4) = 48$ , but  $(2 \times 4)! = 8! = 40320$ .  $\binom{5}{1} = \frac{5!}{1! \times 4!} = \frac{1 \times 2 \times 3 \times 4 \times 5}{1 \times (1 \times 2 \times 3 \times 4)} = \frac{5}{1} = 5$ .

- In general:  $\binom{n}{1} = n$  and  $\binom{n}{0} = 1$ . (<sup>6</sup><sub>2</sub>) =  $\frac{6!}{2! \times 4!} = \frac{4! \times 5 \times 6}{2! \times 4!} = \frac{5 \times 6}{2!} = 15$ . (<sup>27</sup><sub>22</sub>) =  $\frac{27!}{22! \times 5!} = \frac{22! \times 23 \times 24 \times 25 \times 26 \times 27}{5! \times 22!} = \frac{23 \times 24 \times 25 \times 26 \times 27}{120}$ .

## **Binomial Experiments**

A Bernoulli trial is a random experiment with two possible outcomes, "success" and "failure". Let *p* denote the probability of a success.

A **binomial experiment** consists of *n* repeated *independent* Bernoulli trials, each with the same probability of success, *p*, such as:

- female/male births (perahps not truly independent, but often treated as such);
- satisfactory/defective items on a production line;
- sampling with replacement with two types of item,
- etc.

## **Probability Mass Function**

In a binomial experiment of n independent events, each with probability of success p, the number of successes X is a discrete random variable that follows a **binomial distribution** with parameters (n, p):

$$f(x) = P(X = x) = {n \choose x} p^x (1-p)^{n-x}$$
, for  $x = 0, 1, 2, ..., n$ .

This is often abbreviated to " $X \sim \mathfrak{B}(n, p)$ ".

If 
$$X \sim \mathfrak{B}(1, p)$$
, then  $P(X = 0) = 1 - p$  and  $P(X = 1) = p$ , so

$$E[X] = (1 - p) \cdot 0 + p \cdot 1 = p.$$

## **Expectation and Variance**

If  $X \sim \mathfrak{B}(n, p)$ , it can be shown that

$$\mathbf{E}[X] = \sum_{x=0}^{n} x P(X = x) = np$$

and

$$\operatorname{Var}[X] = \operatorname{E}\left[ (X - np)^2 \right] = \sum_{x=0}^n (x - np)^2 \cdot P(X = x) = np(1 - p)$$

(we will eventually see an easier way to derive these formulas by interpreting *X* as a sum of discrete random variables).

Recognizing that certain situations can be modeled *via* a distribution whose p.m.f. and c.d.f. are already known can simplify computations.

## Examples

Suppose that water samples taken in some well-defined region have a 10% probability of being polluted. If 12 samples are selected independently, then it is reasonable to model the number X of polluted samples as *B*(12, 0.1).

Find

1. E[X] and Var[X];

**2**. 
$$P(X = 3);$$

3.  $P(X \le 3)$ .

1. If  $X \sim \mathfrak{B}(n, p)$ , then

$$E[X] = np$$
 and  $Var[X] = np(1-p)$ .

With n = 12 and p = 0.1, we obtain

$$\begin{split} \mathbf{E}[X] &= 12 \times 0.1 = 1.2; \\ \mathrm{Var}[X] &= 12 \times 0.1 \times 0.9 = 1.08 \,. \end{split}$$

2. By definition,

$$P(X=3) = {\binom{12}{3}} (0.1)^3 (0.9)^9 \approx 0.0852.$$

3. By definition,

$$P(X \le 3) = \sum_{x=0}^{3} P(X = x)$$
$$= \sum_{x=0}^{3} {\binom{12}{x}} (0.1)^{x} (0.9)^{12-x}.$$

This sum can be computed directly, however, for  $X \sim \mathfrak{B}(12, 0.1)$ ,  $P(X \le 3)$  can also be read directly from tabulated values (as in Figure 6.7):

12	0	0.2824	0.0687	0.0138	0.0022	0.0002	0.0000			
	1	0.6590	0.2749	0.0850	0.0196	0.0032	0.0003	0.0000		
	2	0.8891	0.5583	0.2528	0.0834	0.0193	0.0028	0.0002		
	3	0.9744	0.7946	0.4925	0.2253	0.0730	0.0153	0.0017	0.0000	
	4	0.9957	0.9274	0.7237	0.4382	0.1938	0.0573	0.0095	0.0006	
	5	0.9995	0.9806	0.8822	0.6652	0.3872	0.1582	0.0386	0.0039	0.0000
	6	0.9999	0.9961	0.9614	0.8418	0.6128	0.3348	0.1178	0.0194	0.0005
	7	1.0000	0.9994	0.9905	0.9427	0.8062	0.5618	0.2763	0.0726	0.0043
	8		0.9999	0.9983	0.9847	0.9270	0.7747	0.5075	0.2054	0.0256
	9		1.0000	0.9998	0.9972	0.9807	0.9166	0.7472	0.4417	0.1109
1	0			1.0000	0.9997	0.9968	0.9804	0.9150	0.7251	0.3410
1	1				1.0000	0.9998	0.9978	0.9862	0.9313	0.7176
1	2					1.0000	1.0000	1.0000	1.0000	1.0000

**Figure 6.7:** Tabulated c.d.f. values for the binomial distribution with n = 12 [source unknown].

The appropriate value  $\approx 0.9744$  can be found in the group corresponding to n = 12, in the row corresponding to x = 3, and in the column corresponding to p = 0.1. The table can also be used to compute

 $P(X = 3) = P(X \le 3) - P(X \le 2) = 0.9744 - 0.8891 \approx 0.0853.$ 

• An airline sells 101 tickets for a flight with 100 seats. Each passenger with a ticket is known to have a probability p = 0.97 of showing up for their flight. What is the probability of 101 passengers showing up (and the airline being caught overbooking)? Make appropriate

assumptions. What if the airline sells 125 tickets?

Let *X* be the number of passengers that show up. We want to compute P(X > 100).

If all passengers show up independently of one another,<sup>21</sup> we can model  $X \sim \Re(101, 0.97)$  and

$$P(X > 100) = P(X = 101)$$
  
=  $\binom{101}{101} (0.97)^{101} (0.03)^0 \approx 0.046.$ 

If the airline sells n = 125 tickets, we can model the situation with the binomial distribution  $\Re(125, 0.97)$ , so that

$$P(X > 100) = 1 - P(X \le 100)$$
$$= 1 - \sum_{x=0}^{100} {\binom{125}{x}} (0.97)^x (0.03)^{125-x}$$

This sum is harder to compute directly, but is very nearly 1 (try it with R, say).<sup>22</sup>

We can evaluate related probabilities in R *via* the base functions rbinom(), dbinom(), etc., whose parameters are n, size, and prob.

We can draw an observation X from a binomial distribution  $\mathfrak{B}(11, 0.2)$  in R as follows:

rbinom(1, size=11, prob=0.2)

## [1] 5

We could also replicate the process 1000 times (and extract the empirical expectation and variance):

```
v<- rbinom(1000,size=11, prob=0.2)
mean(v)
var(v)</pre>
```

[1] 2.236[1] 1.794098

The histogram of the sample is shown below.

brks = min(v):max(v)
hist(v, breaks = brks)

22: Do these results match your intuition?

21: No families or late bus?

## Histogram of v



If we change the parameters of the distribution  $(\mathfrak{B}(19.0.7))$ , we get a different looking histogram (and a different expectation and variance).

```
v<- rbinom(1000,size=19, prob=0.7)
mean(v)
var(v)</pre>
```

[1] 13.308
[1] 4.253389

brks = min(v):max(v)
hist(v, breaks = brks)



## 6.2.4 Geometric Distributions

Now consider a sequence of Bernoulli trials, with probability p of success at each step. Let the **geometric** random variable X denote the number of steps before the first success occurs. Its p.m.f. is given by

$$f(x) = P(X = x) = (1 - p)^{x - 1}p, \quad x = 1, 2, \dots$$

and we denote it by  $X \sim \text{Geo}(p)$ . For this r.v., we have

$$E[X] = \frac{1}{p}$$
 and  $Var[X] = \frac{1-p}{p^2}$ .

## Examples

• A fair 6-sided die is thrown until it shows a 6. What is the probability that 5 throws are required?

If 5 throws are required, we have to compute P(X = 5), where  $X \sim \text{Geo}(1/6)$ :

$$P(X = 5) = (1 - p)^{5-1}p = (5/6)^4(1/6) \approx 0.0804.$$

In the example above, how many throws would you expect to need?

It's fairly simple:  $E[X] = \frac{1}{1/6} = 6.^{23}$ 

## 6.2.5 Negative Binomial Distributions

Consider now a sequence of Bernoulli trials, with probability p of success at each step. Let the **negative binomial** random variable X denote the number of steps before the rth success occurs. Its p.m.f. is given by

$$f(x) = P(X = x) = {\binom{x-1}{r-1}}(1-p)^{x-r}p^r, \quad x = r, r+1, \dots$$

and we denote it by  $X \sim \text{NegBin}(p, r)$ . For this r.v., we have

$$E[X] = \frac{r}{p}$$
 and  $Var[X] = \frac{r(1-p)}{p^2}$ .

#### Examples

• A fair 6–sided die is thrown until it three 6's are rolled. What is the probability that 5 throws are required?

If 5 throws are required, we have to compute P(X = 5), where  $X \sim \text{NegBin}(1/6, 3)$ :

$$P(X=5) = {\binom{5-1}{3-1}}(1-p)^{5-3}p^3 = {\binom{4}{2}}(5/6)^2(1/6)^3 \approx 0.0193.$$

In the example above, how many throws would you expect to need?

This one is also fairly simple:  $E[X] = \frac{3}{1/6} = 18$ .

23: Understand, however, that this **does not mean** that we obtain get a 6 every 6 throws.

## 6.2.6 Poisson Distributions

Let us say we are counting the number of "changes" that occur in a continuous interval of time or space.<sup>24</sup>

We have a **Poisson process** with rate  $\lambda$ , denoted by  $\mathcal{P}(\lambda)$ , if:

- the number of changes occurring in non-overlapping intervals are independent;
- 2. the probability of exactly one change in a short interval of length *h* is approximately  $\lambda h$ , and
- 3. The probability of 2+ changes in a sufficiently short interval is essentially 0.

Assume that an experiment satisfies the above properties. Let *X* be the number of changes in a **unit interval**.<sup>25</sup> What is P(X = x), for x = 0, 1, ...? We get to the answer by first partition the unit interval into *n* disjoint sub-intervals of length 1/n. Then,

- 1. by the second condition, the probability of one change occurring in one of the sub-intervals is approximately  $\lambda/n$ ;
- 2. by the third condition, the probability of 2+ changes is  $\approx$  0, and
- 3. by the first condition, we have a sequence of *n* Bernoulli trials with probability  $p = \lambda/n$ .

Therefore,

$$f(x) = P(X = x) \approx \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$
$$= \frac{\lambda^x}{x!} \cdot \underbrace{\frac{n!}{(n-x)!} \cdot \frac{1}{n^x}}_{\text{term 1}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\text{term 2}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\text{term 3}}.$$

Letting  $n \to \infty$ , we obtain

$$P(X = x) = \lim_{n \to \infty} \frac{\lambda^x}{x!} \cdot \underbrace{\frac{n!}{(n-x)!} \cdot \frac{1}{n^x}}_{\text{term 1}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\text{term 2}} \cdot \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\text{term 3}}$$
$$= \frac{\lambda^x}{x!} \cdot 1 \cdot \exp(-\lambda) \cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \dots$$

Let *X* ~  $\mathcal{P}(\lambda)$ . Then it can be shown that

 $E[X] = \lambda$  and  $Var[X] = \lambda$ ;

the mean and the variance of a Poisson random variable are identical!

We can compute related probabilities in R *via* the base functions rpois(), dpois(), etc., with required parameters n and lambda. We start by drawing a sample of size 1 from  $\mathcal{P}(13)$ , say, in R as follows:<sup>26</sup>

rpois(1,lambda=13)

24: Such as # of defects on a production line over a 1 hr period, # of customers that arrive at a teller over a 15 min interval, etc.

25: This could be 1 day, or 15 minutes, or 10 years, etc.

26: No seed has been specified, so it is conceivable that your results would be different.

[1] 18

Next, we sample independently 500 times; this yields an empirical expectation and variance.

```
u<-rpois(500,lambda=13)
head(u)
mean(u)
var(u)
```

[1] 13 12 14 12 18 9
[1] 12.874
[1] 12.92798

The sample's histogram is shown below.



## Examples

• A traffic flow is typically modeled by a Poisson distribution. It is known that the traffic flowing through an intersection is 6 cars/minute, on average. What is the probability of no cars entering the intersection in a 30 second period?

Note that 6 cars/min = 3 cars/30 sec. Thus  $\lambda$  = 3, and we need to compute

$$P(X=0) = \frac{3^0 e^{-3}}{0!} = \frac{e^{-3}}{1} \approx 0.0498.$$

• A hospital needs to schedule night shifts in the maternity ward. It is known that there are 3000 deliveries per year; if these happened randomly round the clock,<sup>27</sup> we would expect 1000 deliveries between the hours of midnight and 8.00 a.m., a time when much of the staff is off-duty.

27: Is this a reasonable assumption?

It is thus important to ensure that the night shift is sufficiently staffed to allow the maternity ward to cope with the workload on any particular night, or at least, on a high proportion of nights.

The average number of deliveries per night

$$\lambda = 1000/365.25 \approx 2.74.$$

If the daily number *X* of night deliveries follows a Poisson process  $\mathcal{P}(\lambda)$ , we can compute the probability of delivering x = 0, 1, 2, ... babies on each night.

For a Poisson distribution, the p.m.f. values f(x) are obtained *via* dpois() in R.<sup>28</sup>

We start by setup the Poisson distribution parameters and the distribution's range.<sup>29</sup>

lambda = 2.74 x=0:10

The p.m.f. and c.d.f. are shown below:

pmf=dpois(x,lambda)
cdf=ppois(x,lambda)
data.frame(x,pmf,cdf)

х	pmf	cdf
0	0.0645703	0.0645703
1	0.1769228	0.2414931
2	0.2423842	0.4838773
3	0.2213775	0.7052548
4	0.1516436	0.8568984
5	0.0831007	0.9399991
6	0.0379493	0.9779484
7	0.0148544	0.9928029
8	0.0050876	0.9978905
9	0.0015489	0.9994394
10	0.0004244	0.9998638

Here are the p.m.f. and c.d.f. plots:

28: For a general distribution, replace the r in the rxxxxx(...) random number generators by d: dxxxxx(...).

29: In theory, it goes to infinity, but we have got to stop somewhere in practice.



• If the maternity ward wants to prepare for the greatest possible traffic on 80% of the nights, how many deliveries should be expected?

We seek an x for which

$$P(X \le x - 1) \le 0.80 \le P(X \le x).$$

Let's plot the height F(x) = 0.8 on the c.d.f.:

```
plot(c(1,x),c(0,cdf), type="s", col=2,
    main="Poisson CDF", xlab="x", ylab="F(x)=P(X<=x)")
abline(h=0:1, col=4)
abline(h=0.8, col=1)
```



The y = 0.8 line crosses the CMF at x = 4; let's evaluate  $F(3) = P(X \le 3)$  and  $F(4) = P(X \le 4)$  to confirm that  $F(3) \le 0.8 \le F(4)$ .

ppois(3,lambda)
ppois(4,lambda)

[1] 0.8568984

30: Note that this is different than asking how many deliveries are expected nightly (namely, E[X] = 2.74).

Thus, if the hospital prepares for 4 deliveries a night, they will be ready for the worst on at least 80% of the nights.<sup>30</sup>

• On how many nights in the year would 5 or more deliveries be expected?

We need to evaluate

 $365.25 \cdot P(X \ge 5) = 365.25(1 - P(X \le 4)).$ 

365.25\*(1-ppois(4,2.74))

[1] 52.26785

Thus, on roughly 14% of the nights.

• Over the course of one year, what is the greatest number of deliveries expected on any night?

We are looking for the largest value of *x* s.t.  $365.25 \cdot P(X = x) \ge 1.^{31}$ The expected number of nights with each number of deliveries can be computed using:

```
nights=c()
for(j in 0:10){
    nights[j+1]=365.25*dpois(j,lambda)
}
rbind(0:10,nights)
```

[,1] [,2] [,3] [,4] 0.00000 1.00000 2.00000 3.00000 nights 23.58432 64.62103 88.53082 80.85815

[,5] [,6] [,7] [,8] 4.00000 5.00000 6.00000 7.000000 nights 55.38783 30.35253 13.86099 5.425587

[,9] [,10] [,11] 8.000000 9.000000 10.0000000 nights 1.858264 0.565738 0.1550122

The largest index is:

max(which(nights>1))-1

## [1] 8

Indeed, for larger values of *x*, we have  $365.25 \cdot P(X = x) < 1$ .

365.25\*dpois(8,lambda) 365.25\*dpois(9,lambda)

[1] 1.858264[1] 0.565738

31: If  $365.25 \cdot P(X = x) < 1$ , then the probability of that number of deliveries is too low to expect that we would ever see it during the year.

## 6.2.7 Other Discrete Distributions

There are numerous commonly-used discrete distributions [5]:

- the Rademacher distribution, which takes values 1 and -1, each with probability 1/2;
- the beta binomial distribution, which describes the number of successes in a series of independent Bernoulli experiments with heterogeneity in the success probability;
- the discrete uniform distribution, where all elements of a finite set are equally likely (balanced coin, unbiased die, first card of a well-shuffled deck, etc.);
- the hypergeometric distribution, which describes the number of successes in the first *m* of a series of *n* consecutive Bernoulli experiments, if the total number of successes is known;
- the Poisson binomial distribution, which describes the number of successes in a series of independent Bernoulli experiments with different success probabilities;
- **Benford's Law**, which describes the frequency of the first digit of many naturally occurring data.
- Zipf's Law, which describes the frequency of words in the English language;
- the beta negative binomial distribution, which describes the number of failures needed to obtain *r* successes in a sequence of independent Bernoulli experiments;
- etc.

# 6.3 Continuous Distributions

How do we approach probabilities where there there are **uncountably infinitely many possible outcomes**, such as one might encounter if *X* represents the height of an individual in the population, for instance (e.g., the outcomes reside in a continuous interval)? What is the probability that a randomly selected person is about 6 feet tall, say?

## 6.3.1 Continuous Random Variables

In the discrete case, the probability mass function  $f_X(x) = P(X = x)$  was the main object of interest. In the continuous case, the analogous role is played by the **probability density function** (p.d.f.), still denoted by  $f_X(x)$ , but there is a major difference with discrete r.v.:

$$f_X(x) \neq P(X=x).$$

The **(cumulative) distribution function** (c.d.f.) of any such random variable *X* is also still defined by

$$F_X(x) = P(X \le x),$$

viewed as a function of a real variable *x*; however  $P(X \le x)$  is not simply computed by adding a few terms of the form  $P(X = x_i)$ .

Note as well that

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F_X(x) = 1.$$

We can describe the **distribution** of the random variable *X* via the following relationship between  $f_X(x)$  and  $F_X(x)$ :<sup>32</sup>

 $f_X(x) = \frac{d}{dx} F_X(x).$ 

### Area Under the Curve

For any a < b, we have

$$\{X \le b\} = \{X \le a\} \cup \{a < X \le b\},\$$

so that

$$P(X \le a) + P(a < X \le b) = P(X \le b)$$

and thus

$$P(a < X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a) = \int_a^b f_X(x) \, dx$$

## **Probability Density Function**

The **probability density function** (p.d.f.) of a continuous random variable *X* is an **integrable** function  $f_X : X(S) \to \mathbb{R}$  such that:

•  $f_X(x) > 0$  for all  $x \in X(\mathcal{S})$  and  $\lim_{x \to \pm \infty} f_X(x) = 0$ ;

• 
$$\int_{\Omega} f_X(x) dx = 1;$$

• for any event  $A = (a, b) = \{X \mid a < X < b\},$ 

$$P(A) = P((a,b)) = \int_a^b f_X(x) \, dx,$$

and the **cumulative distribution function** (c.d.f.)  $F_X$  is given by

$$F_X(x) = P(X \le x) = \int_{-\infty}^x f_X(t) dt.$$

Unlike discrete distributions, the **endpoints** do not affect the probability computations for continuous distributions: for any *a*, *b*,

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b),$$

all taking the value

$$F_X(b) - F_X(a) = \int_a^b f(x) \, dx.$$

Furthermore, for any *x*,

$$P(x < X) = 1 - P(X \le x) = 1 - F_X(x) = 1 - \int_{-\infty}^{x} f_X(t) dt;$$

32: In the continuous case, probability is simply an application of calculus!

and for any *a*,

$$P(X = a) = P(a \le X \le a) = \int_{a}^{a} f_X(x) \, dx = 0.$$

That last result explains why it is pointless to speak of the probability of a random variable taking on a specific value in the continuous case; rather, we are interested in **ranges** of values.

## Examples

• Assume that *X* has the following p.d.f.:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0\\ x/2 & \text{if } 0 \le x \le 2\\ 0 & \text{if } x > 2 \end{cases}$$

Note that  $\int_0^2 f(x) dx = 1$ . The corresponding c.d.f. is given by:

$$F_X(x) = P(X \le x) = \int_{-\infty}^x f_X(t) dt$$
  
= 
$$\begin{cases} 0 & \text{if } x < 0\\ 1/2 \cdot \int_0^x t \, dt = x^2/4 & \text{if } 0 < x < 2\\ 1 & \text{if } x \ge 2 \end{cases}$$

The p.d.f. and the c.d.f. for this r.v. are shown in Figure 6.8.



**Figure 6.8:** P.d.f. and c.d.f. for the continuous r.v. *X* defined above.

• What is the probability of the event A = {X | 0.5 < X < 1.5} if X is the r.v. above?</p>

We need to evaluate

$$P(A) = P(0.5 < X < 1.5) = F_X(1.5) - F_X(0.5)$$
$$= \frac{(1.5)^2}{4} - \frac{(0.5)^2}{4} = \frac{1}{2}.$$





**Figure 6.9:** P.d.f. and c.d.f. for the continuous r.v. *X* defined above, with event *A*.

• What is the probability of the event *B* = {*X* | *X* = 1}?

We need to evaluate

 $P(B) = P(X = 1) = P(1 \le X \le 1) = F_X(1) - F_X(1) = 0.$ 

This is not unexpected: even though  $f_X(1) = 0.5 \neq 0$ , P(X = 1) = 0, as we saw earlier.

• Assume that, for  $\lambda > 0$ , *X* has the following p.d.f.:

.

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \ge 0\\ 0 & \text{if } x < 0 \end{cases}$$

Verify that  $f_X$  is a p.d.f. for all  $\lambda > 0$ , and compute the probability that X > 10.2.

That  $f_X$  is a p.d.f. is obvious; the only work goes into showing that

$$\int_{-\infty}^{\infty} f(x) dx = \int_{0}^{\infty} \lambda \exp(-\lambda x) dx = \lim_{b \to \infty} \int_{0}^{b} \lambda \exp(-\lambda x) dx$$
$$= \lim_{b \to \infty} \lambda \left[ \frac{\exp(-\lambda x)}{-\lambda} \right]_{0}^{b} = \lim_{b \to \infty} \left[ -\exp(-\lambda x) \right]_{0}^{b}$$
$$= \lim_{b \to \infty} \left[ -\exp(-\lambda b) + \exp(0) \right] = 1.$$

The corresponding c.d.f. is given by:

$$F_X(x;\lambda) = P_\lambda(X \le x) = \int_{-\infty}^x f_X(t) dt = \begin{cases} 0 & \text{if } x < 0\\ \lambda \int_0^x \exp(-\lambda t) dt & \text{if } x \ge 0 \end{cases}$$
$$= \begin{cases} 0 & \text{if } x < 0\\ [-\exp(-\lambda t)]_0^x & \text{if } x \ge 0 \end{cases} = \begin{cases} 0 & \text{if } x < 0\\ 1 - \exp(-\lambda x) & \text{if } x \ge 0 \end{cases}$$

Then

$$P_{\lambda}(X > 10.2) = 1 - F_X(10.2; \lambda) = 1 - [1 - \exp(-10.2\lambda)] = \exp(-10.2\lambda)$$

is a function of the **distribution parameter**  $\lambda$  itself:

λ	0.002	0.02	0.2	2	20	200
$P_{\lambda}(X > 10.2)$	0.9798	0.8155	0.13	$1.38 \times 10^{-9}$	$2.54 \times 10^{-89}$	$\approx 0$







**Figure 6.10:** P.d.f. and c.d.f. for the r.v. *X* with  $\lambda = 0.2$ .



**Figure 6.11:** Probability of X > 10.2 (in blue), for X with  $\lambda = 0.2$ .

**Figure 6.12:** Probability of X > 10.2, for X with  $\lambda = 2$ ; the probability is so small  $(1.38 \times 10^{-9})$  that it cannot even be made out in the p.d.f. (blue area).

33: This is not a general property of distributions, however, but a property of this specific family of distributions. Note that in all cases, the **shape** of the p.d.f. and the c.d.f. are the same, although the spike when  $\lambda = 2$  is much higher than that when  $\lambda = 0.2$  – why must that be the case?.<sup>33</sup>

## 6.3.2 Expectation of a Continuous Random Variables

For a continuous random variable *X* with p.d.f.  $f_X(x)$ , the **expectation** of *X* is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx \, .$$

For any function h(X), we can also define

$$\operatorname{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) \, dx \, .$$

## Examples

• Find E[X] and E[X<sup>2</sup>] in the first example, above. we need to evaluate

$$\begin{split} \mathbf{E}[X] &= \int_{-\infty}^{\infty} x f_X(X) \, dx = \int_0^2 x f_X(x) \, dx \\ &= \int_0^2 \frac{x^2}{2} \, dx = \left[\frac{x^3}{6}\right]_{x=0}^{x=2} = \frac{4}{3}; \\ \mathbf{E}[X^2] &= \int_0^2 \frac{x^3}{2} \, dx = 2. \end{split}$$

• Note that the **expectation need not exist**. Compute the expectation of the random variable *X* with p.d.f.

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

let's verify that  $f_X(x)$  is indeed a p.d.f.:

$$\int_{-\infty}^{\infty} f_X(x) \, dx = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1+x^2} \, dx$$
$$= \frac{1}{\pi} [\arctan(x)]_{-\infty}^{\infty} = \frac{1}{\pi} \left[ \frac{\pi}{2} + \frac{\pi}{2} \right] = 1.$$

We can also easily see that

$$F_X(x) = P(X \le x) = \int_{-\infty}^x f_X(t) dt$$
  
=  $\frac{1}{\pi} \int_{-\infty}^x \frac{1}{1+t^2} dt = \frac{1}{\pi} \arctan(x) + \frac{1}{2}$ ,

so that  $P(X \le 3) = \frac{1}{\pi} \arctan(3) + \frac{1}{2}$ , say (see Figure 6.13). The expectation of *X* is

$$\operatorname{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{-\infty}^{\infty} \frac{x}{\pi (1+x^2)} \, dx.$$



**Figure 6.13:** P.d.f. and c.d.f. for the Cauchy distribution, with area under the curve F(3).

If this improper integral exists, then it needs to be equal **both** to

$$\underbrace{\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)} dx + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)} dx}_{\text{candidate 1}}$$

and to the Cauchy principal value

$$\underbrace{\lim_{a\to\infty}\int_{-a}^{a}\frac{x}{\pi(1+x^2)}\,dx}_{\text{candidate 2}}.$$

But it is straightforward to find an antiderivative of  $\frac{x}{\pi(1+x^2)}$ . Set  $u = 1 + x^2$ . Then du = 2xdx and  $xdx = \frac{du}{2}$ , and we obtain

$$\int \frac{x}{\pi(1+x^2)} \, dx = \frac{1}{2\pi} \int u \, du = \frac{1}{2\pi} \ln|u| = \frac{1}{2\pi} \ln(1+x^2).$$

Then the candidate 2 integral reduces to

$$\lim_{a \to \infty} \left[ \frac{\ln(1+x^2)}{2\pi} \right]_{-a}^a = \lim_{a \to \infty} \left[ \frac{\ln(1+a^2)}{2\pi} - \frac{\ln(1+(-a)^2)}{2\pi} \right] = \lim_{a \to \infty} 0 = 0;$$

while the candidate 1 integral reduces to

$$\left[\frac{\ln(1+x^2)}{2\pi}\right]_{-\infty}^{0} + \left[\frac{\ln(1+x^2)}{2\pi}\right]_{0}^{\infty} = 0 - (\infty) + \infty - 0 = \infty - \infty$$

34: Actually, this is not quite true: the integral for candidate 1 is undetermined of the form  $\infty - \infty$ ; usually, when we reach this point in calculus, we have to use some other approach, such as de l'Hôpital's rule, to reduce the expression to a determinate form. The real reason why the mean does not exist is because the value of the integral for candidate 1 depends on how we approach  $-\infty$  and  $\infty$  for each of the constituents. For instance, if the integral exists, we should also have

$$\int_{-\infty}^{\infty} x f_X(x) \, dx = \lim_{a \to \infty} \int_{-a}^{2a} x f_X(x) \, dx.$$

In the Cauchy case, that second integral can be shown to take on the value  $\ln 2/\pi$ , which is different from the principal value 0; hence, the integral does not exist, which is to say, the mean of the Cauchy r.v. does not exist.

which is **undefined**. Thus E[X] cannot not exist, as it would have to be both equal to 0 and be undefined simultaneously.<sup>34</sup>

#### Mean and Variance

Similarly to the discrete case, the **mean** of *X* is defined to be E[X], and the **variance** and **standard deviation** of *X* are, as before,

. .

$$Var[X] \stackrel{\text{def}}{=} E\left[ (X - E[X])^2 \right] = E[X^2] - E^2[X],$$
  
SD[X] =  $\sqrt{Var[X]}.$ 

As in the discrete case, if *X*, *Y* are continuous random variables, and  $a, b \in \mathbb{R}$ , then  $\Gamma[x] + h \Sigma[y] + h \Sigma[y]$ 

$$E[aY + bX] = aE[Y] + bE[X]$$
  
Var[a + bX] = b<sup>2</sup>Var[X]  
SD[a + bX] = |b|SD[X]

The interpretations of the mean as a measure of **centrality** and of the variance as a measure of **dispersion** still apply in the continuous case.

For the time being, however, we cannot easily compute the variance of a sum X + Y, unless X and Y are **independent** random variables:

$$Var[X + Y] = Var[X] + Var[Y]$$

## 6.3.3 Normal Distributions

A **very** important example of a continuous distribution is that provided by the special probability distribution function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The corresponding cumulative distribution function is denoted by

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \phi(t) \, dt \, .$$

A random variable *Z* with this c.d.f. is said to have a **standard normal distribution**, denoted by  $Z \sim \mathcal{N}(0, 1)$ .



**Figure 6.14:** P.d.f. and c.d.f. for the standard normal distribution.

#### Standard Normal Random Variable

The expectation and variance of  $Z \sim \mathcal{N}(0, 1)$  are

$$E[Z] = \int_{-\infty}^{\infty} z \,\phi(z) \,dz = \int_{-\infty}^{\infty} z \,\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \,dz = 0,$$
$$Var[Z] = \int_{-\infty}^{\infty} z^2 \,\phi(z) \,dz = 1,$$
$$SD[Z] = \sqrt{Var[Z]} = \sqrt{1} = 1.$$

Other quantities of interest include:

$$\Phi(0) = P(Z \le 0) = \frac{1}{2}, \quad \Phi(-\infty) = 0, \quad \Phi(\infty) = 1,$$
  
$$\Phi(1) = P(Z \le 1) \approx 0.8413, \quad \text{etc.}$$

## Normal Random Variables

Let  $\sigma > 0$  and  $\mu \in \mathbb{R}$ . If  $Z \sim \mathcal{N}(0, 1)$  and  $X = \mu + \sigma Z$ , then

$$\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1).$$

Thus, the c.d.f. of *X* is given by

$$F_X(x) = P(X \le x) = P(\mu + \sigma Z \le x) = P\left(Z \le \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right);$$

its p.d.f. must then be

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right)$$

Any random variable X with this c.d.f./p.d.f. satisfies

$$E[X] = \mu + \sigma E[Z] = \mu,$$
  

$$Var[X] = \sigma^{2}Var[Z] = \sigma^{2},$$
  

$$SD[X] = \sigma$$

and is said to be **normal with mean**  $\mu$  **and variance**  $\sigma^2$ , denoted by  $X \sim \mathcal{N}(\mu, \sigma^2)$ . As it happens, every general normal *X* can be obtained by a linear transformation of the standard normal *Z*.

Traditionally, probability computations for normal distributions are done with tables which compile values of the standard normal distribution c.d.f., such as the one found in [4] or at <u>ztable.net</u> C<sup>2</sup>. With the advent of freely-available **statistical software**, the need for tabulated values had decreased.<sup>35</sup>

In R, the standard normal c.d.f.  $F_Z(z) = P(Z \le z)$  can be computed with the function pnorm(z) – for instance, pnorm(0)=0.5.<sup>36</sup>

#### Examples

• Let *Z* represent the standard normal random variable. Then:

35: Although it would still be a good idea to learn how to read and use them.

36: In the examples that follow, whenever  $P(Z \le a)$  is evaluated for some *a*, the value is found either by consulting a table or using pnorm.

37: In theory, this cannot be the true model as this would imply that some of the wait times could be negative, but it may nevertheless be an acceptable assumption in practice.

38: The statement from the previous sidenote applies here as well – we will assume that this is understood from this point onward.

- 1.  $P(Z \le 0.5) = 0.6915$ 2. P(Z < -0.3) = 0.38213.  $P(Z > 0.5) = 1 - P(Z \le 0.5) = 1 - 0.6915 = 0.3085$
- 4. P(0.1 < Z < 0.3) = P(Z < 0.3) P(Z < 0.1) = 0.0781
- 5. P(-1.2 < Z < 0.3) = P(Z < 0.3) P(Z < -1.2) = 0.5028
- Suppose that the waiting time (in minutes) in a coffee shop at 9am is normally distributed with mean 5 and standard deviation 0.5.<sup>37</sup> What is the probability that the waiting time for a customer is at most 6 minutes?

Let *X* denote the waiting time. Then  $X \sim \mathcal{N}(5, 0.5^2)$  and the **standardised random variable** is a standard normal:

$$Z = \frac{X-5}{0.5} \sim \mathcal{N}(0,1) \,.$$

The desired probability is

$$P(X \le 6) = P\left(\frac{X-5}{0.5} \le \frac{6-5}{0.5}\right)$$
$$= P\left(Z \le \frac{6-5}{0.5}\right) = \Phi\left(\frac{6-5}{0.5}\right)$$
$$= \Phi(2) = P(Z \le 2) \approx 0.9772.$$

• Suppose that bottles of beer are filled in such a way that the actual volume of the liquid content (in mL) varies randomly according to a normal distribution with  $\mu = 376.1$  and  $\sigma = 0.4$ .<sup>38</sup> What is the probability that the volume in any randomly selected bottle is less than 375mL?

Let *X* denote the volume of the liquid in the bottle. Then

$$X \sim \mathcal{N}(376.1, 0.4^2) \implies Z = \frac{X - 376.1}{0.4} \sim \mathcal{N}(0, 1).$$

The desired probability is thus

$$P(X < 375) = P\left(\frac{X - 376.1}{0.4} < \frac{375 - 376.1}{0.4}\right)$$
$$= P\left(Z < \frac{-1.1}{0.4}\right)$$
$$= P(Z \le -2.75) = \Phi(-2.75) \approx 0.003$$

- If  $Z \sim \mathcal{N}(0, 1)$ , for which values *a*, *b* and *c* do:
  - 1.  $P(Z \le a) = 0.95$ ?

From the table (or R) we see that

$$P(Z \le 1.64) \approx 0.9495, P(Z \le 1.65) \approx 0.9505.$$

Clearly we must have 1.64 < a < 1.65; a linear interpolation provides a decent guess at  $a \approx 1.645$ .

This level of precision is usually not necessary - it is often suf-

ficient to simply present the interval estimate:  $a \in (1.64, 1.65)$ 2.  $P(|Z| \le b) = P(-b \le Z \le b) = 0.99$ ?

Note that

$$P(-b \le Z \le b) = P(Z \le b) - P(Z < -b)$$

However the p.d.f.  $\phi(z)$  is **symmetric** about z = 0, which means that

$$P(Z < -b) = P(Z > b) = 1 - P(Z \le b),$$

and so that

$$\begin{split} P\left(-b \leq Z \leq b\right) &= P(Z \leq b) - \left[1 - P(Z \leq b)\right] \\ &= 2P(Z \leq b) - 1 \end{split}$$

In the question,  $P(-b \le Z \le b) = 0.99$ , so that

$$2P(Z \le b) - 1 = 0.99 \implies P(Z \le b) = \frac{1 + 0.99}{2} = 0.995.$$

Consulting the table we see that

 $P(Z \le 2.57) \approx 0.9949, P(Z \le 2.58) \approx 0.9951;$ 

a linear interpolation suggests that  $b \approx 2.575$ .

3.  $P(|Z| \ge c) = 0.01$ ?

Note that  $\{|Z| \ge c\} = \{|Z| < c\}^c$ , so we need to find *c* such that

 $P(|Z| < c) = 1 - P(|Z| \ge c) = 0.99.$ 

But this is equivalent to

 $P(-c < Z < c) = P(-c \le Z \le c) = 0.99$ 

as  $|x| < y \Leftrightarrow -y < x < y$ , and P(Z = c) = 0 for all *c*. This problem was solved in part b); set  $c \approx 2.575$ .

Normally distributed numbers can be generated by rnorm() in R, which accepts three parameters: n, mean, and sd. The default parameter values are mean=0 and sd=1.

We can draw a single number from  $\mathcal{N}(0, 1)$  as follows:<sup>39</sup>

rnorm(1)

[1] -0.2351372

We can generate a histogram of a sample of size 500, say, from  $\mathcal{N}(0, 1)$  as follows:

z<-rnorm(500)
hist(z)</pre>

39: Note: no seed is provided, so results may vary.

Histogram of z



A histogram with 20 bins is shown below:

brks = seq(min(z),max(z),(max(z)-min(z))/20)
hist(z, breaks = brks)

Histogram of z



For normal distributions with mean  $\mu$  and standard deviation  $\sigma$ , we need to modify the call to rnorm().

For instance, we can draw 5000 observations from  $\mathcal{N}(-2, 3^2)$  using the following code:

w<-rnorm(5000, sd=3, mean=-2)
mean(w)
sd(w)</pre>

[1] -1.943782
[1] 2.920071

A histogram with 50 bins is displayed below:

```
brks = seq(min(w),max(w),(max(w)-min(w))/50)
hist(w, breaks = brks)
```



## **6.3.4 Exponential Distributions**

Assume that cars arrive according to a **Poisson process with rate**  $\lambda$ , that is, the number of cars arriving within a fixed unit time period is a Poisson random variable with parameter  $\lambda$ .

Over a period of time x, we would then expect the number of arrivals N to follow a Poisson process with parameter  $\lambda x$ . Let X be the wait time to the first car arrival. Then

$$P(X > x) = 1 - P(X \le x) = P(N = 0) = \exp(-\lambda x).$$

We say that *X* follows an **exponential distribution**  $Exp(\lambda)$ :

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0\\ 1 - e^{-\lambda x} & \text{for } 0 \le x \end{cases} \text{ and } f_X(x) = \begin{cases} 0 & \text{for } x < 0\\ \lambda e^{-\lambda x} & \text{for } 0 \le x \end{cases}$$

Note that  $f_X(x) = F'_X(x)$  for all x.

If  $X \sim \text{Exp}(4)$ , then  $P(X < 0.5) = F_X(0.5) = 1 - e^{-4(0.5)} \approx 0.865$  is the area of the shaded region in Figure 6.15.

## Properties

## If $X \sim \text{Exp}(\lambda)$ , then:

• 
$$\mu = \mathbb{E}[X] = 1/\lambda$$
, since  

$$\mu = \int_0^\infty x\lambda e^{-\lambda x} \, dx = \left[-\frac{\lambda x + 1}{\lambda}e^{-\lambda x}\right]_0^\infty = \left[0 + \frac{\lambda(0) + 1}{\lambda}e^{-0}\right] = \frac{1}{\lambda};$$



**Figure 6.15:** P.d.f. and c.d.f. for the exponential distribution. with parameter *lambda* = 4 [source unknown].

•  $\sigma^2 = \operatorname{Var}[X] = 1/\lambda^2$ , since

$$\sigma^2 = \int_0^\infty (x - \mathbb{E}[X])^2 \,\lambda e^{-\lambda x} \,dx = \int_0^\infty \left(x - \frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x} \,dx$$
$$= \left[-\frac{\lambda^2 x^2 + 1}{\lambda^2} e^{-\lambda x}\right]_0^\infty = \left[0 + \frac{\lambda^2 (0)^2 + 1}{\lambda^2} e^{-0}\right] = \frac{1}{\lambda^2};$$

• and P(X > s + t | X > t) = P(X > s), for all s, t > 0, since

$$P(X > s + t \mid X > t) = \frac{P(X > s + t \text{ and } X > t)}{P(X > t)}$$
$$= \frac{P(X > s + t)}{P(X > t)} = \frac{1 - F_X(s + t)}{1 - F_X(t)}$$
$$= \frac{\exp(-\lambda(s + t))}{\exp(-\lambda t)}$$
$$= \exp(-\lambda s) = P(X > s).$$

Among continuous r.v., only exponential distributions satisfy this **memoryless** property; geometric distributions are the only memoryless discrete r.v., which makes, In a sense,  $\text{Exp}(\lambda)$  the continuous counterpart of Geo(p).

**Example** The lifetime of a certain type of light bulb follows an exponential distribution whose mean is 100 hours (i.e.  $\lambda = 1/100$ ).

• What is the probability that a light bulb will last at least 100 hours?

Since  $X \sim \text{Exp}(1/100)$ , we have

 $P(X > 100) = 1 - P(X \le 100) = \exp(-100/100) \approx 0.37.$ 

• Given that a light bulb has already been burning for 100 hours, what is the probability that it will last at least 100 hours more?

We seek P(X > 200 | X > 100). By the memory-less property,

 $P(X > 200 \mid X > 100) = P(X > 200 - 100) = P(X > 100) \approx 0.37.$ 

The manufacturer wants to guarantee that their light bulbs will last at least *t* hours. What should *t* be in order to ensure that 90% of the light bulbs will last longer than *t* hours?

We need to find *t* such that P(X > t) = 0.9. In other words, we are looking for *t* such that

$$0.9 = P(X > t) = 1 - P(X \le t) = 1 - F_X(t) = e^{-0.01t},$$

that is,

 $\ln 0.9 = -0.01t \implies t = -100 \ln 0.9 \approx 10.5$  hours.

Exponentially distributed numbers are generated by rexp() in R, with required parameters n and rate.

We can draw from Exp(100) as follows:<sup>40</sup>

rexp(1,100)

40: This is the last time we mention that these are seedless (pseudo-)random numbers.

[1] 0.0009430804

If we repeat the process 1000 times, the empirical mean and variance are:

q<-rexp(1000,100)
mean(q)
var(q)</pre>

[1] 0.01029523
[1] 0.000102973

The histogram is displayed below:

hist(q)



## 6.3.5 Gamma Distributions

Assume that cars arrive according to a Poisson process with rate  $\lambda$ . Recall that if *X* is the time to the first car arrival, then  $X \sim \text{Exp}(\lambda)$ .

If *Y* is the wait time to the *r*th arrival, then *Y* follows a **Gamma distribution** with parameters  $\lambda$ , *r*, denoted  $Y \sim \Gamma(\lambda, r)$ , for which the p.d.f. is

$$f_Y(y) = \begin{cases} 0 & \text{for } y < 0\\ \frac{y^{r-1}}{06(r-1)!} \lambda^r e^{-\lambda y} & \text{for } y \ge 0 \end{cases}$$

The c.d.f.  $F_Y(y)$  exists – it is the area under  $f_Y$  from 0 to y – but it cannot be expressed with elementary functions.

We can also show that

$$\mu = \mathbb{E}[Y] = \frac{r}{\lambda}$$
 and  $\sigma^2 = \operatorname{Var}[Y] = \frac{r}{\lambda^2}$ .

#### Examples

Suppose that an average of 30 customers per hour arrive at a shop in accordance with a Poisson process, that is to say, λ = 1/2 customers arrive on average every minute. What is the probability that the shopkeeper will wait more than 5 minutes before both of the first two customers arrive?

Let *Y* denote the wait time in minutes until the second customer arrives. Then  $Y \sim \Gamma(1/2, 2)$  and

$$P(Y > 5) = \int_{5}^{\infty} \frac{y^{2-1}}{(2-1)!} (1/2)^2 e^{-y/2} \, dy = \int_{5}^{\infty} \frac{y e^{-y/2}}{4} \, dy$$
$$= \frac{1}{4} \left[ -2y e^{-y/2} - 4e^{-y/2} \right]_{5}^{\infty} = \frac{7}{2} e^{-5/2} \approx 0.287.$$

Telephone calls arrive at a switchboard at a mean rate of λ = 2 per minute, according to a Poisson process. Let Y be the waiting time until the 5th call arrives. What is the p.d.f., the mean, and the variance of Y?

We have

$$f_Y(y) = \frac{2^5 y^4}{4!} e^{-2y}, \text{ for } 0 \le y < \infty,$$
$$E[Y] = \frac{5}{2}, \quad Var[Y] = \frac{5}{4}.$$

The Gamma distribution can be extended to cases where r > 0 is not an integer by replacing (r - 1)! by

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} \, dt$$

The exponential and the  $\chi^2$  distributions (we will discuss the latter later) are special cases of the Gamma distribution:  $\text{Exp}(\lambda) = \Gamma(\lambda, 1)$  and  $\chi^2(r) = \Gamma(1/2, r)$ .

Gamma distributed numbers are generated by rgamma(), with required parameters n, shape, and scale.

We can draw from a  $\Gamma(2, 3)$  distribution, for example, using:

rgamma(1, shape=2, scale=1/3)

[1] 2.249483

This can be repeated 1000 times, say, and we get the empirical mean and variance:

```
q<-rgamma(1000,shape=2, scale=1/3)
mean(q)
var(q)</pre>
```

[1] 0.6663675[1] 0.2205931

The corresponding histogram is displayed below:



## 6.3.6 Approximation of the Binomial Distribution

If  $X \sim \mathfrak{B}(n, p)$  then we may interpret X as a sum of **independent and identically distributed** random variables

$$X = I_1 + I_2 + \cdots + I_n$$
 where each  $I_i \sim \mathfrak{B}(1, p)$ .

Thus, according to the **Central Limit Theorem**, <sup>41</sup> for large n we have

$$\frac{X-np}{\sqrt{np(1-p)}} \stackrel{\text{approx}}{\sim} \mathcal{N}(0,1);$$

for large *n* if  $X \stackrel{\text{exact}}{\sim} \mathfrak{B}(n, p)$  then  $X \stackrel{\text{approx}}{\sim} \mathcal{N}(np, np(1-p))$ .

41: We will have more to say on this crucial topic in Section 6.5.

## Normal Approximation with Continuity Correction

When  $X \sim \mathfrak{B}(n, p)$ , we know that  $\mathbb{E}[X] = np$  and  $\operatorname{Var}[X] = np(1-p)$ . If n is large, we may approximate X by a normal random variable in the following way:

$$P(X \le x) = P(X < x + 0.5) = P\left(Z < \frac{x - np + 0.5}{\sqrt{np(1 - p)}}\right)$$

and

$$P(X \ge x) = P(X > x - 0.5) = P\left(Z > \frac{x - np - 0.5}{\sqrt{np(1 - p)}}\right)$$

The **continuity correction terms** are the corresponding  $\pm 0.5$  in the expressions – they are required.

**Example** Suppose  $X \sim \mathfrak{B}(36, 0.5)$ . Provide a normal approximation to the probability  $P(X \le 12)$ .<sup>42</sup>

The expectation and the variance of a binomial r.v. are known:

$$E[X] = 36(0.5) = 18$$
 and  $Var[X] = 36(0.5)(1 - 0.5) = 9$ 

and so

$$P(X \le 12) = P\left(\frac{X - 18}{3} \le \frac{12 - 18 + 0.5}{3}\right)$$
  
$$\stackrel{\text{norm.approx'n}}{\approx} \Phi(-1.83) \stackrel{\text{table}}{\approx} 0.033$$

#### **Computing Binomial Probabilities**

There are thus at least four ways of computing (or approximating) binomial probabilities:

- using the **exact formula** if  $X \sim \mathcal{B}(n, p)$ , then we have  $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$  for each x = 0, 1, ..., n;
- using **tables** if  $n \le 15$  and p is one of 0.1, ..., 0.9, then the corresponding c.d.f. can be found in many textbooks (we must first express the desired probability in terms of the c.d.f.  $P(X \le x)$ ), such as in

$$P(X < 3) = P(X \le 2);$$
  

$$P(X = 7) = P(X \le 7) - P(X \le 6);$$
  

$$P(X > 7) = 1 - P(X \le 7);$$
  

$$P(X \ge 5) = 1 - P(X \le 4), \text{ etc.}$$

- using statistical software (pbinom() in R, say), and
- using the **normal approximation** when np and n(1-p) are both  $\geq 5$ :

$$P(X \le x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right);$$
$$P(X \ge x) \approx 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

42: The binomial probabilities are not typically available in textbooks (or online) for n = 36, although they could be computed directly in R, such as with pbinom(12,26,0.5)=0.0326.

## 6.3.7 Other Continuous Distributions

Some other common continuous distributions are listed in [5]:

- the Beta distribution, a family of 2-parameter distributions with one mode and which is useful to estimate success probabilities (special cases: uniform, arcsine, PERT distributions);
- the logit-normal distribution on (0, 1), which is used to model proportions;
- the Kumaraswamy distribution, which is used in simulations in lieu of the Beta distribution (as it has a closed form c.d.f.);
- the triangular distribution, which is typically used as a subjective description of a population for which there is only limited sample data (it is based on a knowledge of the minimum and maximum and a guess of the mode);
- the **chi-squared** distribution, which is the sum of the squares of *n* independent normal random variables, is used in goodness-of-fit tests in statistics;
- the *F*-distribution, which is the ratio of two chi-squared random variables, used in the analysis of variance;
- the Erlang distribution is the distribution of the sum of k independent and identically distributed exponential random variables, and it is used in queueing models (it is a special case of the Gammma distribution);
- the Pareto distribution, which is used to describe financial data and critical behavior;
- Student's T statistic, which arise when estimating the mean of a normally-distributed population in situations where the sample size is small and the population's standard deviation is unknown;
- the logistic distribution, whose cumulative distribution function is the logistic function;
- the log-normal distribution, which describing variables that are the product of many small independent positive variables;
- etc.

# **6.4** Joint Distributions

Let *X*, *Y* be two continuous random variables. The **joint probability distribution function** (joint p.d.f.) of *X*, *Y* is a function f(x, y) satisfying:

1. 
$$f(x, y) \ge 0$$
, for all  $x, y$ ;  
2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ , and  
3.  $P(A) = \iint_A f(x, y) dx dy$ , where  $A \subseteq \mathbb{R}^2$ .

For a discrete variable, the properties are the same, except that we replace integrals by sums, and we add a property to the effect that  $f(x, y) \le 1$  for all x, y.

Property 3 implies that P(A) is the **volume** of the solid over the region A in the xy plane bounded by the surface z = f(x, y).

#### Examples

- Roll a pair of unbiased dice. For each of the 36 possible outcomes, let X denote the smaller roll, and Y the larger roll (taken from [1]).
  - 1. How many outcomes correspond to the event

$$A = \{(X = 2, Y = 3)\}?$$

The rolls (3, 2) and (2, 3) both give rise to event *A*.

2. What is P(A)?

There are 36 possible outcomes, so  $P(A) = \frac{2}{36} \approx 0.0556$ .

3. What is the joint p.m.f. of *X*, *Y*?

Only one outcome, (X = a, Y = a), gives rise to the event  $\{X = Y = a\}$ . For every other event  $\{X \neq Y\}$ , two outcomes do the trick: (X, Y) and (Y, X). The joint p.m.f. is thus

$$f(x, y) = \begin{cases} 1/36 & 1 \le x = y \le 6\\ 2/36 & 1 \le x < y \le 6 \end{cases}$$

The first property is automatically satisfied, as is the third (by construction). There are only 6 outcomes for which X = Y, all the remaining outcomes (of which there are 15) have X < Y.

Thus,

$$\sum_{x=1}^{6} \sum_{y=x}^{6} f(x, y) = 6 \cdot \frac{1}{36} + 15 \cdot \frac{2}{36} = 1.$$

4. Compute P(X = a) and P(Y = b), for a, b = 1, ..., 6.

For every a = 1, ..., 6,  $\{X = a\}$  corresponds to the following union of events:

$$\{X = a, Y = a\} \cup \{X = a, Y = a + 1\} \cup \dots \cup \{X = a, Y = 6\}.$$

These events are mutually exclusive, so that

$$P(X = a) = \sum_{y=a}^{6} P(\{X = a, Y = y\})$$
  
=  $\frac{1}{36} + \sum_{y=a+1}^{6} \frac{2}{36} = \frac{1}{36} + \frac{2(6-a)}{36}, \quad a = 1, \dots, 6.$ 

Similarly, we get

$$P(Y = b) = \frac{1}{36} + \frac{2(b-6)}{36}, \ b = 1, \dots, 6.$$

These **marginal probabilities** can be found in the margins of the p.m.f.

5. Compute P(X = 3 | Y > 3),  $P(Y \le 3 | X \ge 4)$ .

The notation suggests how to compute these **conditional probabilities**:

$$P(X = 3 | Y > 3) = \frac{P(X = 3 \cap Y > 3)}{P(Y > 3)}$$
$$P(Y = 3 | X \ge 4) = \frac{P(Y = 3 \cap X \ge 4)}{P(X \ge 4)}$$

The region corresponding to  $P(Y > 3) = \frac{27}{36}$  is shaded in red (see Figure 6.16); the region corresponding to  $P(X = 3) = \frac{7}{36}$  is shaded in blue. The region corresponding to

$$P(X = 3 \cap Y > 3) = \frac{6}{36}$$

is the intersection of the regions:

$$P(X = 3 | Y > 3) = \frac{6/36}{27/36} = \frac{6}{27} \approx 0.2222.$$

As  $P(Y \le 3 \cap X \ge 4) = 0$ ,  $P(Y \le 3 \mid X \ge 4) = 0$ .



**Figure 6.16:** Conditional and marginal probabilities in the dice example [1].

6. Are *X* and *Y* independent?

Why didn't we simply use the multiplicative rule to compute

$$P(X = 3 \cap Y > 3) = P(X = 3)P(Y > 3)?$$

It's because X and Y are **not independent**, that is, it is not always the case that

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all allowable *x*, *y*. Indeed,  $P(X = 1, Y = 1) = \frac{1}{36}$ , but

$$P(X=1)P(Y=1) = \frac{11}{36} \cdot \frac{1}{36} \neq \frac{1}{36},$$

so *X* and *Y* are **dependent**.<sup>43</sup>

43: This is often the case when the domain of the joint p.d.f./p.m.f. is not rectangular.

- There are 8 similar chips in a bowl: three marked (0, 0), two marked (1, 0), two marked (0, 1) and one marked (1, 1). A player selects a chip at random and is given the sum of the two coordinates, in dollars (taken from [1]).
  - 1. What is the joint probability mass function of  $X_1$ , and  $X_2$ ?

Let  $X_1$  and  $X_2$  represent the coordinates; we have

$$f(x_1, x_2) = \frac{3 - x_1 - x_2}{8}, \quad x_1, x_2 = 0, 1$$

2. What is the expected pay-off for this game?

The pay-off is simply  $X_1 + X_2$ . The expected pay-off is thus

$$E[X_1 + X_2] = \sum_{x_1=0}^{1} \sum_{x_2=1}^{0} (x_1 + x_2) f(x_1, x_2)$$
$$= 0 \cdot \frac{3}{8} + 1 \cdot \frac{2}{8} + 1 \cdot \frac{2}{8} + 2 \cdot \frac{1}{8}$$
$$= 0.75.$$

• Let *X* and *Y* have joint p.d.f.

$$f(x, y) = 2, \quad 0 \le y \le x \le 1.$$

1. What is the support of f(x, y)?

The support is the set  $S = \{(x, y) : 0 \le y \le x \le 1\}$ , a triangle in the *xy* plane bounded by the *x*-axis, the line *x* = 1, and the line *y* = *x*.

The support is the blue triangle shown in Figure 6.17.





2. What is  $P(0 \le X \le 0.5, 0 \le Y \le 0.5)$ ?

We need to evaluate the integral over the shaded area:

$$P(0 \le X \le 0.5, 0 \le Y \le 0.5) = P(0 \le X \le 0.5, 0 \le Y \le X)$$
$$= \int_0^{0.5} \int_0^x 2\,dy dx = \int_0^{0.5} [2y]_{y=0}^{y=x} dx = \int_0^{0.5} 2x\,dx = 1/4.$$

3. What are the marginal probabilities P(X = x) and P(Y = y)?

For  $0 \le x \le 1$ , we get

$$P(X = x) = \int_{-\infty}^{\infty} f(x, y) \, dy = \int_{y=0}^{y=x} 2 \, dy = [2y]_{y=0}^{y=x} = 2x,$$

and for  $0 \le y \le 1$ ,

$$P(Y = y) = \int_{-\infty}^{\infty} f(x, y) \, dx = \int_{x=y}^{x=1} 2 \, dx = [2x]_{x=y}^{x=1} = 2 - 2y.$$

4. Compute E[X], E[Y],  $E[X^2]$ ,  $E[Y^2]$ , and E[XY].

We have

$$\begin{split} \mathrm{E}[X] &= \iint_{S} xf(x,y) \, dA = \int_{0}^{1} \int_{0}^{x} 2x \, dy dx \\ &= \int_{0}^{1} [2xy]_{y=0}^{y=x} \, dx = \int_{0}^{1} 2x^{2} \, dx = \left[\frac{2}{3}x^{3}\right]_{0}^{1} = \frac{2}{3}; \\ \mathrm{E}[Y] &= \iint_{S} yf(x,y) \, dA = \int_{0}^{1} \int_{y}^{1} 2y \, dx dy \\ &= \int_{0}^{1} [2xy]_{x=y}^{x=1} \, dy = \int_{0}^{1} (2y - 2y^{2}) \, dy = \left[y^{2} - \frac{2}{3}y^{3}\right]_{0}^{1} = \frac{1}{3}; \\ \mathrm{E}[X^{2}] &= \iint_{S} x^{2}f(x,y) \, dA = \int_{0}^{1} \int_{0}^{x} 2x^{2} \, dy dx \\ &= \int_{0}^{1} [2x^{2}y]_{y=0}^{y=x} \, dx = \int_{0}^{1} 2x^{3} \, dx = \left[\frac{1}{2}x^{4}\right]_{0}^{1} = \frac{1}{2}; \\ \mathrm{E}[Y^{2}] &= \iint_{S} y^{2}f(x,y) \, dA = \int_{0}^{1} \int_{y}^{1} 2y^{2} \, dx dy \\ &= \int_{0}^{1} [2xy^{2}]_{x=y}^{x=1} \, dy = \int_{0}^{1} (2y - 2y^{3}) \, dy = \left[\frac{2}{3}y^{3} - \frac{1}{2}y^{4}\right]_{0}^{1} = \frac{1}{6}; \\ \mathrm{E}[XY] &= \iint_{S} xyf(x,y) = \int_{0}^{1} \int_{0}^{x} 2xy \, dy dx \\ &= \int_{0}^{2} [xy^{2}]_{y=0}^{y=x} = \int_{0}^{1} x^{2} \, dx = \left[\frac{x^{4}}{4}\right]_{0}^{1} = \frac{1}{4}. \end{split}$$

5. Are *X* and *Y* independent?

They are not, as the **support** of the joint p.d.f. is not rectangular.

The **covariance** of two random variables *X* and *Y* can give some indication of how they depend on one another:

$$\operatorname{Cov}(X,Y) = \operatorname{E}[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y].$$

When X = Y, the covariance reduces to the variance.<sup>44</sup> In the last example,

for instance, we have:  $\operatorname{Var}[X] = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$ ,  $\operatorname{Var}[X] = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{18}$ , and  $\operatorname{Cov}(X, Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{36}$ .<sup>45</sup>

44: Note that the covariance could be negative, unlike the variance.

45: We will use the covariance again in Chapters 8 and 10.

In R, we can generate a **multivariate joint normal** *via* MASS's mvrnorm(), whose required paramters are n, a mean vector mu and a covariance matrix Sigma.

We look at two standard bivariate joint normals.

mu1 = c(0,0); mu2 = c(-3,12)
Sigma1 = matrix(c(1,0,0,1),2,2)
Sigma2 = matrix(c(110,15,15,3),2,2)

We sample 1000 observations from each joint normal.

```
library(MASS)
al<-mvrnorm(1000,mul,Sigmal)
al<-data.frame(al)
a2<-mvrnorm(1000,mu2,Sigma2)
a2<-data.frame(a2)</pre>
```

What would you expect to see when we plot the data? In the first case, the covariance matrix is the identity (**diagonal**), so we expect the blob to be circular; in the second case, we have a **non-diagonal** covariance matrix, which stretches the blob.<sup>46</sup>

```
library(ggplot2)
library(hexbin)
qplot(X1, X2, data=a1, geom="hex")
qplot(X1, X2, data=a, geom="hex") +
    ylim(-40,40) + xlim(-40,40)
```



46: The blob will have a "positive" slope since Cov(X, Y) = 15 > 0.

# 6.5 Central Limit Theorem and Sampling Distributions

In this section, we introduce one of the fundamental results of probability theory and statistical analysis.

## 6.5.1 Sampling Distributions

A **population** is a set of similar items which of interest in relation to some questions or experiments.

In some situations, it is impossible to observe the entire set of observations that make up a population – perhaps the entire population is too large to query, or some units are out-of-reach.

In these cases, we can only hope to infer the behaviour of the entire population by considering a **sample** (subset) of the population.

Suppose that  $X_1, ..., X_n$  are *n* **independent** random variables, each having the same c.d.f. *F*, i.e.they are **identically distributed**. Then,  $\{X_1, ..., X_n\}$  is a **random sample** of size *n* from the population, with c.d.f. *F*.

Any function of such a random sample is called a **statistic** of the sample; the probability distribution of a statistic is called a **sampling distribution**.

Recall the linear properties of the expectation and the variance: if *X* is a random variable and  $a, b \in \mathbb{R}$ , then

$$E[a + bX] = a + bE[X],$$
  
Var [a + bX] = b<sup>2</sup>Var[X],  
SD[a + bX] = |b|SD[X].

## Sum of Independent Random Variables

For any random variables *X* and *Y*, we have

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y].$$

In general,

$$Var[X + Y] = Var[X] + 2Cov(X, Y) + Var[Y];$$

if **in addition** *X* and *Y* are **independent**, then

$$Var[X + Y] = Var[X] + Var[Y].$$

More generally, if  $X_1, X_2, \ldots, X_n$  are **independent**, then

$$\operatorname{E}\left[\sum_{i=1}^{n} X_{i}\right] = \sum_{i=1}^{n} \operatorname{E}[X_{i}] \quad \text{and} \quad \operatorname{Var}\left[\sum_{i=1}^{n} X_{i}\right] = \sum_{i=1}^{n} \operatorname{Var}[X_{i}].$$

#### Independent and Identically Distributed Random Variables

A special case of the above occurs when all of  $X_1, \ldots, X_n$  have **exactly the same distribution**. In that case we say they are **independent and identically distributed**, which is traditionally abbreviated to "**iid**".

If 
$$X_1, \ldots, X_n$$
 are iid, and

$$E[X_i] = \mu$$
 and  $Var[X_i] = \sigma^2$  for  $i = 1, ..., n_i$ 

then

$$\operatorname{E}\left[\sum_{i=1}^{n} X_{i}\right] = n\mu \quad \text{and} \quad \operatorname{Var}\left[\sum_{i=1}^{n} X_{i}\right] = n\sigma^{2}.$$

#### Examples

• A random sample of size 100 is taken from a population with mean 50 and variance 0.25. Find the expected value and variance of the **sample total**.

This problem translates to "if  $X_1, \ldots, X_{100}$  are iid with  $E[X_i] = \mu = 50$  and  $Var[X] = \sigma^2 = 0.25$  for  $i = 1, \ldots, 100$ , find  $E[\tau]$  and  $Var[\tau]$  for

$$\tau = \sum_{i=1}^n X_i."$$

According to the iid formulas,

$$E\left[\sum_{i=1}^{n} X_{i}\right] = 100\mu = 5000, \quad Var\left[\sum_{i=1}^{n} X_{i}\right] = 100\sigma^{2} = 25.$$

The mean value of potting mix bags weights is 5 kg, with standard deviation 0.2. If a shop assistant carries 4 bags (selected independently from the stock) then what is the expected value and standard deviation of the total weight carried?

There is an implicit "population" of bag weights. Let  $X_1, X_2, X_3, X_4$ be iid with  $E[X_i] = \mu = 5$ ,  $SD[X_i] = \sigma = 0.2$  and  $Var[X_i] = \sigma^2 = 0.2^2 = 0.04$  for i = 1, 2, 3, 4. Let  $\tau = X_1 + X_2 + X_3 + X_4$ .

According to the iid formulas,

 $E[\tau] = n\mu = 4 \cdot 5 = 20,$   $Var[\tau] = n\sigma^2 = 4 \cdot 0.04 = 0.16.$ 

Thus,  $SD[\tau] = \sqrt{0.16} = 0.4$ .

#### Sample Mean

The **sample mean** is a typical statistic of interest:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

If  $X_1, \ldots, X_n$  are iid with  $E[X_i] = \mu$  and  $Var[X_i] = \sigma^2$  for all  $i = 1, \ldots, n$ , then

$$E\left[\overline{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}X_{i}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}X_{i}\right] = \frac{1}{n}(n\mu) = \mu$$
$$Var\left[\overline{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n}X_{i}\right] = \frac{1}{n^{2}}Var\left[\sum_{i=1}^{n}X_{i}\right] = \frac{1}{n^{2}}(n\sigma^{2}) = \frac{\sigma^{2}}{n}$$

**Example** A set of scales returns the true weight of the object being weighed plus a random error with mean 0 and standard deviation 0.1 g. Find the standard deviation of the average of 9 measurements of an object.

Suppose the object has true weight  $\mu$ . The "random error" indicates that each measurement i = 1, ..., 9 is written as  $X_i = \mu + Z_i$  where  $E[Z_i] = 0$  and  $SD[Z_i] = 0.1$  and the  $Z_i$ 's are iid.

The  $X_i$ 's are iid with  $E[X_i] = \mu$  and  $SD[X_i] = \sigma = 0.1$ . If we average  $X_1, \ldots, X_n$  (with n = 9) to get  $\overline{X}$ , then

$$\operatorname{E}\left[\overline{X}\right] = \mu \text{ and } \operatorname{SD}\left[\overline{X}\right] = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{9}} = \frac{1}{30} \approx 0.033$$

We do not need to know the **actual** distribution of the  $X_i$ ; only  $\mu$  and  $\sigma^2$  are required to compute  $E[\overline{X}]$  and  $Var[\overline{X}]$ .

#### Sum of Independent Normal Random Variables

Another interesting case occurs when we have **multiple independent normal** random variables on the same experiment.

Suppose  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$  for i = 1, ..., n, and all the  $X_i$  are independent. We already know that

$$E[\tau] = E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n] = \mu_1 + \dots + \mu_n;$$
  

$$Var[\tau] = Var[X_1 + \dots + X_n] = Var[X_1] + \dots + Var[X_n] = \sigma_1^2 + \dots + \sigma_n^2.$$

It turns out that, under these hypotheses,  $\tau$  is **also normally distributed**, i.e.

$$\tau = \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mathbf{E}[\tau], \operatorname{Var}[\tau]) = \mathcal{N}\left(\mu_1 + \dots + \mu_n, \sigma_1^2 + \dots + \sigma_n^2\right).$$

Thus, if  $\{X_1, \ldots, X_n\}$  is a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ , then  $\sum_{i=1}^n X_i$  and  $\overline{X}$  are also normal, which, combined with the above work, means that

$$\sum_{i=1}^{n} X_{i} \sim \mathcal{N}\left(n\mu, n\sigma^{2}\right) \quad \text{and} \quad \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^{2}}{n}\right) \,.$$

**Example** Suppose that the population of students' weights is normal with mean 75 kg and standard deviation 5 kg. If 16 students are picked at random, what is the distribution of the (random) total weight  $\tau$ ? What is the probability that the total weight exceeds 1250 kg?