# Probability and Applications 6




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 $\mathcal{S}$. A sample space can be **discrete** or **continuous**.

An **event** is a collection of outcomes from the sample space $\mathcal{S}$. Events will be denoted by $A$, $B$, $E_1$, $E_2$, etc.

#### Examples

- Toss a fair coin – the corresponding (discrete) sample space is $\mathcal{S} = \{\text{Head}, \text{Tail}\}$.
- Roll a die – the corresponding (discrete) sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, 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 $\mathcal{S} = (0, \infty)$, the positive half line, and various events by subsets of $\mathcal{S}$, such as
  - sample is less than 1.5 grams: $(0, 1.5)$;
  - sample exceeds 5 grams: $(5, \infty)$.

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

- the **union** $A \cup B$ of $A$ and $B$ are all outcomes in $\mathcal{S}$ contained in either $A$ or $B$;
- the **intersection** $A \cap B$ of $A$ and $B$ are all outcomes in $\mathcal{S}$ contained in both $A$ and $B$;
- the **complement** $A^c$ of $A$ (sometimes denoted $\overline{A}$ or $-A$) is the set of all outcomes in $\mathcal{S}$ that are **not** in $A$.

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

1: Events can be represented graphically using Venn diagrams – mutually exclusive events are those which do not have a common intersection.

#### 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.

- 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 $n_1$ possibilities at stage 1;
- regardless of the 1st outcome there are $n_2$ possibilities at stage 2,
- ...
- regardless of the previous outcomes, there are $n_k$ 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 $1, \ldots, n$. 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

$$\underbrace{n \times n \times \cdots \times n}_{r \text{ stages}} = n^r$$

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

#### 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

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

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

### Factorial Notation

For a positive integer $n$, write

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

There are two possibilities:

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

$$\begin{aligned} {}_nP_r &= \frac{n(n-1)\cdots(n-r+1)}{} \frac{(n-r)\cdots 1}{(n-r)\cdots 1} \\ &= \frac{n!}{(n-r)!} = n \times \cdots \times (n-r+1). \end{aligned}$$

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

$${}_nP_r = \frac{n!}{(n-r)!}, \quad \text{for all } r \leq n.$$

**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 $\{0, 1, \ldots, 9\}$?

  **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 ${}_nC_r$. We can derive the expression for ${}_nC_r$ by noting that the following two processes are equivalent:

- take an **ordered** sample of size $r$ (there are ${}_nP_r$ ways to do this);
- take an **unordered** sample of size $r$ (there are ${}_nC_r$ ways to do this) **and then** rearrange (permute) the objects in the sample (there are $r!$ ways to do this).

Thus

$${}_nP_r = {}_nC_r \times r! \implies {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n-r)!\,r!} = \binom{n}{r};$$

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

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

There is a variety of binomial coefficient identities, such as

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

### 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 $\mathcal{S} = \{\text{Head}, \text{Tail}\}$, i.e., $N = 2$. The probability of observing a Head on a toss is thus $\frac{1}{2}$.
- 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
$$\text{Prob(number is a multiple of 3)} = P(A) = \frac{2}{6} = \frac{1}{3}.$$
- The probabilities of seeing an even/odd number are:
$$\text{Prob}\{\text{even}\} = P(\{2, 4, 6\}) = \frac{3}{6} = \frac{1}{2};$$
$$\text{Prob}\{\text{prime}\} = P(\{2, 3, 5\}) = 1 - P(\{1, 4, 6\}) = \frac{1}{2}.$$
- 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.[2]

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

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 [37] and Chapter 26 for details) and the **propensity interpretation**.[3]

3: Introducing causality as a mechanism.

#### Axioms of Probability

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

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

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

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 $\mathcal{S} = A \cup A^c$, and $A$ and $A^c$ are mutually exclusive, then

$$1 \overset{\mathbf{A2}}{=} P(\mathcal{S}) = P(A \cup A^c) \overset{\mathbf{A4}}{=} P(A) + P(A^c)$$
$$\implies P(A^c) = 1 - P(A).$$

#### 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 $A \cap B = \varnothing$. 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 $W = \{\text{the ball is white}\}$, $R = \{\text{the ball is red}\}$ and $B = \{\text{the ball is black}\}$. Then

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

and $P(W \text{ or } R) = 7/8$.

#### 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 $P(A \text{ fails}) = 0.2$, $P(B \text{ fails}) = 0.3$ and $P(\text{both } A \text{ and } B \text{ fail}) = 0.15$. Find $P(\text{at least one of } A \text{ and } B \text{ fails})$ and $P(\text{neither } A \text{ nor } B \text{ fails})$.

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

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

If $A$, $B$ are mutually exclusive, $P(A \cap B) = P(\varnothing) = 0$ 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:

$$\begin{aligned} P(A \cup B \cup C) = &P(A) + P(B) + P(C) \\ &- P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C). \end{aligned}$$

### 6.1.6 Conditional Probability and Independent Events

Any two events $A$ and $B$ satisfying

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

are said to be **independent**.[4] When events are not independent, we say that they are **dependent** or **conditional**.

4: This is a purely mathematical definition, but it agrees with the intuitive notion of independence in simple examples.

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):

$$\begin{aligned} \varnothing = P(A \cap B) = P(A) \times P(B) &\implies P(A) = 0 \text{ or } P(B) = 0 \\ &\implies A = \varnothing \text{ or } B = \varnothing. \end{aligned}$$

#### Examples

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

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

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

  $$\begin{aligned} P(A) &= P(\{HH\}) + P(\{HT\}) - P(\{HH\} \cap \{HT\}) \\ &= \frac{1}{4} + \frac{1}{4} - P(\varnothing) = \frac{1}{2} - 0 = \frac{1}{2}. \end{aligned}$$

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

Axiom **A2**, $P(\mathcal{S}) = P(1) + \cdots + P(6) = 1$, 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,[5] then

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

  5: Is it clear what is meant by "independent tosses' '?

- 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.[6]
  The next step is to decide what type engine failure will force a plane down:[7]

  - A 2-engine plane will be forced down if both engines fail – the probability is $p^2$;
  - 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 ${}_3C_2 = \frac{3!}{2!1!} = 3$ 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 $p^3$.

    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 $0 \leq p \leq 1$).

  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.

- (Taken from [38]) 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.

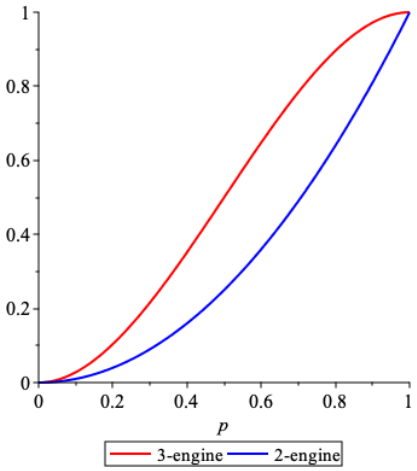



**Figure 6.1:** Failure probability for the 2-engine and 3-engine planes.

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).

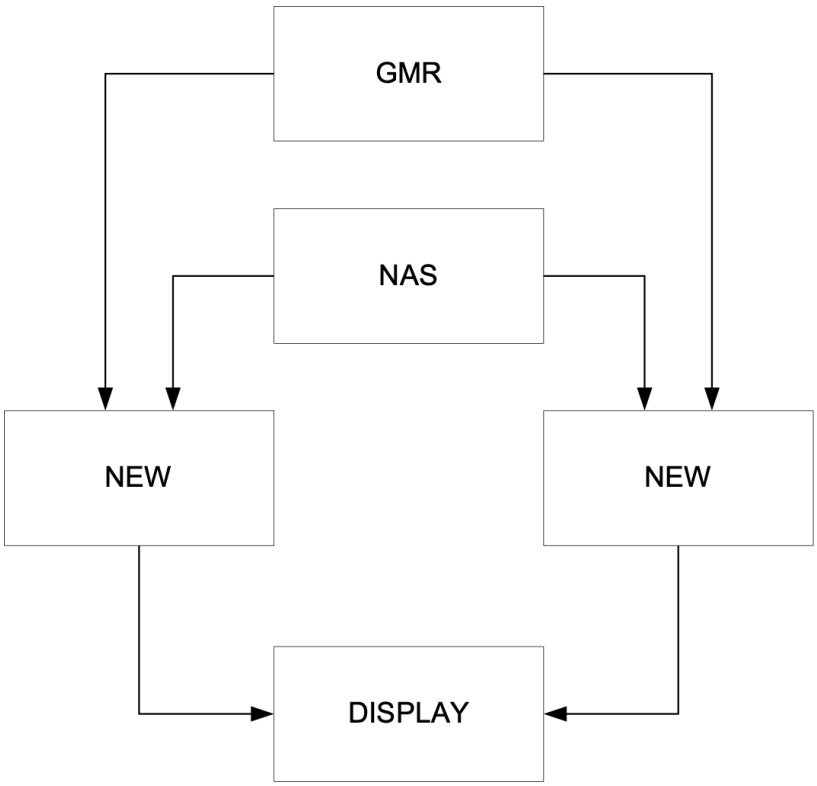



**Figure 6.2:** Conceptual model of air traffic control security system.

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 $P = P(\text{failure})$. We assume that $R_{\text{GMR}} = R_{\text{NAS}} = 0.9999$,[8] and that the components' failure probabilities are independent.

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

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

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

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

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

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 $R_{\text{NEW}}$.

| $R_{\text{NEW}}$ | 0.1 | 0.2 | 0.5 | 0.75 | 0.99 | 0.999 | 0.9999 | 0.99999 |
|---|---|---|---|---|---|---|---|---|
| $\rho$ (%) | 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).[9] If the new module is **as reliable as** GMR and NAS, that is, if

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

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

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

**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.

| | **O** | **A** | **B** | **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:

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

  $A \cap B = A$. 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.

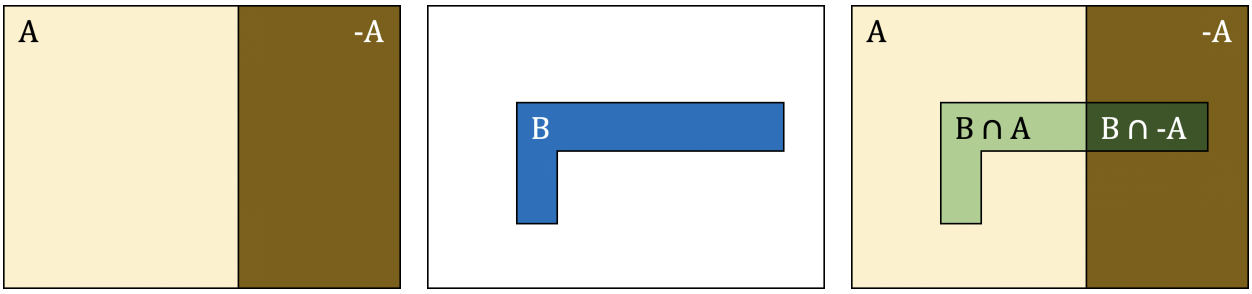



**Figure 6.3:** Decomposition of *B via A*.

**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 $\varnothing \neq A \neq \mathcal{S}$, we have

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

so that

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

This generalizes as follows: if $A_1, ...A_k$ are **mutually exclusive** and **exhaustive** (i.e., $A\_i \cap A\_j = \varnothing$ for all $i \neq j$ and $A_1 \cup .... \cup A_k = \mathcal{S}$), then for any event $B$

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

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

The blood types $\{\mathbf{O}, \mathbf{A}, \mathbf{B}, \mathbf{AB}\}$ 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,

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

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.[10]

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

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:

- $P(\text{hypothesis})$ is the **prior** – the probability of the hypothesis being true prior to the experiment;
- $P(\text{hypothesis} \mid \text{data})$ is the **posterior** – the probability of the hypothesis being true once the experimental data is taken into account;
- $P(\text{data} \mid \text{hypothesis})$ 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**.

**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 $A_1, \ldots A_k$ are **mutually exclusive** and **exhaustive** events, then for any event $B$ and for each $1 \leq i \leq k$,

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

**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,

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

     so that

     $$\begin{aligned} P(D) &= P(D \mid \mathrm{S}) \times P(\mathrm{S}) + P(D \mid \mathrm{M}) \times P(\mathrm{M}) + P(D \mid \mathrm{Pa}) \times P(\mathrm{Pa}) \\ &= 0.12 \cdot 0.5 + 0.15 \cdot 0.3 + 0.25 \cdot 0.2 \\ &= 0.155 = 15.5\%. \end{aligned}$$

  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 $P(\mathrm{M} \mid D)$, $P(\mathrm{S} \mid D)$, and $P(\mathrm{Pa} \mid D)$ (and not the priors!), computed using Bayes' Theorem:

     $$\begin{aligned} P(\mathrm{S} \mid D) &= \tfrac{P(D|\mathrm{S})P(\mathrm{S})}{P(D)} = \tfrac{0.12 \times 0.5}{0.155} \approx 38.7\% \\ P(\mathrm{M} \mid D) &= \tfrac{P(D|\mathrm{M})P(\mathrm{M})}{P(D)} = \tfrac{0.15 \times 0.3}{0.155} \approx 29.0\% \\ P(\mathrm{Pa} \mid D) &= \tfrac{P(D|\mathrm{Pa})P(\mathrm{Pa})}{P(D)} = \tfrac{0.25 \times 0.2}{0.155} \approx 32.3\% \end{aligned}$$

     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).[11]

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

- [**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?

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

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

12: Note that the prior is $P(\mathrm{D}) = 1/3$.

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

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

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

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

  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.

14: Note that the prior is $P(\mathrm{D}) = 1/3$.

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

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 of 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
```

Le 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         B              B
2         B              B
3         A              B
4         C              C
5         A              C
6         A              A
```

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)
```

```
  locations player.guesses open.door
1         B              B         C
2         B              B         C
3         A              B         C
4         C              C         A
5         A              C         B
6         A              A         B
```

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
```

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)
}
```

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)
```

```
  locations player.guesses open.door no.switch.win switch.door switch.win
1         B              B         C           TRUE           A      FALSE
2         B              B         C           TRUE           A      FALSE
3         A              B         C          FALSE           A       TRUE
4         C              C         A           TRUE           B      FALSE
5         A              C         B          FALSE           A       TRUE
6         A              A         B           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.

## 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**.[16]

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

### 6.2.1 Random Variables and Distributions

Recall that, for any random "experiment", the set of all possible outcomes is denoted by $\mathcal{S}$. A **random variable** (r.v.) is a function $X : \mathcal{S} \to \mathbb{R}$, which is to say, it is a rule that associates a (real) number to every outcome of the experiment; $\mathcal{S}$ is the **domain** of the r.v. $X$ and $X(\mathcal{S}) \subseteq \mathbb{R}$ 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**,[17] we say that $X$ is a **discrete r.v.** and the p.d.f. is called a **probability mass function** (p.m.f.).

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.

#### 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 $X(\mathcal{S}) = \{x_i\}$, then we can define the events $A_i = \{s \in \mathcal{S} : X(s) = x_i\}$ :

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

#### Properties

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

- $0 < f(x) \leq 1$ 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 $a, b \in \mathbb{R}$,

$$P(a < X) = 1 - P(X \leq a) = 1 - F(a)$$
$$P(X < b) = P(X \leq b) - P(X = b) = F(b) - f(b)$$

- for any $a, b \in \mathbb{R}$,

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

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

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

**Examples**

- Flip a fair coin – the outcome space is $\mathcal{S} = \{\text{Head}, \text{Tail}\}$. Let $X : S \to \mathbb{R}$ be defined by $X(\text{Head}) = 1$ and $X(\text{Tail}) = 0$. Then $X$ is a discrete random variable.[18]

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

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

- Roll a fair die – the outcome space is $\mathcal{S} = \{1, \ldots, 6\}$. Let $X : \mathcal{S} \to \mathbb{R}$ be defined by $X(i) = i$ for $i = 1, \ldots, 6$. 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

  $$f(i) = P(X = i) = 1/6, \text{ for } i = 1, \ldots, 6,$$
  $$f(x) = 0 \text{ for all other } x.$$

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

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

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

  $$\begin{aligned} P(3 \le X \le 5) &= P(X = 3) + P(X = 4) + P(X = 5) \\ &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}, \end{aligned}$$

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

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

- The number of calls received over a specific time period, $X$, is a discrete random variable, with potential values $0, 1, 2, \ldots$.

- 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 ♦ or ♥, and all of the 5 cards in the hands are ♠ or ♣. 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 $x$ ♦ or ♥ in the hand, and $\binom{26}{5-x}$ ways of having $5 - x$ ♠ and ♣ 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 $0 < f(x) \leq 1$ for all $x$ and

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

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

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

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

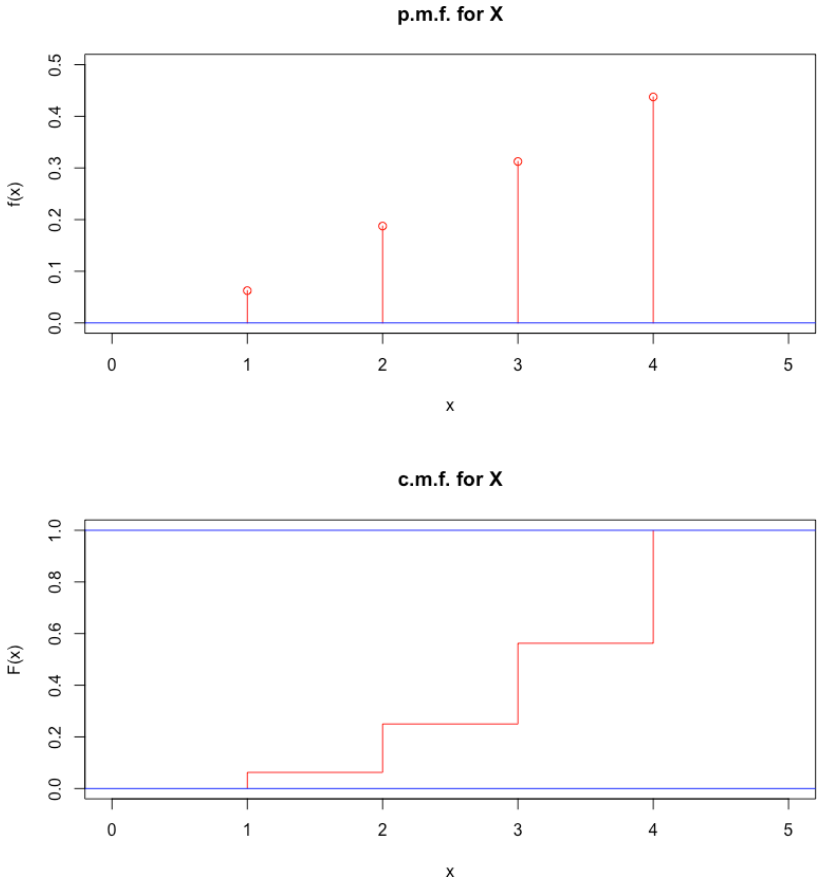



**Figure 6.5:** P.m.f. and c.m.f. for the discrete r.v. $X$ defined in the last example.

### 6.2.2 Expectation of a Discrete Random Variable

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

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

where the sum extends over all values of $x$ taken by $X$.[19] The definition can be extended to a general function of $X$:

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

As an important example, note that

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

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

#### Examples

- What is the expectation on the roll $Z$ of 6−sided die?

  If the die is fair, then

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

- 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.[20] The expected value of $X$ is thus

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

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

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

$$\mathrm{E}[Z^2] = \sum_z z^2 P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} z^2 = \frac{1}{6}(1^2 + \cdots + 6^2) = \frac{91}{6}$$

$$\begin{aligned}\mathrm{E}[(Z - 3.5)^2] &= \sum_{z=1}^{6} (z - 3.5)^2 \times P(Z = z) = \frac{1}{6} \sum_{z=1}^{6} (z - 3.5)^2 \\ &= \frac{(1 - 3.5)^2 + \cdots + (6 - 3.5)^2}{6} = \frac{35}{12}.\end{aligned}$$

#### 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,

$$\mu_Z = \mathrm{E}[Z] = 3.5$$

Note that in the final example, we could have written

$$\mathrm{E}[(Z-3.5)^2] = \mathrm{E}[(Z-\mathrm{E}[Z])^2].$$

This is an important quantity associated to a random variable $X$, its **variance** $\mathrm{Var}[X]$.

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

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu_X)^2] = \sum_x (x-\mu_X)^2 P(X=x)\\ &= \sum_x \left(x^2 - 2x\mu_X + \mu_X^2\right) f(x)\\ &= \sum_x x^2 f(x) - 2\mu_X \sum_x x f(x) + \mu_X^2 \sum_x f(x)\\ &= \mathrm{E}[X^2] - 2\mu_X\mu_X + \mu_X^2 \cdot 1\\ &= \mathrm{E}[X^2] - \mu_X^2.\end{aligned}$$

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

### Standard Deviation

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

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

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.

| $x$ | $P(X=x)$ | $y$ | $P(Y=y)$ |
|---|---|---|---|
| −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 $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ and

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

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).

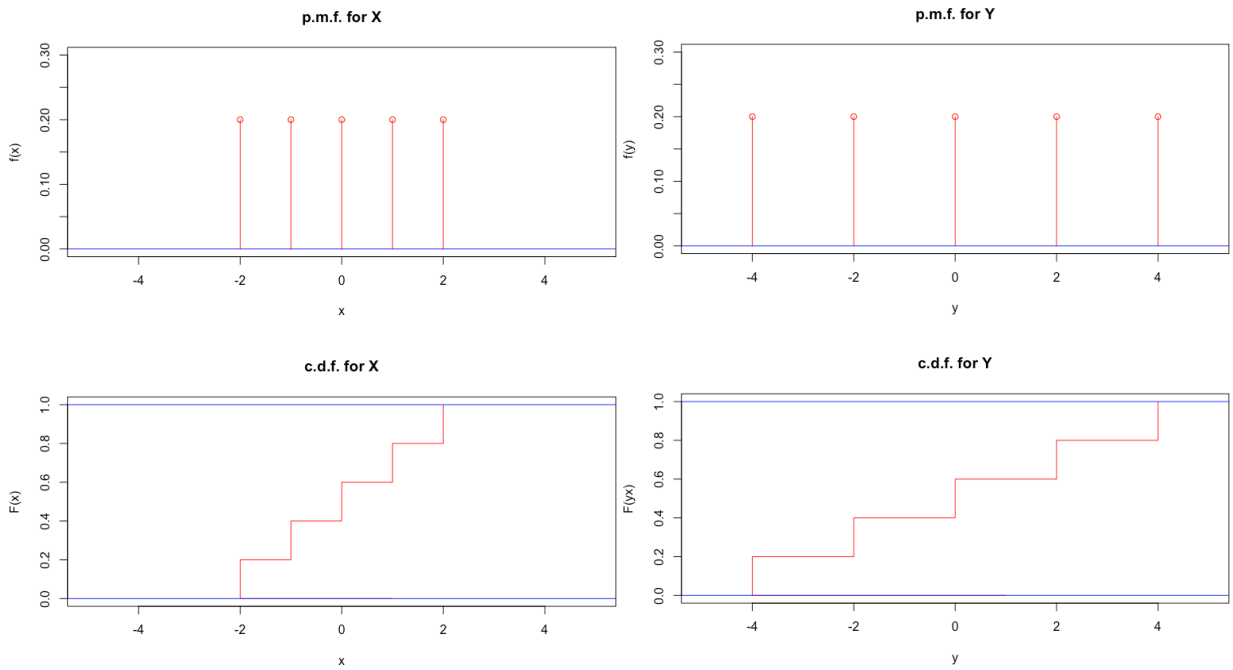



**Figure 6.6:** R.v. $X$ (left) and $Y$ (right) for two uniform distributions, as defined in the example.

#### Properties

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

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

### 6.2.3 Binomial Distributions

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

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

#### 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 $\binom{n}{0} = 1$.
- $\binom{6}{2} = \frac{6!}{2!\times 4!} = \frac{4!\times 5\times 6}{2!\times 4!} = \frac{5\times 6}{2} = 15$.
- $\binom{27}{22} = \frac{27!}{22!\times 5!} = \frac{22!\times 23\times 24\times 25\times 26\times 27}{5!\times 22!} = \frac{23\times 24\times 25\times 26\times 27}{120}$.

#### 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 $(n, p)$:

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

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

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

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

#### Expectation and Variance

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

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

and

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

(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 $\mathcal{B}(12, 0.1)$.

  Find

  1. $\mathrm{E}[X]$ and $\mathrm{Var}[X]$;
  2. $P(X = 3)$;
  3. $P(X \leq 3)$.

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

$$\mathrm{E}[X] = np \quad \text{and} \quad \mathrm{Var}[X] = np(1-p).$$

With $n = 12$ and $p = 0.1$, we obtain

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

2. By definition,

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

3. By definition,

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

This sum can be computed directly, however, for $X \sim \mathcal{B}(12, 0.1)$, $P(X \le 3)$ 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 |
| | 10 | | | 1.0000 | 0.9997 | 0.9968 | 0.9804 | 0.9150 | 0.7251 | 0.3410 |
| | 11 | | | | 1.0000 | 0.9998 | 0.9978 | 0.9862 | 0.9313 | 0.7176 |
| | 12 | | | | | 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 $P(X > 100)$.

21: No families or late bus?

If all passengers show up independently of one another,[21] we can model $X \sim \mathcal{B}(101, 0.97)$ and

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

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

$$\begin{aligned} P(X > 100) &= 1 - P(X \leq 100) \\ &= 1 - \sum_{x=0}^{100} \binom{125}{x}(0.97)^x(0.03)^{125-x}. \end{aligned}$$

22: Do these results match your intuition?

This sum is harder to compute directly, but is very nearly 1 (try it with R, say).[22]

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 $\mathcal{B}(11, 0.2)$ in R as follows:

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

```
[1] 5
```

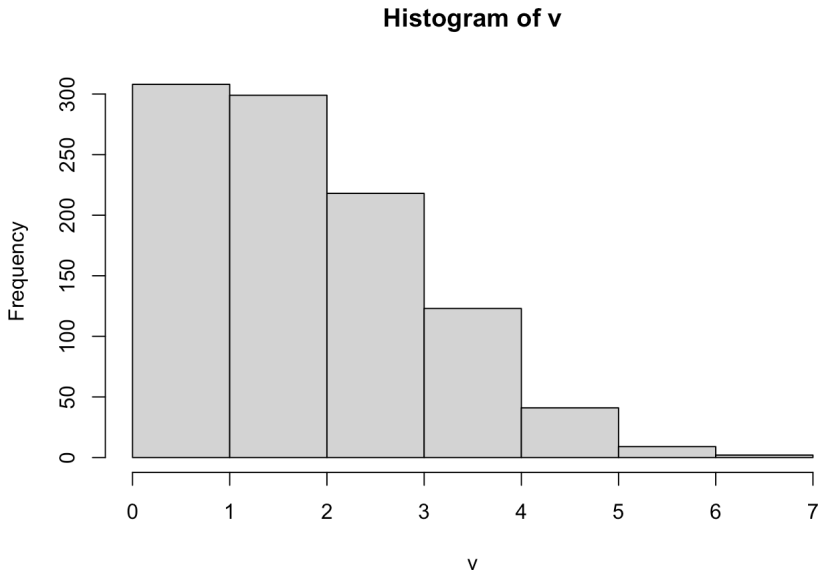
We could also replicate the process 1000 times (and extract the empirical expectation and variance):

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

```
[1] 2.236
[1] 1.794098
```

The histogram of the sample is shown below.

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

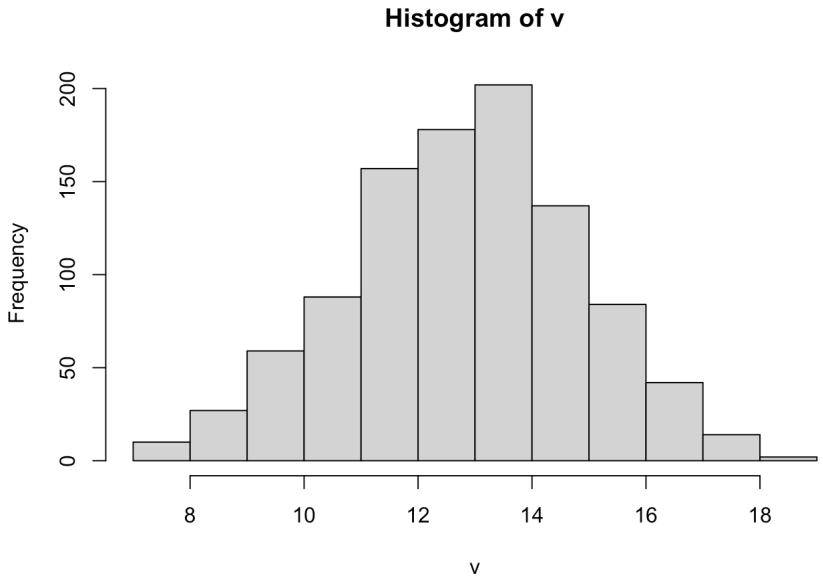



If we change the parameters of the distribution ($\mathcal{B}(19.0.7)$), we get a different looking histogram (and a different expectation and variance).

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

```
[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

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

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

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

**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: $\text{E}[X] = \frac{1}{1/6} = 6$.[23]

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

### 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 $r$th success occurs. Its p.m.f. is given by

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

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

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

**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: $\text{E}[X] = \frac{3}{1/6} = 18$.

### 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.[24]

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

1. 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**.[25] What is $P(X = x)$, for $x = 0, 1, \ldots$? 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,

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

Letting $n \to \infty$, we obtain

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

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

$$\mathrm{E}[X] = \lambda \quad \text{and} \quad \mathrm{Var}[X] = \lambda;$$

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:[26]

```
rpois(1,lambda=13)
```

```
[1] 18
```

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.

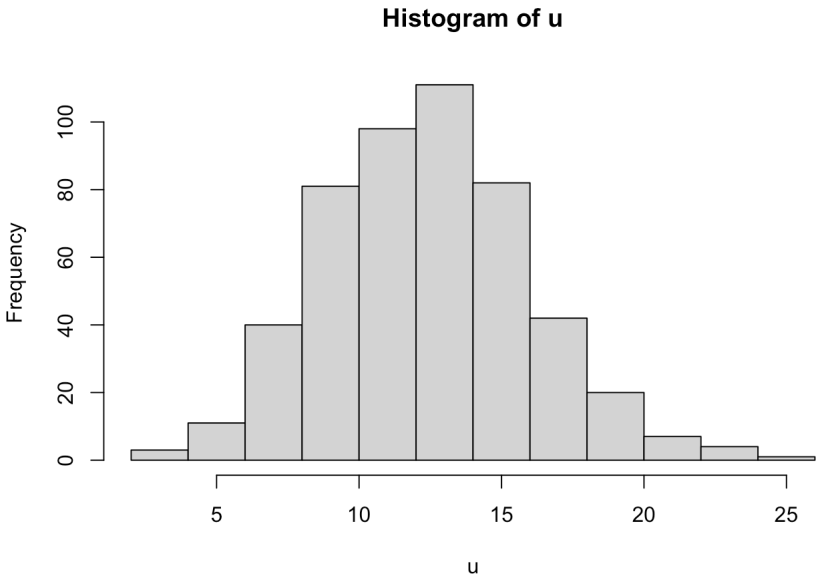
Next, we sample independently 500 times; this yields an empirical expectation and variance.

```
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.

```
hist(u)
```




**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,[27] 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 $x = 0, 1, 2, \ldots$ babies on each night.

For a Poisson distribution, the p.m.f. values $f(x)$ are obtained *via* `dpois()` in R.[28]

We start by setup the Poisson distribution parameters and the distribution's range.[29]

```
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)
```

```
x   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
```

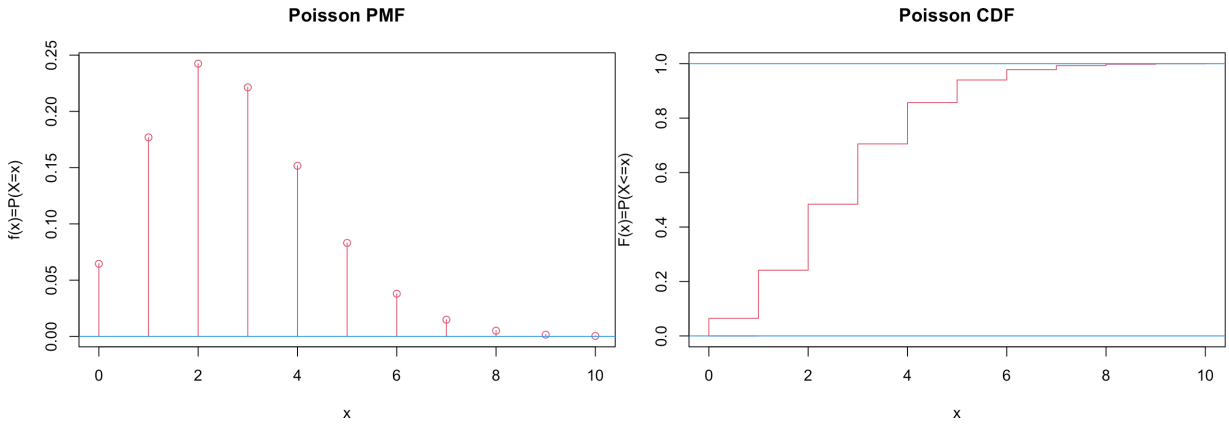
Here are the p.m.f. and c.d.f. plots:

```
plot(x,pmf, type="h", col=2, main="Poisson PMF",
     xlab="x", ylab="f(x)=P(X=x)")
points(x,pmf, col=2)
abline(h=0, col=4)

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)
```

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.



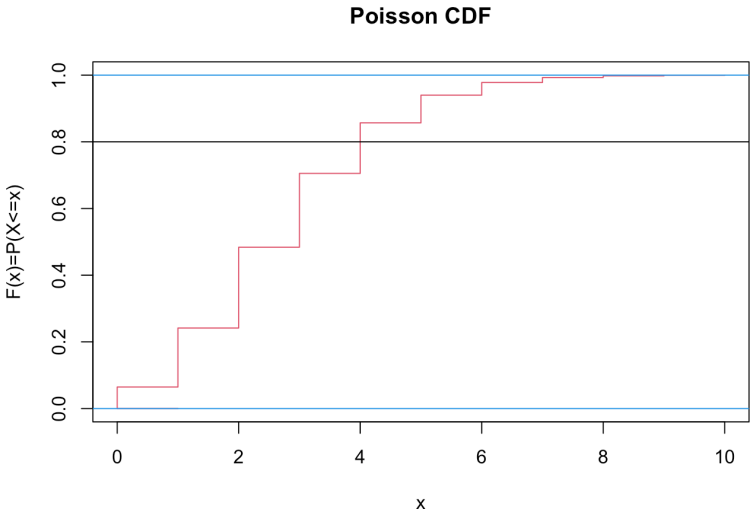

- If the maternity ward wants to prepare for the greatest possible traffic on 80% of the nights, how many deliveries should be expected?

  We seek an $x$ for which

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

  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 $F(3) = P(X \leq 3)$ and $F(4) = P(X \leq 4)$ to confirm that $F(3) \leq 0.8 \leq F(4)$.

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

```
[1] 0.7052548
[1] 0.8568984
```

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

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

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

  We need to evaluate

  $$365.25 \cdot P(X \geq 5) = 365.25(1 - P(X \leq 4)).$$

```
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. $365.25 \cdot P(X = x) \geq 1$.[31] The expected number of nights with each number of deliveries can be computed using:

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.

```
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
```

### 6.2.7 Other Discrete Distributions

There are numerous commonly-used discrete distributions [39]:

- 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

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

viewed as a function of a real variable $x$; however $P(X \leq x)$ 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 $F_X(x)$:[32]

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

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

**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 $f_X : X(\mathcal{S}) \to \mathbb{R}$ such that:

- $f_X(x) > 0$ for all $x \in X(\mathcal{S})$ and $\lim_{x\to\pm\infty} f_X(x) = 0$;
- $\int_{\mathcal{S}} f_X(x)\,dx = 1$;
- 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 $x$,

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

and for any $a$,

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\,dx = 0.$$

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.:

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

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

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

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

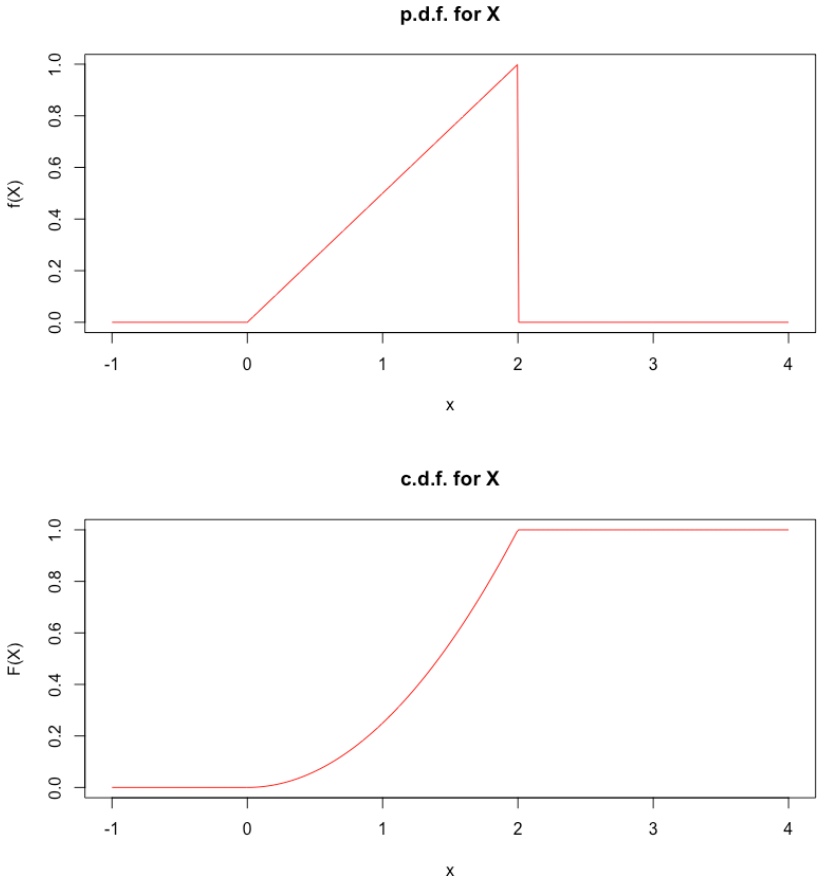



**Figure 6.8:** P.d.f. and c.d.f. for the continuous r.v. $X$ defined above.

- What is the probability of the event $A = \{X \mid 0.5 < X < 1.5\}$ if $X$ is the r.v. above?

  We need to evaluate

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

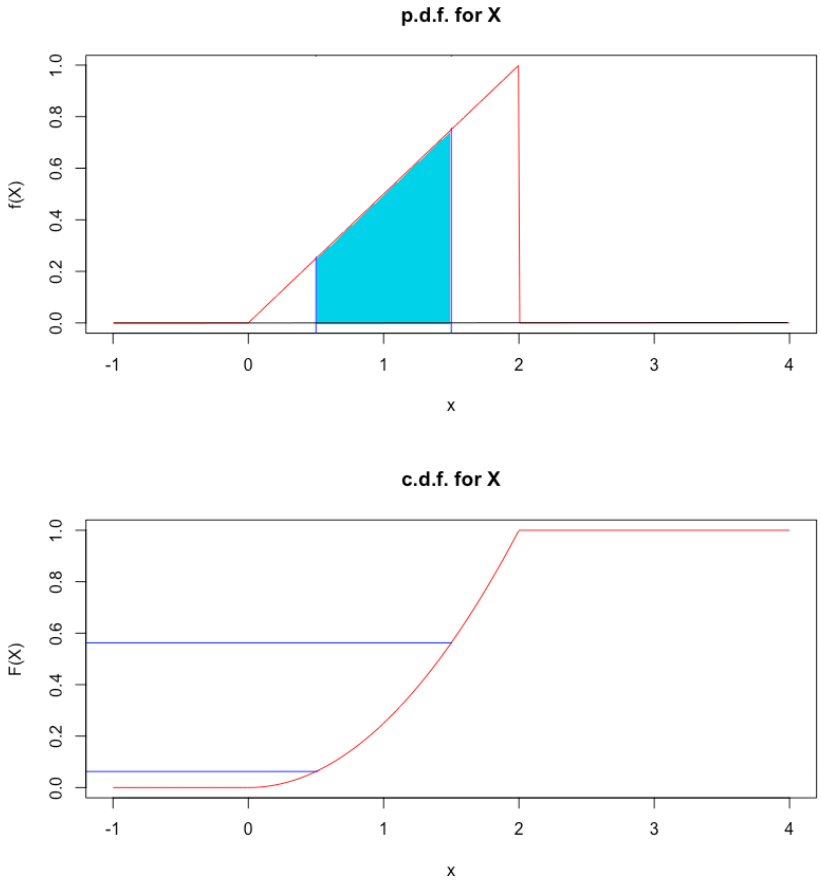



**Figure 6.9:** P.d.f. and c.d.f. for the continuous r.v. $X$ defined above, with event $A$.

- What is the probability of the event $B = \{X \mid X = 1\}$?

  We need to evaluate

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

  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.:

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

  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:

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

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:

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

For $\lambda = 0.2$, for instance, the p.d.f. and c.d.f. are:

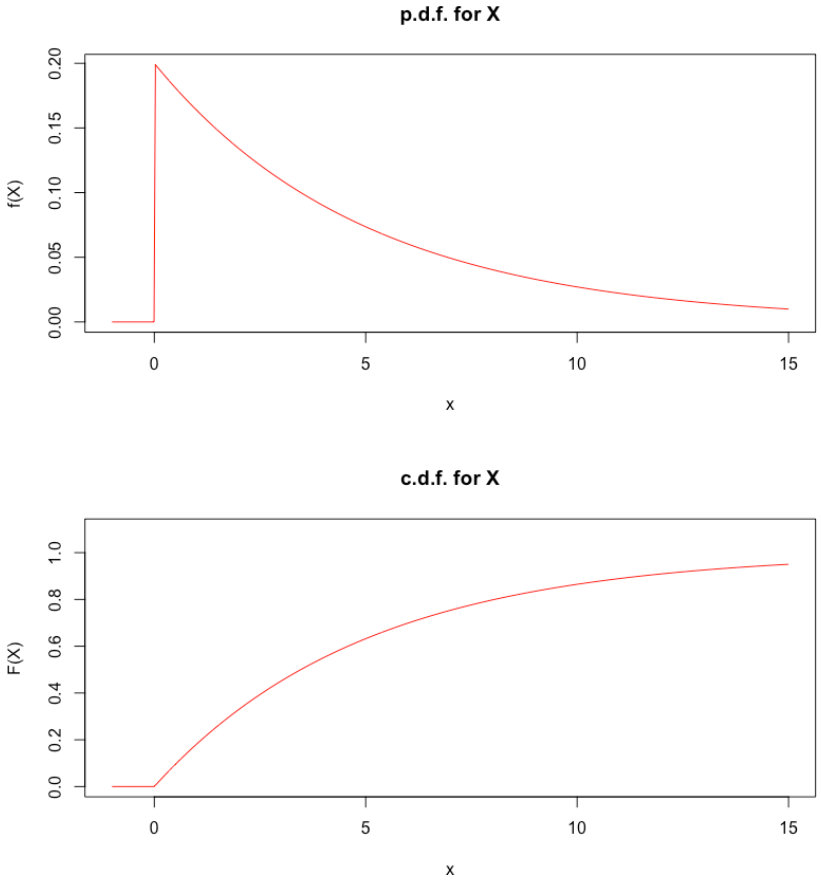



**Figure 6.10:** P.d.f. and c.d.f. for the r.v. $X$ with $\lambda = 0.2$.

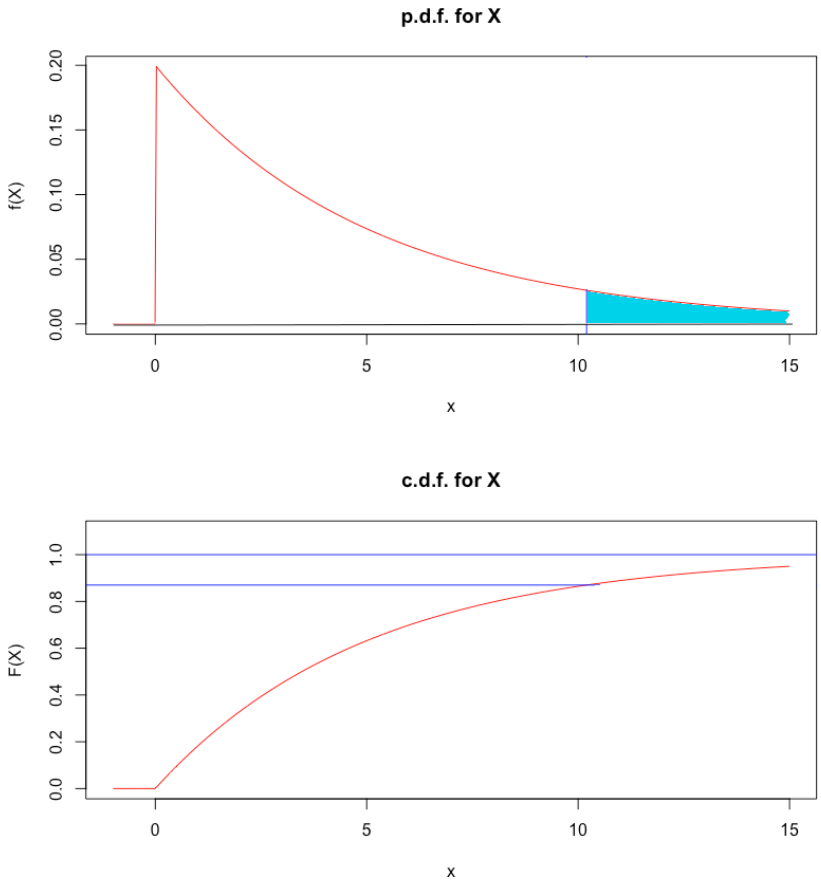



**Figure 6.11:** Probability of $X > 10.2$ (in blue), for $X$ with $\lambda = 0.2$.

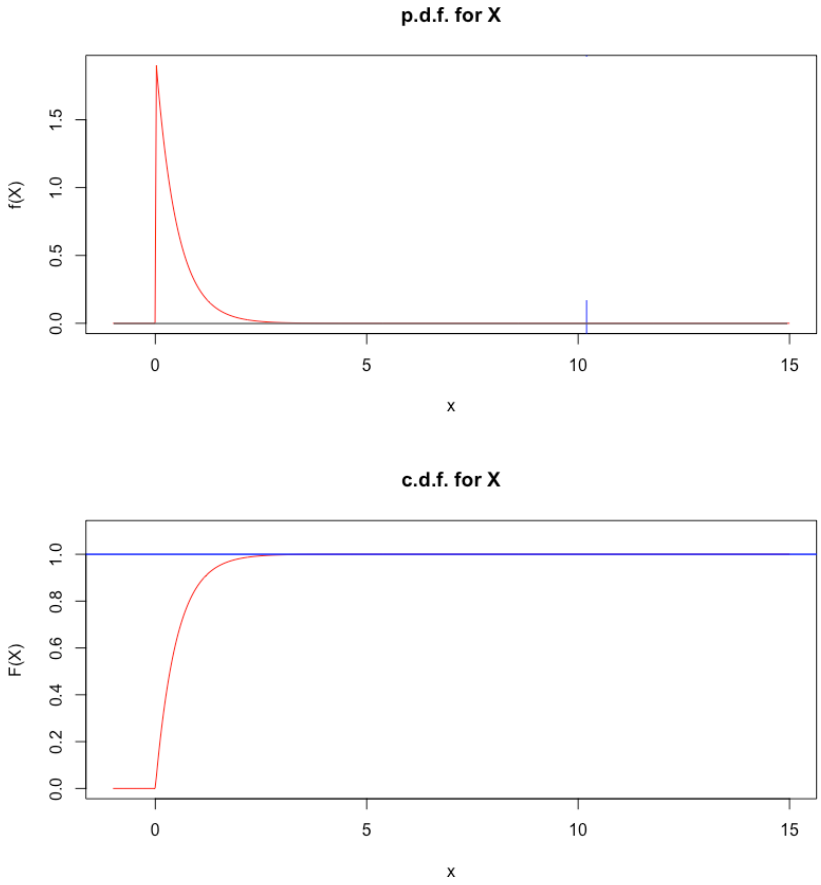



**Figure 6.12:** Probability of $X > 10.2$, for $X$ with $\lambda = 2$; the probability is so small ($1.38 \times 10^{-9}$) that it cannot even be made out in the p.d.f. (blue area).

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?.[33]

33: This is not a general property of distributions, however, but a property of this specific family of distributions.

### 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

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

#### Examples

- Find $\mathrm{E}[X]$ and $\mathrm{E}[X^2]$ in the first example, above.
  we need to evaluate

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

- 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.:

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

  We can also easily see that

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

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

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

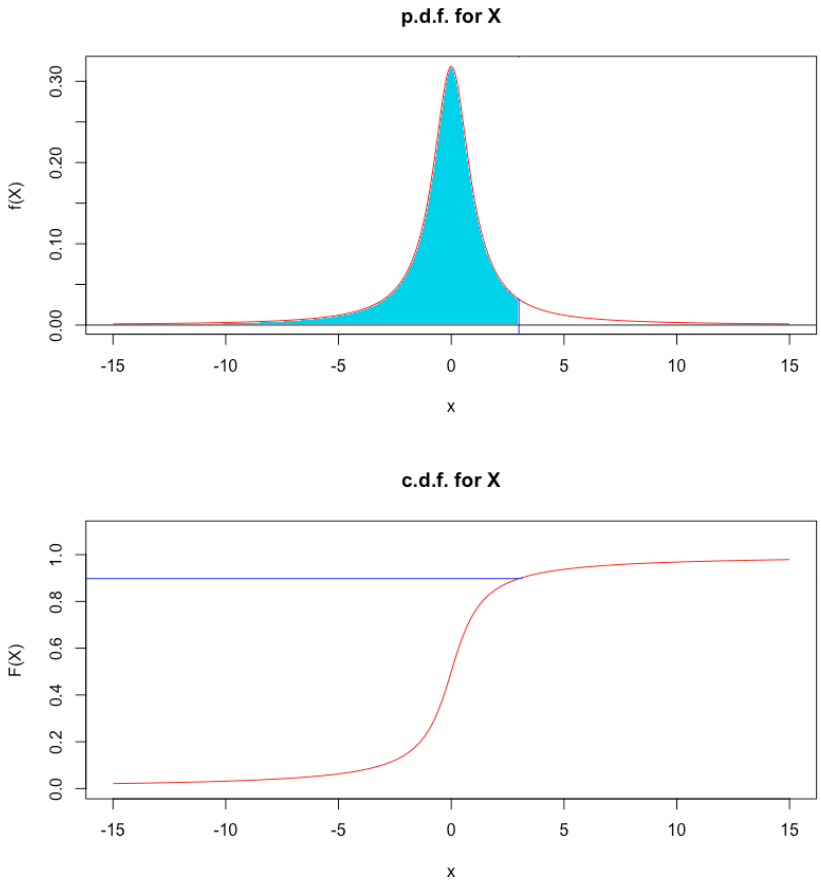



**Figure 6.13:** P.d.f. and c.d.f. for the Cauchy distribution, with area under the curve $F(3)$.

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 $du = 2x\,dx$ and $x\,dx = \frac{du}{2}$, 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

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

which is **undefined**. Thus $\mathrm{E}[X]$ cannot not exist, as it would have to be both equal to 0 and be undefined simultaneously.[34]

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.

### Mean and Variance

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

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

As in the discrete case, if $X, Y$ are continuous random variables, and $a, b \in \mathbb{R}$, then

$$\mathrm{E}[aY + bX] = a\mathrm{E}[Y] + b\mathrm{E}[X]$$
$$\mathrm{Var}[a + bX] = b^2\mathrm{Var}[X]$$
$$\mathrm{SD}[a + bX] = |b|\mathrm{SD}[X]$$

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:

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

## 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 $Z \sim \mathcal{N}(0, 1)$.

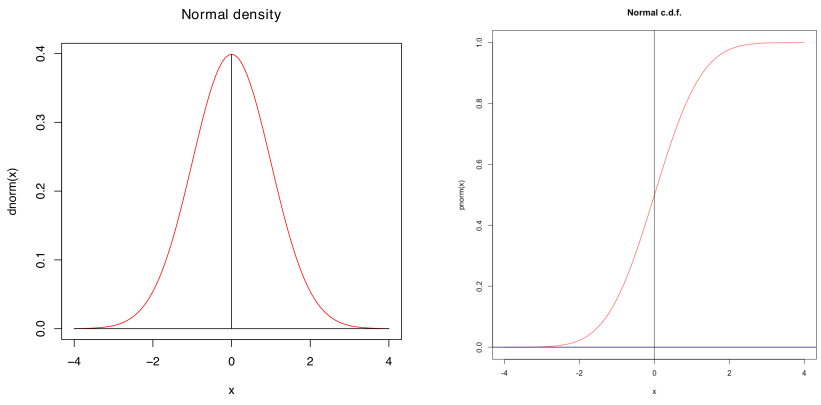



**Figure 6.14:** P.d.f. and c.d.f. for the standard normal distribution.

### Standard Normal Random Variable

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

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

Other quantities of interest include:

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

### 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

$$\begin{aligned}
\mathrm{E}[X] &= \mu + \sigma\mathrm{E}[Z] = \mu,\\
\mathrm{Var}[X] &= \sigma^2\mathrm{Var}[Z] = \sigma^2,\\
\mathrm{SD}[X] &= \sigma
\end{aligned}$$

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 [40] or at ztable.net. With the advent of freely-available **statistical software**, the need for tabulated values had decreased.[35]

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`.[36]

#### 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`.

1. $P(Z \leq 0.5) = 0.6915$
2. $P(Z < -0.3) = 0.3821$
3. $P(Z > 0.5) = 1 - P(Z \leq 0.5) = 1 - 0.6915 = 0.3085$
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.[37] 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

  $$\begin{aligned} P(X \leq 6) &= P\left(\frac{X-5}{0.5} \leq \frac{6-5}{0.5}\right) \\ &= P\left(Z \leq \frac{6-5}{0.5}\right) = \Phi\left(\frac{6-5}{0.5}\right) \\ &= \Phi(2) = P(Z \leq 2) \approx 0.9772. \end{aligned}$$

- 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 $\sigma = 0.4$.[38] 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

  $$\begin{aligned} P(X < 375) &= P\left(\frac{X-376.1}{0.4} < \frac{375-376.1}{0.4}\right) \\ &= P\left(Z < \frac{-1.1}{0.4}\right) \\ &= P(Z \leq -2.75) = \Phi(-2.75) \approx 0.003\,. \end{aligned}$$

- If $Z \sim \mathcal{N}(0, 1)$, for which values $a$, $b$ and $c$ do:

  1. $P(Z \leq a) = 0.95$?

     From the table (or R) we see that

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

     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-

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.

ficient to simply present the interval estimate: $a \in (1.64, 1.65)$

2. $P(|Z| \leq b) = P(-b \leq Z \leq b) = 0.99$?

   Note that

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

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

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

   and so that

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

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

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

   Consulting the table we see that

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

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

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

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

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

   But this is equivalent to

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

   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:[39]

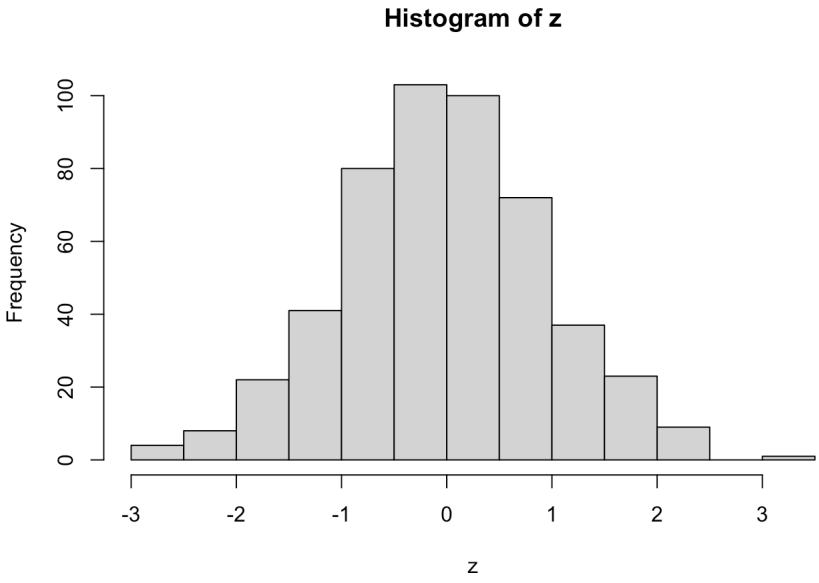
39: Note: no seed is provided, so results may vary.

```
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)
```




A histogram with 20 bins is shown below:

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


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

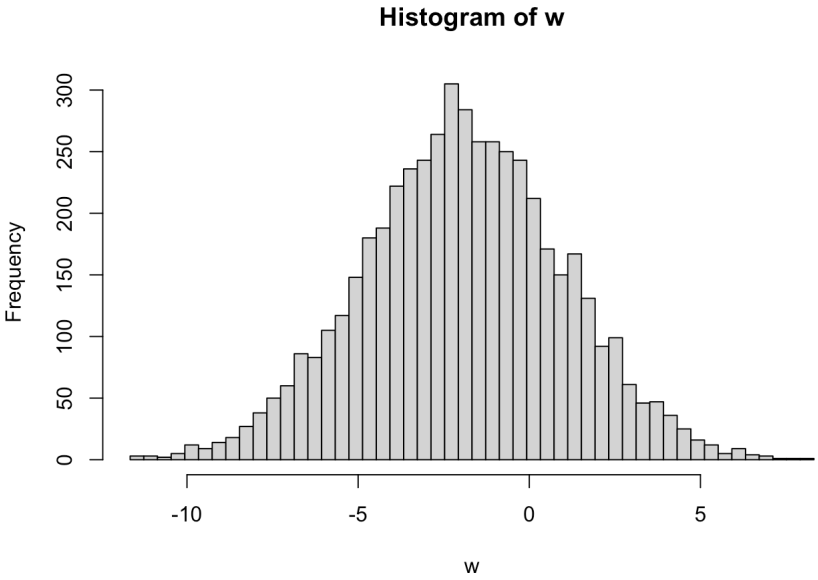
For instance, we can draw 5000 observations from $\mathcal{N}(-2, 3^2)$ using the following code:

```
w<-rnorm(5000, sd=3, mean=-2)
mean(w)
sd(w)
```

```
[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

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

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

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

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:

- $\mu = \text{E}[X] = 1/\lambda$, since

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

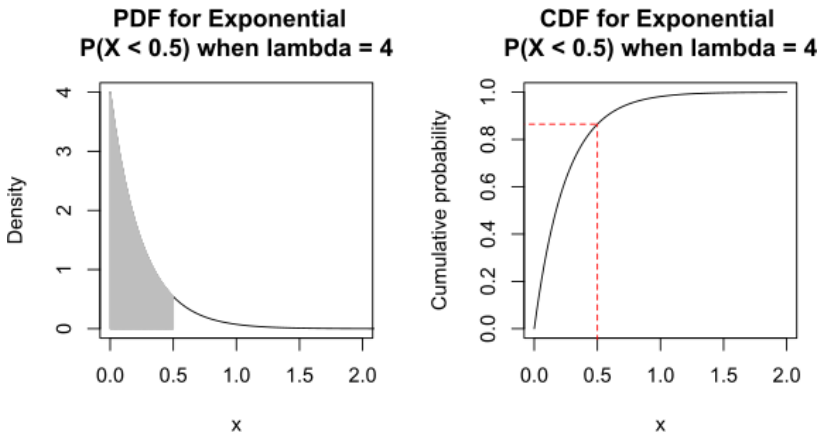



**Figure 6.15:** P.d.f. and c.d.f. for the exponential distribution. with parameter *lambda* = 4 [source unknown].

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

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

- and $P(X > s + t \mid X > t) = P(X > s)$, 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, 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

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

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

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

that is,

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

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

We can draw from Exp(100) as follows:[40]

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

```
rexp(1,100)
```

```
[1] 0.0009430804
```

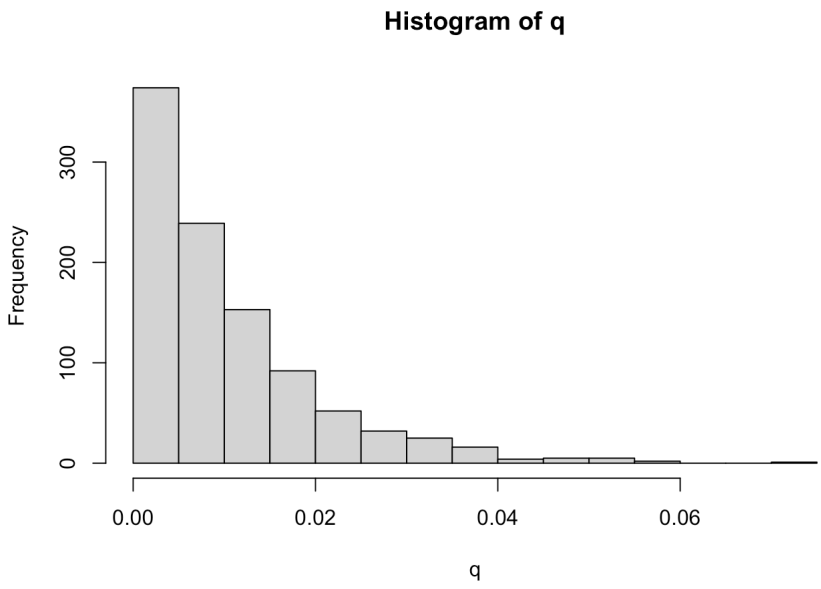
If we repeat the process 1000 times, the empirical mean and variance are:

```
q<-rexp(1000,100)
mean(q)
var(q)
```

```
[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

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

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

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

#### Examples

- Suppose that an average of 30 customers per hour arrive at a shop in accordance with a Poisson process, that is to say, $\lambda = 1/2$ 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

  $$P(Y > 5) = \int_5^\infty \frac{y^{2-1}}{(2-1)!}(1/2)^2 e^{-y/2}\, dy = \int_5^\infty \frac{ye^{-y/2}}{4}\, dy$$
  $$= \frac{1}{4}\left[-2ye^{-y/2} - 4e^{-y/2}\right]_5^\infty = \frac{7}{2}e^{-5/2} \approx 0.287.$$

- Telephone calls arrive at a switchboard at a mean rate of $\lambda = 2$ 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

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

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
```

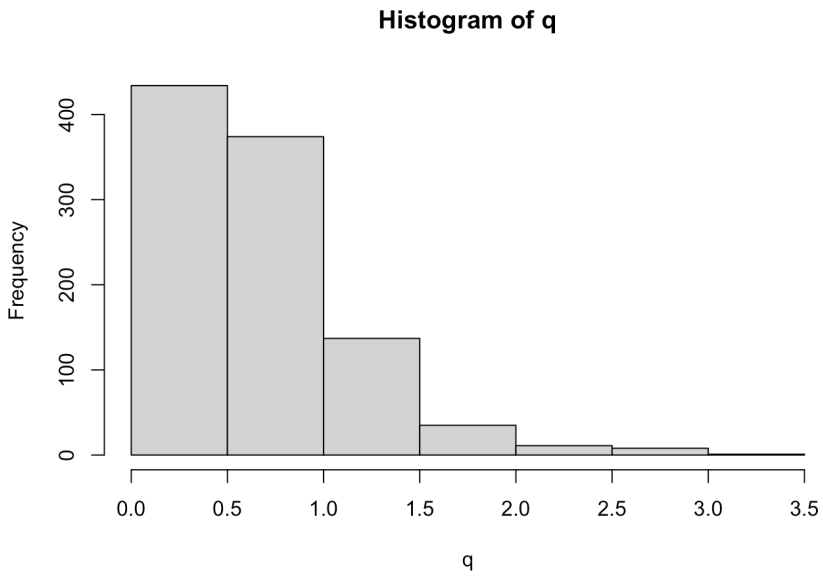
This can be repeated 1000 times, say, and we get the empirical mean and variance:

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

```
[1] 0.6663675
[1] 0.2205931
```

The corresponding histogram is displayed below:

```
hist(q)
```




### 6.3.6 Approximation of the Binomial Distribution

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

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

Thus, according to the **Central Limit Theorem**,[41] for large $n$ we have

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

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

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

### Normal Approximation with Continuity Correction

When $X \sim \mathcal{B}(n,p)$, we know that $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = np(1-p)$. 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 $X \sim \mathcal{B}(36, 0.5)$. Provide a normal approximation to the probability $P(X \le 12)$.[42]

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`.

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

$$\mathrm{E}[X] = 36(0.5) = 18 \quad \text{and} \quad \mathrm{Var}[X] = 36(0.5)(1-0.5) = 9,$$

and so

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

### 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 $x = 0, 1, \ldots, n$;
- using **tables** – if $n \le 15$ and $p$ is one of $0.1, \ldots, 0.9$, 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 $\ge 5$:
$$P(X \le x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right);$$
$$P(X \ge x) \approx 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

### 6.3.7 Other Continuous Distributions

Some other common continuous distributions are listed in [39]:

- 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. $f(x, y) \geq 0$, for all $x$, $y$;
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$, and
3. $P(A) = \iint_A f(x, y)\, dxdy$, 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 $f(x, y) \leq 1$ 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 [41]).

  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

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

  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 $a, b = 1, \ldots, 6$.

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

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

  These events are mutually exclusive, so that

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

  Similarly, we get

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

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

  5. Compute $P(X = 3 \mid Y > 3)$, $P(Y \leq 3 \mid X \geq 4)$.

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

$$P(X = 3 \mid Y > 3) = \frac{P(X = 3 \cap Y > 3)}{P(Y > 3)}$$
$$P(Y = 3 \mid X \geq 4) = \frac{P(Y = 3 \cap X \geq 4)}{P(X \geq 4)}$$

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:

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

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

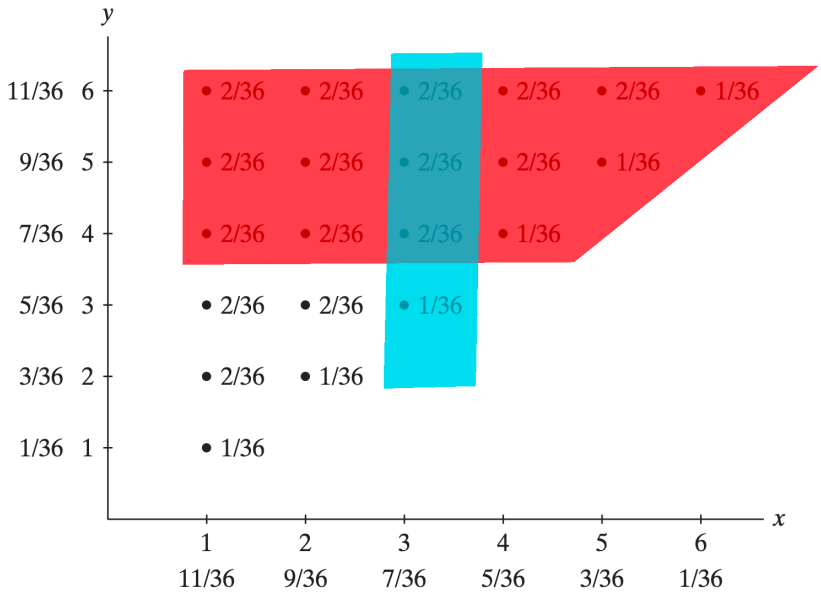



**Figure 6.16:** Conditional and marginal probabilities in the dice example [41].

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**.[43]

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 [41]).

  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

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

- 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 $y = 1$, and the line $y = x$.

     The support is the blue triangle shown in Figure 6.17.

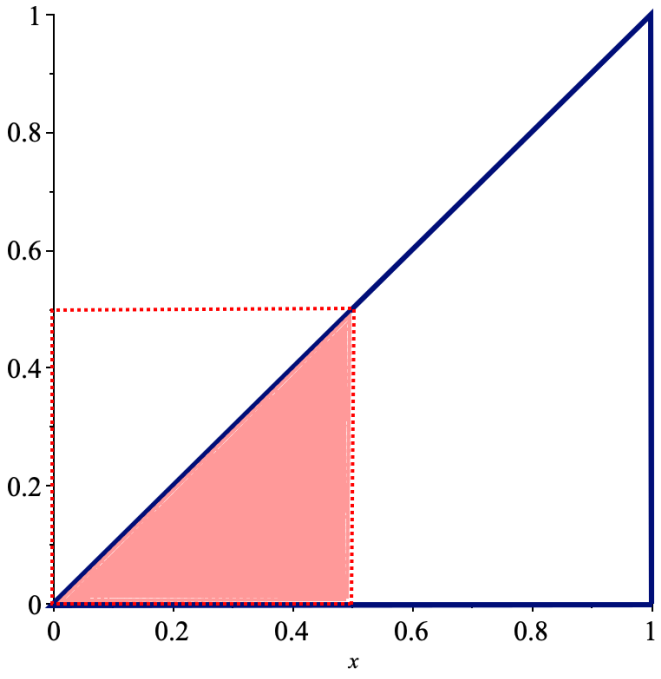



**Figure 6.17:** Support for the joint distribution of $X$ and $Y$ in the above example.

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:

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

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 $\mathrm{E}[X]$, $\mathrm{E}[Y]$, $\mathrm{E}[X^2]$, $\mathrm{E}[Y^2]$, and $\mathrm{E}[XY]$.

   We have

$$\begin{aligned} \mathrm{E}[X] &= \iint_S x f(x, y)\, dA = \int_0^1 \int_0^x 2x\, dy dx \\ &= \int_0^1 [2xy]_{y=0}^{y=x}\, dx = \int_0^1 2x^2\, dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}; \\ \mathrm{E}[Y] &= \iint_S y f(x, y)\, dA = \int_0^1 \int_y^1 2y\, dx dy \\ &= \int_0^1 [2xy]_{x=y}^{x=1}\, dy = \int_0^1 (2y - 2y^2)\, dy = \left[y^2 - \frac{2}{3}y^3\right]_0^1 = \frac{1}{3}; \\ \mathrm{E}[X^2] &= \iint_S x^2 f(x, y)\, dA = \int_0^1 \int_0^x 2x^2\, dy dx \\ &= \int_0^1 \left[2x^2 y\right]_{y=0}^{y=x}\, dx = \int_0^1 2x^3\, dx = \left[\frac{1}{2}x^4\right]_0^1 = \frac{1}{2}; \\ \mathrm{E}[Y^2] &= \iint_S y^2 f(x, y)\, dA = \int_0^1 \int_y^1 2y^2\, dx dy \\ &= \int_0^1 \left[2xy^2\right]_{x=y}^{x=1}\, dy = \int_0^1 (2y - 2y^3)\, dy = \left[\frac{2}{3}y^3 - \frac{1}{2}y^4\right]_0^1 = \frac{1}{6}; \\ \mathrm{E}[XY] &= \iint_S xy f(x, y) = \int_0^1 \int_0^x 2xy\, dy dx \\ &= \int_0^2 \left[xy^2\right]_{y=0}^{y=x} = \int_0^1 x^2\, dx = \left[\frac{x^4}{4}\right]_0^1 = \frac{1}{4}. \end{aligned}$$

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:

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

When $X = Y$, the covariance reduces to the variance.[44] In the last example, for instance, we have: $\text{Var}[X] = \frac{1}{2} - (\frac{2}{3})^2 = \frac{1}{18}$, $\text{Var}[X] = \frac{1}{6} - (\frac{1}{3})^2 = \frac{1}{18}$, and $\text{Cov}(X,Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{36}$.[45]

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

45: We will use the covariance again in Chapters 8 and 11.

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)
```

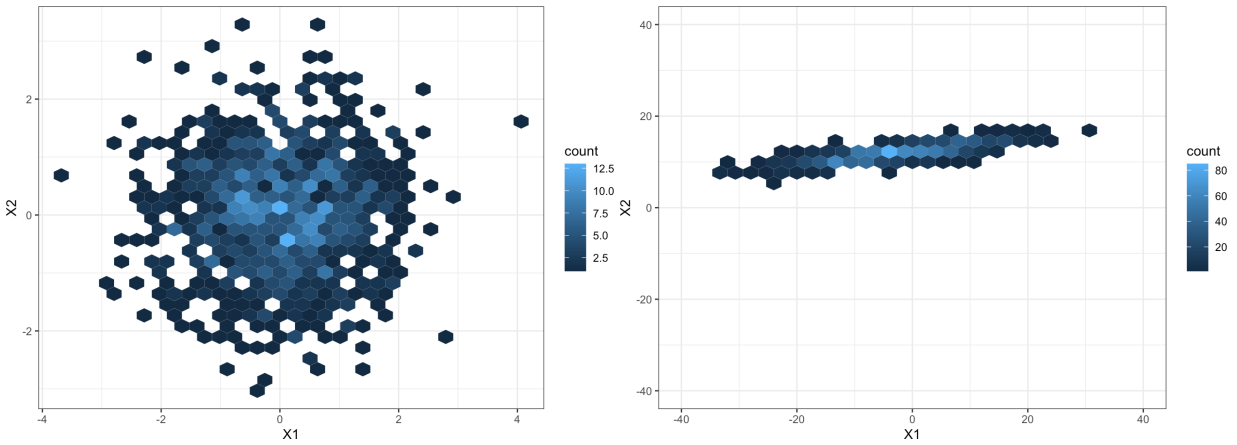
We sample 1000 observations from each joint normal.

```
library(MASS)
a1<-mvrnorm(1000,mu1,Sigma1)
a1<-data.frame(a1)
a2<-mvrnorm(1000,mu2,Sigma2)
a2<-data.frame(a2)
```

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.[46]

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

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



## 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 $X_1, \ldots, X_n$ are $n$ **independent** random variables, each having the same c.d.f. $F$, i.e.they are **identically distributed**. Then, $\{X_1, \ldots, X_n\}$ 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

$$\begin{aligned} \mathrm{E}[a + bX] &= a + b\mathrm{E}[X], \\ \mathrm{Var}[a + bX] &= b^2\mathrm{Var}[X], \\ \mathrm{SD}[a + bX] &= |b|\mathrm{SD}[X]. \end{aligned}$$

#### Sum of Independent Random Variables

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

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

In general,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + 2\mathrm{Cov}(X, Y) + \mathrm{Var}[Y];$$

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

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

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

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

### 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

$$\mathrm{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2 \quad \text{for } i = 1, \ldots, n,$$

then

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

#### 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 $\mathrm{E}[X_i] = \mu = 50$ and $\mathrm{Var}[X] = \sigma^2 = 0.25$ for $i = 1, \ldots, 100$, find $\mathrm{E}[\tau]$ and $\mathrm{Var}[\tau]$ for

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

  According to the iid formulas,

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

- 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 $\mathrm{E}[X_i] = \mu = 5$, $\mathrm{SD}[X_i] = \sigma = 0.2$ and $\mathrm{Var}[X_i] = \sigma^2 = 0.2^2 = 0.04$ for $i = 1, 2, 3, 4$. Let $\tau = X_1 + X_2 + X_3 + X_4$.

  According to the iid formulas,

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

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

### 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 $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ for all $i = 1, \ldots, n$, then

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}(n\mu) = \mu$$

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}.$$

**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 $i = 1, \ldots, 9$ is written as $X_i = \mu + Z_i$ where $\mathrm{E}[Z_i] = 0$ and $\mathrm{SD}[Z_i] = 0.1$ and the $Z_i$'s are iid.

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

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

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

### Sum of Independent Normal Random Variables

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

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

$$\mathrm{E}[\tau] = \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = \mu_1 + \cdots + \mu_n;$$
$$\mathrm{Var}[\tau] = \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \sigma_1^2 + \cdots + \sigma_n^2.$$

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

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

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?

If $X_1, \ldots, X_{16}$ are iid as $\mathcal{N}(75, 25)$, then the sum $\tau = X_1 + \cdots + X_{16}$ is also normally distributed with

$$\tau = \sum_{i=1}^{16} X_i \sim \mathcal{N}(16 \cdot 75, 16 \cdot 25) = \mathcal{N}(1200, 400), \quad \text{and}$$

$$Z = \frac{\tau - 1200}{\sqrt{400}} \sim \mathcal{N}(0, 1).$$

Thus,

$$P(\tau > 1250) = P\left(\frac{\tau - 1200}{\sqrt{400}} > \frac{1250 - 1200}{20}\right) = P(Z > 2.5) = 1 - P(Z \leq 2.5)$$
$$\approx 1 - 0.9938 = 0.0062\,.$$

### 6.5.2 Central Limit Theorem

Suppose that a professor has been teaching a course for the last 20 years. For every cohort during that period, the mid-term exam grades of all the students have been recorded. Let $X_{i,j}$ be the grade of student $i$ in year $j$. Looking back on the class lists, they find that

$$\mathrm{E}[X_{i,j}] = 56 \quad \text{and} \quad \mathrm{SD}[X_{i,j}] = 11.$$

This year, there are 49 students in the class. What should the professor expect for the class mid-term exam average?

Of course, the professor cannot predict any of the student grades or the class average with absolute certainty, but they could try the following approach:

1. simulate the results of the class of 49 students by generating sample grades $X_{1,1}, \ldots, X_{1,49}$ from a **normal** distribution $\mathcal{N}(65, 15^2)$;
2. compute the sample mean for the sample and record it as $\overline{X}_1$;
3. repeat steps 1-2 $m$ times and compute the standard deviation of the sample means $\overline{X}_1, \ldots, \overline{X}_m$;
4. plot the histogram of the sample means $\overline{X}_1, \ldots, \overline{X}_m$.

What do you think is going to happen?

**Central Limit Theorem:** if $\overline{X}$ is the mean of a random sample of size $n$ taken from a population with mean $\mu$ and finite variance $\sigma^2$, then

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

as $n \to \infty$. More precisely, this is a **limiting** result. If we view the **standardization**

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

as functions of $n$, we have, for each $z$,

$$\lim_{n\to\infty} P\left(Z_n \leq z\right) = \Phi(z) \quad \text{and} \quad P\left(Z_n \leq z\right) \approx \Phi(z), \text{ if } n \text{ is large enough,}$$

**whether the original $X_i$'s are normal or not**.

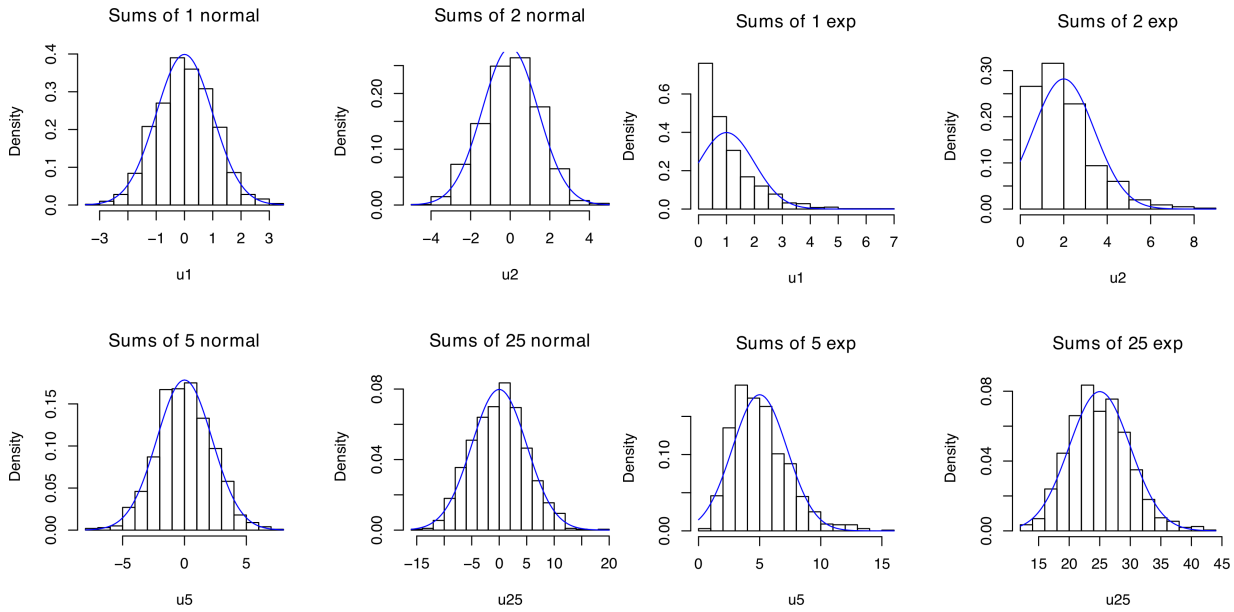



**Figure 6.18:** Illustration of the central limit theorem with a normal underlying distribution and with an exponential underlying distribution [source unknown].

### Examples

- The examination scores in an university course have mean 56 and standard deviation 11. In a class of 49 students, what is the probability that the average mark is below 50? What is the probability that the average mark lies between 50 and 60?

  Let the marks be $X_1, \ldots, X_{49}$ and assume the performances are independent. According to the central limit theorem,

  $$\overline{X} = (X_1 + X_2 + \cdots + X_{49})/49,$$

  with $\mathrm{E}[\overline{X}] = 56$ and $\mathrm{Var}[\overline{X}] = 11^2/49$. We thus have

  $$P(\overline{X} < 50) \approx P\left(Z < \frac{50 - 56}{11/7}\right) = P(Z < -3.82) = 0.0001$$

  and

  $$P(50 < \overline{X} < 60) \approx P\left(\frac{50 - 56}{11/7} < Z < \frac{60 - 56}{11/7}\right)$$
  $$= P(-3.82 < Z < 2.55) = \Phi(2.55) - \Phi(-3.82) = 0.9945.$$

  Note that this says nothing about whether the scores are normally distributed or not, only that the average scores follow an approximate normal distribution.[47]

- Systolic blood pressure readings for pre-menopausal, non-pregnant women aged $35 - 40$ have mean 122.6 standard deviation 11 mm Hg. An independent sample of 25 women is drawn from this target population and their blood pressure is recorded. What is the probability that the average blood pressure is greater than 125 mm Hg? How would the answer change if the sample size increases to 40?

47: If the scores did arise from a normal distribution, the $\approx$ would be replaced by a $=$.

According to the CLT, $\overline{X} \sim \mathcal{N}(122.6, 121/25)$, approximately. Thus

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{25}}\right) = P(Z > 1.09) = 1 - \Phi(1.09) = 0.14.$$

However, if the sample size is 40, then

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{40}}\right) = 0.08.$$

Increasing the sample size reduces the probability that the average is far from the expectation of each original measurement.

- Suppose that we select a random sample $X_1, \ldots, X_{100}$ from a population with mean 5 and variance 0.01. What is the probability that the difference between the sample mean of the random sample and the mean of the population exceeds 0.027?

  According to the CLT, we know that, approximately, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has standard normal distribution. The desired probability is thus

$$\begin{aligned} P &= P(|\overline{X} - \mu| \geq 0.027) \\ &= P(\overline{X} - \mu \geq 0.027 \text{ or } \mu - \overline{X} \geq 0.027) \\ &= P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) + P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \leq \frac{-0.027}{0.1/\sqrt{100}}\right) \\ &\approx P\left(Z \geq 2.7\right) + P\left(Z \leq -2.7\right) \\ &= 2P\left(Z \geq 2.7\right) \approx 2(0.0035) = 0.007. \end{aligned}$$

In the next example, we illustrate how to use the CLT with R.

**Example** A large freight elevator can transport a maximum of 9800 lbs. Suppose a load containing 49 boxes must be transported. From experience, the weight of boxes follows a distribution with mean $\mu = 205$ lbs and standard deviation $\sigma = 15$ lbs. Estimate the probability that all 49 boxes can be safely loaded onto the freight elevator and transported.

We are given $n = 49$, $\mu = 205$, and $\sigma = 15$. Let us further assume that the boxes all come from different sources, which is to say, the boxes' weight $x_i$, $i = 1, \ldots, 49$, are independent of one another.

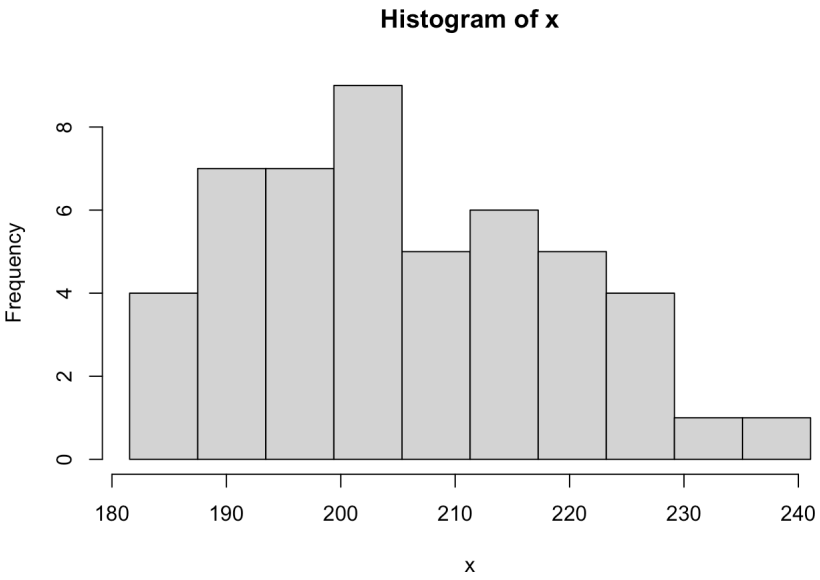
To get a sense of the task's feasibility, we simulate a few scenarios. Note that the problem makes no mention of the type of distribution that the weights follow.

To start, we assume that the weights are normally distributed.

```
set.seed(0) # to ensure replicability
x<-rnorm(49,mean=205,sd=15)
```

The histogram shows a distribution which is roughly normal.

```
brks = seq(min(x),max(x),(max(x)-min(x))/10)
hist(x, breaks = brks)
```




The elevator can transport up to 9800 lbs; the $n = 49$ boxes can be transported if their total weight

$$T = 49w = x_1 + \cdots + x_{49},$$

where $w = \overline{x}$, is less than 9800 lbs. In mathematical terms, we are interested in the value of the probability $P(T < 9800)$.

For the sample x from above, we get:

```
(T<-sum(x))
```

```
[1] 10066.36
```

That specific group of 49 boxes would be too heavy to carry in one trip. But perhaps we were simply unlucky – perhaps another group of boxes would have been light enough. Let us try again, but with a different group of boxes.

```
set.seed(999)
(T=sum(rnorm(49,mean=205,sd=15)))
```

```
[1] 9852.269
```

It's closer, but still no cigar. However, two tries are not enough to establish a trend and to estimate $P(T < 9800)$.

Next, we write a little function to help us find an estimate of the probability. The idea is simple: if we were to try a large number of random combinations of 49 boxes, the proportion of the attempts for which the total weight $T$ falls below 9800 is (hopefully?) going to approximate $P(T < 9800)$.

```
estimate_T.normal <- function(n, T.threshold, mean, sd, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(rnorm(n,mean=mean,sd=sd))<T.threshold){
      a=a+1
    }
  }
  estimate_T.normal <- a/num.tries
}
```

What kind of inputs are these meant to be? What does this code do? Note that running this cell will **compile** the function `estimate_T.normal()`, but that it still needs to be called with appropriate inputs to provide an estimate for $P(T < 9800)$.

We try the experiment (`num.tries`) 10, 100, 1000, 10000, 100000, and 1000000 times, with `n=49`, `T.threshold=9800`, `mu=205`, and `sigma=15`.

```
(c(estimate_T.normal(49,9800,205,15,10),
estimate_T.normal(49,9800,205,15,100),
estimate_T.normal(49,9800,205,15,1000),
estimate_T.normal(49,9800,205,15,10000),
estimate_T.normal(49,9800,205,15,100000),
estimate_T.normal(49,9800,205,15,1000000)))
```

```
[1] 0.00000 0.01000 0.00700 0.00990 0.00973 0.00975
```

We cannot say too much from such a simple set up, but it certainly seems as though we should expect success about 1% of the time.

That is a low probability, which suggests that 49 may be too many boxes for the elevator to work correctly, in general, but perhaps that is only the case because we assumed normality. What happens if we used other distributions with the same characteristics, such as $U(179.02, 230.98)$ or $\Lambda(5.32, 0.0054)$?[48]

48: How would we verify that these distributions indeed have the right characteristics? How would we determine the appropriate parameters in the first place?

Let us write new functions `estimate_T.unif()` and `estimate_T.lnormf()` to repeat the previous work with those two distributions.

```
estimate_T.unif <- function(n, T.threshold, min, max, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(runif(n,min=min,max=max))<T.threshold){
      a=a+1
    }
  }
estimate_T.unif <- a/num.tries
}

estimate_T.lnorm <- function(n, T.threshold, meanlog, sdlog, num.tries){
  a=0
  for(j in 1:num.tries){
    if(sum(rlnorm(n,meanlog=meanlog,sdlog=sdlog))<T.threshold){
```

```
      a=a+1
    }
  }
estimate_T.lnorm <- a/num.tries
}
```

For the uniform distribution, we obtain:

```
(c(estimate_T.unif(49,9800,179.02,230.98,10),
 estimate_T.unif(49,9800,179.02,230.98,100),
 estimate_T.unif(49,9800,179.02,230.98,1000),
 estimate_T.unif(49,9800,179.02,230.98,10000),
 estimate_T.unif(49,9800,179.02,230.98,100000),
 estimate_T.unif(49,9800,179.02,230.98,1000000)))
```

```
[1] 0.000000 0.010000 0.008000 0.007900 0.010230 0.009613
```

For the log-normal distribution, we obtain:

```
(c(estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),10),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),100),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),1000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),10000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),100000),
 estimate_T.lnorm(49,9800,5.32,sqrt(0.0054),1000000)))
```

```
[1] 0.000000 0.000000 0.006000 0.009500 0.009060 0.009184
```

Under all three distributions, it appears as though $P(T < 9800)$ converges to a value near 1%, even though the three distributions are very different. That might be surprising at first glance, but it is really a consequence of the **Central Limit Theorem**.

In effect, we are estimating $P(T < 9800) = P(w < 9800/49) = P(w < 200)$, where $w$ is the mean weight of the boxes.

According to the CLT, the distribution of $w$ is approximately normal with mean $\mu = 205$ and variance $\sigma^2/n = 15^2/49$, even if the weights themselves were not normally distributed.

By subtracting the mean of $w$ and dividing by the standard deviation we obtain a new random variable $z$ which is approximately the standard unit normal, i.e.

$$P(w < 200) \approx P\left(z < \frac{200 - 205}{15/7}\right).$$

But

```
(200-205)/(15/7)
```

```
[1] -2.333333
```

Thus, $P(w < 200) \approx P(z < -2.33)$ and we need to find the probability that the standard normal p.d.f. is smaller than $-2.33$.

This can be calculated with the `pnorm()` function:

```
pnorm(-2.33, mean=0, sd=1)
```

```
[1] 0.009903076
```

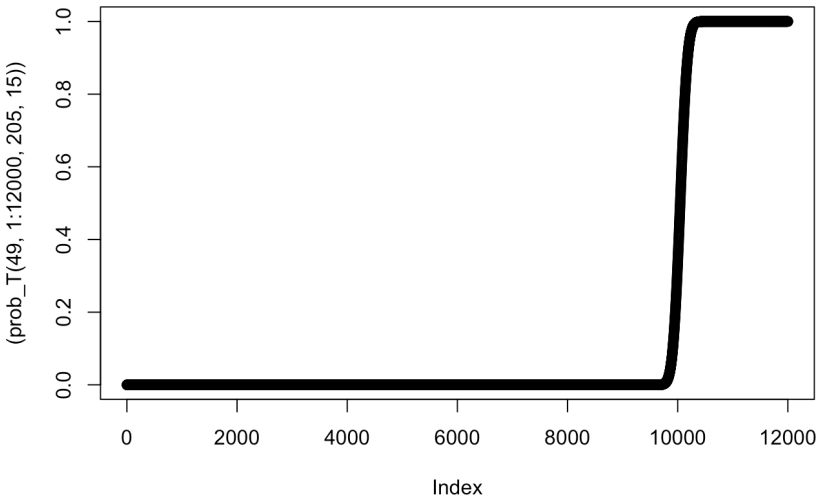
Hence, $P(T < 9800) \approx 0.0099$, which means that it is highly unlikely that the 49 boxes can be transported in the elevator all at once.

**Example** What elevator threshold would be required to reach a probability of success of 10%? 50%? 75%?

The following routine approximates the probability in question without resorting to simulating the weights (that is, independently of the underlying distribution of weights) for given `n`, `threshold`, `mean`, and `sd`. Can you figure out what `pnorm()` is doing?

```
prob_T <- function(n,threshold,mean,sd){
  prob_T=pnorm((threshold/n - mean)/(sd/sqrt(n)),0,1)
}

plot((prob_T(49,1:12000,205,15)))
```




We can find the desired thresholds by calling:

```
max(which(prob_T(49,1:12000,205,15)<0.1))
max(which(prob_T(49,1:12000,205,15)<0.5))
max(which(prob_T(49,1:12000,205,15)<0.75))
```

```
[1] 9910
[1] 10044
[1] 10115
```

### 6.5.3 Sampling Distributions (Reprise)

We now revisit sampling distributions in a some specific contexts.

#### Difference Between Two Means

Statisticians are often interested in the difference between various populations; a result akin to the CLT provides guidance in that area.

**Theorem:** let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean $\mu_1$ and variance $\sigma_1^2$, and $\{Y_1, \ldots, Y_m\}$ be another random sample, independent of $X$, from a population with mean $\mu_2$ and variance $\sigma_2^2$.

If $\overline{X}$ and $\overline{Y}$ are the respective sample means, then

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

has standard normal distribution $\mathcal{N}(0,1)$ as $n, m \to \infty$.[49]

[49]: Like the CLT, this is a **limiting** result.

**Example** Two different machines are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes.

The variances of these weights is identical, $\sigma^2 = 1$. Each machine produces a sample of 36 boxes, and the weights are recorded. What is the probability that the difference between the respective averages is less than 0.2, assuming that the true means are identical?

We have $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = 1$, $n = m = 36$. The desired probability is

$$\begin{aligned}
&P\left(|\overline{X} - \overline{Y}| < 0.2\right) = P\left(-0.2 < \overline{X} - \overline{Y} < 0.2\right) \\
&= P\left(\frac{-0.2 - 0}{\sqrt{1/36 + 1/36}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{1/36 + 1/36}} < \frac{0.2 - 0}{\sqrt{1/36 + 1/36}}\right) \\
&= P(-0.8485 < Z < 0.8485) \\
&\approx \Phi(0.8485) - \Phi(-0.8485) \approx 0.6.
\end{aligned}$$

#### Sample Variance $S^2$

When the underlying variance is unknown (which is usually the case in practice), it must be approximated by the sample variance.

**Theorem:** let $\{X_1, \ldots, X_n\}$ be a random sample taken from a normal population with mean $\sigma^2$, and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

be the sample variance. The statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

follows a **chi-squared distribution with $\nu = n - 1$ degrees of freedom** (d.f.), where $\chi^2(\nu) = \Gamma(1/2, \nu)$.

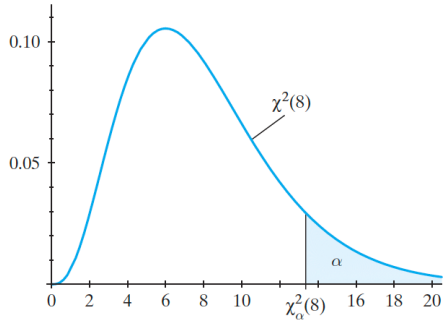



**Figure 6.19:** Chi-squared distribution with 8 degrees of freedom [41].

**Notation:** for $0 < \alpha < 1$ and $\nu \in \mathbb{N}^*$, $\chi^2_\alpha(\nu)$ is the **critical value** for which

$$P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha\,,$$

where $\chi^2 \sim \chi^2(\nu)$ follows a chi-squared distribution with $\nu$ degrees of freedom.

The values of $\chi^2_\alpha(\nu)$ can be found in various textbook tables, or by using `R` or specialized online calculators.

For instance, when $\nu = 8$ and $\alpha = 0.95$, we compute $\chi^2_{0.95}(8)$ *via*

```
qchisq(0.95, df=8,lower.tail = FALSE)
```

```
[1] 2.732637
```

Thus $P(\chi^2 > 2.732) = 0.95$, where $\chi^2 \sim \chi^2(8)$, i.e., $\chi^2$ has a chi-squared distribution with $\nu = 8$ degrees of freedom.

In other words, 95% of the area under the curve of the probability density function of $\chi^2(8)$ is found to the right of 2.732.

### Sample Mean With Unknown Population Variance

Suppose that $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2(\nu)$. If $Z$ and $V$ are independent, then the distribution of the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

is a **Student $t-$distribution with $\nu$ degrees of freedom**, which we denote by $T \sim t(\nu)$.[50]

**Theorem:** let $X_1, \ldots, X_n$ be independent normal random variables with mean $\mu$ and standard deviation $\sigma$. Let $\overline{X}$ and $S^2$ be the sample mean and sample variance, respectively. Then the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

follows a **Student $t-$distribution with $\nu = n - 1$ degrees of freedom**.

50: The probability density function of $t(\nu)$ is

$$f(x) = \frac{\Gamma(\nu/2 + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1 + x^2/\nu)^{\nu/2+1/2}}.$$

Using the same notation as with the chi-squared distribution, let $t_\alpha(\nu)$ represent the **critical $t$-value** above which we find an area under the p.d.f. of $t(\nu)$ equal to $\alpha$ , i.e.

$$P(T > t_\alpha(\nu)) = \alpha \,,$$

where $T \sim t(\nu)$.

For all $\nu$, the Student $t$-distribution is a symmetric distribution around zero, so we have $t_{1-\alpha}(\nu) = -t_\alpha(\nu)$. The critical values can be found in tables, or by using the R function `qt()`.

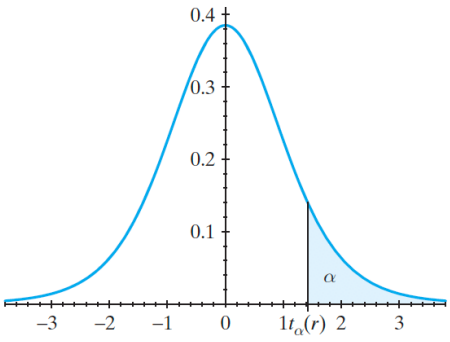



**Figure 6.20:** Student $t-$distribution with $r$ degrees of freedom [41].

If $T \sim t(\nu)$, then for any $0 < \alpha < 1$, we have

$$\begin{aligned} P\left(|T| < t_{\alpha/2}(\nu)\right) &= P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) \\ &= P\left(T < t_{\alpha/2}(\nu)\right) - P\left(T < -t_{\alpha/2}(\nu)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > -t_{\alpha/2}(\nu)\right)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > t_{1-\alpha/2}(\nu)\right)\right) \\ &= 1 - \alpha/2 - (1 - (1 - \alpha/2)) = 1 - \alpha. \end{aligned}$$

Consequently,

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha \,.$$

We can show that $t(\nu) \to \mathcal{N}(0,1)$ as $\nu \to \infty$; intuitively, this makes sense because the estimate $S$ gets better at estimating $\sigma$ when $n$ increases.

**Example** In R, we can see that when $T \sim t(8)$,

```
qt(0.025, df=8, lower.tail=FALSE)
```

```
[1] 2.306004
```

Thus, $P\left(T > 2.306\right) = 0.025$, which implies

$$P\left(T < -2.306\right) = 0.025$$

, so $t_{0.025}(8) = 2.306$ and

$$\begin{aligned} P(|T| \leq 2.306) &= P(-2.306 \leq T \leq 2.306) \\ &= 1 - P(T < -2.306) - P(T > 2.306) \\ &= 1 - 2P(T < -2.306) = 0.95\,. \end{aligned}$$

The Student $t-$distribution will be useful when the time comes to compute confidence intervals and to do hypothesis testing (see Chapter **??**).

### $F-$Distributions

Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$. If $U$ and $V$ are independent, then the random variable

$$F = \frac{U/\nu_1}{V/\nu_2}$$

follows an **$F$-distribution with $\nu_1$ and $\nu_2$ degrees of freedom**, which we denote by $F \sim F(\nu_1, \nu_2)$.

The probability density function of $F(\nu_1, \nu_2)$ is

$$f(x) = \frac{\Gamma(\nu_1/2 + \nu_2/2)(\nu_1/\nu_2)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1 + x\nu_1/\nu_2)^{\nu_1/2+\nu_2/2}}, \quad x \geq 0.$$

**Theorem:** if $S_1^2$ and $S_2^2$ are the sample variances of independent random samples of size $n$ and $m$, respectively, taken from normal populations with variances $\sigma_1^2$ and $\sigma_2^2$ , then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

follows an **$F$-distribution with $\nu_1 = n - 1$, $\nu_2 = m - 1$ d.f.**

**Notation:** for $0 < \alpha < 1$ and $\nu_1, \nu_2 \in \mathbb{N}^*$, $f_\alpha(\nu_1, \nu_2)$ is the **critical value** for which $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$ where $F \sim F(\nu_1, \nu_2)$. Critical values can be found in tables, or by using the `R` function `qf()`.

It can be shown that

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)};$$

for instance, since

```
qf(0.95, df1=6, df2=10, lower.tail=FALSE)
```

```
[1] 0.2463077
```

Thus,

$$f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10, 6)} = \frac{1}{4.06} = 0.246\,.$$

These distributions play a role in linear regression and ANOVA models (see Chapters 8 and 10).

## 6.6 Exercises

1. Two events each have probability 0.2 of occurring and are independent. What is the probability that neither occur?
2. Two events each have probability 0.2 and are mutually exclusive. What is the probability that neither occur?
3. A smoke-detector system has two parts, $A$ and $B$. If smoke occurs then the item $A$ detects it with probability 0.95, the item $B$ detects it with probability 0.98 whereas both of them detect it with probability 0.94. What is the probability that the smoke is undetected?
4. Let $A_1, A_2, A_3$ denote the events that the field goal is made by player $1, 2, 3$, respectively. Assume independence and $P(A_1) = 0.5$, $P(A_2) = 0.7$, $P(A_3) = 0.6$. Compute the probability that exactly 1 player is successful.
5. In a group of 16 candidates, 7 are chemists and 9 are physicists. In how many ways can one choose a group of 5 candidates with 2 chemists and 3 physicists?
6. A theorem of combinatorics states that the number of permutations of $n$ objects in which $n_1$ are alike of kind 1, $n_2$ are alike of kind 2, ..., and $n_r$ are alike of kind $r$ (that is, $n = n_1 + n_2 + \cdots + n_r$) is

$$\frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_r!}.$$

   Find the number of different words that can be formed by rearranging the letters in the following words.

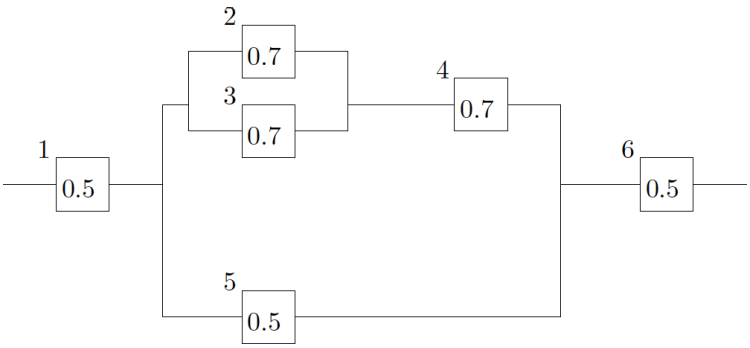
   a) NORMAL
   b) HHTTTT
   c) LLEWELLYN
   d) KITCHISSIPPI

7. A class consists of 490 engineering and 510 science students. The students are divided according to their marks:

| | Passed | Failed |
|---|---|---|
| Eng. | 430 | 60 |
| Sci. | 410 | 100 |

   If one person is selected randomly, what is the probability that they failed if they were an engineering student?
8. A company which produces a particular drug has two factories, $A$ and $B$. 70% of the drugs are made in factory $A$, 30% in factory $B$. If 95% of the drugs produced by factory $A$ meet standards while only 75% of those produced by factory $B$ do so, what is the probability that a random dose meets standards?
9. A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease; in 436 cases the test result was positive. The test was also given to a random sample of 500 patients without the disease; only in 5 cases was the result positive. In Canada 11.3% of the population aged 65+ have Alzheimer's disease. Find the probability that a person has the disease given that their test was positive.

10. Twelve items are independently sampled from a production line. If the probability that any given item is defective is 0.1, what is the probability of at most two defectives in the sample?
11. A student can solve 6 problems from a list of 10. For an exam 8 questions are selected at random from the list. What is the probability that the student will solve exactly 5 problems?
12. Consider the following system with six components. We say that it is functional if there exists a path of functional components from left to right. The probability of each component functions is shown. Assume that the components function or fail independently. What is the probability that the system operates?



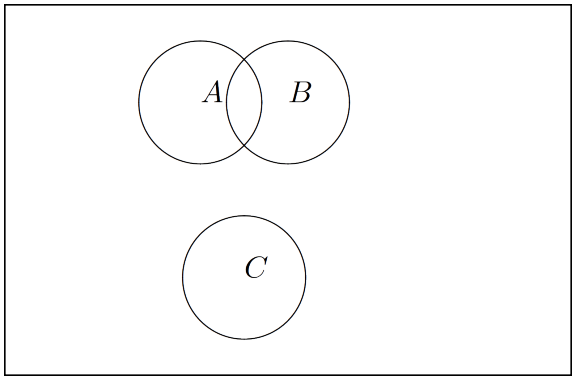

13. Pieces of aluminum are classified according to the finishing of the surface and according to the finishing of edge. The results from 85 samples are summarized as follows:

| | **Edge** | |
|---|---|---|
| **Surface** | excellent | good |
| excellent | 60 | 5 |
| good | 16 | 4 |

Let $A$ denote the event that a selected piece has an "excellent" surface, and let $B$ denote the event that a selected piece has an "excellent" edge. If samples are elected randomly, determine the following probabilities:

a) $P(A)$ b) $P(B)$ c) $P(A^c)$ d) $P(A \cap B)$ e) $P(A \cup B)$ f) $P(A^c \cup B)$

14. Three events are shown in the Venn diagram below.



Shade the region corresponding to the following events:

a) $A^c\ (A \cap B) \cup (A \cap B^c)$
b) $(A \cap B) \cup C$
c) $(B \cup C)^c$
d) $(A \cap B)^c \cup C$

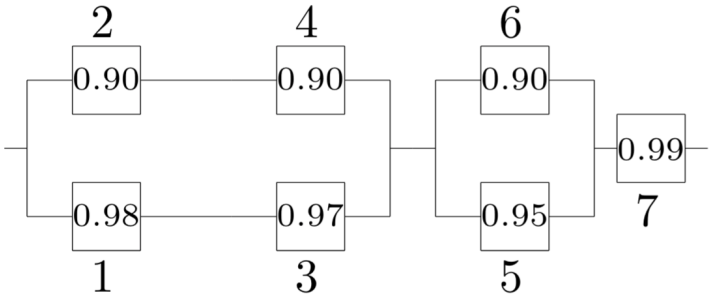
15. If $P(A) = 0.1$, $P(B) = 0.3$, $P(C) = 0.3$, and events $A, B, C$ are mutually exclusive, determine the following probabilities:

a) $P(A \cup B \cup C)$
b) $P(A \cap B \cap C)$
c) $P(A \cap B)$
d) $P((A \cup B) \cap C)$
e) $P(A^c \cap B^c \cap C^c)$
f) $P[(A \cup B \cup C)^c]$

16. The probability that an electrical switch, which is kept in dryness, fails during the guarantee period, is 1%. If the switch is humid, the failure probability is 8%. Assume that 90% of switches are kept in dry conditions, whereas remaining 10% are kept in humid conditions.

a) What is the probability that the switch fails during the guarantee period?
b) If the switch failed during the guarantee period, what is the probability that it was kept in humid conditions?

17. The following system operates only if there is a path of functional device from left to the right. The probability that each device functions is as shown. What is the probability that the circuit operates?




Assume independence.

18. An inspector working for a manufacturing company has a 95% chance of correctly identifying defective items and 2% chance of incorrectly classifying a good item as defective. The company has evidence that 1% of the items it produces are nonconforming (defective).

a) What is the probability that an item selected for inspection is classified as defective?
b) If an item selected at random is classified as non defective, what is the probability that it is indeed good?

19. Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit).

a) How many different 5−card poker hands can be drawn from the deck?
b) How many different 13−card bridge hands can be drawn from the deck?
c) What is the probability of an all-spade 5−card poker hand?
d) What is the probability of a flush (5−cards from the same suit)?
e) What is the probability that a 5−card poker hand contains exactly 3 Kings and 2 Queens?
f) What is the probability that a 5−card poker hand contains exactly 2 Kings, 2 Queens, and 1 Jack?

20. Students on a boat send messages back to shore by arranging seven coloured flags on a vertical flagpole.

a) If they have 4 orange flags and 3 blue flags, how many messages can they send?
b) If they have 7 flags of different colours, how many messages can they send?
c) If they have 3 purple flags, 2 red flags, and 4 yellow flags, how many messages can they send?

21. The Stanley Cup Finals of hockey or the NBA Finals in basketball continue until either the representative team form the Western Conference or from the Eastern Conference wins 4 games. How many different orders are possible ($WWEEEE$ means that the Eastern team won in 6 games) if the series goes

a) 4 games?
b) 5 games?
c) 6 games?
d) 7 games?

22. Consider an ordinary 52-card North American playing deck (4 suits, 13 cards in each suit), from which cards are drawn at random and without replacement, until 3 spades are drawn.

a) What is the probability that there are 2 spades in the first 5 draws?
b) What is the probability that a spade is drawn on the 6th draw given that there were 2 spades in the first 5 draws?
c) What is the probability that 6 cards need to be drawn in order to obtain 3 spades?
d) All the cards are placed back into the deck, and the deck is shuffled. 4 cards are then drawn from. What is the probability of having drawn a spade, a heart, a diamond, and a club, in that order?

23. A student has 5 blue marbles and 4 white marbles in his left pocket, and 4 blue marbles and 5 white marbles in his right pocket. If they transfer one marble at random from their left pocket to his right pocket, what is the probability of them then drawing a blue marble from their right pocket?
24. An insurance company sells a number of different policies; among these, 60% are for cars, 40% are for homes, and 20% are for both. Let $A_1, A_2, A_3, A_4$ represent people with only a car policy, only a home policy, both, or neither, respectively. Let $B$ represent the

event that a policyholder renews at least one of the car or home policies.

a) Compute $P(A_1)$, $P(A_2)$, $P(A_3)$, and $P(A_4)$.
b) Assume $P(B \mid A_1) = 0.6$, $P(B \mid A_2) = 0.7$, $P(B \mid A_3) = 0.8$. Given that a client selected at random has a car or a home policy, what is the probability that they will renew one of these policies?

25. An urn contains four balls numbered 1 through 4. The balls are selected one at a time, without replacement. A match occurs if ball $m$ is the $m$th ball selected. Let the event $A_i$ denote a match on the $i$th draw, $i = 1, 2, 3, 4$.

a) Compute $P(A_i)$, $i = 1, 2, 3, 4$.
b) Compute $P(A_i \cap A_j)$, $i, j = 1, 2, 3, 4$, $i \neq j$.
c) Compute $P(A_i \cap A_j \cap A_k)$, $i, j, k = 1, 2, 3, 4$, $i \neq j$, $i \neq k$, $j \neq k$.
d) What is the probability of at least 1 match?

26. The probability that a company's workforce has at least one accident in a given month is $(0.01)k$, where $k$ is the number of days in the month. Assume that the number of accidents is independent from month to month. If the company's year starts on January 1, what is the probability that the first accident occurs in April?
27. A Pap smear is a screening procedure used to detect cervical cancer. Let $T^-$ and $T^+$ represent the events that the test is negative and positive, respectively, and let $C$ represent the event that the person tested has cancer. The false negative rate for this test when the patient has the cancer is 16%; the false positive test for this test when the patient does not have cancer is 19%. In North America, the rate of incidence for this cancer is roughly 8 out of 100,000 women. Based on these numbers, is a Pap smear an effective procedure? What factors influence your conclusion?
28. Of three different fair dice, one each is given to Elowyn, Llewellyn, and Gwynneth. They each roll it. Let $E = \{$Elowyn rolls a 1 or a 2$\}$, $LL = \{$Llewellyn rolls a 3 or a 4$\}$, and $G = \{$Gwynneth rolls a 5 or a 6$\}$ be events.

a) What are the probabilities of each of $E$, $LL$, and $G$ occurring?
b) What are the probabilities of any two of $E$, $LL$, and $G$ occurring simultaneously?
c) What is the probability of all three of the events occurring simultaneously?
d) What is the probability of at least one of $E$, $LL$, or $G$ occurring?

29. Over the course of two baseball seasons, player $A$ obtained 126 hits in 500 at-bats in Season 1, and 90 hits in 300 at-bats in Season 2; player $B$, on the other hand, obtained 75 hits in 300 at-bats in Season 1, and 145 hits in 500 at-bats in Season 2. A player's batting average is the number of hits they obtain divided by the number of at-bats.

a) Which player has the best batting average in Season 1? In Season 2?
b) Which player has the best batting average over the 2-year period?
c) Can you explain what is happening here?

30. A stranger comes to you and shows you what appears to be a normal coin, with two distinct sides: Heads ($H$) and Tails ($T$). They flip the coin 4 times and record the following sequence of tosses: $HHHH$.

    a) What is the probability of obtaining this specific sequence of tosses? What assumptions do you make along the way in order to compute the probability? What is the probability that the next toss will be a $T$.
    b) The stranger offers you a bet: they will toss the coin another time; if the toss is $T$, they give you 100$, but if it is $H$, you give them 10$. Would you accept the bet (if you are not morally opposed to gambling)?
    c) Now the stranger tosses the coin 60 times and records $60 \times H$ in a row: $H \cdots H$. They offer you the same bet. Do you accept it?
    d) What if they offered 1000$ instead? $1,000,000$$?

31. An experiment consists in selecting a bowl, and then drawing a ball from that bowl. Bowl $B_1$ contains two red balls and four white balls; bowl $B_2$ contains one red ball and two white balls; and bowl $B_3$ contains five red balls and four white balls. The probabilities for selecting the bowls are not uniform: $P(B_1) = 1/3$, $P(B_2) = 1/6$, and $P(B_3) = 1/2$, respectively.

    a) What is the probability of drawing a red ball $P(R)$?
    b) If the experiment is conducted and a red ball is drawn, what is the probability that the ball was drawn from bowl $B_1$? $B_2$? $B_3$?

32. Two companies $A$ and $B$ consider making an offer for road construction. Company $A$ submits a proposal. The probability that $B$ submits a proposal is $1/3$. If $B$ does not submit the proposal, the probability that $A$ gets the job is $3/5$. If $B$ submits the proposal, the probability that $A$ gets the job is $1/3$. What is the probability that $A$ will get the job?
33. In a box of 50 fuses there are 8 defective ones. We choose 5 fuses randomly (without replacement). What is the probability that all 5 fuses are not defective?
34. The sample space of a random experiment is $\{a, b, c, d, e, f\}$ and each outcome is equally likely. A random variable is defined as follows

| outcome | $a$ | $b$ | $c$ | $d$ | $e$ | $f$ |
|---|---|---|---|---|---|---|
| $X$ | 0 | 0 | 1.5 | 1.5 | 2 | 3 |

    Determine the probability mass function of $X$. Determine the following probabilities:

    a) $P(X = 1.5)$
    b) $P(0.5 < X < 2.7)$
    c) $P(X > 3)$
    d) $P(0 \leq X < 2)$
    e) $P(X = 0 \text{ or } 2)$

35. Determine the mean and the variance of the random variable defined in the previous question.

36. $X$ has **uniform distribution** on a set of values $\{X_1, \ldots, X_k\}$ if

$$P(X = X_i) = \frac{1}{k}, \qquad i = 1, \ldots, k.$$

The thickness measurements of a coating process are **uniformly distributed** with values 0.15, 0.16, 0.17, 0.18, 0.19. Determine the mean and variance of the thickness measurements. Is this result compatible with a uniform distribution?

37. Samples of rejuvenated mitochondria are mutated in 1% of cases. Suppose 15 samples are studied and that they can be considered to be independent (from a mutation standpoint). Determine the following probabilities:

a) no samples are mutated;
b) at most one sample is mutated, and
c) more than half the samples are mutated.

Use the CDF table for the $\mathcal{B}(n, p)$, with $n = 15$ and $p = 0.99$:

| $r$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $P(X \leq r)$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| $r$ | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| $P(X \leq r)$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0004 | 0.0096 | 0.1399 | 1.0000 |

38. Samples of 20 parts from a metal punching process are selected every hour. Typically, 1% of the parts require re-work. Let $X$ denote the number of parts in the sample that require re-work. A process problem is suspected if $X$ exceeds its mean by more than three standard deviations.

a) What is the probability that there is a process problem?
b) If the re-work percentage increases to 4%, what is the probability that $X$ exceeds 1?
c) If the re-work percentage increases to 4%, what is the probability that $X$ exceeds 1 in at least one of the next five sampling hours?

39. In a clinical study, volunteers are tested for a gene that has been found to increase the risk for a particular disease. The probability that the person carries a gene is 0.1.

a) What is the probability that 4 or more people will have to be tested in order to detect 1 person with the gene?
b) How many people are expected to be tested in order to detect 1 person with the gene?
c) How many people are expected to be tested in order to detect 2 people with the gene?

40. The number of failures of a testing instrument from contaminated particles on the product is a Poisson random variable with a mean of 0.02 failure per hour.

a) What is the probability that the instrument does not fail in an 8−hour shift?
b) What is the probability of at least 1 failure in a 24−hour day?

41. Use R to generate a sample from a binomial distribution and from a Poisson distribution (select parameters as you wish). Use R to compute the sample means and sample variances. Compare these values to population means and population variances.

42. A container of 100 light bulbs contains 5 bad bulbs. We draw 10 bulbs without replacement. Find the probability of drawing at least 1 defective bulb.
43. Let $X$ be a discrete random variable with range $\{0, 1, 2\}$ and probability mass function (p.m.f.) given by $f(0) = 0.5$, $f(1) = 0.3$, and $f(2) = 0.2$. What are the expected value and variance of $X$?
44. A factory employs several thousand workers, of whom 30% are not from an English-speaking background. If 15 members of the union executive committee were chosen from the workers at random, evaluate the probability that exactly 3 members of the committee are not from an English-speaking background.
45. Assuming the context of the previous questions, what is the probability that a majority of the committee members do not come from an English-speaking background?
46. In a video game, a player is confronted with a series of opponents and has an 80% probability of defeating each one. Success with any opponent (that is, defeating the opponent) is independent of previous encounters. The player continues until defeated. What is the probability that the player encounters at least three opponents?
47. Assuming the context of the previous question, how many encounters is the player expected to have?
48. From past experience it is known that 3% of accounts in a large accounting company are in error. The probability that exactly 5 accounts are audited before an account in error is found, is:
49. A receptionist receives on average 2 phone calls per minute. Assume that the number of calls can be modeled using a Poisson random variable. What is the probability that he does not receive a call within a 3−minute interval?
50. Roll a 4−sided die twice, and let $X$ equal the larger of the two outcomes if they are different and the common value if they are the same. Find the p.m.f. and the c.d.f. of $X$.
51. Compute the mean and the variance of $X$ as defined in the previous question, as well as $\mathrm{E}[X(5 - X)]$.
52. A basketball player is successful in 80% of her (independent) free throw attempts. Let $X$ be the minimum number of attempts in order to succeed 10 times. Find the p.m.f. of $X$ and the probability that $X = 12$.
53. Let $X$ be the minimum number of independent trials (each with probability of success $p$) that are needed to observe $r$ successes. The p.m.f. of $X$ is

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

The mean and variance of $X$ are

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

Compute the mean minimum number of independent free throw attempts required to observe 10 successful free throws if the probability of success at the free thrown line is 80%. What about the standard deviation of $X$?

54. If $n \geq 20$ and $p \leq 0.05$, it can be shown that the binomial distribution with $n$ trials and an independent probability of success $p$ can be approximated by a Poisson distribution with parameter $\lambda = np$:

$$\frac{(np)^x e^{-np}}{x!} \approx \binom{n}{x} p^x (1-p)^{n-x}.$$

A manufacturer of light bulbs knows that 2% of its bulbs are defective. What is the probability that a box of 100 bulbs contains exactly at most 3 defective bulbs? Use the Poisson approximation to estimate the probability.

55. Consider a discrete random variable $X$ which has a uniform distribution over the first positive $m$ integers, i.e.

$$f(x) = P(X = x) = \frac{1}{m}, \quad x = 1, \ldots, m,$$

and $f(x) = 0$ otherwise. Compute the mean and the variance of $X$. For what values of $m$ is $\mathrm{E}[X] > \mathrm{Var}[X]$?

56. Assume that arrivals of small aircrafts at an airport can be modeled by a Poisson random variable with an average of 1 aircraft per hour.

    a) What is the probability that more than 3 aircrafts arrive within an hour?
    b) Consider 15 consecutive and disjoint 1−hour intervals. What is the probability that in none of these intervals we have more than 3 aircraft arrivals?
    c) What is the probability that exactly 3 aircrafts arrive within 2 hours?

57. In a group of ten students, each student has a probability of 0.7 of passing the exam. What is the probability that exactly 7 of them will pass an exam?

58. A company's warranty states that the probability that a new swimming pool requires some repairs within the 1st year is 20%. What is the probability, that the sixth sold pool is the first one which requires some repairs within the 1st year?

59. Consider the following R output:

```
> pbinom(16,100,0.25)
[1] 0.02111062
> pbinom(30,100,0.25)
[1] 0.8962128
> pbinom(32,100,0.25)
[1] 0.9554037
> pbinom(15,100,0.25)
[1] 0.01108327
> pbinom(17,100,0.25)
[1] 0.03762626
> pbinom(31,100,0.25)
[1] 0.9306511
```

Let $X \sim \mathcal{B}(n, p)$ with $n = 100$ and $p = 0.25$. Using the R output above, calculate $P(16 \leq X \leq 31)$.

60. Consider a random variable $X$ with probability density function given by

$$f(x) = \begin{cases} 0 & \text{if } x \leq -1 \\ 0.75(1-x^2) & \text{if } -1 \leq x < 1 \\ 0 & \text{if } x \geq 1 \end{cases}$$

What is the expected value and the standard deviation of $X$?

61. A random variable $X$ has a cumulative distribution function (c.d.f.)

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

What is the mean value of $X$?

62. Let $X$ be a random variable with p.d.f. $f(x) = \frac{3}{2}x^2$ for $-1 \leq x \leq 1$, and $f(x) = 0$ otherwise. Find $P(X^2 \leq 0.25)$.
63. In the inspection of tin plate produced by a continuous electrolytic process, 0.2 imperfections are spotted per minute, on average. Find the probability of spotting at least 2 imperfections in 5 minutes. Assume that we can model the occurrences of imperfections as a Poisson process.
64. If $X \sim \mathcal{N}(0, 4)$, find $P(|X| \geq 2.2)$.
65. If $X \sim \mathcal{N}(10, 1)$, what value of $k$ yields $P(X \leq k) = 0.701944$?
66. The time it takes a supercomputer to perform a task is normally distributed with mean 10 milliseconds and standard deviation 4 milliseconds. What is the probability that it takes more than 18.2 milliseconds to perform the task? (use the normal table or R).
67. Let $X$ be a random variable. What is the value of $b$ (where $b$ is not a function of $X$) which minimizes $\mathrm{E}[(X-b)^2]$?
68. The time to reaction to a visual signal follows a normal distribution with mean 0.5 seconds and standard deviation 0.035 seconds.

    a) What is the probability that time to react exceeds 1 second?
    b) What is the probability that time to react is between 0.4 and 0.5 seconds?
    c) What is the time to reaction that is exceeded with probability of 0.9?

69. Refer to the situation described in the aircraft question above.

    a) What is the length of the interval such that the probability of having no arrival within this interval is 0.1?
    b) What is the probability that one has to wait at least 3 hours for the arrival of 3 aircrafts?
    c) What is the mean and variance of the waiting time for 3 aircrafts?

70. Assume that $X$ is normally distributed with mean 10 and standard deviation 3. In each case, find the value $x$ such that:

    a) $P(X > x) = 0.5$
    b) $P(X > x) = 0.95$
    c) $P(x < X < 10) = 0.2$
    d) $P(-x < X - 10 < x) = 0.95$
    e) $P(-x < X - 10 < x) = 0.99$

71. Let $X \sim \mathrm{Exp}(\lambda)$ with mean 10. Find $P(X > 30 \mid X > 10)$.

72. Consider a random variable $X$ with the following probability density function:

$$f(x) = \begin{cases} 0 & \text{if } x \le -1 \\ \frac{3}{4}(1-x^2) & \text{if } -1 < x < 1 \\ 0 & \text{if } x \ge 1 \end{cases}$$

What is the value of $P(X \le 0.5)$?

73. A receptionist receives on average 2 phone calls per minute. If the number of calls follows a Poisson process, what is the probability that the waiting time for call will be greater than 1 minute?
74. A company manufactures hockey pucks. It is known that their weight is normally distributed with mean 1 and standard deviation 0.05. The pucks used by the NHL must weigh between 0.9 and 1.1. What is the probability that a randomly chosen puck can be used by NHL?
75. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ for the dice example above. Are $X$ and $Y$ independent?
76. Find $\text{Var}[X_1]$, $\text{Var}[X_2]$, and $\text{Cov}(X_1, X_2)$ for the chip example above. Are $X_1$ and $X_2$ independent?
77. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{x+y}{21}, \quad x = 1, 2, 3, \quad y = 1, 2.$$

78. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{xy^2}{30}, \quad x = 1, 2, 3, \quad y = 1, 2.$$

Are $X$ and $Y$ independent?

79. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.m.f.

$$f(x, y) = \frac{xy^2}{13}, \quad (x, y) = (1, 1), (1, 2), (2, 2)$$

Are $X$ and $Y$ independent?

80. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ have joint p.d.f.

$$f(x, y) = \frac{3}{2}x^2(1 - |y|), \quad -1 < x < 1, \quad -1 < y < 1.$$

Are $X$ and $Y$ independent?

81. Find $\text{Var}[X]$, $\text{Var}[Y]$, and $\text{Cov}(X, Y)$ if $X$ and $Y$ follow

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

82. Suppose that samples of size $n = 25$ are selected at random from a normal population with mean 100 and standard deviation 10. What is the probability that sample mean falls in the interval

$$(\mu_{\overline{X}} - 1.8\sigma_{\overline{X}}, \mu_{\overline{X}} + 1.0\sigma_{\overline{X}})?$$

83. The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean $\mu = 8.2$ minutes and standard deviation $\sigma = 1.5$ minutes. Suppose that a random sample of $n = 49$ customers is taken. Compute the approximate

probability that the average waiting time for these customers is:

a) Less than 10 min.
b) Between 5 and 10 min.
c) Less than 6 min.

84. A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and standard deviation of 8. A second random sample of size $n_2 = 9$ is taken independently from another normal population with mean 70 and standard deviation of 12. Let $\overline{X}_1$ and $\overline{X}_2$ be the two sample means. Find

a) The probability that $\overline{X}_1 - \overline{X}_2$ exceeds 4.
b) The probability that $3.5 < \overline{X}_1 - \overline{X}_2 < 5.5$.

85. Using R, illustrate the central limit theorem by generating $M = 300$ samples of size $n = 30$ from:

a) a normal random variable with mean 10 and variance 0.75;
b) a binomial random variable with 3 trials and probability of success 0.3

Repeat the same procedure for samples of size $n = 200$. What do you observe?

86. Suppose that the weight in pounds of a North American adult can be represented by a normal random variable with mean 150 lbs and variance 900 $\text{lbs}^2$. An elevator containing a sign "Maximum 12 people'' can safely carry 2000 lbs. What is the probability that 12 North American adults will not overload the elevator?
87. Let $X_1, \cdots, X_{50}$ be an independent random sample from a Poisson distribution with mean 1. Set $Y = X_1 + \cdots + X_{50}$. Find an approximation of the probability $P(48 \leq Y \leq 52)$.
88. A new type of electronic flash for cameras will last an average of 5000 hours with a standard deviation of 500 hours. A quality control engineer intends to select a random sample of 100 of these flashes and use them until they fail. What is the probability that the mean life time of the sample of 100 flashes will be less than 4928 hours?
89. Assume that random variables $\{X_1, \ldots, X_8\}$ follow a normal distribution with mean 2 and variance 24. Independently, assume that random variables $\{Y_1, \ldots, X_{16}\}$ follow a normal distribution with mean 1 and variance 16. Let $\overline{X}$ and $\overline{Y}$ be the corresponding sample means. What is $P(\overline{X} + \overline{Y} > 4)$?
90. Suppose that $X_1 \sim \mathcal{N}(3, 4)$ and $X_2 \sim \mathcal{N}(3, 45)$. Given that $X_1$ and $X_2$ are independent random variables, what is a good approximation of $P(X_1 + X_2 > 9.5)$?
91. Consider a sample $\{X_1, \ldots, X_{10}\}$ from a normal population $X_i \sim \mathcal{N}(4, 9)$. Denote by $\overline{X}$ and $S^2$ the sample mean and the sample variance, respectively. Find $c$ such that

$$P\left(\frac{\overline{X} - 4}{S/\sqrt{10}} \leq c\right) = 0.99.$$