COURSE NOTES
STATS 210
Statistical Theory

INGENIO·ET·LABORE

The
University
of Auckland

Department of Statistics
University of Auckland

Contents

1. Probability **3**

1.1 Sample Space, Events, Probabilities, and Random Variables 3

1.2 Bernoulli trials and the Binomial Distribution 10

1.3 Conditional probability 16

1.4 Statistical independence 21

1.5 Bayes' Theorem 23

1.6 The Partition Theorem (Law of Total Probability) 24

1.7 Extra practice and reference 28

1.8 Probability Reference List 31

2. Foundations of Statistical Inference **33**

2.1 Introduction 33

2.2 Hypothesis testing 35

2.3 Example: Presidents and deep-sea divers 42

2.4 Example: Birthdays and sports professionals 49

2.5 Likelihood and estimation 52

2.6 Random numbers and histograms 64

2.7 Expectation 67

2.8 Variable transformations 73

2.9 Variance 79

2.10 Mean and Variance of the Binomial($\boldsymbol{n, p}$) distribution 84

2.11 Mean and Variance of Estimators 88

3. Modelling with Discrete Probability Distributions **92**

3.1 Binomial distribution 92

3.2 Geometric distribution 93

3.3 Negative Binomial distribution 98

3.4 Hypergeometric distribution: sampling without replacement 101

3.5 Poisson distribution 104

3.6 Likelihood and log-likelihood for $\boldsymbol{n}$ independent observations 110

3.7 Subjective modelling 113

3.8 Statistical regression modelling 115

4. Continuous Random Variables **123**

4.1 Introduction 123

4.2 The probability density function 125

4.3 The Exponential distribution 136

4.4 Likelihood and estimation for continuous random variables 141
4.5 Hypothesis tests 143
4.6 Expectation and variance 145
4.7 Exponential distribution mean and variance 147
4.8 The Uniform distribution 151
4.9 The Change of Variable Technique: finding the distribution of $\boldsymbol{g(X)}$ 153
4.10 Change of variable for non-monotone functions: non-examinable 158
4.11 The Gamma distribution 160
4.12 The Beta Distribution: non-examinable 163

5. The Normal Distribution and the Central Limit Theorem 164
5.1 The Normal Distribution 164
5.2 The Central Limit Theorem (CLT) 168
5.3 Confidence intervals 174

6. Wrapping Up 178
6.1 Estimators — the good, the bad, and the estimator PDF 178
6.2 Hypothesis tests: in search of a distribution 183

Chapter 1: Probability Essentials

In this chapter we review the essential concepts of probability that will be needed as building blocks for the rest of the course.

1.1 Sample Space, Events, Probabilities, and Random Variables

First of all, **everything about probability starts with a sample space.** Probabilities have no meaning without reference to a sample space, and the values of probabilities change according to which sample space they relate to. Understanding the role and importance of the sample space is one of the most important steps in mastering probability and statistical theory.

Definition: A **random experiment** is an experiment whose outcome is not known until it is observed.

- *A random experiment describes a situation with an unpredictable, or random, outcome.*

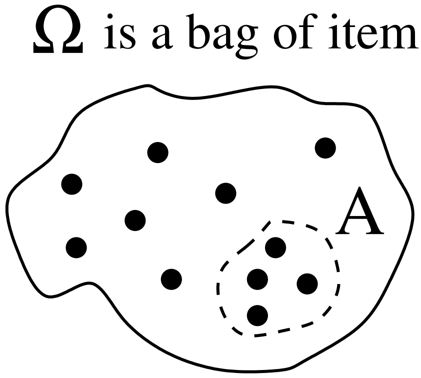
Definition: A **sample space**, Ω, is a set of outcomes of a random experiment. Every possible outcome is included in one, and only one, element of Ω.

- Ω *is a collection of all the things that could happen.*
- Ω *is a set. This means we can use the language of set theory, e.g.* $\cap$ *and* $\cup$.

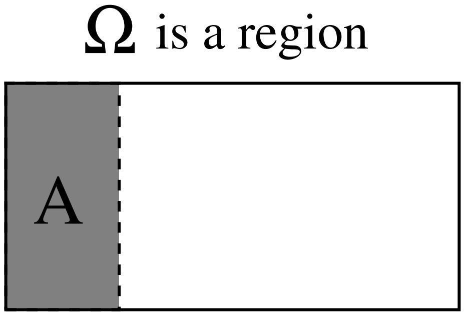
Definition: An **event**, A, is also a collection of outcomes. It is a subset of Ω.

- *An event* A *is 'something that could happen'.*
- *An event* A *is a set of specific outcomes we are interested in.*
- *The formal definition of an event* A *is a subset of the sample space:* $A \subseteq \Omega$.
- *Just like* Ω, A *is also a set. This means we can use the language of set theory, e.g. for two events* A *and* B *we talk about* $A \cap B$, $A \cup B$, $\overline{A}$, *and so on.*
- *It makes no sense to talk about events unless we have first defined the random experiment and the sample space. This is not always as easy as it sounds!*

It is helpful to conceptualise sample spaces and events in pictures.



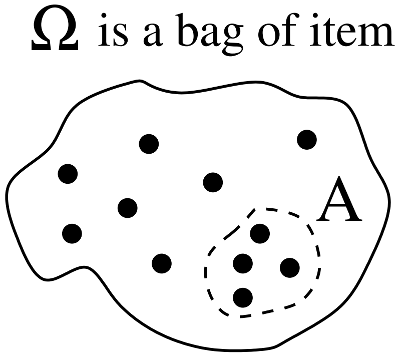
Event A is a smaller bag of items



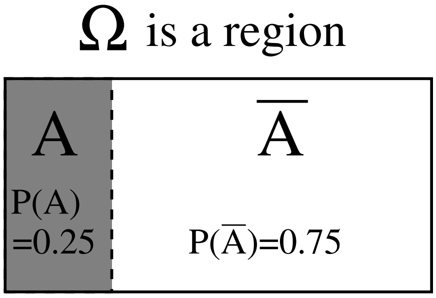
Event A is a subregion

Probability

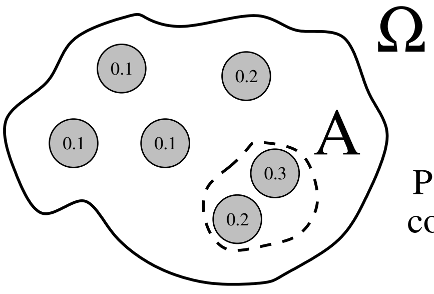
The idea of probability is to *attach a number to every item or event in* Ω *that reflects how likely the event is to occur.*



P(A)=4/11 if all items are equally likely



Probability is represented by AREA



P(A)=0.5 even though bag A contains only 2 out of 6 items

Question: What random experiments are we implicitly assuming here?

- Ω as a bag of items? *Pick an item at random from the bag.*
- Ω as a region? *Select a point at random from the region.*

Formal probability definition: the three axioms

As the pictures imply, the idea of probability is to allocate a number to every subset of Ω that reflects how likely we are to obtain an outcome in this subset. Imagine that all of Ω is given a cake: the idea of probability is to ***distribute*** a piece of cake to each item in Ω. Some items might get more cake than others, reflecting that the corresponding events are more likely to occur. This is why we talk about ***probability distributions:*** *a probability distribution describes how much 'cake' is given to each subset of Ω.*

Definition: A **probability distribution** allocates an amount of probability to every possible subset of Ω.

This idea is formalized in the following three ***Axioms,*** which constitute the *definition* of a probability distribution. A rule for allocating probability to subsets of Ω is a valid probability distribution if and only if it satisfies the following three axioms or conditions.

Axiom 1: $\mathbb{P}(\Omega) = 1$.

- *This means that the total amount of 'cake' available is 1.*
- *It also makes it clear that probability depends on the sample space, Ω, and has no meaning unless Ω is defined first.*

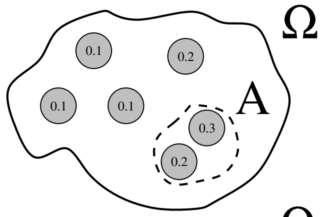
Axiom 2: $0 \leq \mathbb{P}(A) \leq 1$ for all events A.

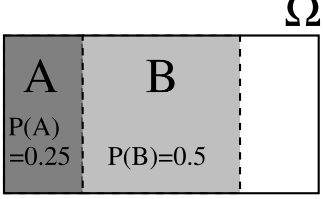
- *This says that probability is always a number between 0 and 1.*

Axiom 3: If $A_1, A_2, \ldots, A_n$ are **mutually exclusive** events, (no overlap), then

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n).$$

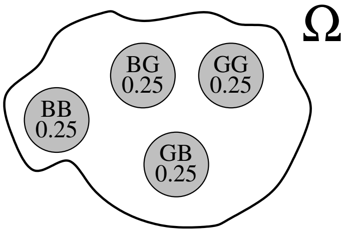
- *This says that if you have non-overlapping sets, the amount of cake they have in total is the sum of their individual amounts.*
- *This axiom is the reason why we can say that* $\mathbb{P}(A) = 0.3 + 0.2 = 0.5$ *in the bag diagram.*
- *In the region diagram,* $\mathbb{P}(A \cup B)$ =$0.25 + 0.5 = 0.75$.





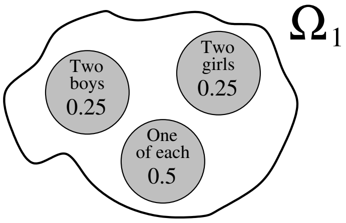
Examples of probability distributions

Suppose we are interested in the composition of a two-child family in terms of number of girls and boys. Assuming each child is equally likely to be a boy or a girl, there are *four equally-likely outcomes:*

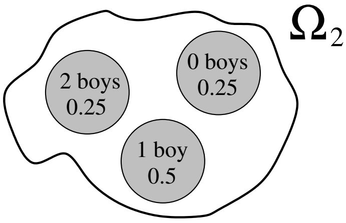


So if we pick a two-child family at random, we have probability $1/4 = 0.25$ of getting each of the outcomes BB, BG, GB, and GG.

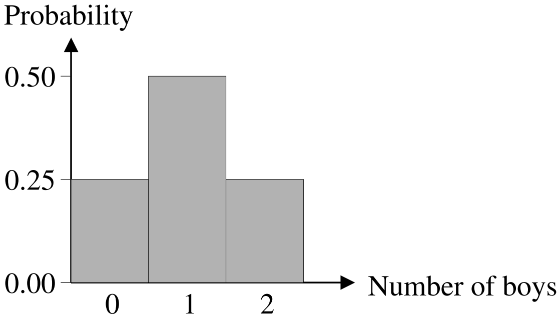
Now suppose we don't care *what order* the children are in: we only care *how many of each sex* are in the family. We could choose to represent this by a second sample space, Ω_1, in which the outcomes are no longer equally likely:



This is cumbersome to write down — especially if we consider listing the options for families of more than two children. Instead, we can be more efficient if we describe the outcomes by *counting the number of boys: 0, 1, or 2, giving* $\Omega_2 = \{0, 1, 2\}$*:*

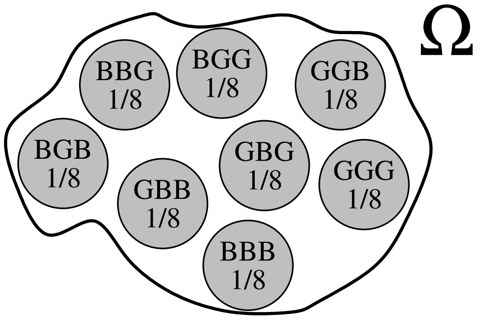


Now think of a different way of picturing Ω_2 that is easier to extend:

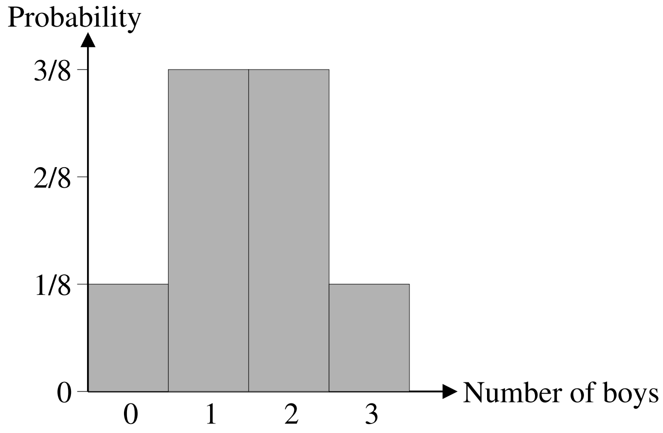


This representation is more like the 'region' image of Ω we used earlier, where probabilities are represented by *areas.*
It also has the huge advantage of being a flexible, graphical display.
Question: where would you draw Ω_2?

If we move to three-child families, we quickly see the advantage of our graphical depiction of Ω:



(a) Representation of the probability distribution if child-order is of interest.



(b) Representation of the probability distribution if we count the number of boys.

We can see that the numerical expression of outcomes (0, 1, 2, or 3 boys) is much more succinct than describing all combinations, BBB, BBG, BGB, . . . , GGG, as long as we do not care about the order that children occur in the family. However, ***we have to take account of all the different orderings when we calculate probabilities:***

$$\mathbb{P}(\textit{2 boys}) = \mathbb{P}(BBG) + \mathbb{P}(BGB) + \mathbb{P}(GBB) = 3 \times \frac{1}{8} = \frac{3}{8}.$$

Random variables

The idea above of converting an outcome described in words (e.g. BBG) into a numeric summary of the outcome (e.g. 2 boys) is the definition of a ***random variable.*** Instead of writing $\mathbb{P}$(2 boys) above, we give the unknown numerical outcome a capital letter, say X.

X is called a ***variable*** because it is a *variable number,* and it is called ***random*** because we don't know what value it will take until we make an observation of our random experiment. For example, if I pick a three-child family at random, I might observe $X = 0$, $X = 1$, $X = 2$, or $X = 3$ boys.

In essence, *a random variable is a numeric summary of the outcome of a random experiment.*

In formal language, a random variable is a mapping from Ω to the real numbers: $X : \Omega \to \mathbb{R}$. For example, for outcome BBG (a member of Ω), the number of boys is 2, so we can write $X(BBG) = 2$. However, we usually use a more succinct notation and just say that our outcome is $X = 2$.

Everything you need to know about random variables

- Random variables always have *CAPITAL LETTERS, e.g. X or Y.*

  Understand the capital letter to mean that X denotes a quantity that will take on values at random.

- The term 'random variable' is often abbreviated to *rv.*

- You can think of a random variable simply as *the name of a mechanism for generating random real numbers.*

  In the example above, X generates random numbers 0, 1, 2, or 3 by picking a 3-child family at random and counting how many boys are in it.

- Each possible ***value*** of a random variable has a probability associated with it. In the example above, where X is the number of boys in a three-child family:

$$\mathbb{P}(X=0)=\frac{1}{8}; \quad \mathbb{P}(X=1)=\frac{3}{8}; \quad \mathbb{P}(X=2)=\frac{3}{8}; \quad \mathbb{P}(X=3)=\frac{1}{8}.$$

- If we want to refer to a generic, unspecified, value of a random variable, we use a *lower-case letter, such as x or y.*

  For example, we might be interested in finding a formula for $\mathbb{P}(X=x)$ for all values $x=0,1,2,3$.

Differences between X, x, $\{X=x\}$, and $\mathbb{P}(X=x)$

It is very important to understand this standard notation and how it is used.

- X (capital letter) is a ***random variable:*** a mechanism for generating random real numbers. It is mainly used as a *name — just like your own name.*

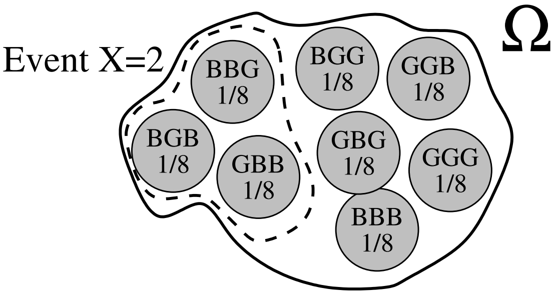
  If we say $X=2$, it is a bit like saying, 'Susan is in the kitchen.' It tells us a current observation of the random variable that we have called X.

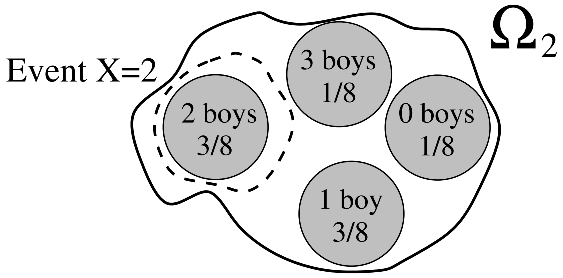
- x (lower-case letter) is a ***real number*** like 2 or 3. It is used to indicate an unspecified value that X might take.

- $\{X=x\}$, often written just as $X=x$, is an ***event:*** it is a *thing that happens.*

  For example, $X=2$ is the event that we count 2 boys when we pick a three-child family at random.

Because $X = 2$ is an event, it is a *set: a subset of the sample space.*





Crucially, *use set notation to combine expressions like* $X = 2$*.*

- $\mathbb{P}(X = x)$ is a ***real number:*** it is a number between 0 and 1. *Use operations like + and * to combine probabilities: for example,* $\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1)$.

When talking about ***events,*** like $\{X = x\}$, use set notation like $\cap$ and $\cup$.

When talking about ***probabilities,*** like $\mathbb{P}(X = x)$, use ordinary addition and multiplication $+$ and $\times$, just as you would for any other real numbers.

Right	**Wrong**
$X = 2 \cup X = 3$ *Event that X takes the value 2 OR 3*	$X = 2 + X = 3$
$\mathbb{P}(X = 2 \cup X = 3)$ *Probability of the event that X is 2 OR 3*	$\mathbb{P}(X = 2) \cup \mathbb{P}(X = 3)$
$\mathbb{P}(X = 2) + \mathbb{P}(X = 3)$ *Probability of the event that X is 2 OR 3*	$\mathbb{P}(X = 2 + X = 3)$
$X \leq 2 \cap X > 1$ *Event that X takes a value that is* *BOTH ≤ 2 AND > 1* *(the value $X = 2$ is the only possibility)*	$X \leq 2 \times X > 1$
$\mathbb{P}(X \leq 2 \cap X > 1)$ *Probability that X is BOTH ≤ 2 AND > 1*	$\mathbb{P}(X \leq 2) \cap \mathbb{P}(X > 1)$

1.2 Bernoulli trials and the Binomial Distribution

The Binomial distribution is one of the simplest probability distributions. We shall use it extensively throughout the course for illustrating statistical concepts.

The Binomial distribution *counts the number of successes in a fixed number n of independent trials, where each trial has two possible outcomes: Success with probability p, and Failure with probability $1 - p$.*

Such trials are called ***Bernoulli trials,*** named after the 17th century Swiss mathematician Jacques Bernoulli.

Definition: A sequence of **Bernoulli trials** is a sequence of independent trials where each trial has two possible outcomes, denoted Success and Failure, and the probability of Success stays constant at p for all trials.

Examples: (1) Repeated tossing of a fair coin: *'Success' = 'Head';* $p = \mathbb{P}(\textit{Head}) = 0.5$.
(2) Repeated rolls of a fair die: $p = \mathbb{P}(\textit{Get a 6}) = 1/6$.

Jacques Bernoulli, and his brother Jean, were bitter rivals. They both studied mathematics secretly, against their father's will. Their father wanted Jacques to be a clergyman and Jean to be a merchant.

Note: Saying the trials are ***independent*** means that *they do not influence each other.* Thus, whether the current trial yields a Success or a Failure is not influenced by the outcomes of any previous trials. For example, you are ***not*** more likely to have a win after a run of losses: the previous outcomes simply have no influence.

Definition: The random variable Y is called a **Bernoulli random variable** if *it takes only two values, 0 and 1. We write* $Y \sim$ *Bernoulli*(p), *where* $p = \mathbb{P}(Y = 1)$.

Definition: For any random variable Y, we define the **probability function** of Y to be the function $f_Y(y) = \mathbb{P}(Y = y)$.

The probability function of the Bernoulli random variable is:

$$f_Y(y) = \mathbb{P}(Y = y) = \begin{cases} p & \textit{if } y = 1 \quad \textit{(Success)} \\ 1 - p & \textit{if } y = 0 \quad \textit{(Failure)} \end{cases}$$

We often write the probability function in table format:

y	0	1
$\mathbb{P}(Y = y)$	$1 - p$	p

Binomial distribution

The Binomial distribution describes the outcome from a fixed number, n, of Bernoulli trials. For example:

- X is the number of boys in a 3-child family: $n = 3$ *trials (children);* $p = \mathbb{P}(\textit{Boy}) = 0.5$ *for each child.*
- X is the number of 6's obtained in 10 rolls of a die: $n = 10$ *trials (die rolls);* $p = \mathbb{P}(\textit{Get a 6}) = 1/6$ *for each roll.*

Definition: Let X be the number of successes obtained in n independent Bernoulli trials, each of which has probability of success p.

Then X has the **Binomial distribution with parameters n and p.**

We write $X \sim \textit{Binomial}(n, p)$*, or* $X \sim \textit{Bin}(n, p)$.

The Binomial distribution counts the number of **successes** in a **fixed number** of Bernoulli trials.

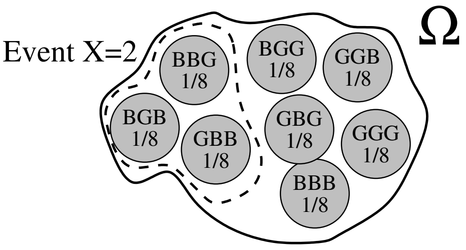
If $X \sim \text{Binomial}(n, p)$, then $X = x$ if there are x successes in the n trials. We don't care what order the successes occur in — in other words, we don't care *which* of the trials are successes and which are failures. However, we do have to bear in mind all the different orderings when we calculate the probabilities of the distribution.

Take the example of X = number of boys in a 3-child family, so $X \sim \textit{Binomial}(n = 3, p = 0.5)$.

If we want to calculate $\mathbb{P}(X = 2)$, we have to take account of all the different ways that we can achieve $X = 2$:

$\mathbb{P}(X = 2) = \mathbb{P}(BBG) + \mathbb{P}(BGB) + \mathbb{P}(GBB)$.

In this case, there are 3 ways of getting the outcome we are interested in: 2 boys and 1 girl. How would we calculate the number of ways in general?



There are 3 trials (children), and we need to choose 2 of them to be boys. The number of ways of choosing 2 trials from 3 is:

$$^3C_2 = \binom{3}{2} = \frac{3!}{(3-2)!\,2!} = \frac{3 \times 2 \times 1}{1 \times (2 \times 1)} = 3.$$

Question: How many ways are there of achieving 6 boys in a 10-child family?
Answer:

$$^{10}C_6 = \binom{10}{6} = \frac{10!}{(10-6)!\,6!} = 210 \qquad \text{— use calculator button } {}^nC_r\,.$$

Question: How many ways are there of achieving x successes in n trials?
Answer:

$$^{n}C_x = \binom{n}{x} = \frac{n!}{(n-x)!\,x!}\,.$$

Question: If each trial has probability p of being a success, what is the probability of getting the precise outcome $SFFSF$ from $n = 5$ trials?

Answer: $p \times (1-p) \times (1-p) \times p \times (1-p) = p^2(1-p)^3$. *This will be the same whatever order the successes and failures are in. But it only gives the probability for one ordering.*

Question: What is the probability of **one** ordering that contains x successes and $n - x$ failures?

Answer: $p^x(1-p)^{n-x}$.

Question: So what is the overall probability of achieving x successes in n trials: $\mathbb{P}(X = x)$ when $X \sim \text{Binomial}(n, p)$?

Answer: *(Number of orderings)* $\times$ *(probability of each ordering)* $= \binom{n}{x}p^x(1-p)^{n-x}$.

This gives the **probability function for the Binomial distribution:**

Let $X \sim \text{Binomial}(n, p)$. The probability function for X is:

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

Note: 1. Importantly, $f_X(x) = 0$ *if x is not one of the values* $0, 1, \ldots, n$.
The correct way to write the range of values is $x = 0, \ldots, n$.

— Writing $x \in [0, n]$ is **wrong,** because this includes decimals like 0.4.

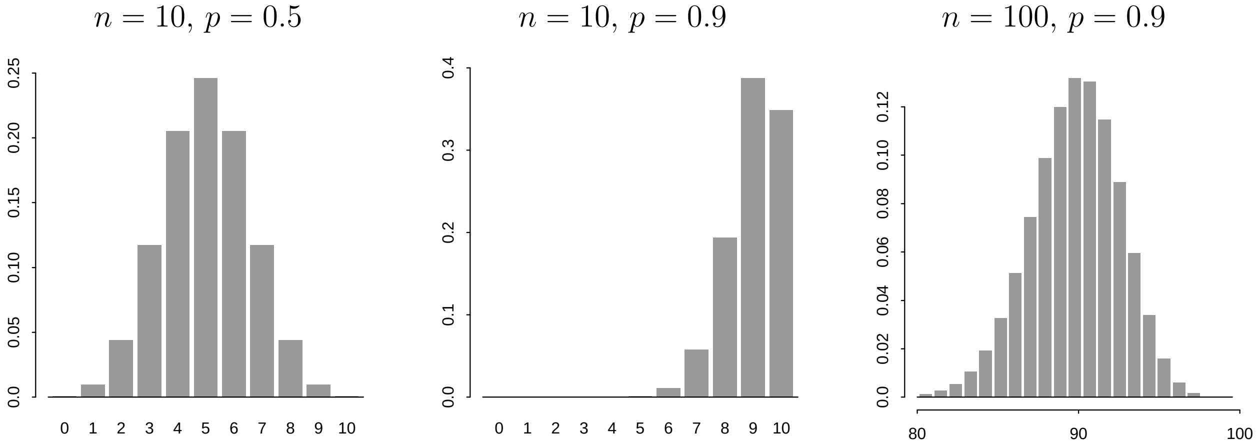
— Writing $x = 0, 1, \ldots$ is **wrong,** because the range of values must stop at n: you can't have more than n successes in n trials.

2. $f_X(x)$ means, *'the probability function belonging to the r.v. I've named X'.*
Use a capital X in the subscript and a lower-case x as the argument.

Shape of the Binomial distribution

The shape of the Binomial distribution depends upon the values of n and p. For small n, the distribution is almost symmetrical for values of p close to 0.5, but highly skewed for values of p close to 0 or 1. As n increases, the distribution becomes more and more symmetrical, and there is noticeable skew only if p is very close to 0 or 1.

The probability functions for various values of n and p are shown below.



Sum of independent Binomial random variables:

If X and Y are *independent*, and $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Binomial}(m, p)$, then

$$X + Y \sim \textit{Bin}(n + m, p).$$

This is because X counts the number of successes out of n trials, and Y counts the number of successes out of m trials: so overall, $X + Y$ counts the total number of successes out of $n + m$ *trials.*

Note: X and Y must both share *the same value of* p.

Binomial random variable as a sum of Bernoulli random variables

It is often useful to express a $\text{Binomial}(n, p)$ random variable as the sum of n $\text{Bernoulli}(p)$ random variables. If $Y_i \sim \text{Bernoulli}(p)$ for $i = 1, 2, \ldots, n$, and if $Y_1, Y_2, \ldots, Y_n$ are independent, then:

$$X = Y_1 + Y_2 + \ldots + Y_n \sim \text{Binomial}(n, p).$$

This is because X and $Y_1 + \ldots + Y_n$ both represent *the number of successes in* n *independent trials, where each trial has success probability* p.

Cumulative distribution function, $F_X(x)$

We have defined the *probability function,* $f_X(x)$, *as* $f_X(x) = \mathbb{P}(X = x)$.

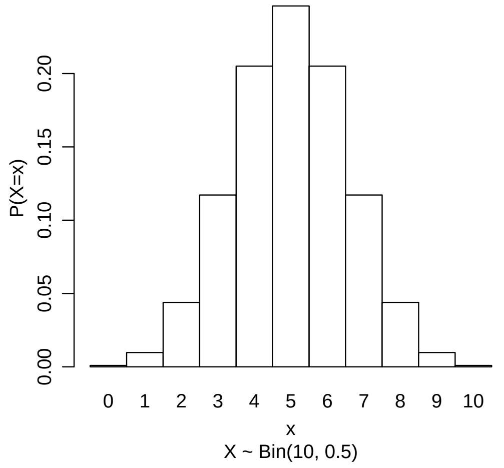
Another function that is widely used is the *cumulative distribution function*, or CDF, written as $F_X(x)$.

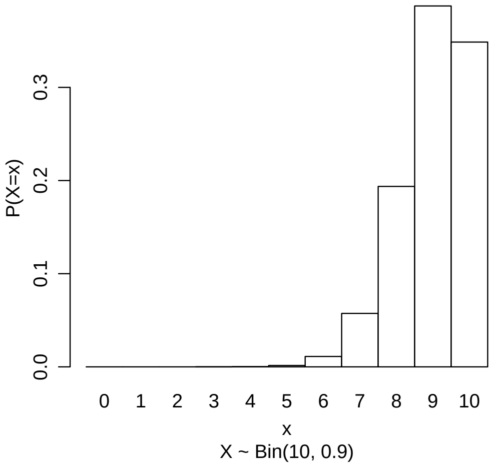
Definition: *The cumulative distribution function, or CDF, is*

$$\boxed{F_X(x) = \mathbb{P}(X \leq x) \ \ \textit{for} \ \ -\infty < x < \infty}$$

The cumulative distribution function $F_X(x)$ as a probability sweeper

The cumulative distribution function, $F_X(x)$, *sweeps up all the probability up to and including the point* x.





Clearly,

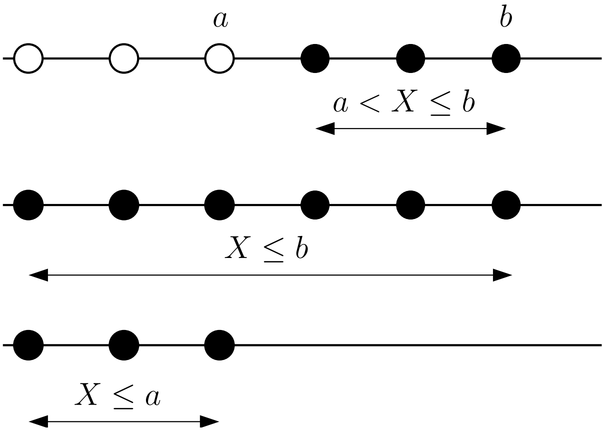
$$\boxed{F_X(x) = \sum_{y \leq x} f_X(y)\,.}$$

Using the cumulative distribution function to find probabilities

$$\boxed{\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad \textit{if}\, b > a.}$$

Proof that $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$:

$$\begin{array}{lrcl} & \mathbb{P}(X \leq b) & = & \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b) \\ \textit{So} & F_X(b) & = & F_X(a) + \mathbb{P}(a < X \leq b) \\ \Rightarrow & F_X(b) - F_X(a) & = & \mathbb{P}(a < X \leq b). \end{array}$$



Warning: endpoints

Be careful of endpoints and the difference between $\leq$ and $<$. For example,

$$\mathbb{P}(X < 10) = \mathbb{P}(X \leq 9) = F_X(9).$$

Warning!
End of interval

Examples: Let $X \sim \text{Binomial}(100, 0.4)$. In terms of $F_X(x)$, what is:

1. $\mathbb{P}(X \leq 30)$? $\quad F_X(30)$.
2. $\mathbb{P}(X < 30)$? $\quad F_X(29)$.
3. $\mathbb{P}(X \geq 56)$?

$$1 - \mathbb{P}(X < 56) = 1 - \mathbb{P}(X \leq 55) = 1 - F_X(55).$$

4. $\mathbb{P}(X > 42)$?

$$1 - \mathbb{P}(X \leq 42) = 1 - F_X(42).$$

5. $\mathbb{P}(50 \leq X \leq 60)$?

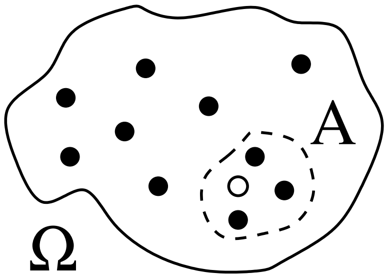
$$\mathbb{P}(X \leq 60) - \mathbb{P}(X \leq 49) = F_X(60) - F_X(49).$$

1.3 Conditional probability

We have mentioned that ***probability*** depends upon the sample space, Ω:

$\mathbb{P}(\Omega) = 1$, *so the symbol* $\mathbb{P}$ *is only defined relative to a particular sample space* Ω.

Conditional probability is about *changing the sample space.*
In particular, conditional probability is about ***reducing the sample space to a smaller one.***



Look at Ω on the right. Pick a ball at random. All 11 balls are equally likely to be picked. What is the probability of selecting the white ball?

$\mathbb{P}(\textit{white ball}) = \frac{1}{11}$.

Now suppose we select a ball only from within the smaller bag A. Recall that A is a subset of Ω, so in probability language, A is an *event.*

What is the probability of selecting the white ball, if we pick only from the balls in bag A?

$\mathbb{P}(\textit{white ball if we select only from } A) = \frac{1}{4}$.

We use a shorthand notation to write this down:

$$\mathbb{P}(\textit{white ball if we select only from } A) = \mathbb{P}(\textit{white ball} \,|\, A) = \frac{1}{4} \,.$$

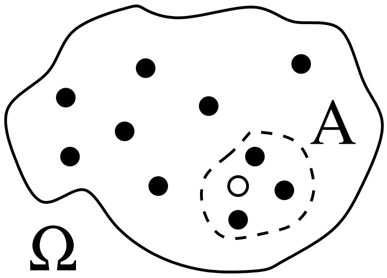
We read this as, 'probability of the white ball ***given*** A', or 'probability of selecting the white ball from ***within*** A'.

$\mathbb{P}(\text{white ball} \,|\, A)$ is called a ***conditional probability,*** and we say we have *conditioned on event* A.

Note: The vertical bar in $\mathbb{P}(\text{white ball} \,|\, A)$ is vertical: $|$.
Do not write a conditional probability as $\mathbb{P}(W/A)$ *or* $\mathbb{P}(W\backslash A)$*: it is* $\mathbb{P}(W \,|\, A)$.

What we have done is to ***reduce the sample space*** from Ω, which was a bag containing 11 equally-likely items, to a smaller bag A which contains only 4 equally-likely items.

But A is still a bag of items — so A is a valid sample space in its own right. When we write $\mathbb{P}(W \,|\, A)$, we have ***changed the sample space*** from Ω to A.



Define event $W = \{\textit{pick white ball}\}$.

We have said: $\mathbb{P}(W) = \frac{1}{11}$.

This means $\mathbb{P}(W \textit{ from within } \Omega) = \frac{1}{11}$,

where we recall that the symbol $\mathbb{P}$ is defined relative to Ω because $\mathbb{P}(\Omega) = 1$.

Now if we reduce to selecting only from the balls in A, we write: $\mathbb{P}(W \,|\, A) = \frac{1}{4}$.

Question: What is $\mathbb{P}(A \,|\, A)$?

Answer: $\mathbb{P}(A \,|\, A) = 1$, *because if we select items from within A, we are definitely going to select* **something** *in A.*

The conditional probability $\mathbb{P}(W \,|\, A)$ means ***the probability of event W, when selecting only from within set A.***

Read it as 'probability of event W, given event A',
or 'probability of event W ***from within the set A***.'

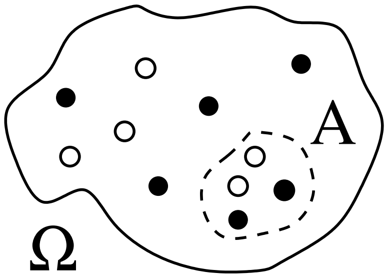
It is equivalent to ***changing the sample space from Ω to A.***

The notation $\mathbb{P}(W \,|\, A)$ is like saying,
'$\mathbb{P}(W)$ when my symbol $\mathbb{P}$ is defined relative to A instead of to Ω.'

Formula for conditional probability

Suppose we have several white balls in Ω, instead of just one. As before, we pick a ball at random and event W is the event that we select a white ball.

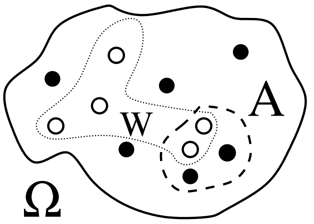
Question: What is $\mathbb{P}(W)$?



Answer: $\mathbb{P}(W)$ refers to the probability within the whole sample space Ω, so $\mathbb{P}(W) = \frac{5}{11}$.

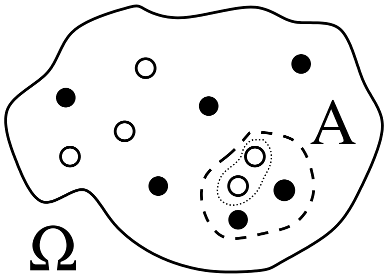
Question: What is $\mathbb{P}(W \,|\, A)$?

Answer: $\mathbb{P}(W \,|\, A)$ refers to the probability within bag A only, so $\mathbb{P}(W \,|\, A) = \frac{2}{4}$.



Question: Can you see why $\mathbb{P}(W \,|\, A) = \dfrac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)}$?

It is obvious from the diagram that $\mathbb{P}(W \mid A) = \frac{2}{4}$.



The probability of W, when selecting from bag A only, is the probability contained in the small dotted bag as a fraction of the probability in the dashed bag.

The small dotted bag represents the set $W \cap A$.

The dashed bag represents the set A.

Thus, the probability of W when selecting from within A is:

$$\mathbb{P}(W \mid A) = \frac{\textit{probability in the dotted bag}}{\textit{probability in the dashed bag}} = \frac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)}\,.$$

This reasoning gives us our formal definition of conditional probability.

Definition: Let A and B be two events on a sample space Ω. The **conditional probability of event B, given event A**, is written $\mathbb{P}(B \mid A)$, and defined as

$$\boxed{\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}\,.}$$

Read $\mathbb{P}(B \mid A)$ as *"probability of B, given A"*, or *"probability of B within A"*.

Note: $\mathbb{P}(B \mid A)$ *gives* $\mathbb{P}(B$ *and* A *, from within the set of A's only*$)$.

$\mathbb{P}(B \cap A)$ *gives* $\mathbb{P}(B$ *and* A *, from the whole sample space* Ω*)*.

Follow this reasoning carefully. It is important to understand why conditional probability is the probability of the intersection within the new sample space.

Conditioning on event A means ***changing the sample space*** to A.

Think of $\mathbb{P}(B \mid A)$ as the chance of getting a B, from the set of A's only.

The notation $\mathbb{P}(B \mid A)$ is good because it emphasises that the ***denominator*** of the proportion is A. In a sense, $\mathbb{P}(B \mid A)$ is asking for event B as a ***fraction*** of event A.

Language of conditional probability

Conditional probability corresponds to ***changing the sample space.*** This means it affects *the set we are picking FROM, when we calculate the probability that its members satisfy a certain event.*

Suppose we are picking a person at random from this class (Ω). Event A is that the person has dark hair, and event B is that the person has blue eyes.

- $\mathbb{P}(B)$ means we want the probability of picking someone who satisfies B *when they are picked from the whole sample space,* Ω*.*
- $\mathbb{P}(B \mid A)$ means the probability of picking someone who satisfies B *when they are picked only from set* A *(dark hair people): it is the probability of* B *WITHIN* A*.*
- $\mathbb{P}(B \cap A)$ means the probability of picking someone who satisfies *BOTH* B *AND* A*, when they are picked from the whole sample space,* Ω*.*

This means you can easily identify which probabilities are conditional and which are intersections by looking to see *who we are picking FROM.* Recall:

$\Omega = \{\text{people in this class}\}$; $A = \{\text{dark-haired people}\}$; $B = \{\text{blue-eyed people}\}$

Define also a random variable $X =$ number of GenEd papers a person has passed. At the University of Auckland, most students have to complete two GenEd (General Education) papers as part of their undergraduate degree. The GenEd papers can be completed at any time during the degree. Nearly everyone in this class will satisfy one of the events $X = 0$, $X = 1$, or $X = 2$.

Define further events: $F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students, e.g. exchange students, COPs, } \ldots\}$.

Exercise: Translate the following statements into probability notation. Assume in all cases we are picking a person at random from this class.

- Probability a person has dark hair and blue eyes: $\mathbb{P}(A \cap B)$.
- Probability a dark-haired person has blue eyes: $\mathbb{P}(B \mid A)$.
- Probability a person has passed two GenEd papers: $\mathbb{P}(X = 2)$.
- Probability a second-year has passed two GenEd papers: $\mathbb{P}(X = 2 \mid S)$.
- Probability a first-year has passed two GenEd papers: $\mathbb{P}(X = 2 \mid F)$.
- Probability a dark-haired first-year has passed one or two GenEd papers: $\mathbb{P}(X = 1 \cup X = 2 \mid F \cap A) = \mathbb{P}(X = 1 \mid F \cap A) + \mathbb{P}(X = 2 \mid F \cap A)$.

Trick for checking conditional probability calculations:

A useful trick for checking a conditional probability expression is to *replace the conditioned set by Ω, and see whether the expression is still true.*

The conditioned set is just another sample space, so probabilities $\mathbb{P}(\cdot \,|\, A)$ should behave exactly like ordinary probabilities $\mathbb{P}(\cdot)$, as long as ***all*** the probabilities are conditioned on the same event A.

Question: Is $\mathbb{P}(B \,|\, A) + \mathbb{P}(\overline{B} \,|\, A) = 1$?

Answer: *Replace A by Ω: this gives*

$$\mathbb{P}(B \,|\, \Omega) + \mathbb{P}(\overline{B} \,|\, \Omega) = \mathbb{P}(B) + \mathbb{P}(\overline{B}) = 1.$$

So, yes, $\mathbb{P}(B \,|\, A) + \mathbb{P}(\overline{B} \,|\, A) = 1$ for any other sample space A too.

Question: Is $\mathbb{P}(B \,|\, A) + \mathbb{P}(B \,|\, \overline{A}) = 1$?

Answer: *Try to replace the conditioning set by Ω: we can't! There are two conditioning sets: A and $\overline{A}$.*

The expression is NOT true. It doesn't make sense to try to add together probabilities from two different sample spaces.

The Multiplication Rule

For any events A and B, $\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B)\mathbb{P}(B) = \mathbb{P}(B \,|\, A)\mathbb{P}(A).}$

Proof: *Immediate from the definitions:*

$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \Rightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B)\mathbb{P}(B)\,,$$

$$\textit{and } \mathbb{P}(B \,|\, A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \quad \Rightarrow \quad \mathbb{P}(B \cap A) = \mathbb{P}(A \cap B) = \mathbb{P}(B \,|\, A)\mathbb{P}(A). \qquad \square$$

1.4 Statistical independence

Events A and B are said to be ***independent*** if they *have no influence on each other.*

For example, in the previous section, would you expect the following pairs of events to be statistically independent?

$\Omega = \{\text{people in this class}\}$; $A = \{\text{dark-haired people}\}$; $B = \{\text{blue-eyed people}\}$
$F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students}\}$;
and random variable X = number of GenEd papers passed.

- A and B? *Probably not: dark-haired people in this class might be more likely to have brown eyes, so less likely to have blue eyes?*
- A and F? *Yes, these are probably independent.*
- F and S? *Definitely not independent. No-one can be in BOTH first year AND second year, so each event STOPS the other one from happening. This is a very strong dependence.*
- Events $X = 2$ and F? *Probably not independent: first-years are less likely to have passed two GenEd papers than second-years.*

To give a formal definition of statistical independence, we need a notion of what it means for two events to have ***no influence*** on each other:

- A has no influence on B if $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.
- B has no influence on A if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.
- So A and B have ***no influence on each other if both*** $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ ***and*** $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

However, it is untidy to have a definition with two statements to check. It would be better to have a definition with just one statement.

Using the multiplication rule:

- If $\mathbb{P}(B \mid A) = \mathbb{P}(B)$, then $\mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A) = \mathbb{P}(B)\mathbb{P}(A)$.
- If $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, then $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B)$.

So both statements imply that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. What about the other way around? Suppose that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. What does that imply about $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$?

- If $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then
$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A).$$
- Similarly, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then
$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).$$

So the ***single*** statement $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, implies ***both*** statements $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ ***and*** $\mathbb{P}(B \mid A) = \mathbb{P}(A)$. Likewise, either of these two statements implies $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We can therefore use this single statement as our definition of statistical independence.

Definition: Events A and B are **statistically independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition: If there are more than two events, we say events $A_1, A_2, \ldots, A_n$ are **mutually independent** if
$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\ldots\mathbb{P}(A_n), \quad \text{AND}$$
the same multiplication rule holds for every subcollection of the events too.

Independence for random variables

Random variables are independent if *they have no influence on each other.* That is, random variables X and Y are independent if, whatever the outcome of X, it has no influence on the outcome of Y.

Definition: Random variables X and Y are **statistically independent** if
$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$
for all *possible values* x *and* y.

We usually replace the cumbersome notation $\mathbb{P}(\{X = x\} \cap \{Y = y\})$ by the simpler notation $\mathbb{P}(X = x,\ Y = y)$.

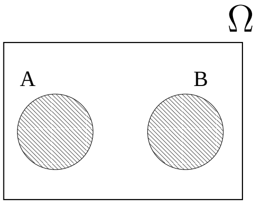
From now on, we will use the following notations interchangeably:
$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x \textit{ AND } Y = y) = \mathbb{P}(X = x, Y = y).$$

Thus X *and* Y *are independent if and only if*
$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \textit{for ALL possible values } x,\ y.$$

Independence in pictures

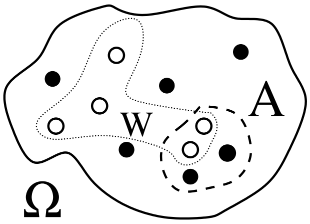
It is very difficult to draw a picture of statistical independence.

Are events A and B statistically independent?



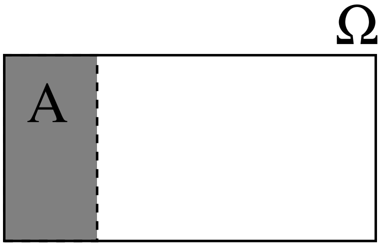
No, they are NOT independent.
Events A and B can't happen together.
They STOP each other from happening.
This is STRONG dependence — high influence.

Are events W and A statistically independent?



No, they are NOT independent.
$\mathbb{P}(W \mid A) = 2/4$, *but* $\mathbb{P}(W) = 5/11$.
So $\mathbb{P}(W \mid A) \neq \mathbb{P}(W)$, *so they are NOT independent.*

Question: How ***would*** you convey independence between events A and B on a diagram? Where would you draw event B?



[Hint: think of the formula $\mathbb{P}(B \mid A) = \mathbb{P}(B)$,
and what this means if we represent probabilities by ***areas.***]

1.5 Bayes' Theorem

Rev. Thomas Bayes (1702–1761), English clergyman and founder of Bayesian Statistics.

Bayes' Theorem follows directly from the multiplication rule. It shows how to invert the conditioning in conditional probabilities, i.e. how to express $\mathbb{P}(B \mid A)$ in terms of $\mathbb{P}(A \mid B)$.

Consider $\mathbb{P}(B \cap A) = \mathbb{P}(A \cap B)$.

Apply the multiplication rule to each side:

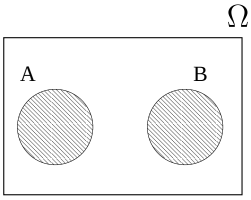
$$\mathbb{P}(B \mid A)\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B).$$

Thus
$$\boxed{\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.}$$

1.6 The Partition Theorem (Law of Total Probability)

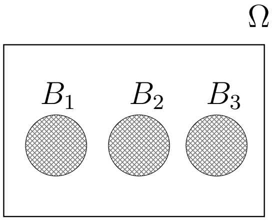
Definition: Events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$.

This means events A and B cannot happen together. If A happens, it excludes B from happening, and vice-versa.



If A and B are mutually exclusive, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
For all other A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Definition: Any number of events $B_1, B_2, \ldots, B_k$ are **mutually exclusive** if every pair of the events is mutually exclusive: ie. $B_i \cap B_j = \emptyset$ *for all* i, j *with* $i \neq j$.



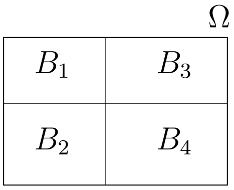
Definition: A **partition** of Ω is a *collection of mutually exclusive events whose union is* Ω.

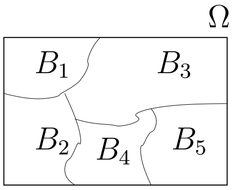
That is, sets $B_1, B_2, \ldots, B_k$ form a partition of Ω if

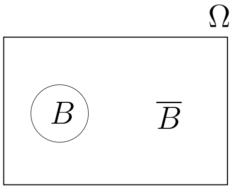
$$\begin{aligned} B_i \cap B_j &= \emptyset \text{ for all } i, j \text{ with } i \neq j, \\ \underline{\textbf{and}} \quad \bigcup_{i=1}^{k} B_i &= B_1 \cup B_2 \cup \ldots \cup B_k = \Omega. \end{aligned}$$

> $B_1, \ldots, B_k$ form a partition of Ω if they *have no overlap and collectively cover all possible outcomes.*

Examples:

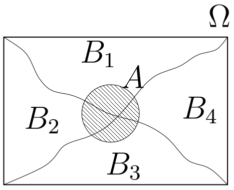






Partitioning an event *A*

Any set A can be partitioned: it doesn't have to be Ω.
In particular, if $B_1, \ldots, B_k$ form a partition of Ω, then $(A \cap B_1), \ldots, (A \cap B_k)$ form a partition of A.



Theorem 1.6: The Partition Theorem (Law of Total Probability)

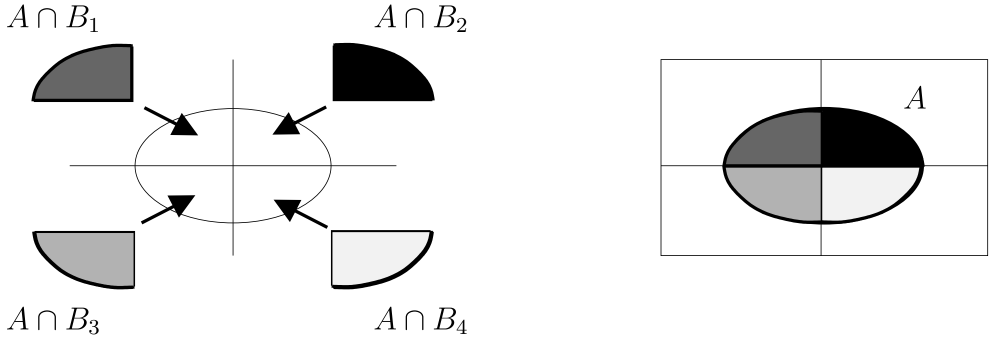
Let $B_1, \ldots, B_m$ form a partition of Ω. Then for any event A,

$$\boxed{\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}$$

Both formulations of the Partition Theorem are very widely used, but especially the conditional formulation $\sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$.

The Partition Theorem in pictures

The Partition Theorem is easy to understand because it simply states that “the whole is the sum of its parts.”



$$\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \mathbb{P}(A \cap B_3) + \mathbb{P}(A \cap B_4).$$

So:

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\mathbb{P}(B_1) + \mathbb{P}(A \mid B_2)\mathbb{P}(B_2) + \mathbb{P}(A \mid B_3)\mathbb{P}(B_3) + \mathbb{P}(A \mid B_4)\mathbb{P}(B_4).$$

Examples of conditional probability and partitions

Tom gets the bus to campus every day. The bus is on time with probability 0.6, and late with probability 0.4.

The sample space can be written as $\Omega = \{\text{bus journeys}\}$. We can formulate events as follows:

$$T = \{\text{on time}\} \qquad\qquad L = \{\text{late}\}$$

From the information given, the events have probabilities:

$$\mathbb{P}(T) = 0.6\,; \qquad\qquad \mathbb{P}(L) = 0.4.$$

(a) Do the events T and L form a partition of the sample space Ω? Explain why or why not.

Yes: they cover all possible journeys (probabilities sum to 1), and there is no overlap in the events by definition.

The buses are sometimes crowded and sometimes noisy, both of which are problems for Tom as he likes to use the bus journeys to do his Stats assignments. When the bus is on time, it is crowded with probability 0.5. When it is late, it is crowded with probability 0.7. The bus is noisy with probability 0.8 when it is crowded, and with probability 0.4 when it is not crowded.

(b) Formulate events C and N corresponding to the bus being crowded and noisy. Do the events C and N form a partition of the sample space? Explain why or why not.

Let $C = \{$ *crowded* $\}$, $N = \{$ *noisy* $\}$.
C *and* N *do NOT form a partition of* Ω. *It is possible for the bus to be noisy when it is crowded, so there must be some overlap between* C *and* N.

(c) Write down probability statements corresponding to the information given above. Your answer should involve two statements linking C with T and L, and two statements linking N with C.

$$\mathbb{P}(C \mid T) = 0.5; \qquad \mathbb{P}(C \mid L) = 0.7.$$
$$\mathbb{P}(N \mid C) = 0.8; \qquad \mathbb{P}(N \mid \overline{C}) = 0.4.$$

(d) Find the probability that the bus is crowded.

$$\begin{aligned} \mathbb{P}(C) &= \mathbb{P}(C \mid T)\mathbb{P}(T) + \mathbb{P}(C \mid L)\mathbb{P}(L) \qquad \textit{(Partition Theorem)} \\ &= 0.5 \times 0.6 + 0.7 \times 0.4 \\ &= 0.58. \end{aligned}$$

(e) Find the probability that the bus is noisy.

$$\begin{aligned} \mathbb{P}(N) &= \mathbb{P}(N \mid C)\mathbb{P}(C) + \mathbb{P}(N \mid \overline{C})\mathbb{P}(\overline{C}) \qquad \textit{(Partition Theorem)} \\ &= 0.8 \times 0.58 + 0.4 \times (1 - 0.58) \\ &= 0.632. \end{aligned}$$

1.7 Extra practice and reference

The following sections include some extra reading and examples taken from the old Stats 210 notes (pre-2015) before this material became a prerequisite for taking the course.

1. Probability of a union

The union operator, $A \cup B$, means A ***OR*** B ***OR both.*** For any events A and B on a sample space Ω:

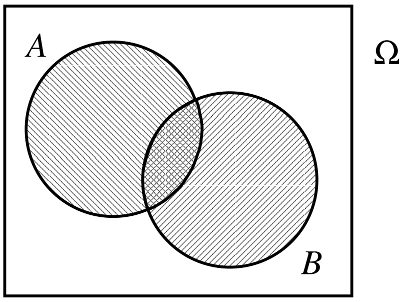
$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).}$$

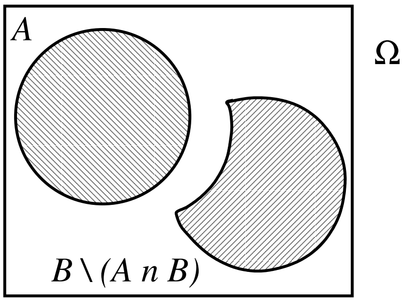
For three or more events: e.g. for any events A, B, and C on Ω:

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

Explanation

To understand the formula, think of the Venn diagrams:



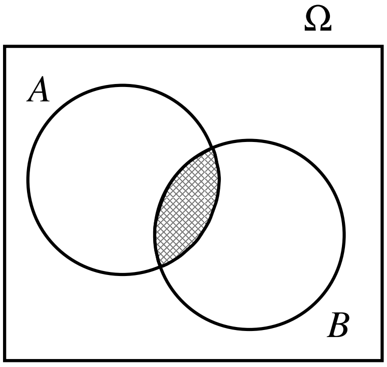


When we add $\mathbb{P}(A) + \mathbb{P}(B)$, we *add the intersection twice.*

So we have to *subtract the intersection once to get* $\mathbb{P}(A \cup B)$*:* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Alternatively, think of $A \cup B$ as *two disjoint sets: all of* A*, and the bits of* B *without the intersection. So* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \Big\{ \mathbb{P}(B) - \mathbb{P}(A \cap B) \Big\}$.

2. Probability of an intersection

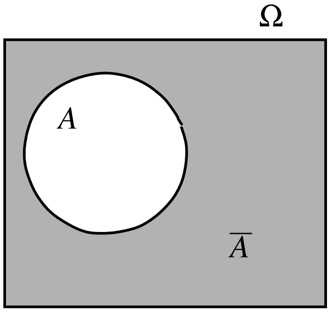


The intersection operator, $A \cap B$, means *both A AND B together.*
There is no easy formula for $\mathbb{P}(A \cap B)$.

We might be able to use *statistical independence:* ***if*** A and B are independent, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

If A and B are not statistically independent, we usually use *conditional probability:* $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$ *for any events A and B. It is usually easier to find a conditional probability than an intersection.*

3. Probability of a complement



The complement of A is written $\overline{A}$ and denotes *everything in Ω that is not in A.*

Clearly, $$\boxed{\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A).}$$

Examples of basic probability calculations

An Australian survey asked people what sort of car they would like if they could choose any car at all. 13% of respondents had children and chose a large car. 12% of respondents did not have children and chose a large car. 33% of respondents had children.

Find the probability that a respondent:
(a) chose a large car;
(b) either had children or chose a large car (or both).

First define the sample space: $\Omega = \{$ *respondents* $\}$. *Formulate events:*

Let $C = \{$ *has children* $\}$ $\quad \overline{C} = \{$ *no children* $\}$

$L = \{$ *chooses large car* $\}$.

Next write down *__all__* *the information given:*

$$\begin{aligned}\mathbb{P}(C) &= 0.33\\ \mathbb{P}(C \cap L) &= 0.13\\ \mathbb{P}(\overline{C} \cap L) &= 0.12.\end{aligned}$$

(a) Asked for $\mathbb{P}(L)$*.*

$$\begin{aligned}\mathbb{P}(L) &= \mathbb{P}(L \cap C) + \mathbb{P}(L \cap \overline{C}) \qquad \textit{(Partition Theorem)}\\ &= \mathbb{P}(C \cap L) + \mathbb{P}(\overline{C} \cap L)\\ &= 0.13 + 0.12\\ &= 0.25. \qquad \mathbb{P}\textit{(chooses large car)} = 0.25.\end{aligned}$$

(b) Asked for $\mathbb{P}(L \cup C)$*.*

$$\begin{aligned}\mathbb{P}(L \cup C) &= \mathbb{P}(L) + \mathbb{P}(C) - \mathbb{P}(L \cap C) \qquad \textit{(formula for probability of a union)}\\ &= 0.25 + 0.33 - 0.13\\ &= 0.45.\end{aligned}$$

Example 2: Facebook statistics for New Zealand university students aged between 18 and 24 suggest that 22% are interested in music, while 34% are interested in sport. Define the sample space $\Omega = \{\text{NZ university students aged 18 to 24}\}$. Formulate events: $M = \{\text{interested in music}\}$, $S = \{\text{interested in sport}\}$.

(a) What is $\mathbb{P}(\overline{M})$?

(b) What is $\mathbb{P}(M \cap S)$?

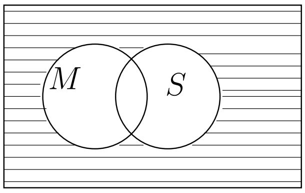
Information given: $\mathbb{P}(M) = 0.22$ $\quad \mathbb{P}(S) = 0.34$.

(a)

$$\begin{aligned}\mathbb{P}(\overline{M}) &= 1 - \mathbb{P}(M)\\ &= 1 - 0.22\\ &= 0.78.\end{aligned}$$

(b) We can not calculate $\mathbb{P}(M \cap S)$ *from the information given.*

(c) Given the further information that 48% of the students are interested in neither music nor sport, find $\mathbb{P}(M \cup S)$ and $\mathbb{P}(M \cap S)$.



Information given: $\mathbb{P}(\overline{M \cup S}) = 0.48$.

Thus

$$\begin{aligned}\mathbb{P}(M \cup S) &= 1 - \mathbb{P}(\overline{M \cup S}) \\ &= 1 - 0.48 \\ &= 0.52.\end{aligned}$$

Probability that a student is interested in music, or sport, or both.

$$\begin{aligned}\mathbb{P}(M \cap S) &= \mathbb{P}(M) + \mathbb{P}(S) - \mathbb{P}(M \cup S) \qquad \textit{(probability of a union)} \\ &= 0.22 + 0.34 - 0.52 \\ &= 0.04.\end{aligned}$$

Only 4% of students are interested in both *music and sport.*

(d) Find the probability that a student is interested in music, but *not* sport.

$$\begin{aligned}\mathbb{P}(M \cap \overline{S}) &= \mathbb{P}(M) - \mathbb{P}(M \cap S) \qquad \textit{(Partition Theorem)} \\ &= 0.22 - 0.04 \\ &= 0.18.\end{aligned}$$

1.8 Probability Reference List

The following properties hold for all events A, B, and C on a sample space Ω.

- $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$. *$\emptyset$ is the 'empty set': the event with no outcomes.*
- $0 \le \mathbb{P}(A) \le 1$: *probabilities are always between 0 and 1.*
- **Complement:** $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.
- **Probability of a union:** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

  For three events A, B, C:

  $$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)\,.$$

  If A and B are **mutually exclusive**, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

- **Conditional probability:** $\mathbb{P}(A \,|\, B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.
- **Multiplication rule:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \,|\, B)\mathbb{P}(B) = \mathbb{P}(B \,|\, A)\mathbb{P}(A)$.
- **The Partition Theorem:** if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

  $$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i) \quad \text{for any event } A.$$

  As a special case, B and $\overline{B}$ partition Ω, so:

  $$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B}) \\ &= \mathbb{P}(A \,|\, B)\mathbb{P}(B) + \mathbb{P}(A \,|\, \overline{B})\mathbb{P}(\overline{B}) \quad \text{for any } A,\ B. \end{aligned}$$

- **Bayes' Theorem:** $\mathbb{P}(B \,|\, A) = \dfrac{\mathbb{P}(A \,|\, B)\mathbb{P}(B)}{\mathbb{P}(A)}$.

  More generally, if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

  $$\mathbb{P}(B_j \,|\, A) = \frac{\mathbb{P}(A \,|\, B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{m} \mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)} \quad \text{for any } j.$$

- **Chains of events:** for any events A_1, A_2, A_3,

  $$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_2 \cap A_1)\,.$$

- **Statistical independence:** events A and B are **independent** if and only if

  $$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\,.$$

  Alternatively, either of the following statements is necessary and sufficient for A and B to be independent: $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$ and $\mathbb{P}(B \,|\, A) = \mathbb{P}(B)$.

- **Manipulating conditional probabilities:**

  If $\mathbb{P}(B) > 0$, then we can treat $\mathbb{P}(\cdot \,|\, B)$ just like $\mathbb{P}$: for example,

  ★ if A_1 and A_2 are mutually exclusive, then

  $$\begin{aligned} \mathbb{P}(A_1 \cup A_2 \,|\, B) &= \mathbb{P}(A_1 \,|\, B) + \mathbb{P}(A_2 \,|\, B) \\ \text{compare with the usual formula,} \quad \mathbb{P}(A_1 \cup A_2) &= \mathbb{P}(A_1) + \mathbb{P}(A_2). \end{aligned}$$

  ★ if $A_1, \ldots, A_m$ partition the sample space Ω, then

  $$\mathbb{P}(A_1 \,|\, B) + \mathbb{P}(A_2 \,|\, B) + \ldots + \mathbb{P}(A_m \,|\, B) = 1;$$

  ★ $\mathbb{P}(A \,|\, B) = 1 - \mathbb{P}(\overline{A} \,|\, B)$ for any A.

Note: it is ***not*** generally true that $\mathbb{P}(A \,|\, B) = 1 - \mathbb{P}(A \,|\, \overline{B})$.

Chapter 2: Foundations of Statistical Inference

2.1 Introduction

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. The word ***inference*** means 'conclusions' or 'decisions'. Statistical inference is about drawing conclusions and making decisions based on observed data.

Data, or observations, typically arise from some *underlying process.* It is the underlying process we are interested in, not the observations themselves. Sometimes we call the underlying process the *population* or *mechanism* of interest.

The data are only a *sample* from this population or mechanism. We cannot possibly observe every outcome of the process, so we have to make do with the sample that we have observed.

The data give us *imperfect insight* into the population of interest. The role of statistical inference is *to use this imperfect data to draw conclusions about the population of interest, while simultaneously giving an honest reflection of the uncertainty in our conclusions.*

Example 1: Tossing a coin.

- ***Population:*** *all possible tosses of this coin.*
- ***Sample:*** *a small number of observed tosses, e.g. 10 observed tosses.*
- ***What do we want to make inference about?*** We might be interested in the probability of getting a Head on each toss. In particular, we might be interested in whether the coin is fair ($\mathbb{P}(\text{Head}) = 0.5$) or has been fiddled.

Example 2: Political polling: how many people will vote for the NZ Labour Party?

- ***Population:*** *all eligible voters in New Zealand.*
- ***Sample:*** *a random sample of voters, e.g. 1000.*
- ***What do we want to make inference about?*** We want to know the support for Labour among ***all*** voters, but this is too expensive to carry out except on election-night itself. Instead we aim to ***deduce*** the support for Labour by asking a smaller number of voters, while simultaneously reporting upon our uncertainty (margin of error).

In the next two chapters we meet several important concepts in statistical inference. We will illustrate them with ***discrete random variables***, then introduce ***continuous random variables*** in Chapter 4 and show how the same ideas still apply.

1. **Hypothesis testing:**

   - I toss a coin ten times and get nine heads. How unlikely is that? Can we continue to believe that the coin is ***fair*** when it produces nine heads out of ten tosses?

2. **Likelihood and estimation:**

   - Suppose we know that our random variable is (say) Binomial$(10, p)$, for some p, but we don't know the value of p. We will see how to ***estimate*** the value of p using maximum likelihood estimation.

3. **Expectation and variance of a random variable:**

   - The ***expectation*** of a random variable is the value it takes ***on average.***
   - The ***variance*** of a random variable measures how much the random variable ***varies about its average.***

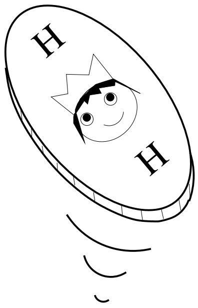
   These are used to report how accurate and reliable our ***estimation procedure*** is. Does it give the right answer ***on average?*** How much does it ***vary*** about its average?

4. **Modelling:**

   - We have a situation in real life that we know is random. But what does the randomness ***look*** like? Is it highly variable, or little variability? Does it sometimes give results much *higher* than average, but never give results much *lower* (long-tailed distribution)? We will see how different probability distributions are suitable for different circumstances. Choosing a probability distribution to fit a situation is called ***modelling.***

2.2 Hypothesis testing

You have probably come across the idea of hypothesis tests, p-values, and significance in other courses. Common hypothesis tests include t-tests and chi-squared tests. However, hypothesis tests can be conducted in much simpler circumstances than these. The concept of the hypothesis test is at its easiest to understand with the Binomial distribution in the following example. All other hypothesis tests throughout statistics are based on the same idea.



Example: **Weird Coin?**

I toss a coin 10 times and get 9 heads. How weird is that?

What is 'weird'?

- Getting 9 heads out of 10 tosses: we'll call this *weird.*
- Getting 10 heads out of 10 tosses: *even more weird!*
- Getting 8 heads out of 10 tosses: *less weird.*
- Getting 1 head out of 10 tosses: *same as getting 9 tails out of 10 tosses: just as weird as 9 heads if the coin is fair.*
- Getting 0 heads out of 10 tosses: *same as getting 10 tails: more weird than 9 heads if the coin is fair.*

Set of weird outcomes

If our coin is fair, the outcomes that are ***as weird or weirder*** than 9 heads are:

9 heads, 10 heads, 1 head, 0 heads.

So how weird is 9 heads or worse, if the coin is fair?

Define X =*#heads out of 10 tosses.*

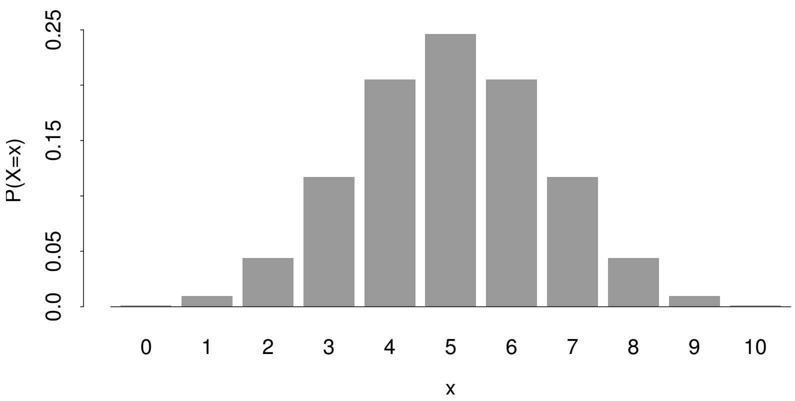
Distribution of X, if the coin is fair: $X \sim$ *Binomial*$(n = 10, p = 0.5)$.

Probability of observing something at least as weird as 9 heads, if the coin is fair:

We can add the probabilities of all the outcomes that are ***at least as weird*** as 9 heads out of 10 tosses, assuming that the coin is fair.

$$\mathbb{P}(X=9)+\mathbb{P}(X=10)+\mathbb{P}(X=1)+\mathbb{P}(X=0) \quad \textit{where} \quad X \sim \textit{Binomial}(10, 0.5).$$

Probabilities for Binomial($n = 10$, $p = 0.5$)



For $X \sim \text{Binomial}(10, 0.5)$, we have:

$$\mathbb{P}(X=9)+\mathbb{P}(X=10)+\mathbb{P}(X=1)+\mathbb{P}(X=0) =$$

$$\begin{aligned} & \binom{10}{9}(0.5)^9(0.5)^1 + \binom{10}{10}(0.5)^{10}(0.5)^0 + \\ & \binom{10}{1}(0.5)^1(0.5)^9 + \binom{10}{0}(0.5)^0(0.5)^{10} \\ = \;& 0.00977 + 0.00098 + 0.00977 + 0.00098 \\ = \;& 0.021. \end{aligned}$$

Is this weird?

Yes, it is quite weird. If we had a fair coin and tossed it 10 times, we would only expect to see something as extreme as 9 heads on about *2.1% of occasions.*

Is the coin fair?

Obviously, we can't say. It might be: after all, on 2.1% of occasions that you toss a fair coin 10 times, you do get something as weird as 9 heads or more.

However, 2.1% is a small probability, so it is still very unusual for a fair coin to produce something as weird as what we've seen. If the coin really was fair, it would be very unusual to get 9 heads or more.

We can deduce that, *EITHER we have observed a very unusual event with a fair coin, OR the coin is not fair.*

In fact, this gives us *some evidence that the coin is not fair.*

The value 2.1% *measures the strength of our evidence. The smaller this probability, the more evidence we have.*

Formal hypothesis test

We now formalize the procedure above. Think of the steps:

- We have a question that we want to answer: *Is the coin fair?*

- There are two alternatives:
  1. The coin is fair.
  2. The coin is not fair.

- Our observed information is X, the number of heads out of 10 tosses. We write down the distribution of X *if the coin is fair:*
  $X \sim$ *Binomial*$(10, 0.5)$.

- We calculate the probability of observing something *AT LEAST AS EXTREME as our observation,* $X = 9$, *if the coin is fair: prob=0.021.*

- The probability is small (2.1%). We conclude that this is unlikely with a fair coin, so *we have observed some evidence that the coin is NOT fair.*

Null hypothesis and alternative hypothesis

We express the steps above as two competing hypotheses.

Null hypothesis: *the first alternative, that the coin IS fair.*

We expect to believe the null hypothesis unless we see convincing evidence that it is wrong.

Alternative hypothesis: *the second alternative, that the coin is NOT fair.*

In hypothesis testing, we often use this same formulation.

- The null hypothesis is *specific.*

  It specifies an exact distribution for our observation: $X \sim \textit{Binomial}(10, 0.5)$.

- The alternative hypothesis is *general.*

  It simply states that the null hypothesis is wrong. It does not say what the *right* answer is.

We use H_0 *and* H_1 to denote the null and alternative hypotheses respectively.

The null hypothesis is H_0 : *the coin is fair.*
The alternative hypothesis is H_1 : *the coin is NOT fair.*

To set up the test, we write:

$$\textit{Number of heads, } X \sim \textit{Binomial}(10, p),$$

and

$$\begin{aligned} H_0 &: \; p = 0.5 \\ H_1 &: \; p \neq 0.5. \end{aligned}$$

Think of 'null hypothesis' as meaning the 'default': the hypothesis we will accept unless we have a good reason not to.

p-values

In the hypothesis-testing framework above, we always *measure evidence AGAINST the null hypothesis.*

That is, we believe that our coin is fair unless we see convincing evidence otherwise.

We measure the strength of evidence against H_0 using the *p-value.*

In the example above, the p-value was $p = 0.021$.

A p-value of 0.021 represents *quite strong evidence against the null hypothesis.*

It states that, if the null hypothesis is TRUE, we would only have *a 2.1% chance of observing something as extreme as 9 heads or tails.*

Some people might even see this as strong enough evidence to decide that the null hypothesis is not true, but this is generally an over-simplistic interpretation.

In general, the p-value is *the probability of observing something AT LEAST AS EXTREME AS OUR OBSERVATION, if H_0 is TRUE.*

This means that *SMALL p-values represent STRONG evidence against H_0.*

Small p-values mean Strong evidence.
Large p-values mean **Little** evidence.

Note: Be careful not to confuse the term p-value, which is 0.021 in our example, with the Binomial probability p. Our hypothesis test is designed to test whether the Binomial probability is $p = 0.5$. To test this, we calculate the p-value of 0.021 as a measure of the strength of evidence ***against*** the hypothesis that $p = 0.5$.

Interpreting the hypothesis test

There are different schools of thought about how a p-value should be interpreted.

- Most people agree that the p-value is a useful measure of the ***strength of evidence against the null hypothesis***. The smaller the p-value, the stronger the evidence against H_0.

- Some people go further and use an ***accept/reject framework.*** Under this framework, the null hypothesis H_0 should be *rejected* if the p-value is less than 0.05 (say), and *accepted* if the p-value is greater than 0.05.

- In this course we use the ***strength of evidence*** interpretation. The p-value measures how far out our observation lies in the tails of the distribution specified by H_0. We do not talk about accepting or rejecting H_0. This decision should usually be taken in the context of other scientific information.

  However, as a rule of thumb, we consider that p-values of 0.05 and less start to suggest that the null hypothesis is doubtful.

Statistical significance

You have probably encountered the idea of ***statistical significance*** in other courses.

Statistical significance refers to the p-value.

The result of a hypothesis test is ***significant at the 5% level*** if the p-value is *less than 0.05.*

This means that *the chance of seeing what we did see (9 heads), or more, is less than 5% if the null hypothesis is true.*

Saying the test is ***significant*** is a quick way of saying that there is evidence against the null hypothesis, usually at the 5% level.

In the coin example, we can say that our test of $H_0 : p = 0.5$ against $H_1 : p \neq 0.5$ *is significant at the 5% level, because the p-value is 0.021 which is* < 0.05.

This means:

- *we have some evidence that* $p \neq 0.5$.

It does ***not*** mean:

- the difference between p and 0.5 is *large,* or
- the difference between p and 0.5 is *important in practical terms.*

Statistically significant means that *we have evidence, in OUR sample, that p is different from 0.5. It says NOTHING about the SIZE, or the IMPORTANCE, of the difference.*

"Substantial evidence of a difference", not "Evidence of a substantial difference."

Beware!

The p-value gives the *probability of seeing something as weird as what we did see, if H_0 is true.*

This means that *5% of the time, we will get a p-value* < 0.05 *WHEN H_0 IS TRUE!!*

Similarly, about once in every thousand tests, we will get a p-value < 0.001, when H_0 is true!

A small p-value does NOT mean that H_0 is definitely wrong.

One-sided and two-sided tests

The test above is a *two-sided test.* This means that we considered it *just as weird to get 9 tails as 9 heads.*

If we had a good reason, ***before*** tossing the coin, to believe that the binomial probability could ***only*** be $= 0.5$ or > 0.5, i.e. that it would be ***impossible*** to have $p < 0.5$, then we could conduct a one-sided test: $H_0 : p = 0.5$ *versus* $H_1 : p > 0.5$.

This would have the effect of halving the resultant p-value.

2.3 Example: Presidents and deep-sea divers

Men in the class: would you like to have daughters? Then become a deep-sea diver, a fighter pilot, or a heavy smoker.

Would you prefer sons? Easy!
Just become a US president.

Numbers suggest that men in different professions tend to have more sons than daughters, or the reverse. Presidents have sons, fighter pilots have daughters. But is it real, or just chance? We can use hypothesis tests to decide.

The facts

- The 44 US presidents from George Washington to Barack Obama have had a total of 153 children, comprising 88 sons and only 65 daughters: a sex ratio of 1.4 sons for every daughter.
- Two studies of deep-sea divers revealed that the men had a total of 190 children, comprising 65 sons and 125 daughters: a sex ratio of 1.9 daughters for every son.

Could this happen by chance?

Is it possible that the men in each group *really had a 50-50 chance of producing sons and daughters?*

This is the same as the question in Section 2.2.

For the presidents: *If I tossed a coin 153 times and got only 65 heads, could I continue to believe that the coin was fair?*

For the divers: If I tossed a coin *190* times and got only *65* heads, could I continue to believe that the coin was fair?

Hypothesis test for the presidents

We set up the competing hypotheses as follows.

Let X be the number of daughters out of 153 presidential children.

Then $X \sim$ Binomial$(153, p)$, where p is the probability that each child is a daughter.

Null hypothesis: $H_0 : p = 0.5.$

Alternative hypothesis: $H_1 : p \neq 0.5.$

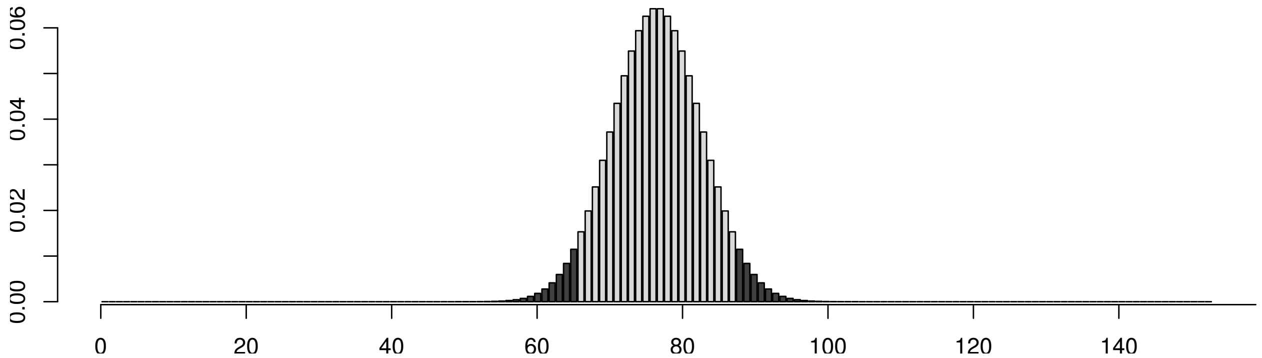
***p*-value:** *We need the probability of getting a result AT LEAST AS EXTREME as $X = 65$ daughters, if H_0 is true and p really is 0.5.*

Which results are at least as extreme as $X = 65$?

$X = 0, 1, 2, \ldots, 65$, *for even fewer daughters.*

$X = (153 - 65), \ldots, 153$, *for too many daughters, because we would be just as surprised if we saw ≤ 65 sons, i.e. $\geq (153 - 65) = 88$ daughters.*

Probabilities for $X \sim$ Binomial$(n = 153, p = 0.5)$



Calculating the p-value

The p-value for the president problem is given by

$\mathbb{P}(X \leq 65) + \mathbb{P}(X \geq 88)$ *where* $X \sim$ *Binomial*$(153, 0.5)$.

In principle, we could calculate this as

$$\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \ldots + \mathbb{P}(X = 65) + \mathbb{P}(X = 88) + \ldots + \mathbb{P}(X = 153)$$

$$= \binom{153}{0}(0.5)^0(0.5)^{153} + \binom{153}{1}(0.5)^1(0.5)^{152} + \ldots$$

This would take a lot of calculator time! Instead, we use a computer with a package such as R.

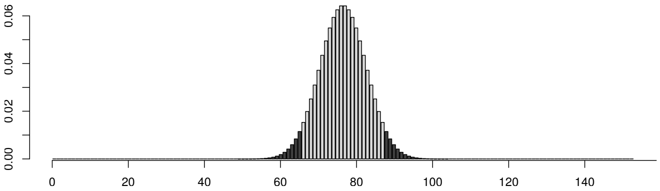
R command for the p-value

The R command for calculating the *lower-tail* p*-value for the Binomial(*$n = 153, p = 0.5$*) distribution is*

```
pbinom(65, 153, 0.5).
```

Typing this in R gives:

```
> pbinom(65, 153, 0.5)
[1] 0.03748079
```



This gives us the *lower-tail* p*-value only:*
$\mathbb{P}(X \leq 65) = 0.0375$.

To get the overall p-value:

Multiply the lower-tail p*-value by 2:*

$$2 \times 0.0375 = 0.0750.$$

In R:

```
> 2 * pbinom(65, 153, 0.5)
[1] 0.07496158
```

This works because the upper-tail p-value, by definition, is always going to be the same as the lower-tail p-value. The upper tail gives us the probability of finding something *equally surprising* at the opposite end of the distribution.

Note: The R command `pbinom` is equivalent to the *cumulative distribution function* for the Binomial distribution:

$$\begin{aligned} \texttt{pbinom(65, 153, 0.5)} &= \mathbb{P}(X \leq 65) \quad \textit{where } X \sim \textit{Binomial}(153, 0.5) \\ &= F_X(65) \qquad \textit{for } X \sim \textit{Binomial}(153, 0.5). \end{aligned}$$

The overall p-value in this example is $2 \times F_X(65)$.

Note: In the R command `pbinom(65, 153, 0.5)`, the order that you enter the numbers 65, 153, and 0.5 is important. If you enter them in a different order, you will get an error. An alternative is to use the longhand command `pbinom(q=65, size=153, prob=0.5)`, in which case you can enter the terms in any order.

Summary: are presidents more likely to have sons?

Back to our hypothesis test. Recall that X was the number of daughters out of 153 presidential children, and $X \sim \text{Binomial}(153, p)$, where p is the probability that each child is a daughter.

Null hypothesis: $H_0 : p = 0.5$.

Alternative hypothesis: $H_1 : p \neq 0.5$.

***p*-value:** $2 \times F_X(65) = 0.075$.

What does this mean?

The p-value of 0.075 means that, *if the presidents really were as likely to have daughters as sons, there would only be 7.5% chance of observing something as unusual as only 65 daughters out of the total 153 children.*

This is slightly unusual, but not very unusual.

We conclude that *there is no real evidence that presidents are more likely to have sons than daughters. The observations are compatible with the possibility that there is no difference.*

Does this mean presidents are equally likely to have sons and daughters? ***No:*** *the observations are also compatible with the possibility that there is a difference. We just don't have enough evidence either way.*

Hypothesis test for the deep-sea divers

For the deep-sea divers, there were 190 children: 65 sons, and 125 daughters.

Let X be the *number of sons out of 190 diver children.*

Then $X \sim$ *Binomial*$(190, p)$, *where* p *is the probability that each child is a son.*

Note: We could just as easily formulate our hypotheses in terms of daughters instead of sons. Because `pbinom` is defined as a lower-tail probability, however, it is usually easiest to formulate them in terms of the *low* result (sons).

Null hypothesis: $H_0 : p = 0.5$.

Alternative hypothesis: $H_1 : p \neq 0.5$.

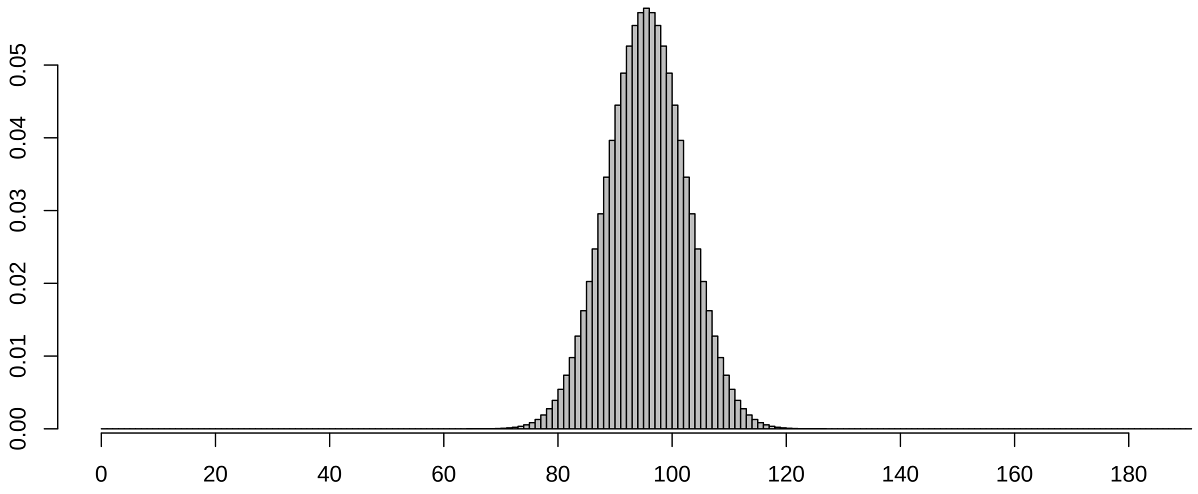
p-value: *Probability of getting a result AT LEAST AS EXTREME as* $X = 65$ *sons, if* H_0 *is true and* p *really is 0.5.*

Results at least as extreme as $X = 65$ are:

$X = 0, 1, 2, \ldots, 65$, *for even fewer sons.*

$X = (190-65), \ldots, 190$, *for the equally surprising result in the opposite direction (too many sons).*

Probabilities for $X \sim$ Binomial$(n = 190, p = 0.5)$



R command for the p-value

p-*value* = 2×`pbinom(65, 190, 0.5)`.

Typing this in R gives:

```
> 2*pbinom(65, 190, 0.5)
[1] 1.603136e-05
```

This is 0.000016, or a little more than *one chance in 100 thousand.*

We conclude that *it is extremely unlikely that this observation could have occurred by chance, if the deep-sea divers had equal probabilities of having sons and daughters.*

We have *very strong evidence that deep-sea divers are more likely to have daughters than sons. The data are not really compatible with H_0.*

What next?

p-values are often badly used in science and business. They are regularly treated as the end point of an analysis, after which no more work is needed. Many scientific journals insist that scientists quote a p-value with every set of results, and often only p-values less than 0.05 are regarded as 'interesting'. The outcome is that some scientists do every analysis they can think of until they finally come up with a p-value of 0.05 or less.

A good statistician will recommend a different attitude. ***It is very rare in science for numbers and statistics to tell us the full story.***

Results like the p-value should be regarded as an investigative *starting point*, rather than the final conclusion. *Why* is the p-value small? What possible *mechanism* could there be for producing this result?

If you were a medical statistician and you gave me a p-value, I would ask you for a mechanism.

Don't accept that Drug A is better than Drug B *only* because the p-value says so: find a biochemist who can explain what Drug A does that Drug B doesn't. Don't accept that sun exposure is a cause of skin cancer on the basis of a p-value alone: find a mechanism by which skin is damaged by the sun.

Why might divers have daughters and presidents have sons?

Deep-sea divers are thought to have more daughters than sons because the underwater work at high atmospheric pressure lowers the level of the hormone testosterone in the men's blood, which is thought to make them more likely to conceive daughters. For the presidents, your guess is as good as mine . . .

2.4 Example: Birthdays and sports professionals

Have you ever wondered what makes a professional sports player? Talent? Dedication? Good coaching?

Or is it just that they happen to have the right birthday…?

The following text is taken from Malcolm Gladwell's book *Outliers*. It describes the play-by-play for the first goal scored in the 2007 finals of the Canadian ice hockey junior league for star players aged 17 to 19. The two teams are the Tigers and Giants. There's one slight difference … instead of the players' names, we're given their birthdays.

> *March 11 starts around one side of the Tigers' net, leaving the puck for his teammate January 4, who passes it to January 22, who flips it back to March 12, who shoots point-blank at the Tigers' goalie, April 27. April 27 blocks the shot, but it's rebounded by Giants' March 6. He shoots! Tigers defensemen February 9 and February 14 dive to block the puck while January 10 looks on helplessly. March 6 scores!*

Notice anything funny?

Here are some figures. There were 25 players in the Tigers squad, born between 1986 and 1990. Out of these 25 players, 14 of them were born in January, February, or March. Is it believable that this should happen by chance, or do we have evidence that there is a birthday-effect in becoming a star ice hockey player?

Hypothesis test

Let X be the number of the 25 players who are born from January to March.
We need to set up hypotheses of the following form:

Null hypothesis: H_0 : *there is no birthday effect.*

Alternative hypothesis: H_1 : *there is a birthday effect.*

What is the distribution of X under H_0 and under H_1?

Under H_0, there is no birthday effect. So the probability that each player has a birthday in Jan to March is about $1/4$.

(3 months out of a possible 12 months).

Thus the distribution of X under H_0 is $X \sim$ *Binomial*$(25, 1/4)$.

Under H_1, there is a birthday effect, so $p \neq 1/4$.

Our formulation for the hypothesis test is therefore as follows.

Number of Jan to March players, $X \sim$ *Binomial*$(25, p)$.

Null hypothesis: $H_0 : p = 0.25$.

Alternative hypothesis: $H_1 : p \neq 0.25$.

Our observation:

The observed proportion of players born from Jan to March is $14/25 = 0.56$.

This is *more than the* 0.25 *predicted by* H_0.

Is it sufficiently greater than 0.25 to provide evidence against H_0?

Just using our intuition, we can make a guess, but we might be wrong. The answer also depends on the sample size (25 in this case). We need the p-value to measure the evidence properly.

p-value: *Probability of getting a result AT LEAST*
AS EXTREME as $X = 14$ *Jan to March players,*
if H_0 *is true and* p *really is 0.25.*

Results at least as extreme as $X = 14$ are:

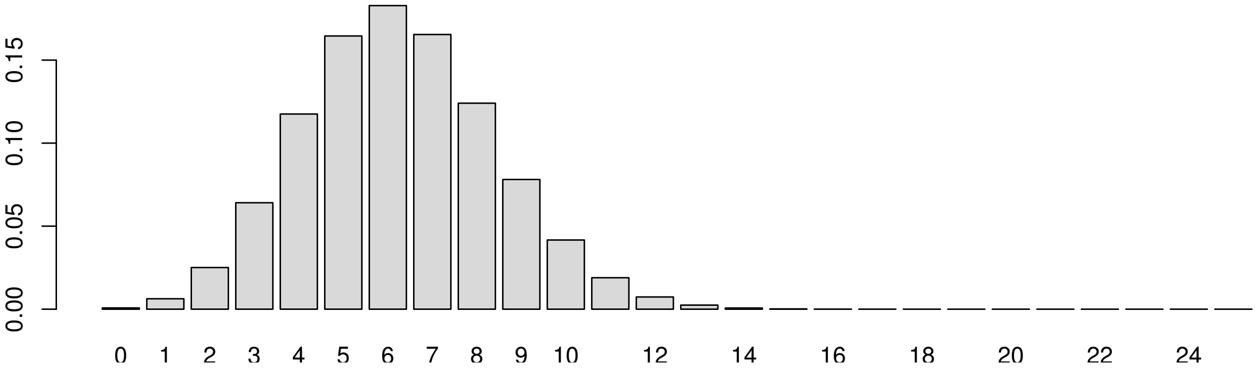
Upper tail: $X = 14, 15, \ldots, 25$, *for even more Jan to March players.*

Lower tail: an equal probability in the opposite direction, for too few Jan to March players.

Note: We do not need to calculate the values corresponding to our lower-tail p-value. It is more complicated in this example than in Section 2.3, because we do not have Binomial probability $p = 0.5$. In fact, the lower tail probability lies somewhere between 0 and 1 player, but it cannot be specified exactly.

We get round this problem for calculating the p-value by *just multiplying the upper-tail p-value by 2.*

Probabilities for $X \sim \text{Binomial}(n = 25, p = 0.25)$



R command for the p-value

We need *twice the UPPER-tail p-value:*

p-value $= 2 \times (1-$`pbinom(13, 25, 0.25)`$)$.
(Recall $\mathbb{P}(X \geq 14) = 1 - \mathbb{P}(X \leq 13)$.)

Typing this in R gives:

```
2*(1-pbinom(13, 25, 0.25))
[1] 0.001831663
```

This p-value is *very small.*

It means that *if there really was no birthday effect, we would expect to see results as unusual as 14 out of 25 Jan to March players less than 2 in 1000 times.*

We conclude that *we have strong evidence that there is a birthday effect in this ice hockey team. Something beyond ordinary chance seems to be going on. The data are barely compatible with H_0.*

Why should there be a birthday effect?

These data are just one example of a much wider - and astonishingly strong - phenomenon. Professional sports players not just in ice hockey, but in soccer, baseball, and other sports have strong birthday clustering. Why?

It's because these sports select talented players for age-class star teams at young ages, about 10 years old. In ice hockey, the cut-off date for age-class teams is January 1st. A 10-year-old born in December is competing against players who are nearly a year older, born in January, February, and March. The age difference makes a big difference in terms of size, speed, and physical coordination. Most of the 'talented' players at this age are simply older and bigger. But there then follow years in which they get the best coaching and the most practice. By the time they reach 17, these players really are the best.

2.5 Likelihood and estimation

So far, the hypothesis tests have only told us whether the Binomial probability p *might be*, or *probably isn't*, equal to the value specified in the null hypothesis. They have told us nothing about the size, or potential importance, of the departure from H_0.

For example, for the deep-sea divers, we found that *it would be very unlikely to observe as many as 125 daughters out of 190 children if the chance of having a daughter really was* $p = 0.5$.

But what does this say about the *actual* value of p?

Remember the p-value for the test was 0.000016. Do you think that:

1. p could be as big as 0.8?

   No idea! The p-value does not tell us.

2. p could be as close to 0.5 as, say, 0.51?

   The test doesn't even tell us this much!
   If there was a huge sample size (number of children), we COULD get a p-value as small as 0.000016 even if the true probability was 0.51.

Common sense, however, gives us a hint. Because there were almost twice as many daughters as sons, my guess is that the probability of a having a daughter is something close to $p = 2/3$. We need some way of formalizing this.

Estimation

The process of using observations to suggest a value for a parameter is called *estimation.*

The value suggested is called the ***estimate*** of the parameter.

In the case of the deep-sea divers, we wish to estimate the probability p that the child of a diver is a daughter. The common-sense estimate to use is

$$p = \frac{\textit{number of daughters}}{\textit{total number of children}} = \frac{125}{190} = 0.658.$$

However, there are many situations where our common sense fails us. For example, what would we do if we had a regression-model situation (see Section 3.8) and wished to specify an alternative form for p, such as

$$p = \alpha + \beta \times (\text{diver age}).$$

How would we estimate the unknown intercept α and slope β, given known information on diver age and number of daughters and sons?

We need a general framework for estimation that can be applied to any situation. The most useful and general method of obtaining parameter estimates is the method of *maximum likelihood estimation.*

Likelihood

Likelihood is one of the most important concepts in statistics.
Return to the deep-sea diver example.

X is the *number of daughters out of 190 children.*

We know that $X \sim \textit{Binomial}(190, p)$,

and we wish to estimate the value of p.

The available data is the observed value of X: $X = 125$.

Suppose for a moment that $p = 0.5$. What is the probability of observing $X = 125$?

When $X \sim$ *Binomial*$(190, 0.5)$,

$$\begin{aligned} \mathbb{P}(X = 125) &= \binom{190}{125}(0.5)^{125}(1-0.5)^{190-125} \\ &= 3.97 \times 10^{-6}. \end{aligned}$$

Not very likely!!

What about $p = 0.6$? What would be the probability of observing $X = 125$ if $p = 0.6$?

When $X \sim$ *Binomial*$(190, 0.6)$,

$$\begin{aligned} \mathbb{P}(X = 125) &= \binom{190}{125}(0.6)^{125}(1-0.6)^{190-125} \\ &= 0.016. \end{aligned}$$

This still looks quite unlikely, but it is almost 4000 times more likely than getting $X = 125$ *when* $p = 0.5$.

So far, we have discovered that *it would be thousands of times more likely to observe* $X = 125$ *if* $p = 0.6$ *than it would be if* $p = 0.5$.

This suggests that $p = 0.6$ *is a better estimate than* $p = 0.5$.

You can probably see where this is heading. If $p = 0.6$ is a better estimate than $p = 0.5$, what if we move p even closer to our common-sense estimate of 0.658?

When $X \sim$ *Binomial*$(190, 0.658)$,

$$\begin{aligned} \mathbb{P}(X = 125) &= \binom{190}{125}(0.658)^{125}(1-0.658)^{190-125} \\ &= 0.061. \end{aligned}$$

This is even more likely than for $p = 0.6$. *So* $p = 0.658$ *is the best estimate yet.*

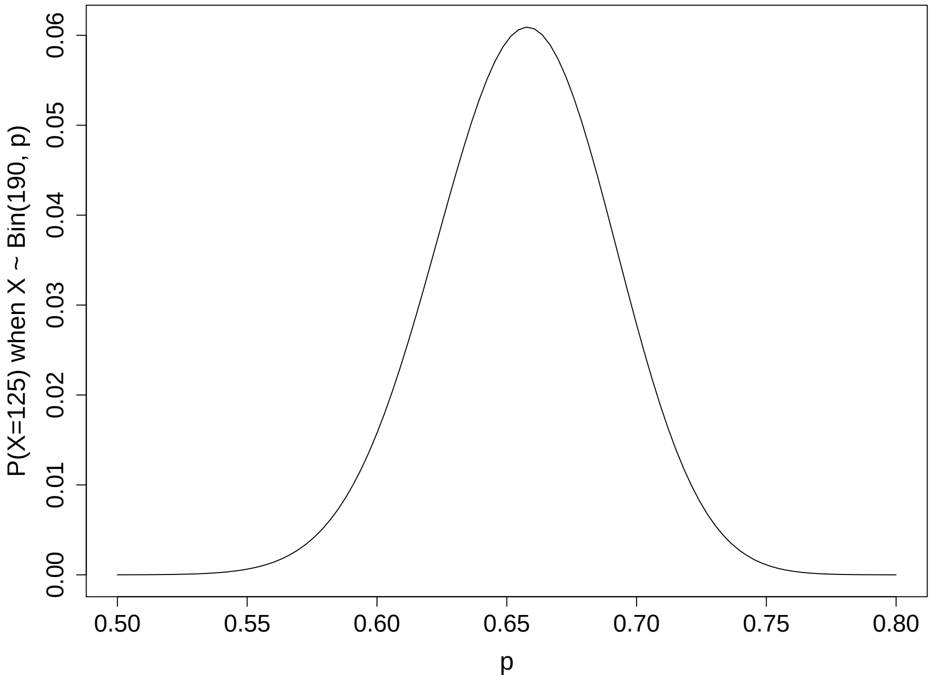
Can we do any better? What happens if we increase p a little more, say to $p = 0.7$?

When $X \sim$ *Binomial*$(190, 0.7)$,

$$\begin{aligned}\mathbb{P}(X = 125) &= \binom{190}{125}(0.7)^{125}(1-0.7)^{190-125} \\ &= 0.028.\end{aligned}$$

This has decreased from the result for $p = 0.658$, *so our observation of 125 is LESS likely under* $p = 0.7$ *than under* $p = 0.658$.

Overall, we can plot a graph showing ***how likely*** our observation of $X = 125$ is under each different value of p.



The graph reaches a *clear maximum. This is a value of* p *at which the observation* $X = 125$ *is MORE LIKELY than at any other value of* p.

This ***maximum likelihood*** value of p is our *maximum likelihood estimate.*

We can see that the maximum occurs somewhere close to our common-sense estimate of $p = 0.658$.

The likelihood function

Look at the graph we plotted overleaf:

Horizontal axis: *The unknown parameter,* p.

Vertical axis: *The probability of our observation,* $X = 125$*, under this value of* p.

This function is called the *likelihood function.*

It is a function of *the unknown parameter* p.

For our *fixed* observation $X = 125$, the likelihood function shows *how LIKELY the observation 125 is for every different value of* p.

The likelihood function is:

$$\begin{aligned} L(p) &= \mathbb{P}(X = 125) \text{ when } X \sim \textit{Binomial}(190, p), \\ &= \binom{190}{125} p^{125}(1-p)^{190-125} \\ &= \binom{190}{125} p^{125}(1-p)^{65} \quad \text{for } 0 < p < 1\,. \end{aligned}$$

This function of p is the curve shown on the graph on page 55.

In general, if our observation were $X = x$ rather than $X = 125$, the likelihood function is *a function of* p *giving* $\mathbb{P}(X = x)$ *when* $X \sim$ *Binomial*$(190, p)$.

We write:

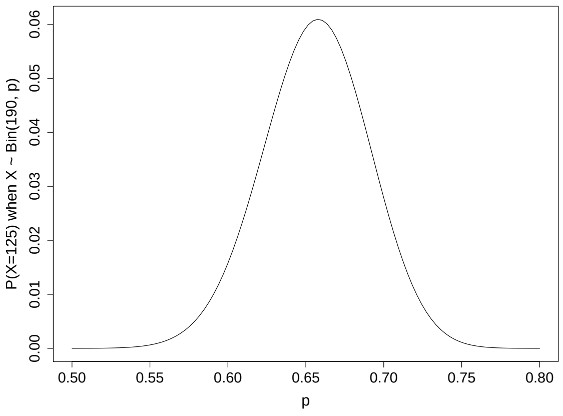
$$\begin{aligned} L(p\,; x) &= \mathbb{P}(X = x) \text{ when } X \sim \textit{Binomial}(190, p), \\ &= \binom{190}{x} p^{x}(1-p)^{190-x}\,. \end{aligned}$$

Difference between the likelihood function and the probability function

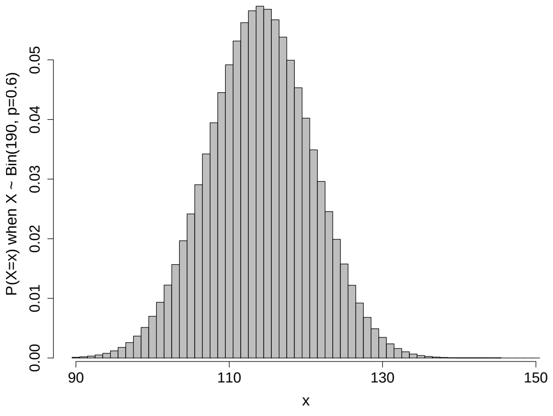
The likelihood function is *a probability of x, but it is a FUNCTION of p.*

The likelihood gives *the probability of a FIXED observation x, for every possible value of the parameter p.*

Compare this with the *probability function*, which is *the probability of every different value of x, for a FIXED value of p.*



Likelihood function, $L(p\,;x)$.
Function of p for fixed x.
Gives $\mathbb{P}(X=x)$ as p changes.
($x = 125$ here, but could be anything.)



Probability function, $f_X(x)$.
Function of x for fixed p.
Gives $\mathbb{P}(X=x)$ as x changes.
($p = 0.6$ here, but could be anything.)

Maximizing the likelihood

We have decided that a sensible parameter estimate for p is the maximum likelihood estimate: *the value of p at which the observation $X = 125$ is more likely than at any other value of p.*

We can find the maximum likelihood estimate using *calculus.*

The likelihood function is

$$L(p\,;125) = \binom{190}{125} p^{125}(1-p)^{65}.$$

We wish to find the value of p that maximizes this expression.

To find the maximizing value of p, *differentiate the likelihood with respect to* p*:*

$$\begin{aligned}\frac{dL}{dp} &= \binom{190}{125} \times \Big\{ 125 \times p^{124} \times (1-p)^{65} + p^{125} \times 65 \times (1-p)^{64} \times (-1) \Big\}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\textit{(Product Rule)}\\ &= \binom{190}{125} \times p^{124} \times (1-p)^{64} \Big\{ 125(1-p) - 65p \Big\}\\ &= \binom{190}{125} p^{124}(1-p)^{64} \Big\{ 125 - 190p \Big\}\ .\end{aligned}$$

The maximizing value of p occurs when

$$\frac{dL}{dp} = 0.$$

This gives:

$$\begin{aligned}\frac{dL}{dp} = \binom{190}{125} p^{124}(1-p)^{64} \Big\{ 125 - 190p \Big\} &= 0\\ \Rightarrow \quad \Big\{ 125 - 190p \Big\} &= 0\\ \Rightarrow \quad p &= \frac{125}{190} = 0.658\,.\end{aligned}$$

For the diver example, the maximum likelihood estimate of 125/190 is *the same as the common-sense estimate (page 53):*

$$p = \frac{\textit{number of daughters}}{\textit{total number of children}} = \frac{125}{190}.$$

This gives us confidence that the method of maximum likelihood is sensible.

The 'hat' notation for an estimate

It is conventional to write the estimated value of a parameter with a 'hat', like this: $\widehat{p}$.

For example,

$$\widehat{p} = \frac{125}{190}\,.$$

The correct notation for the maximization is:

$$\left.\frac{dL}{dp}\right|_{p=\widehat{p}} = 0 \quad \Rightarrow \quad \widehat{p} = \frac{125}{190}\,.$$

Summary of the maximum likelihood procedure

1. Write down the distribution of X in terms of the unknown parameter:

$$X \sim \textit{Binomial}(190, p).$$

2. Write down the observed value of X:

$$\textit{Observed data: } X = 125.$$

3. Write down the likelihood function for this observed value:

$$\begin{aligned} L(p\,; 125) &= \mathbb{P}(X = 125) \textit{ when } X \sim \textit{Binomial}(190, p) \\ &= \binom{190}{125} p^{125}(1-p)^{65} \qquad \textit{for } 0 < p < 1. \end{aligned}$$

4. Differentiate the likelihood with respect to the parameter, and set to 0 for the maximum:

$$\frac{dL}{dp} = \binom{190}{125} p^{124}(1-p)^{64} \Big\{ 125 - 190p \Big\} = 0, \text{ \textit{when} } p = \widehat{p}.$$

   This is the *Likelihood Equation.*

5. Solve for $\widehat{p}$: *From the graph, we can see that* $p = 0$ *and* $p = 1$ *are not maxima.*

$$\therefore \qquad \widehat{p} = \frac{125}{190}\,.$$

This is the *maximum likelihood estimate* (MLE) of p.

Verifying the maximum

Strictly speaking, when we find the maximum likelihood estimate using

$$\left.\frac{dL}{dp}\right|_{p=\widehat{p}} = 0,$$

we should verify that the result is a maximum (rather than a minimum) by showing that

$$\left.\frac{d^2L}{dp^2}\right|_{p=\widehat{p}} < 0\,.$$

In Stats 210, we will be relaxed about this. You will usually be told to assume that the MLE occurs in the interior of the parameter range. Where possible, it is always best to *plot the likelihood function, as on page 55.*

This confirms that the maximum likelihood estimate *exists and is unique.*

In particular, *care must be taken when the parameter has a restricted range like* $0 < p < 1$ *(see later).*

Estimators

For the example above, we had observation $X = 125$, and the maximum likelihood estimate of p was

$$\widehat{p} = \frac{125}{190}\,.$$

It is clear that we could follow through the same working with *any* value of X, which we can write as $X = x$, and we would obtain

$$\widehat{p} = \frac{x}{190}\,.$$

Exercise: Check this by maximizing the likelihood using x instead of 125.

This means that even *before* we have made our observation of X, we can provide *a RULE for calculating the maximum likelihood estimate once X is observed:*

<u>**Rule:**</u> *Let*

$$X \sim \textit{Binomial}(190,\ p).$$

Whatever value of X we observe, the maximum likelihood estimate of p will be

$$\widehat{p} = \frac{X}{190}\,.$$

Note that this expression is now a *random variable: it depends on the random value of X.*

A random variable specifying how an estimate is calculated from an observation is called *an estimator.*

In the example above, *the maximum likelihood estimaTOR of p is*

$$\widehat{p} = \frac{X}{190}\,.$$

The maximum likelihood estimaTE of p, once we have observed that $X = x$, is

$$\widehat{p} = \frac{x}{190}\,.$$

General maximum likelihood estimator for Binomial(n, p)

Take *any* situation in which our observation X has the distribution

$$X \sim \textit{Binomial}(n, p),$$

where n is KNOWN and p is to be estimated.

We make a single observation $X = x$.

Follow the steps on page 59 to find the maximum likelihood estimator for p.

1. Write down the distribution of X in terms of the unknown parameter:

   $$X \sim \textit{Binomial}(n, p).$$

   (n is known.)

2. Write down the observed value of X:

   $$\textit{Observed data: } X = x.$$

3. Write down the likelihood function for this observed value:

   $$\begin{aligned} L(p\,; x) &= \mathbb{P}(X = x) \textit{ when } X \sim \textit{Binomial}(n, p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \qquad \textit{for } 0 < p < 1. \end{aligned}$$

4. Differentiate the likelihood with respect to the parameter, and set to 0 for the maximum:

   $$\frac{dL}{dp} = \binom{n}{x} p^{x-1}(1-p)^{n-x-1} \Big\{ x - np \Big\} = 0, \textit{ when } p = \widehat{p}.$$

   (Exercise)

5. Solve for $\widehat{p}$:

   $$\widehat{p} = \frac{x}{n}\,.$$

This is the *maximum likelihood estimate of p.*

The maximum likelihood estimator of p is

$$\widehat{p} = \frac{X}{n}.$$

(Just replace the x in the MLE with an X, to convert from the estimate to the estimator.)

By deriving the general maximum likelihood estimator for *any* problem of this sort, we can plug in values of n and x to get an instant MLE for any Binomial(n, p) problem in which n is known.

Example: Recall the president problem in Section 2.3. Out of 153 children, 65 were daughters. Let p be the probability that a presidential child is a daughter. What is the maximum likelihood estimate of p?

Solution: *Plug in the numbers* $n = 153$, $x = 65$*:*

the maximum likelihood estimate is

$$\widehat{p} = \frac{x}{n} = \frac{65}{153} = 0.425\,.$$

Note: We showed in Section 2.3 that p *was not significantly different from* 0.5 *in this example.*
However, the MLE of p is definitely different from 0.5.
This comes back to the meaning of *significantly different* in the statistical sense.
Saying that p *is not significantly different from* 0.5 *just means that we can't DISTINGUISH any difference between* p *and* 0.5 *from routine sampling variability.*

We expect that p probably IS different from 0.5, just by a little. The maximum likelihood estimate gives us the *'best'* estimate of p.

Note: We have only considered the class of problems for which $X \sim$ Binomial(n, p) and n is KNOWN. If n is not known, we have a harder problem: we have two parameters, and one of them (n) should only take discrete values $1, 2, 3, \ldots$.
We will not consider problems of this type in Stats 210.

2.6 Random numbers and histograms

We often wish to generate random numbers from a given distribution. Statistical packages like R have custom-made commands for doing this.

To generate (say) 100 random numbers from the Binomial$(n = 190, p = 0.6)$ distribution in R, we use:

```
rbinom(100, 190, 0.6)
```

or in long-hand,

```
rbinom(n=100, size=190, prob=0.6)
```

Caution: the R inputs `n` and `size` are the opposite to what you might expect: `n` gives the required sample size, and `size` gives the Binomial parameter n!

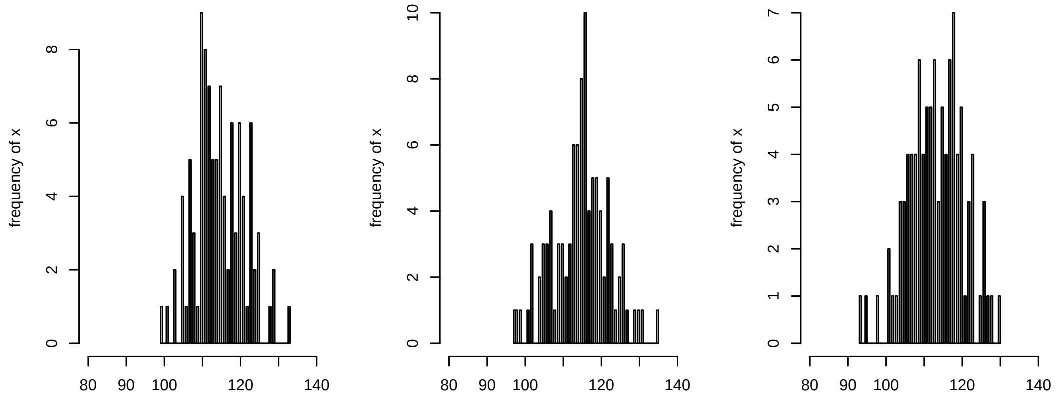
Histograms

The usual graph used to visualise a set of random numbers is the ***histogram.***

The height of each bar of the histogram shows how many of the random numbers fall into the interval represented by the bar.

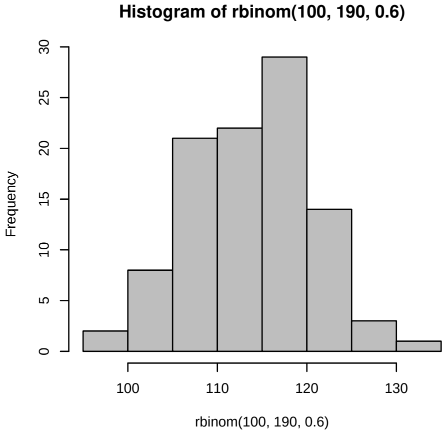
For example, if each histogram bar covers an interval of length 5, and if 24 of the random numbers fall between 105 and 110, then the height of the histogram bar for the interval (105, 110) would be *24.*

Here are histograms from applying the command `rbinom(100, 190, 0.6)` three different times.



Each graph shows *100 random numbers from the Binomial*$(n = 190, p = 0.6)$ *distribution.*

Note: The histograms above have been specially adjusted so that each histogram bar covers an interval of just one integer. For example, the height of the bar plotted at $x = 109$ shows *how many of the 100 random numbers are equal to 109.*



Usually, histogram bars would cover a larger interval, and the histogram would be smoother. For example, on the right is a histogram using the default settings in R, obtained from the command `hist(rbinom(100, 190, 0.6))`.

Each histogram bar covers an interval of *5 integers.*

In all the histograms above, the sum of the heights of all the bars is 100, because there are 100 observations.

Histograms as the sample size increases

Histograms are useful because *they show the approximate shape of the underlying probability function.*

They are also useful for exploring the effect of increasing sample size.

All the histograms below have bars covering an interval of *1 integer.*
They show how the histogram becomes smoother and less erratic as sample size increases.

Eventually, with a large enough sample size, *the histogram starts to look identical to the probability function.*

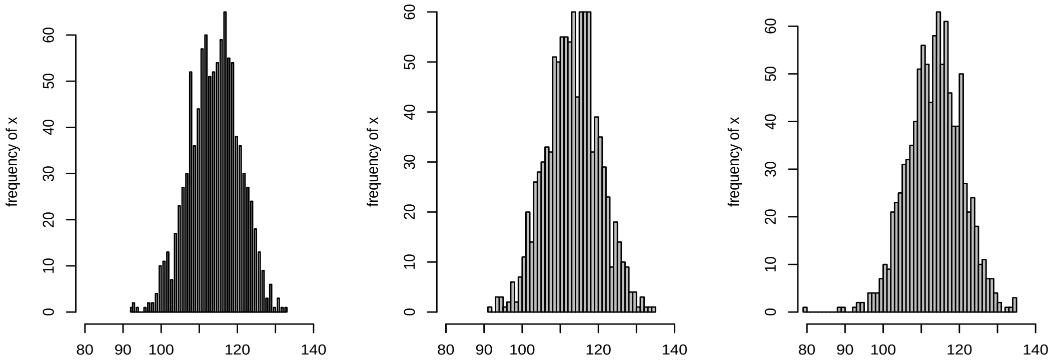
Note: difference between a histogram and the probability function

The histogram plots OBSERVED FREQUENCIES of a set of random numbers.

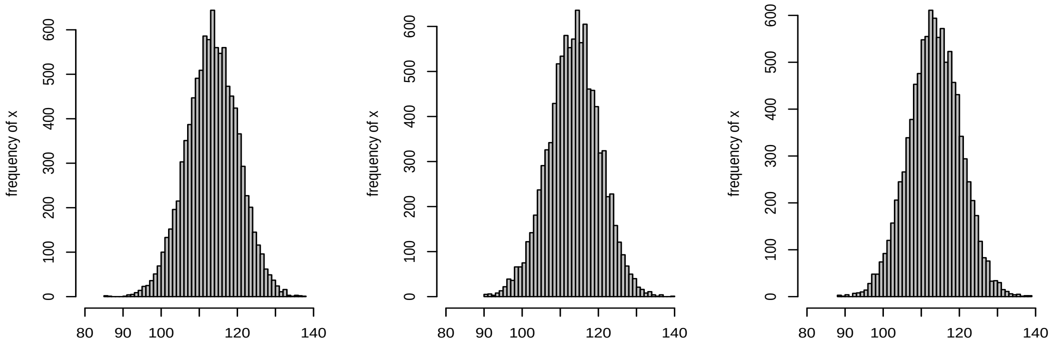
The probability function plots EXACT PROBABILITIES for the distribution.

The histogram *should have the same shape as the probability function, especially as the sample size gets large.*

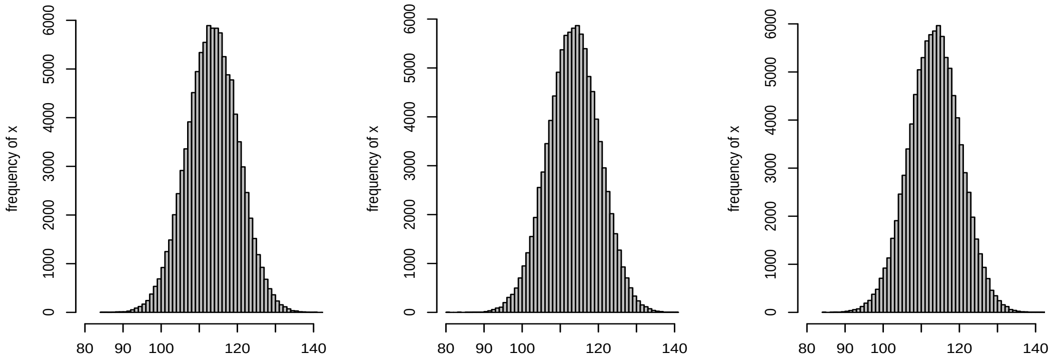
Sample size 1000: `rbinom(1000, 190, 0.6)`



Sample size 10,000: `rbinom(10000, 190, 0.6)`



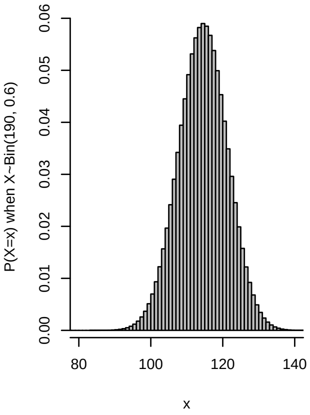
Sample size 100,000: `rbinom(100000, 190, 0.6)`



Probability function for Binomial(190, 0.6):

The probability function is *fixed and exact.*

The histograms become stable in shape and approach the shape of the probability function as sample size gets large.

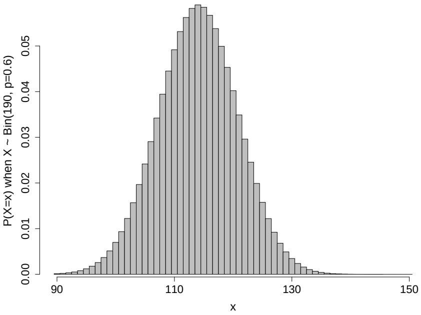


2.7 Expectation

Given a random variable X that measures something, we often want to know *what is the* **average** *value of X?*

For example, here are 30 random observations taken from the distribution $X \sim \text{Binomial}(n = 190, p = 0.6)$:

***R* command:** `rbinom(30, 190, 0.6)`



```
 116 116 117 122 111 112 114 120 112 102
 125 116  97 105 108 117 118 111 116 121
 107 113 120 114 114 124 116 118 119 120
```

The average, or *mean*, of the ***first ten*** values is:

$$\frac{116 + 116 + \ldots + 112 + 102}{10} = 114.2.$$

The mean of the ***first twenty*** values is:

$$\frac{116 + 116 + \ldots + 116 + 121}{20} = 113.8.$$

The mean of the ***first thirty*** values is:

$$\frac{116 + 116 + \ldots + 119 + 120}{30} = 114.7.$$

The answers all seem to be close to *114*. What would happen if we took the average of hundreds of values?

100 values from Binomial(190, 0.6):

R command: `mean(rbinom(100, 190, 0.6))`
Result: `114.86`

Note: You will get a different result every time you run this command.

1000 values from Binomial(190, 0.6):

R command: `mean(rbinom(1000, 190, 0.6))`
Result: `114.02`

1 million values from Binomial(190, 0.6):

R command: `mean(rbinom(1000000, 190, 0.6))`
Result: `114.0001`

The average seems to be *converging to the value 114.*

The larger the sample size, *the closer the average seems to get to 114.*

If we kept going for larger and larger sample sizes, we would keep getting answers closer and closer to 114. This is because *114 is the DISTRIBUTION MEAN: the mean value that we would get if we were able to draw an infinite sample from the Binomial(190, 0.6) distribution.*

This distribution mean is called the *expectation, or expected value, of the Binomial(190, 0.6) distribution.*

It is a *FIXED property of the Binomial(190, 0.6) distribution.* This means it is a *fixed constant: there is nothing random about it.*

Definition: The **expected value**, also called the **expectation** or **mean**, of a discrete random variable X, *can be written as either* $\mathbb{E}(X)$*, or E(X), or* μ_X*, and is given by*

$$\boxed{\mu_X = \mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)\,.}$$

The expected value is a measure of the centre, or average, of the set of values that X can take, weighted according to the probability of each value.

If we took a very large sample of random numbers from the distribution of X*, their average would be approximately equal to* μ_X*.*

Example: Let $X \sim \text{Binomial}(n = 190, p = 0.6)$. What is $\mathbb{E}(X)$?

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{190} x \binom{190}{x} (0.6)^x (0.4)^{190-x}.\end{aligned}$$

Although it is not obvious, the answer to this sum is $n \times p = 190 \times 0.6 = 114$. We will see why in Section 2.10.

Explanation of the formula for expectation

We will move away from the Binomial distribution for a moment, and use a simpler example.

Let the random variable X be defined as $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

X takes only the values 1 and -1. What is the 'average' value of X?

Using $\frac{1+(-1)}{2} = 0$ *would not be useful, because it ignores the fact that* <u>*usually*</u> $X = 1$*, and only occasionally is* $X = -1$.

Instead, think of observing X many times, say 100 times.

Roughly *90* of these 100 times will have $X = 1$.
Roughly *10* of these 100 times will have $X = -1$

The average of the 100 values will be roughly

$$\begin{aligned}&\frac{90 \times 1 + 10 \times (-1)}{100}, \\ &= 0.9 \times 1 + 0.1 \times (-1) \\ (\ &= 0.8.\ \ )\end{aligned}$$

We could repeat this for any sample size.

As the sample gets large, the average of the sample will get ever closer to

$$0.9 \times 1 + 0.1 \times (-1).$$

This is why the distribution mean is given by

$$\mathbb{E}(X) = \mathbb{P}(X = 1) \times 1 + \mathbb{P}(X = -1) \times (-1),$$

or in general,

$$\mathbb{E}(X) = \sum_x \mathbb{P}(X = x) \times x.$$

$\mathbb{E}(X)$ *is a fixed constant giving the average value we would get from a large sample of* X.

Linear property of expectation

Expectation is a ***linear*** operator:

Theorem 2.7: *Let a and b be constants. Then*

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Proof:

Immediate from the definition of expectation.

$$\begin{aligned} \mathbb{E}(aX + b) &= \sum_x (ax + b) f_X(x) \\ &= a \sum_x x f_X(x) + b \sum_x f_X(x) \\ &= a\,\mathbb{E}(X) + b \times 1. \qquad \square \end{aligned}$$

Example: finding expectation from the probability function

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$. Write down the probability function of X and find $\mathbb{E}(X)$.

We have:

$$\mathbb{P}(X = x) = \binom{3}{x}(0.2)^x(0.8)^{3-x} \textit{ for } x = 0, 1, 2, 3.$$

x	*0*	*1*	*2*	*3*
$f_X(x) = \mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Then

$$\begin{aligned}\mathbb{E}(X) = \sum_{x=0}^{3} x f_X(x) &= 0 \times 0.512 + 1 \times 0.384 + 2 \times 0.096 + 3 \times 0.008 \\ &= 0.6.\end{aligned}$$

Note: We have: $\mathbb{E}(X) = 0.6 = 3 \times 0.2$ *for* $X \sim$ *Binomial*$(3, 0.2)$.
We will prove in Section 2.10 that whenever $X \sim \text{Binomial}(n, p)$, then $\mathbb{E}(X) = np$.

Example 2: Let Y be Bernoulli(p) (Section 1.2). That is,

$$Y = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability } 1 - p. \end{cases}$$

Find $\mathbb{E}(Y)$.

y	*0*	*1*
$\mathbb{P}(Y = y)$	$1 - p$	p

$$\mathbb{E}(Y) = 0 \times (1 - p) + 1 \times p = p.$$

Expectation of a sum of random variables: $\mathbb{E}(X+Y)$

For ANY random variables $X_1, X_2, \ldots, X_n$,

$$\boxed{\mathbb{E}\left(X_1 + X_2 + \ldots + X_n\right) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n).}$$

In particular, $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ *for ANY* X *and* Y.

This result holds for ***any*** random variables $X_1, \ldots, X_n$. *It does NOT require* $X_1, \ldots, X_n$ *to be independent.*

We can summarize this important result by saying:

The expectation of a sum
is the sum of the expectations – ALWAYS.

The proof requires multivariate methods, to be studied in later courses.

Note: We can combine the result above with the linear property of expectation. For any constants $a_1, \ldots, a_n$, we have:

$$\mathbb{E}\left(a_1X_1 + a_2X_2 + \ldots + a_nX_n\right) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n).$$

Expectation of a product of random variables: $\mathbb{E}(XY)$

There are two cases when finding the expectation of a product:

1. **General case:**

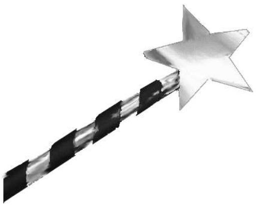
   For general X *and* Y, $\mathbb{E}(XY)$ *is NOT equal to* $\mathbb{E}(X)\mathbb{E}(Y)$.

   We have to find $\mathbb{E}(XY)$ either using their joint probability function (see later), or using their covariance (see later).

2. **Special case:** when X and Y are *INDEPENDENT:*

   When X *and* Y *are INDEPENDENT,* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

2.8 Variable transformations



We often wish to *transform* random variables through a function. For example, given the random variable X, possible transformations of X include:

$$X^2, \qquad \sqrt{X}, \qquad 4X^3, \qquad \ldots$$

We often summarize all possible variable transformations by referring to $Y = g(X)$ *for some function* g.

For discrete random variables, it is very easy to find the probability function for $Y = g(X)$, given that the probability function for X is known. Simply *change all the values and keep the probabilities the same.*

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$, and let $Y = X^2$. Find the probability function of Y.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Thus *the probability function for* $Y = X^2$ *is:*

y	0^2	1^2	2^2	3^2
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

This is because Y *takes the value* 0^2 *whenever* X *takes the value* 0*, and so on.*

Thus the probability that $Y = 0^2$ is *the same as the probability that* $X = 0$*.*

Overall, we would write the probability function of $Y = X^2$ as:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

To transform a discrete random variable, *transform the values and leave the probabilities alone.*

Example 2: Mr Chance hires out giant helium balloons for advertising. His balloons come in three sizes: heights 2m, 3m, and 4m. 50% of Mr Chance's customers choose to hire the cheapest 2m balloon, while 30% hire the 3m balloon and 20% hire the 4m balloon.

The amount of helium gas in cubic metres required to fill the balloons is $h^3/2$, where h is the height of the balloon. Find the probability function of Y, the amount of helium gas required for a randomly chosen customer.

Let X be the height of balloon ordered by a random customer. The probability function of X is:

height, x (m)	*2*	*3*	*4*
$\mathbb{P}(X = x)$	*0.5*	*0.3*	*0.2*

Let Y be the amount of gas required: $Y = X^3/2$.
The probability function of Y is:

gas, y (m^3)	*4*	*13.5*	*32*
$\mathbb{P}(Y = y)$	*0.5*	*0.3*	*0.2*

Transform the values, and leave the probabilities alone.

Expected value of a transformed random variable

We can find the expectation of a transformed random variable just like any other random variable. For example, in Example 1 we had $X \sim \text{Binomial}(3, 0.2)$, and $Y = X^2$.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

and for $Y = X^2$:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

Thus the expectation of $Y = X^2$ is:

$$\begin{aligned}\mathbb{E}(Y) = \mathbb{E}(X^2) &= 0 \times 0.512 + 1 \times 0.384 + 4 \times 0.096 + 9 \times 0.008 \\ &= 0.84.\end{aligned}$$

Note: $\mathbb{E}(X^2)$ ***is NOT the same as*** $\{\mathbb{E}(X)\}^2$. ***Check that*** $\{\mathbb{E}(X)\}^2 = 0.36.$

To make the calculation quicker, we could cut out the middle step of writing down the probability function of Y. Because we transform the values and keep the probabilities the same, we have:

$$\mathbb{E}(X^2) = 0^2 \times 0.512 + 1^2 \times 0.384 + 2^2 \times 0.096 + 3^2 \times 0.008.$$

If we write $g(X) = X^2$, this becomes:

$$\mathbb{E}\{g(X)\} = \mathbb{E}(X^2) = g(0) \times 0.512 + g(1) \times 0.384 + g(2) \times 0.096 + g(3) \times 0.008.$$

Clearly the same arguments can be extended to any function $g(X)$ and any discrete random variable X:

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x).}$$

$$\boxed{\text{Transform the values, and leave the probabilities alone.}}$$

Definition: For any function g and discrete random variable X, the expected value of $g(X)$ is given by

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x) = \sum_x g(x) f_X(x).}$$

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Gas required was $X^3/2$ *from page 74.*
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\begin{aligned}\mathbb{E}\left(\frac{X^3}{2}\right) &= \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x) \\ &= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2 \\ &= 12.45 \ \textit{m}^3 \textit{ gas.}\end{aligned}$$

(b) Mr Chance charges $\$400 \times h$ to hire a balloon of height h. What is his expected earning per customer?

Expected earning is $\mathbb{E}(400X)$.

$$\begin{aligned}\mathbb{E}(400X) &= 400 \times \mathbb{E}(X) \qquad \textit{(expectation is linear)} \\ &= 400 \times (2 \times 0.5 + 3 \times 0.3 + 4 \times 0.2) \\ &= 400 \times 2.7 \\ &= \$1080 \textit{ per customer.}\end{aligned}$$

(c) How much does Mr Chance expect to earn in total from his next 5 customers?

Let $Z_1, \ldots, Z_5$ *be the earnings from the next 5 customers. Each* Z_i *has* $\mathbb{E}(Z_i) = 1080$ *by part (b). The total expected earning is*

$$\begin{aligned}\mathbb{E}(Z_1 + Z_2 + \ldots + Z_5) &= \mathbb{E}(Z_1) + \mathbb{E}(Z_2) + \ldots + \mathbb{E}(Z_5) \\ &= 5 \times 1080 \\ &= \$5400.\end{aligned}$$

Getting the expectation. . .

Suppose $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then $3/4$ of the time, X takes value 3, and $1/4$ of the time, X takes value 8.

So $\mathbb{E}(X) \;=\; \frac{3}{4} \times 3 \;+\; \frac{1}{4} \times 8.$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

What about $\mathbb{E}(\sqrt{X})$?

$$\sqrt{X} = \begin{cases} \sqrt{3} & \text{with probability } 3/4, \\ \sqrt{8} & \text{with probability } 1/4. \end{cases}$$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

$$\mathbb{E}(\sqrt{X}) = \frac{3}{4} \times \sqrt{3} + \frac{1}{4} \times \sqrt{8}.$$

Common mistakes

i) $\mathbb{E}(\sqrt{X}) = \sqrt{\mathbb{E}X} = \sqrt{\frac{3}{4} \times 3 + \frac{1}{4} \times 8}$

Wrong!

ii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4}} \times 3 + \sqrt{\frac{1}{4}} \times 8$

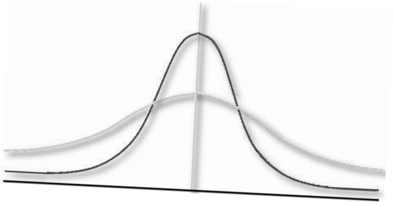
Wrong!

iii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4} \times 3} + \sqrt{\frac{1}{4} \times 8}$

$= \sqrt{\frac{3}{4}} \times \sqrt{3} + \sqrt{\frac{1}{4}} \times \sqrt{8}$

Wrong!

2.9 Variance



Example: Mrs Tractor runs the Rational Bank of Remuera. Every day she hopes to fill her cash machine with enough cash to see the well-heeled citizens of Remuera through the day. She knows that the expected amount of money withdrawn each day is \$50,000. How much money should she load in the machine? \$50,000?

No: \$50,000 is the average, near the centre of the distribution. About half the time, the money required will be GREATER than the average.

How much money should Mrs Tractor put in the machine if she wants to be 99% certain that there will be enough for the day's transactions?

Answer: it depends how much the amount withdrawn *varies above and below its mean.*

For questions like this, we need the study of *variance.*

Variance is the *average squared distance of a random variable from its own mean.*

Definition: The **variance** of a random variable X *is written as either Var*(X) *or* σ_X^2, *and is given by*

$$\boxed{\sigma_X^2 = \text{Var}(X) = \mathbb{E}\left[(X-\mu_X)^2\right] = \mathbb{E}\left[(X-\mathbb{E}X)^2\right].}$$

Similarly, the variance of a function of X is

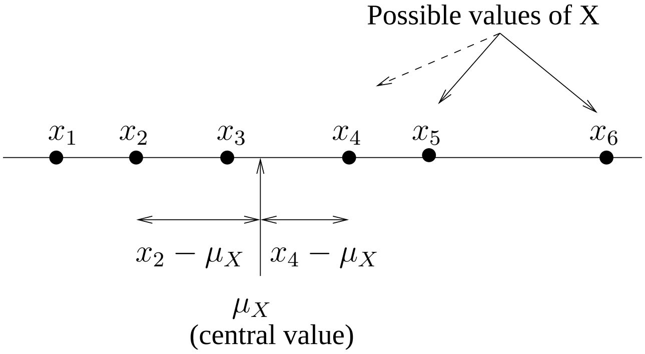
$$\boxed{\text{Var}(g(X)) = \mathbb{E}\left[\Big(g(X) - \mathbb{E}(g(X))\Big)^2\right].}$$

Note: The variance *is the square of the standard deviation of X, so*

$$sd(X) = \sqrt{\text{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

Variance as the average squared distance from the mean

The variance is a measure of how *spread out* are the values that X can take. It is the *average squared distance between a value of X and the central (mean) value, μ_X.*



$$\textit{Var}(X) = \underbrace{\mathbb{E}}_{(2)} \underbrace{[(X - \mu_X)^2]}_{(1)}$$

(1) Take distance from observed values of X to the central point, μ_X. Square it to balance positive and negative distances.

(2) Then take the average over all values X can take: ie. if we observed X many times, find what would be the average squared distance between X and μ_X.

Note: The mean, μ_X, and the variance, σ_X^2, of X are just *numbers: there is nothing random or variable about them.*

Example: Let $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then

$$\begin{aligned} \mathbb{E}(X) = \mu_X &= 3 \times \frac{3}{4} + 8 \times \frac{1}{4} = 4.25 \\ \textit{Var}(X) = \sigma_X^2 &= \frac{3}{4} \times (3 - 4.25)^2 + \frac{1}{4} \times (8 - 4.25)^2 \\ &= 4.6875. \end{aligned}$$

When we observe X, we get either *3 or 8: this is random.*
But μ_X *is fixed at 4.25, and σ_X^2 is fixed at 4.6875, regardless of the outcome of X.*

For a discrete random variable,

$$\boxed{\mathit{Var}(X) = \mathbb{E}\left[(X-\mu_X)^2\right] = \sum_x (x-\mu_X)^2 f_X(x) = \sum_x (x-\mu_X)^2 \mathbb{P}(X=x).}$$

This uses the definition of the expected value of a function of X:

$$\mathit{Var}(X) = \mathbb{E}(g(X)) \;\; \textit{where} \;\; g(X) = (X-\mu_X)^2.$$

Theorem 2.9A: (important)

$$\boxed{\mathit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu_X^2}$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu_X)^2\right] \quad \text{by definition} \\
&= \mathbb{E}[\underbrace{X^2}_{\text{r.v.}} - 2\underbrace{X}_{\text{r.v.}} \underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}}] \\
&= \mathbb{E}(X^2) - 2\mu_X \mathbb{E}(X) + \mu_X^2 \quad \text{by Thm 2.7} \\
&= \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= \mathbb{E}(X^2) - \mu_X^2 . \qquad \square
\end{aligned}$$

Note: $\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = \sum_x x^2\,\mathbb{P}(X=x)$. This is <u>not</u> the same as $(\mathbb{E}X)^2$:

e.g.
$$X = \begin{cases} 3 & \text{with probability } 0.75, \\ 8 & \text{with probability } 0.25. \end{cases}$$

Then $\mu_X = \mathbb{E}X = 4.25$, so $\mu_X^2 = (\mathbb{E}X)^2 = (4.25)^2 = 18.0625$.
But

$$\mathbb{E}(X^2) = \left(3^2 \times \frac{3}{4} + 8^2 \times \frac{1}{4}\right) = 22.75.$$

Thus $\boxed{\mathbb{E}(X^2) \neq (\mathbb{E}X)^2 \textit{ in general.}}$

Theorem 2.9B: If a and b are constants and $g(x)$ is a function, then

i) $\text{Var}(aX+b) = a^2\,\text{Var}(X)$.

ii) $\text{Var}(a\,g(X)+b) = a^2\,\text{Var}\{g(X)\}$.

Proof:

(part (i))

$$\begin{aligned}
\text{Var}(aX+b) &= \mathbb{E}\Big[\{(aX+b)-\mathbb{E}(aX+b)\}^2\Big] \\
&= \mathbb{E}\Big[\{aX+b-a\mathbb{E}(X)-b\}^2\Big] \quad \textit{by Thm 2.7} \\
&= \mathbb{E}\Big[\{aX-a\mathbb{E}(X)\}^2\Big] \\
&= \mathbb{E}\Big[a^2\{X-\mathbb{E}(X)\}^2\Big] \\
&= a^2\mathbb{E}\Big[\{X-\mathbb{E}(X)\}^2\Big] \quad \textit{by Thm 2.7} \\
&= a^2\,\text{Var}(X)\,.
\end{aligned}$$

Part (ii) follows similarly.

Note: These are very different from the corresponding expressions for expectations (Theorem 2.7). Variances are more difficult to manipulate than expectations.

Example: finding expectation and variance from the probability function

Recall Mr Chance's balloons from page 74. The random variable Y is the amount of gas required by a randomly chosen customer. The probability function of Y is:

gas, y (m^3)	4	13.5	32
$\mathbb{P}(Y=y)$	0.5	0.3	0.2

Find Var(Y).

We know that $\mathbb{E}(Y) = \mu_Y = 12.45$ *from page 76.*

First method: *use* $\mathit{Var}(Y) = \mathbb{E}[(Y - \mu_Y)^2]$*:*

$$\begin{aligned} \mathit{Var}(Y) &= (4 - 12.45)^2 \times 0.5 + (13.5 - 12.45)^2 \times 0.3 + (32 - 12.45)^2 \times 0.2 \\ &= 112.47. \end{aligned}$$

Second method: *use* $\mathbb{E}(Y^2) - \mu_Y^2$*: (usually easier)*

$$\begin{aligned} \mathbb{E}(Y^2) &= 4^2 \times 0.5 + 13.5^2 \times 0.3 + 32^2 \times 0.2 \\ &= 267.475. \end{aligned}$$

So $\mathit{Var}(Y) = 267.475 - (12.45)^2 = 112.47$ *as before.*

Variance of a sum of random variables: $\mathrm{Var}(\boldsymbol{X} + \boldsymbol{Y})$

There are two cases when finding the variance of a sum:

1. **General case:**

   For general X and Y,
   $\mathit{Var}(X + Y)$ is NOT equal to $\mathit{Var}(X) + \mathit{Var}(Y)$.

   We have to find $\mathrm{Var}(X + Y)$ using their covariance (see later courses).

2. **Special case:** when X and Y are *INDEPENDENT:*

   When X and Y are INDEPENDENT,
   $\mathit{Var}(X + Y) = \mathit{Var}(X) + \mathit{Var}(Y)$.

2.10 Mean and Variance of the Binomial(n, p) distribution

Let $X \sim$ Binomial(n, p). We have mentioned several times that $\mathbb{E}(X) = np$. We now prove this and the additional result for Var(X).

If $X \sim$ Binomial(n, p), then:

$$\boxed{\begin{aligned} \mathbb{E}(X) &= \mu_X = np \\ \text{Var}(X) &= \sigma_X^2 = np(1-p). \end{aligned}}$$

We often write $q = 1 - p$, *so* Var$(X) = npq$.

Easy proof: X as a sum of Bernoulli random variables

If $X \sim$ Binomial(n, p), then X is the *number of successes out of* n *independent trials, each with* $\mathbb{P}$(*success*) $= p$.

This means that we can write:

$$X = Y_1 + Y_2 + \ldots + Y_n,$$

where each

$$Y_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

That is, Y_i *counts as a 1 if trial* i *is a success, and as a 0 if trial* i *is a failure.*

Overall, $Y_1 + \ldots + Y_n$ *is the total number of successes out of* n *independent trials, which is the same as* X.

Note: Each Y_i is a Bernoulli(p) random variable (Section 1.2).

Now if $X = Y_1 + Y_2 + \ldots + Y_n$, and $Y_1, \ldots, Y_n$ are independent, then:

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \ldots + \mathbb{E}(Y_n) \quad \textit{(does NOT require independence)}, \\ \text{Var}(X) &= \text{Var}(Y_1) + \text{Var}(Y_2) + \ldots + \text{Var}(Y_n) \quad \textit{(DOES require independence)}. \end{aligned}$$

The probability function of each Y_i is:

y	0	1
$\mathbb{P}(Y_i = y)$	$1-p$	p

Thus,

$$\mathbb{E}(Y_i) = 0 \times (1-p) + 1 \times p = p.$$

Also,

$$\mathbb{E}(Y_i^2) = 0^2 \times (1-p) + 1^2 \times p = p.$$

So

$$\begin{aligned} \mathit{Var}(Y_i) &= \mathbb{E}(Y_i^2) - (\mathbb{E}Y_i)^2 \\ &= p - p^2 \\ &= p(1-p). \end{aligned}$$

Therefore:

$$\begin{aligned} \mathbb{E}(X) &= \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \ldots + \mathbb{E}(Y_n) \\ &= p + p + \ldots + p \\ &= n \times p. \end{aligned}$$

And:

$$\begin{aligned} \mathit{Var}(X) &= \mathit{Var}(Y_1) + \mathit{Var}(Y_2) + \ldots + \mathit{Var}(Y_n) \\ &= n \times p(1-p). \end{aligned}$$

Thus we have proved that $\mathbb{E}(X) = np$ and $\text{Var}(X) = np(1-p)$. □

Hard proof: for mathematicians (non-examinable)

We show below how the Binomial mean and variance formulae can be derived directly from the probability function.

$$\mathbb{E}(X) = \sum_{x=0}^{n} x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} x \left(\frac{n!}{(n-x)!x!}\right) p^x (1-p)^{n-x}$$

But $\dfrac{x}{x!} = \dfrac{1}{(x-1)!}$ and also the first term $x f_X(x)$ is 0 when $x = 0$.

So, continuing,

$$\mathbb{E}(X) = \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!} p^x (1-p)^{n-x}$$

Next: make n's into $(n-1)$'s, x's into $(x-1)$'s, wherever possible: e.g.

$$\begin{aligned} n - x &= (n-1) - (x-1), \quad p^x = p \cdot p^{x-1} \\ n! &= n(n-1)! \ \textit{etc.} \end{aligned}$$

This gives,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=1}^{n} \frac{n(n-1)!}{[(n-1)-(x-1)]!(x-1)!} p \cdot p^{(x-1)} (1-p)^{(n-1)-(x-1)} \\ &= \underbrace{np}_{\text{what we want}} \quad \underbrace{\sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)}}_{\text{need to show this sum } = 1} \end{aligned}$$

Finally we let $y = x - 1$ and let $m = n - 1$.
When $x = 1, y = 0$; and when $x = n$, $y = n - 1 = m$.

So

$$\begin{aligned} \mathbb{E}(X) &= np \sum_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y} \\ &= np(p + (1-p))^m \quad \text{(Binomial Theorem)} \end{aligned}$$

$$\mathbb{E}(X) = np, \quad \text{as required.}$$

For $\text{Var}(X)$, use the same ideas again.

For $\mathbb{E}(X)$, we used $\frac{x}{x!} = \frac{1}{(x-1)!}$; so instead of finding $\mathbb{E}(X^2)$, it will be easier to find $\mathbb{E}[X(X-1)] = \mathbb{E}(X^2) - \mathbb{E}(X)$ because then we will be able to cancel $\frac{x(x-1)}{x!} = \frac{1}{(x-2)!}$.

Here goes:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{x=0}^{n} x(x-1)\binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \frac{x(x-1)n(n-1)(n-2)!}{[(n-2)-(x-2)]!(x-2)!x(x-1)} p^2 p^{(x-2)} (1-p)^{(n-2)-(x-2)}
\end{aligned}$$

First two terms ($x = 0$ and $x = 1$) are 0 due to the $x(x-1)$ in the numerator. Thus

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= p^2 n(n-1) \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} (1-p)^{(n-2)-(x-2)} \\
&= n(n-1)p^2 \underbrace{\sum_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y}}_{\text{sum=1 by Binomial Theorem}} \qquad \text{if } \begin{cases} m = n-2, \\ y = x-2. \end{cases}
\end{aligned}$$

$$\text{So} \quad \mathbb{E}[X(X-1)] = n(n-1)p^2 .$$

$$\begin{aligned}
\text{Thus} \quad \text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\
&= \mathbb{E}(X^2) - \mathbb{E}(X) + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\
&= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= np(1-p). \qquad \square
\end{aligned}$$

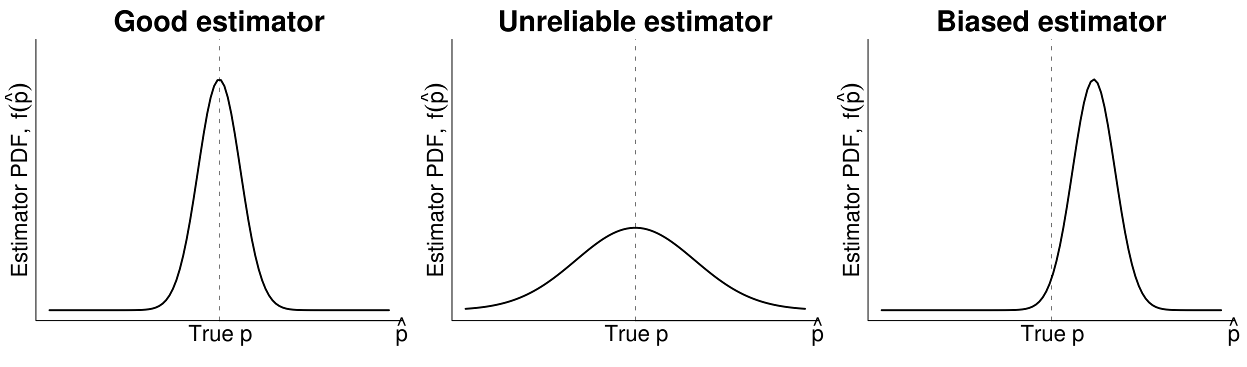
Note the steps: take out $x(x-1)$ and replace n by $(n-2)$, x by $(x-2)$ wherever possible.

2.11 Mean and Variance of Estimators

Perhaps the most important application of mean and variance is in the context of ***estimators:***

- An estimator is a *random variable.*
- It has a ***mean*** and a ***variance.***
- The ***mean*** tells us how ***accurate*** the estimator is: in particular, does it get the right answer on average?
- The ***variance*** tells us how ***reliable*** the estimator is. If the variance is high, it has high spread and we can get estimates a long way from the true answer. Because we don't know what the right answer is, we don't know whether our particular estimate is a good one (close to the true answer) or a bad one (a long way from the true answer). So an estimator with high variance is ***unreliable:*** sometimes it gives bad answers, sometimes it gives good answers; and we don't know which we've got.
  An unreliable estimator is like a friend who often tells lies. Once you find them out, you can't believe ANYTHING they say!

Good estimators and bad estimators



Example: MLE of the Binomial *p* parameter

In Section 2.5 we derived the MLE for the Binomial parameter p.

Reminder: $X \sim \text{Binomial}(n, p)$, where n is KNOWN and p is to be estimated.

Make a single observation $X = x$.

The maximum likelihood estimator of p is $\quad \widehat{p} = \dfrac{X}{n}$.

Why do we convert the estimate x/n to the estimator X/n?

The estimator has a ***mean*** and a ***variance***, and this means *we can study its properties.* Is it accurate? Is it reliable?

Estimator Mean, $\mathbb{E}(\widehat{p})$

$$\begin{aligned}
\mathbb{E}(\widehat{p}) = \mathbb{E}\left(\frac{X}{n}\right) &= \frac{1}{n}\mathbb{E}(X) \\
&= \frac{1}{n} \times np \qquad \textit{because } \mathbb{E}(X) = np \textit{ when } X \sim \textit{Binomial}(n,p) \\
\therefore \quad \mathbb{E}(\widehat{p}) &= p.
\end{aligned}$$

So this estimator ***gets the right answer on average*** — it is ***unbiased.***

Definition: If $\widehat{p}$ is an estimator of the parameter p, then $\widehat{p}$ is ***unbiased*** if $\mathbb{E}(\widehat{p}) = p$.

That is, an unbiased estimator *gets the right answer on average.*

If $\mathbb{E}(\widehat{p}) \neq p$, then $\widehat{p}$ is said to be a *biased estimator.*

If an estimator has a large bias, we probably don't want to use it. However, even if the estimator is *unbiased*, we still need to look at its ***variance*** to decide how *reliable* it is.

Estimator Variance, Var$(\widehat{p})$

We have:

$$\begin{aligned}
\textit{Var}(\widehat{p}) &= \textit{Var}\left(\frac{X}{n}\right) \\
&= \frac{1}{n^2}\textit{Var}(X) \\
&= \frac{1}{n^2} \times np(1-p) \qquad \textit{because Var}(X) = np(1-p) \textit{ for } X \sim\textit{Bin}(n,p) \\
\therefore \quad \textit{Var}(\widehat{p}) &= \frac{p(1-p)}{n}. \qquad\qquad (\star)
\end{aligned}$$

To decide how reliable our estimator $\widehat{p}$ is, we would like to calculate the value of Var$(\widehat{p})$. But Var$(\widehat{p}) = p(1-p)/n$, and *we do not know the true value of p, so we cannot calculate the exact Var$(\widehat{p})$.*

Instead, *we have to ESTIMATE Var*$(\widehat{p})$ *by replacing the unknown* p *in equation* $(\star)$ *by* $\widehat{p}$*.*

We call our ***estimated variance*** $\widehat{\text{Var}}(\widehat{p})$*:*

$$\boxed{\widehat{\text{Var}}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n}.}$$

The ***standard error of*** $\boldsymbol{\widehat{p}}$ is defined as: $se(\widehat{p}) = \sqrt{\widehat{\text{Var}}(\widehat{p})}$.

The ***margin of error*** associated with $\widehat{p}$ is defined as:

$$\boxed{\textit{Margin of error} = 1.96 \times se(\widehat{p}) = 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.}$$

The expression $\widehat{p} \pm 1.96 \times \text{se}(\widehat{p})$ gives an approximate *95% confidence interval for* p *under the Normal approximation.*

This is because the Central Limit Theorem guarantees that $\widehat{p}$ will be approximately Normally distributed when n is large. We will study the Central Limit Theorem and this result in Chapter 5, Section 5.3.

Example: For the deep-sea diver example in Section 2.3, we had $X \sim \text{Binomial}(190, p)$ with observation $X = 125$ daughters out of 190 children.

So

$$\widehat{p} = \frac{X}{n} = \frac{125}{190} = 0.658, \qquad \Rightarrow \qquad se(\widehat{p}) = \sqrt{\frac{0.658 \times (1-0.658)}{190}} = 0.034.$$

For our final answer, we should therefore quote:

$$\widehat{p} = 0.658 \pm 1.96 \times 0.034 = 0.658 \pm 0.067 \qquad \textit{or} \qquad \widehat{p} = 0.658 \quad (0.591, 0.725).$$

Our estimate is fairly precise, although not extremely precise. We are pretty sure that the daughter probability is somewhere between 0.59 and 0.73.

Why do we use the MLE instead of some other estimator?

The MLE is a sensible estimator to use, but we could think of other sensible estimators too. The reason why the MLE is so highly preferred is because it has ***excellent general properties.*** Under mild conditions, and with a large enough sample size, any MLE will be (i) unbiased, (ii) Normally distributed, and (iii) have the minimal possible variance of all estimators. Wow!

Comments about p, $\widehat{p}$, $\mathbb{E}(\widehat{p})$, $\mathrm{Var}(\widehat{p})$, and $\widehat{\mathrm{Var}}(\widehat{p})$

It might seem difficult at first to get to grips with what these quantities are and what they represent. Here are some comments and notes.

- p is a parameter: an unknown but fixed ***number*** that we wish to estimate.
- $\widehat{p}$ is a ***random variable:*** for example, $\widehat{p} = \frac{X}{n}$. It is a particular type of random variable that generates estimates of p, so it is called an ***estimator***.
- $\mathbb{E}(\widehat{p})$ is a number that tells us whether or not our estimator is unbiased. We are mostly interested in $\mathbb{E}(\widehat{p})$ in an abstract sense: for example, if $\mathbb{E}(\widehat{p}) = p$, no matter what p is, then our estimator is unbiased and we are happy. If $\mathbb{E}(\widehat{p}) \neq p$, we want to know how badly wrong it is and whether we should devise a correction factor.

  For example, if we discovered that $\mathbb{E}(\widehat{p}) = \left(\frac{n}{n+1}\right) p$, then we could create a different estimator $\widehat{q}$ such that $\widehat{q} = \left(\frac{n+1}{n}\right) \widehat{p}$. Then we would have,

  $$\mathbb{E}(\widehat{q}) = \left(\tfrac{n+1}{n}\right) \mathbb{E}(\widehat{p}) = \left(\tfrac{n+1}{n}\right) \times \left(\tfrac{n}{n+1}\right) p = p.$$

  So our new estimator $\widehat{q}$ is unbiased for p, but on the downside it also has higher variance than $\widehat{p}$, because $\mathrm{Var}(\widehat{q}) = \left(\frac{n+1}{n}\right)^2 \mathrm{Var}(\widehat{p})$. So we might or might not prefer to use $\widehat{q}$ instead of $\widehat{p}$. As the sample size n grows large, we might prefer to accept a tiny bias with the lower variance and use $\widehat{p}$.
- $\mathrm{Var}(\widehat{p})$ is a number that tells us about the ***reliability*** of our estimator. Unlike $\mathbb{E}(\widehat{p})$ which we care about more as an abstract property, we would like to know the actual numeric value of $\mathrm{Var}(\widehat{p})$ so we can calculate confidence intervals. Confidence intervals quantify our estimator reliability and should be included with our final report.

  Unfortunately, we find that $\mathrm{Var}(\widehat{p})$ depends upon the unknown value p: for example, $\mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}$, so we can't calculate it because we don't know what p is. This is why we use $\widehat{\mathrm{Var}}(\widehat{p})$ described next.
- $\widehat{\mathrm{Var}}(\widehat{p})$ is our best attempt at getting a value for $\mathrm{Var}(\widehat{p})$. We just take the expression for $\mathrm{Var}(\widehat{p})$ and substitute $\widehat{p}$ for the unknown p everywhere. This means that $\widehat{\mathrm{Var}}(\widehat{p})$ is an ***estimator*** for $\mathrm{Var}(\widehat{p})$.

  For example, if $\mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}$, then $\widehat{\mathrm{Var}}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n}$.

  Because $\widehat{\mathrm{Var}}(\widehat{p})$ is a function of the random variable $\widehat{p}$, $\widehat{\mathrm{Var}}(\widehat{p})$ is also a ***random variable.*** Typically, we use it only for calculating its numerical value and transforming this into a standard error and a confidence interval as described on the previous page.

Chapter 3: Modelling with Discrete Probability Distributions

In Chapter 2 we introduced several fundamental ideas: hypothesis testing, likelihood, estimators, expectation, and variance. Each of these was illustrated by the Binomial distribution. We now introduce several other discrete distributions and discuss their properties and usage. First we revise Bernoulli trials and the Binomial distribution.

Bernoulli Trials

A set of Bernoulli trials is a series of trials such that:

i) each trial has only 2 possible outcomes: *Success* and *Failure*;
ii) the probability of success, p, is constant for all trials;
iii) the trials are independent.

Examples: 1) Repeated tossing of a fair coin: each toss is a Bernoulli trial with $\mathbb{P}(\text{success}) = \mathbb{P}(\text{head}) = \frac{1}{2}$.

2) Having children: each child can be thought of as a Bernoulli trial with outcomes $\{\text{girl, boy}\}$ and $\mathbb{P}(\text{girl}) = 0.5$.

3.1 Binomial distribution

Description: $X \sim \text{Binomial}(n, p)$ if X is the ***number of successes out of a fixed number*** n ***of Bernoulli trials, each with*** $\mathbb{P}($***success***$) = p$.

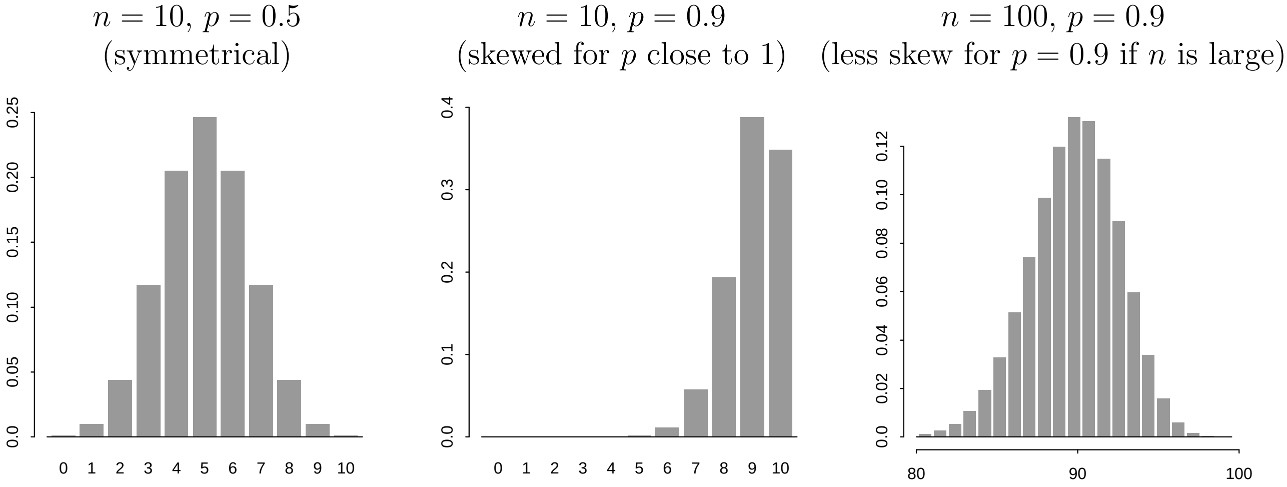
Probability function: $f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ ***for*** $x = 0, 1, \ldots, n$.

Mean: $\mathbb{E}(X) = np$.

Variance: *Var*$(X) = np(1-p)$.

Sum of independent Binomials: If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$, and if X and Y are *independent*, and if X and Y both share the same parameter p, then $X + Y \sim$ ***Binomial***$(n + m,\ p)$.

Shape: Usually symmetrical unless p *is close to 0 or 1.*
Peaks at approximately np.


$n = 10$, $p = 0.5$ (symmetrical)

$n = 10$, $p = 0.9$ (skewed for p close to 1)

$n = 100$, $p = 0.9$ (less skew for $p = 0.9$ if n is large)

3.2 Geometric distribution

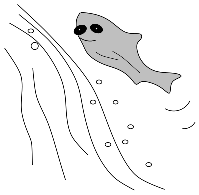
Like the Binomial distribution, the Geometric distribution is defined in terms of a sequence of Bernoulli trials.

- The Binomial distribution counts the *number of <u>successes</u> out of a <u>fixed</u> number of trials.*
- The Geometric distribution counts the *number of <u>trials</u> before the first success occurs.*

This means that the Geometric distribution counts the *number of <u>failures</u> before the first success.*

If every trial has probability p of success, we write: $X \sim$ *Geometric*(p).

Examples: 1) X =number of boys before the first girl in a family:
$X \sim$ *Geometric*$(p = 0.5)$.



2) Fish jumping up a waterfall. On every jump the fish has probability p of reaching the top.
Let X be *the number of <u>failed</u> jumps before the fish succeeds.*
Then $X \sim$ *Geometric*(p).

Properties of the Geometric distribution

i) Description

$X \sim$ Geometric(p) if X is the *number of failures before the first success in a series of Bernoulli trials with* $\mathbb{P}(\textit{success}) = p$.

ii) Probability function

For $X \sim$ Geometric(p),

$$\boxed{f_X(x) = \mathbb{P}(X = x) = (1-p)^x p \text{ for } x = 0, 1, 2, \ldots}$$

Explanation: $\mathbb{P}(X = x) = \underbrace{(1-p)^x}_{\text{need } x \text{ failures}} \quad \times \quad \underbrace{p}_{\text{final trial must be a success}}$

Difference between Geometric and Binomial: For the Geometric distribution, the trials must always occur in the order $\underbrace{FF\ldots F}_{x \text{ failures}}S$.

For the Binomial distribution, failures and successes can occur in any order: e.g. $FF\ldots FS$, $\quad FSF\ldots F$, $\quad SF\ldots F$, etc.

This is why the Geometric distribution has probability function

$$\mathbb{P}(x \text{ failures, } 1 \text{ success}) = (1-p)^x p,$$

while the Binomial distribution has probability function

$$\mathbb{P}(x \text{ failures, } 1 \text{ success}) = \binom{x+1}{x}(1-p)^x p.$$

iii) Mean and variance

For $X \sim$ Geometric(p),

$$\boxed{\begin{aligned} \mathbb{E}(X) &= \frac{1-p}{p} = \frac{q}{p} \\ \text{Var}(X) &= \frac{1-p}{p^2} = \frac{q}{p^2} \end{aligned}}$$

iv) Sum of independent Geometric random variables

If $X_1, \ldots, X_k$ are *independent*, and each $X_i \sim$ Geometric(p), then

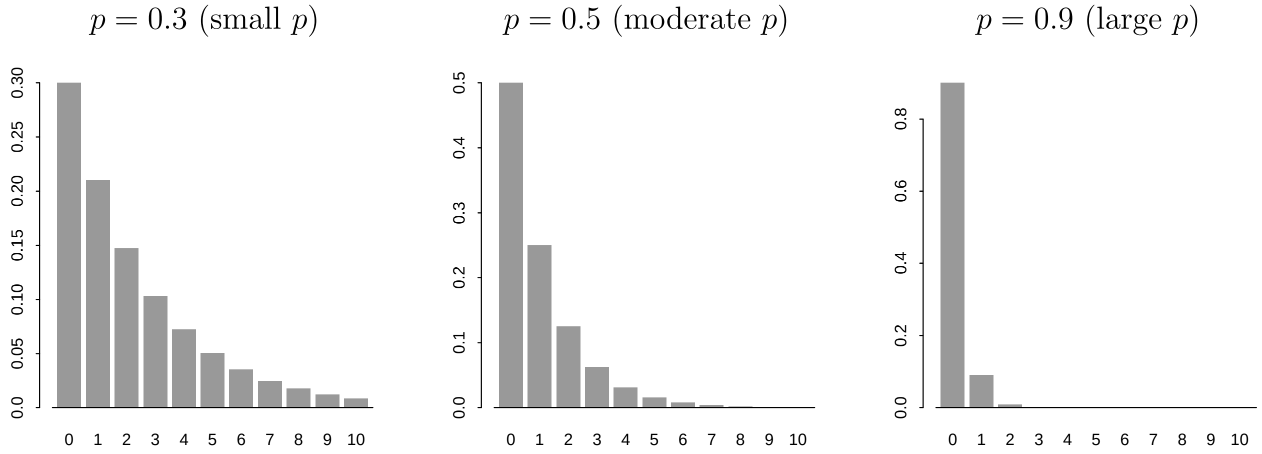
$$X_1 + \ldots + X_k \sim \textit{Negative Binomial}(k, p). \quad \textit{(see later)}$$

v) Shape

Geometric probabilities are always greatest at $x = 0$.
The distribution always has a *long right tail (positive skew).*

The length of the tail depends on p. For small p, there could be many failures before the first success, so the tail is *long.*

For large p, a success is likely to occur almost immediately, so the tail is *short.*



$p = 0.3$ (small p) $p = 0.5$ (moderate p) $p = 0.9$ (large p)

vi) Likelihood

For any random variable, the likelihood function is just the probability function expressed as a function of the unknown parameter. If:

- $X \sim$ Geometric(p);
- p is *unknown;*
- the observed value of X is x*;*

then the likelihood function is: $L(p\,; x) = p(1-p)^x$ *for* $0 < p < 1$.

Example: we observe a fish making 5 *failed* jumps before reaching the top of a waterfall. We wish to estimate the probability of success for each jump.

$$\textit{Then} \quad L(p\,; 5) = p(1-p)^5 \quad \textit{for } 0 < p < 1.$$

Maximize L *with respect to* p *to find the MLE,* $\widehat{p}$*.*

For mathematicians: proof of Geometric mean and variance formulae (non-examinable)

We wish to prove that $\mathbb{E}(X) = \frac{1-p}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$ when $X \sim \text{Geometric}(p)$.

We use the following results:

$$\sum_{x=1}^{\infty} x q^{x-1} = \frac{1}{(1-q)^2} \qquad \text{(for } |q| < 1\text{)}, \tag{3.1}$$

and

$$\sum_{x=2}^{\infty} x(x-1) q^{x-2} = \frac{2}{(1-q)^3} \qquad \text{(for } |q| < 1\text{)}. \tag{3.2}$$

Proof of (3.1) and (3.2):

Consider the infinite sum of a geometric progression:

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} \qquad \text{(for } |q| < 1\text{)}.$$

Differentiate both sides with respect to q:

$$\begin{aligned} \frac{d}{dq}\left(\sum_{x=0}^{\infty} q^x\right) &= \frac{d}{dq}\left(\frac{1}{1-q}\right) \\ \sum_{x=0}^{\infty} \frac{d}{dq}(q^x) &= \frac{1}{(1-q)^2} \\ \sum_{x=1}^{\infty} x q^{x-1} &= \frac{1}{(1-q)^2}, \qquad \text{as stated in (3.1).} \end{aligned}$$

Note that the lower limit of the summation becomes $x = 1$ because the term for $x = 0$ vanishes.

The proof of (3.2) is obtained similarly, by differentiating both sides of (3.1) with respect to q (Exercise).

Now we can find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^{\infty} x\mathbb{P}(X = x) \\
&= \sum_{x=0}^{\infty} xpq^x \qquad \text{(where } q = 1 - p\text{)} \\
&= p\sum_{x=1}^{\infty} xq^x \qquad \text{(lower limit becomes } x = 1 \text{ because term in } x = 0 \text{ is zero)} \\
&= pq\sum_{x=1}^{\infty} xq^{x-1} \\
&= pq\left(\frac{1}{(1-q)^2}\right) \qquad \text{(by equation (3.1))} \\
&= pq\left(\frac{1}{p^2}\right) \qquad \text{(because } 1 - q = p\text{)} \\
&= \frac{q}{p}, \quad \text{as required.}
\end{aligned}$$

For $\mathrm{Var}(X)$, we use

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}\left\{X(X-1)\right\} + \mathbb{E}(X) - (\mathbb{E}X)^2\,. \qquad (\star)$$

Now

$$\begin{aligned}
\mathbb{E}\{X(X-1)\} &= \sum_{x=0}^{\infty} x(x-1)\mathbb{P}(X = x) \\
&= \sum_{x=0}^{\infty} x(x-1)pq^x \qquad \text{(where } q = 1 - p\text{)} \\
&= pq^2\sum_{x=2}^{\infty} x(x-1)q^{x-2} \quad \text{(note that terms below } x = 2 \text{ vanish)} \\
&= pq^2\left(\frac{2}{(1-q)^3}\right) \qquad \text{(by equation (3.2))} \\
&= \frac{2q^2}{p^2}\,.
\end{aligned}$$

Thus by $(\star)$,

$$\mathrm{Var}(X) = \frac{2q^2}{p^2} + \frac{q}{p} - \left(\frac{q}{p}\right)^2 = \frac{q(q+p)}{p^2} = \frac{q}{p^2}\,,$$

as required, because $q + p = 1$.

3.3 Negative Binomial distribution

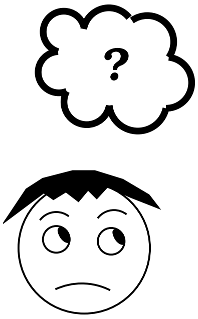
The Negative Binomial distribution is a generalised form of the Geometric distribution:

- the Geometric distribution counts the number of *failures before the first success;*
- the Negative Binomial distribution counts the number of *failures before the k'th success.*

If every trial has probability p of success, we write: $X \sim$ *NegBin*(k, p).

Examples: 1) X =number of boys before the second girl in a family:
$X \sim$ *NegBin*$(k = 2,\ p = 0.5)$.

2) Tom needs to pass 24 papers to complete his degree. He passes each paper with probability p, independently of all other papers. Let X be *the number of papers Tom fails in his degree.*



Then $X \sim$ *NegBin*$(24, p)$.

Properties of the Negative Binomial distribution

i) Description

$X \sim$ NegBin$(k,\ p)$ if X is the *number of failures before the k'th success in a series of Bernoulli trials with $\mathbb{P}$(success) $= p$.*

ii) Probability function

For $X \sim$ NegBin$(k,\ p)$,

$$\boxed{f_X(x) = \mathbb{P}(X = x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \textit{for} \quad x = 0, 1, 2, \ldots}$$

Explanation:

- For $X = x$, we need x failures and k successes.
- The trials stop when we reach the k'th success, so the last trial must be *a success.*
- This leaves x failures and $k-1$ successes to occur in *any order:* a total of $k-1+x$ trials.

For example, if $x = 3$ failures and $k = 2$ successes, we could have:

$$FFFS\underline{S} \quad FFSF\underline{S} \quad FSFF\underline{S} \quad SFFF\underline{S}$$

So:

$$\mathbb{P}(X = x) = \underbrace{\binom{k+x-1}{x}}_{\substack{(k-1) \textit{ successes and } x \textit{ failures} \\ \textit{out of } (k-1+x) \textit{ trials.}}} \times \overbrace{p^k}^{k \textit{ successes}} \times \underbrace{(1-p)^x}_{x \textit{ failures}}$$

iii) Mean and variance

For $X \sim \text{NegBin}(k,\, p)$,

$$\boxed{\begin{aligned} \mathbb{E}(X) &= \frac{k(1-p)}{p} = \frac{kq}{p} \\ \textit{Var}(X) &= \frac{k(1-p)}{p^2} = \frac{kq}{p^2} \end{aligned}}$$

These results can be proved from the fact that the Negative Binomial distribution is obtained as the sum of k independent Geometric random variables:

$$\begin{aligned} X &= Y_1 + \ldots + Y_k, \quad \text{where each } Y_i \sim \text{Geometric}(p), \quad Y_i \text{ indept}, \\ \Rightarrow \quad \mathbb{E}(X) &= k\mathbb{E}(Y_i) = \frac{kq}{p}, \\ \text{Var}(X) &= k\text{Var}(Y_i) = \frac{kq}{p^2}. \end{aligned}$$

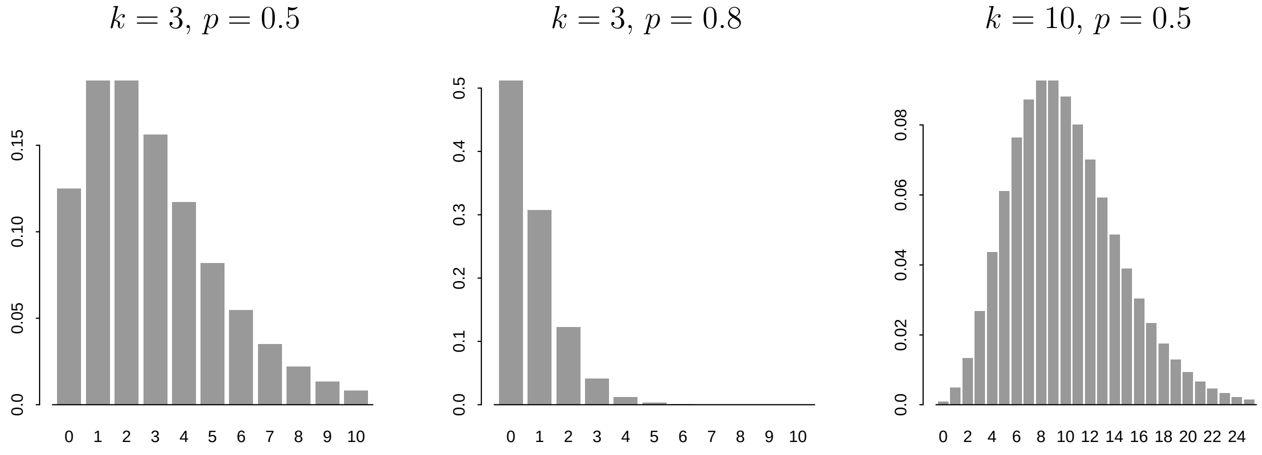
iv) Sum of independent Negative Binomial random variables

If X and Y are *independent,*
and $X \sim \text{NegBin}(k,\, p)$, $Y \sim \text{NegBin}(m,\, p)$, with the same value of p, then

$$X + Y \sim \textit{NegBin}(k+m, p).$$

v) Shape

The Negative Binomial is flexible in shape. Below are the probability functions for various different values of k and p.


$k = 3,\, p = 0.5$ $\quad$ $k = 3,\, p = 0.8$ $\quad$ $k = 10,\, p = 0.5$

vi) Likelihood

As always, the likelihood function is the probability function expressed as a function of the unknown parameters. If:

- $X \sim \text{NegBin}(k,\ p)$;
- k is *known;*
- p is *unknown;*
- the observed value of X is x*;*

then the likelihood function is:

$$L(p\,; x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \textit{for } 0 < p < 1.$$

Example: Tom fails a total of 4 papers before finishing his degree. What is his pass probability for each paper?

X =# failed papers before 24 passed papers: $X \sim$ NegBin$(24, p)$.

Observation: $X = 4$ failed papers.
Likelihood:

$$L(p\,; 4) = \binom{24+4-1}{4} p^{24}(1-p)^4 = \binom{27}{4} p^{24}(1-p)^4 \quad \textit{for } 0 < p < 1.$$

Maximize L with respect to p to find the MLE, $\widehat{p}$.

3.4 Hypergeometric distribution: sampling without replacement

The hypergeometric distribution is used when we *are sampling without replacement from a finite population.*

i) Description

Suppose we have N objects:

- M of the N objects are *special;*
- the other $N - M$ objects are *not special.*

We remove n objects *at random without replacement.*

Let $X =$ *number of the* n *removed objects that are special.*

Then $X \sim$ *Hypergeometric*$(N,\ M,\ n)$.

Example: Ron has a box of Chocolate Frogs. There are 20 chocolate frogs in the box. Eight of them are dark chocolate, and twelve of them are white chocolate.

Ron grabs a random handful of 5 chocolate frogs and stuffs them into his mouth when he thinks that noone is looking. Let X be the number of dark chocolate frogs he picks.

Then $X \sim$ *Hypergeometric*$(N = 20,\ M = 8,\ n = 5)$.

ii) Probability function

For $X \sim \text{Hypergeometric}(N, M, n)$,

$$f_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for $x = \max(0, n + M - N)$ to $x = \min(n, M)$.

Explanation: We need to choose x special objects and $n-x$ other objects.

- Number of ways of selecting x special objects from the M available is: $\binom{M}{x}$.
- Number of ways of selecting $n-x$ other objects from the $N-M$ available is: $\binom{N-M}{n-x}$.
- Total number of ways of choosing x special objects and $(n-x)$ other objects is: $\binom{M}{x} \times \binom{N-M}{n-x}$.
- Overall number of ways of choosing n objects from N is: $\binom{N}{n}$.

Thus:

$$\mathbb{P}(X=x) = \frac{\textit{number of desired ways}}{\textit{total number of ways}} = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} .$$

Note: We need $0 \le x \le M$ (number of special objects),
and $0 \le n-x \le N-M$ (number of other objects).
After some working, this gives us the stated constraint that

$$x = \max(0, n+M-N) \text{ to } x = \min(n, M).$$

Example: What is the probability that Ron selects 3 white and 2 dark chocolates?

X =# dark chocolates. There are $N = 20$ chocolates, including $M = 8$ dark chocolates. We need

$$\mathbb{P}(X=2) = \frac{\binom{8}{2}\binom{12}{3}}{\binom{20}{5}} = \frac{28 \times 220}{15504} = 0.397 .$$

iii) Mean and variance

For $X \sim \text{Hypergeometric}(N, M, n)$,

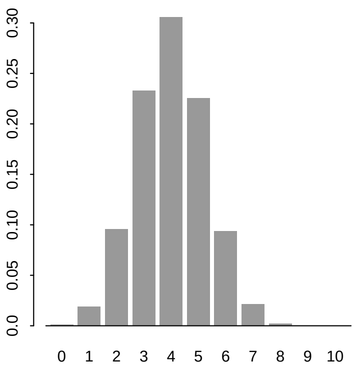
$$\boxed{\begin{aligned} &\mathbb{E}(X) = np \\ &\text{Var}(X) = np(1-p)\left(\frac{N-n}{N-1}\right) \end{aligned}} \qquad \text{where } p = \frac{M}{N}.$$

iv) Shape

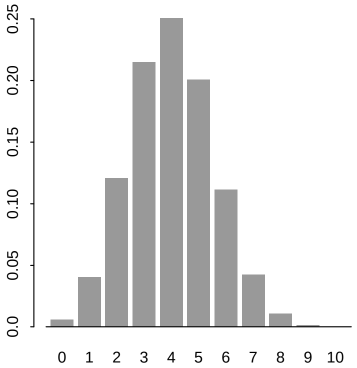
The Hypergeometric distribution is similar to the Binomial distribution when n/N is small, because removing n objects does not change the overall composition of the population very much when n/N is small.

For $n/N < 0.1$ we often approximate the Hypergeometric(N, M, n) distribution by the ***Binomial***$(n, p = \frac{M}{N})$ ***distribution.***

Hypergeometric$(30, 12, 10)$



Binomial$(10, \frac{12}{30})$



Note: The Hypergeometric distribution can be used for opinion polls, because these involve sampling without replacement from a finite population.

The Binomial distribution is used when the population is sampled ***with replacement.***

As noted above, Hypergeometric$(N, M, n) \to$ Binomial$(n, \frac{M}{N})$ as $N \to \infty$.

A note about distribution names

Discrete distributions often get their names from mathematical power series.

- Binomial probabilities sum to 1 because of the Binomial Theorem:

$$\Big(p + (1-p)\Big)^n = \text{<sum of Binomial probabilities>} = 1.$$

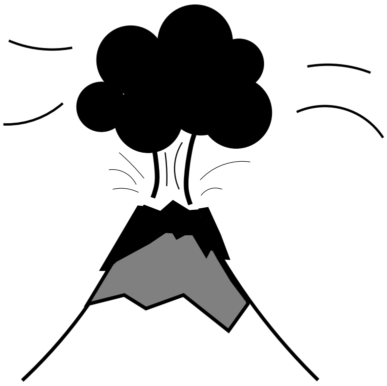
- Negative Binomial probabilities sum to 1 by the Negative Binomial expansion: i.e. the Binomial expansion with a negative power, $-k$:

$$p^k\Big(1 - (1-p)\Big)^{-k} = \text{<sum of NegBin probabilities>} = 1.$$

- Geometric probabilities sum to 1 because they form a Geometric series:

$$p\sum_{x=0}^{\infty}(1-p)^x = \text{<sum of Geometric probabilities>} = 1.$$

3.5 Poisson distribution



When is the next volcano due to erupt in Auckland?

Any moment now, unfortunately!
(give or take 1000 years or so...)

A volcano could happen in Auckland this afternoon, or it might not happen for another 1000 years. Volcanoes are almost impossible to predict: they seem to happen completely at random.

A distribution that counts the *number of random events in a fixed space of time is the Poisson distribution.*

How many cars will cross the Harbour Bridge today? $X \sim$ *Poisson*.
How many road accidents will there be in NZ this year? $X \sim$ *Poisson*.
How many volcanoes will erupt over the next 1000 years? $X \sim$ *Poisson*.

The Poisson distribution arose from a mathematical formulation called the Poisson Process, published by Siméon-Denis Poisson in 1837.

Poisson Process

The Poisson process counts the *number of events occurring in a fixed time or space, when events occur independently and at a constant average rate.*

Example: Let X be the number of road accidents in a year in New Zealand. Suppose that:

i) all accidents are *independent of each other;*

ii) accidents occur at a *constant average rate of λ per year;*

iii) accidents *cannot occur simultaneously.*

Then the number of accidents in a year, X, has the distribution

$$X \sim \textit{Poisson}(\lambda).$$

Number of accidents in one year

Let X be the number of accidents to occur in one year: $X \sim$ *Poisson*(λ).

The probability function for $X \sim \text{Poisson}(\lambda)$ is

$$\boxed{\mathbb{P}(X = x) = \frac{\lambda^x}{x!}e^{-\lambda} \quad \textit{for } x = 0, 1, 2, \ldots}$$

Number of accidents in t years

Let X_t be the number of accidents to occur in time t *years.*

Then $X_t \sim$ *Poisson*(λt),

and

$$\boxed{\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!}e^{-\lambda t} \quad \textit{for } x = 0, 1, 2, \ldots}$$

General definition of the Poisson process

Take any sequence of random events such that:

i) all events are *independent;*

ii) events occur at a *constant average rate of* λ *per unit time;*

iii) events *cannot occur simultaneously.*

Let X_t *be the number of events to occur in time* t.

Then $X_t \sim$ *Poisson*(λt),

and

$$\boxed{\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!}e^{-\lambda t} \quad \textit{for } x = 0, 1, 2, \ldots}$$

Note: For a Poisson process in space, let X_A = *# events in area of size A.*
Then $X_A \sim$ *Poisson*(λA).

Example: X_A = number of raisins in a volume A of currant bun.

Where does the Poisson formula come from?

(Sketch idea, for mathematicians; non-examinable).
The formal definition of the Poisson process is as follows.

Definition: The random variables $\{X_t : t > 0\}$ form a Poisson process with rate λ if:

i) events occurring in any time interval are independent of those occurring in any other disjoint time interval;

ii)

$$\lim_{\delta t \downarrow 0} \left(\frac{\mathbb{P}(\text{exactly one event occurs in time interval}[t, t+\delta t])}{\delta t} \right) = \lambda\,;$$

iii)

$$\lim_{\delta t \downarrow 0} \left(\frac{\mathbb{P}(\text{more than one event occurs in time interval}[t, t+\delta t])}{\delta t} \right) = 0\,.$$

These conditions can be used to derive a partial differential equation on a function known as the *probability generating function* of X_t. The partial differential equation is solved to provide the form $\mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}$.

Poisson distribution

The Poisson distribution is not just used in the context of the Poisson process. It is also used in many other situations, often as a *subjective model* (see Section 3.7). Its properties are as follows.

i) Probability function

For $X \sim \text{Poisson}(\lambda)$,

$$\boxed{f_X(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \textit{for} \quad x = 0, 1, 2, \ldots}$$

The parameter λ is called the ***rate*** of the Poisson distribution.

ii) Mean and variance

The mean and variance of the Poisson(λ) distribution are both λ.

$$\boxed{\mathbb{E}(X) = \textit{Var}(X) = \lambda \quad \textit{when} \quad X \sim \textit{Poisson}(\lambda).}$$

Notes:

1. It makes sense for $\mathbb{E}(X) = \lambda$: by definition, λ is the *average* number of events per unit time in the Poisson process.

2. The variance of the Poisson distribution increases with the mean (in fact, variance = mean). This is often the case in real life: there is more uncertainty associated with larger numbers than with smaller numbers.

iii) Sum of independent Poisson random variables

If X and Y are ***independent***, and $X \sim$ Poisson(λ), $Y \sim$ Poisson(μ), then

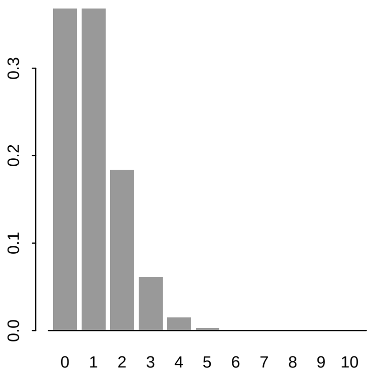
$$X + Y \sim \textit{Poisson}(\lambda + \mu).$$

iv) Shape

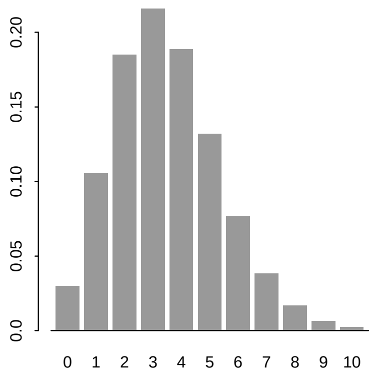
The shape of the Poisson distribution depends upon the value of λ. For small λ, the distribution has positive (right) skew. As λ increases, the distribution becomes more and more symmetrical, until for large λ it has the familiar bell-shaped appearance.

The probability functions for various λ are shown below.

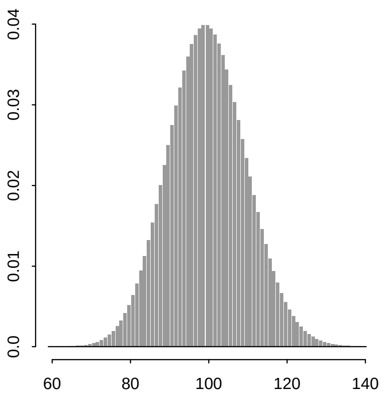
$\lambda = 1$



$\lambda = 3.5$



$\lambda = 100$



v) Likelihood and Estimator Variance

As always, the likelihood function is the probability function expressed as a function of the unknown parameters. If:

- $X \sim \text{Poisson}(\lambda)$;
- λ is *unknown;*
- the observed value of X is x*;*

then the likelihood function is:

$$L(\lambda;\, x) = \frac{\lambda^x}{x!}e^{-\lambda} \quad \textit{for } 0 < \lambda < \infty.$$

Example: 28 babies were born in Mt Roskill yesterday. Estimate λ.

Let X =# babies born in a day in Mt Roskill. Assume that $X \sim$ Poisson(λ).

Observation: $X = 28$ babies.
Likelihood:

$$L(\lambda\,; 28) = \frac{\lambda^{28}}{28!}e^{-\lambda} \quad \textit{for } 0 < \lambda < \infty.$$

Maximize L with respect to λ to find the MLE, $\hat{\lambda}$.
We find that $\hat{\lambda} = x = 28$.
Similarly, the maximum likelihood *estimator* of λ is: $\hat{\lambda} = X$.

Thus the estimator variance is:

Var$(\hat{\lambda})$ = *Var*$(X) = \lambda$*, because $X \sim$ Poisson(λ).*

Because we don't know λ, we have to *estimate* the variance:

$\widehat{\textit{Var}}(\hat{\lambda}) = \hat{\lambda}$.

vi) R command for the p-value:

If $X \sim \text{Poisson}(\lambda)$, then the R command for $\mathbb{P}(X \leq x)$ is *ppois(x, lambda).*

Proof of Poisson mean and variance formulae (non-examinable)

We wish to prove that $\mathbb{E}(X) = \text{Var}(X) = \lambda$ for $X \sim \text{Poisson}(\lambda)$.

For $X \sim \text{Poisson}(\lambda)$, the probability function is $f_X(x) = \dfrac{\lambda^x}{x!}e^{-\lambda}$ for $x = 0, 1, 2, \ldots$

So

$$\begin{aligned}
E(X) = \sum_{x=0}^{\infty} x f_X(x) &= \sum_{x=0}^{\infty} x \left(\frac{\lambda^x}{x!} e^{-\lambda} \right) \\
&= \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} e^{-\lambda} \qquad \text{(note that term for } x = 0 \text{ is 0)} \\
&= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} \quad \text{(writing everything in terms of } x-1\text{)} \\
&= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \qquad \text{(putting } y = x-1\text{)} \\
&= \lambda, \quad \text{because the sum=1 (sum of Poisson probabilities)} .
\end{aligned}$$

So $\mathbb{E}(X) = \lambda$, as required.

For $\text{Var}(X)$, we use:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\
&= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}X)^2 \\
&= \mathbb{E}[X(X-1)] + \lambda - \lambda^2.
\end{aligned}$$

$$\begin{aligned}
\text{But } \mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x}{x!} e^{-\lambda} \\
&= \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} e^{-\lambda} \quad \text{(terms for } x = 0 \text{ and } x = 1 \text{ are 0)} \\
&= \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda} \quad \text{(writing everything in terms of } x-2\text{)} \\
&= \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \qquad \text{(putting } y = x-2\text{)} \\
&= \lambda^2.
\end{aligned}$$

So

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[X(X-1)] + \lambda - \lambda^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda, \quad \text{as required.}
\end{aligned}$$

3.6 Likelihood and log-likelihood for n independent observations

So far, we have seen how to calculate the maximum likelihood estimator in the case of a *single observation made from a distribution:*

- $Y \sim \text{Binomial}(n, p)$ where n is known and p is to be estimated. *Maximum likelihood estimator:* $\widehat{p} = \frac{Y}{n}$.
- $Y \sim \text{Geometric}(p)$. *Maximum likelihood estimator:* $\widehat{p} = \frac{1}{Y+1}$.
- $Y \sim \text{NegBin}(k, p)$ where k is known and p is to be estimated. *Maximum likelihood estimator:* $\widehat{p} = \frac{k}{Y+k}$.
- $Y \sim \text{Poisson}(\lambda)$. *Maximum likelihood estimator:* $\widehat{\lambda} = Y$.

Question: What would we do if we had n independent observations, $Y_1, Y_2, \ldots, Y_n$?

Answer: As usual, the likelihood function is defined as the *probability of the data, expressed as a function of the unknown parameter.*

If the data consist of several independent observations, their probability is gained by *multiplying the individual probabilities together.*

Example: Suppose we have observations $Y_1, Y_2, \ldots, Y_n$ where each $Y_i \sim \text{Poisson}(\lambda)$, and $Y_1, \ldots, Y_n$ are independent. Find the maximum likelihood estimator of λ.

Before we start, what would you guess $\widehat{\lambda}$ to be in this situation?

Solution: *For observations* $Y_1 = y_1, \ldots, Y_n = y_n$, *the likelihood is:*

$$\begin{aligned}
L(\lambda\,; y_1, \ldots, y_n) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) \text{ under parameter } \lambda \\
&= \mathbb{P}(Y_1 = y_1 \cap Y_2 = y_2 \cap \ldots \cap Y_n = y_n) \\
&= \mathbb{P}(Y_1 = y_1)\mathbb{P}(Y_2 = y_2)\ldots\mathbb{P}(Y_n = y_n) \text{ by independence} \\
&= \prod_{i=1}^{n} \left(\frac{\lambda^{y_i}}{y_i\,!} e^{-\lambda} \right) \\
&= \left(\prod_{i=1}^{n} \frac{1}{y_i\,!} \right) \left(e^{-\lambda}\right)^n \lambda^{(y_1+y_2+\ldots+y_n)} \\
&= K e^{-n\lambda} \lambda^{n\overline{y}}\,.
\end{aligned}$$

So $$L(\lambda\,; y_1, \ldots, y_n) = K e^{-n\lambda}\, \lambda^{n\overline{y}},$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and $K = \prod_{i=1}^{n} \frac{1}{y_i\,!}$ is a constant that doesn't depend on λ.

Differentiate $L(\lambda\,; y_1, \ldots, y_n)$ and set to 0 to find the MLE:

$$\begin{aligned} 0 &= \frac{d}{d\lambda} L(\lambda\,; y_1, \ldots, y_n) \\ &= K\left\{-ne^{-n\lambda}\,\lambda^{n\overline{y}} + (n\overline{y})\,e^{-n\lambda}\,\lambda^{(n\overline{y}-1)}\right\} \\ &= Ke^{-n\lambda}\lambda^{(n\overline{y}-1)}\left\{-n\lambda + (n\overline{y})\right\} \\ \Rightarrow &\quad \lambda = \infty, \qquad \lambda = 0, \qquad \textit{or} \quad \lambda = \overline{y}. \end{aligned}$$

If we know that $L(\lambda\,; y_1, \ldots, y_n)$ reaches a unique maximum in $0 < \lambda < \infty$, for example *by reference to a graph,* then we can deduce that the MLE is $\overline{y}$.

So the maximum likelihood estimator is:
$$\widehat{\lambda} = \overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}.$$

Note: When $n = 1$, we get the same result as we had before: $\widehat{\lambda} = \frac{Y_1}{1} = Y_1$.

Log-likelihood

Instead of maximizing the likelihood function L to find the MLE, we often take logs and maximize the log-likelihood function, $\log L$. (***Note:*** $\log = \log_e = \ln$.)

There are several reasons for using the log-likelihood:

1. The logarithmic function $L \mapsto \log(L)$ is ***increasing,*** so the functions $L(\lambda)$ and $\log\{L(\lambda)\}$ will *have the same maximum,* $\widehat{\lambda}$.

2. When there are observations $Y_1, \ldots, Y_n$, the likelihood L is a product. Because $\log(ab) = \log(a) + \log(b)$, the log-likelihood *converts the product into a sum.* It is often easier to differentiate a sum than a product, so the log-likelihood is easier to maximize while still giving the same MLE.

3. If we need to use a computer to calculate and maximize the likelihood, there will often be numerical problems with computing the likelihood product, whereas the log-likelihood sum can be accurately calculated.

Example: Suppose we have observations $Y_1, Y_2, \ldots, Y_n$ where each $Y_i \sim \text{Poisson}(\lambda)$, and $Y_1, \ldots, Y_n$ are independent, as before. Use the ***log-likelihood function*** to find the maximum likelihood estimator of λ, and show that you get the same answer $\widehat{\lambda} = \overline{Y}$ as we obtained by maximizing the likelihood function directly.

Solution: *For observations $Y_1 = y_1, \ldots, Y_n = y_n$, the likelihood is:*

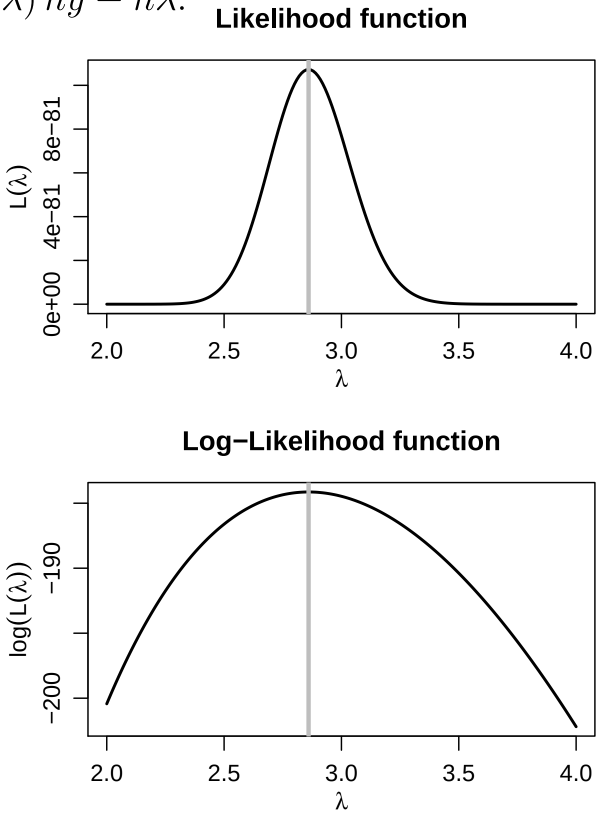
$$\begin{aligned}
L(\lambda\,; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \left(\frac{\lambda^{y_i}}{y_i\,!} e^{-\lambda} \right) \quad \textit{(by independence)} \\
\textit{So } \log\{L(\lambda\,; y_1, \ldots, y_n)\} &= \sum_{i=1}^{n} \log \left(\frac{\lambda^{y_i}}{y_i\,!} e^{-\lambda} \right) \\
&= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{y_i\,!} \right) + \log\left(\lambda^{y_i}\right) + \log\left(e^{-\lambda}\right) \right\} \\
&= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{y_i\,!} \right) + y_i \log(\lambda) + (-\lambda) \right\} \\
&= K' + \log(\lambda) \sum_{i=1}^{n} y_i - n\lambda \qquad \textit{where } K' \textit{ is a constant} \\
&= K' + \log(\lambda)\, n\overline{y} - n\lambda.
\end{aligned}$$

Differentiate and set to 0 for the MLE:

$$\begin{aligned}
0 &= \frac{d}{d\lambda} \log\{L(\lambda\,; y_1, \ldots, y_n)\} \\
0 &= \frac{d}{d\lambda} \{K' + \log(\lambda)\, n\overline{y} - n\lambda\} \\
\Rightarrow 0 &= \frac{n\overline{y}}{\lambda} - n \\
\Rightarrow \widehat{\lambda} &= \overline{y},
\end{aligned}$$

assuming a unique maximum in $0 < \lambda < \infty$.

So the MLE is $\widehat{\lambda} = \overline{Y}$ as before.



$L(\lambda)$ and $\log\{L(\lambda)\}$ for $n = 100$, $\overline{y} = 2.86$.

3.7 Subjective modelling

Most of the distributions we have talked about in this chapter are *exact* models for the situation described. For example, the Binomial distribution describes *exactly* the distribution of the number of successes in n Bernoulli trials.

However, there is often no exact model available. If so, we will use a ***subjective model.***

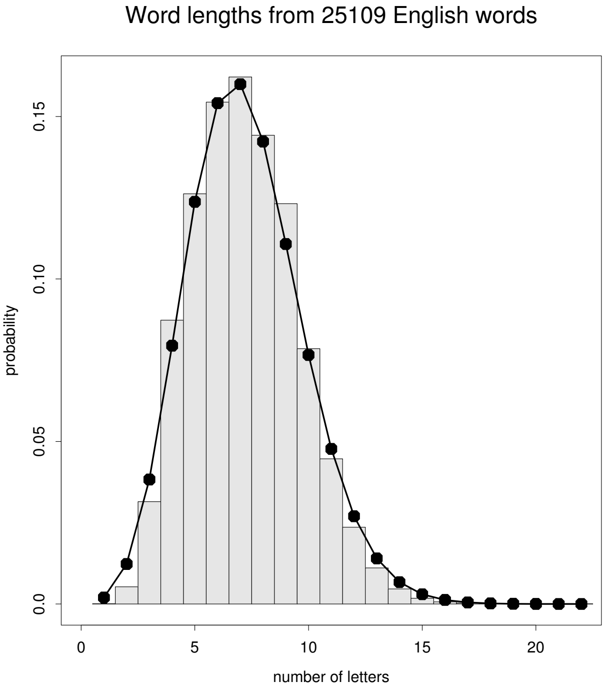
In a subjective model, we pick a probability distribution to describe a situation ***just because it has properties that we think are appropriate to the situation, such as the right sort of symmetry or skew, or the right sort of relationship between variance and mean.***

Example: Distribution of word lengths for English words.

Let $Y =$ ***number of letters in an English word chosen at random from the dictionary.***

If we plot the frequencies on a barplot, we see that ***the shape of the distribution is roughly Poisson.***

English word lengths: $Y - 1 \sim \text{Poisson}(6.22)$



The Poisson probabilities (with λ estimated by maximum likelihood) are plotted as points overlaying the barplot.

We need to use $Y \sim 1 + \text{Poisson}$ because Y cannot take the value 0.

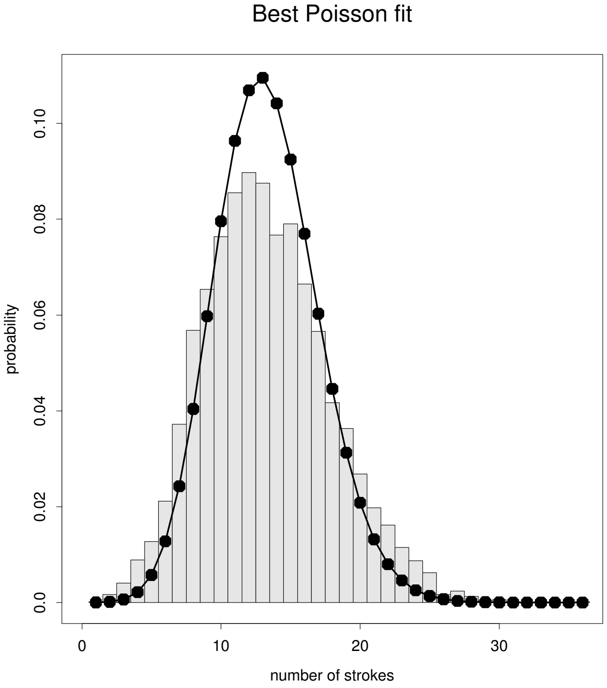
The fit of the Poisson distribution is ***quite good.***

In this example we can not say that the Poisson distribution represents the number of events in a fixed time or space: *instead, it is being used as a subjective model for word length.*

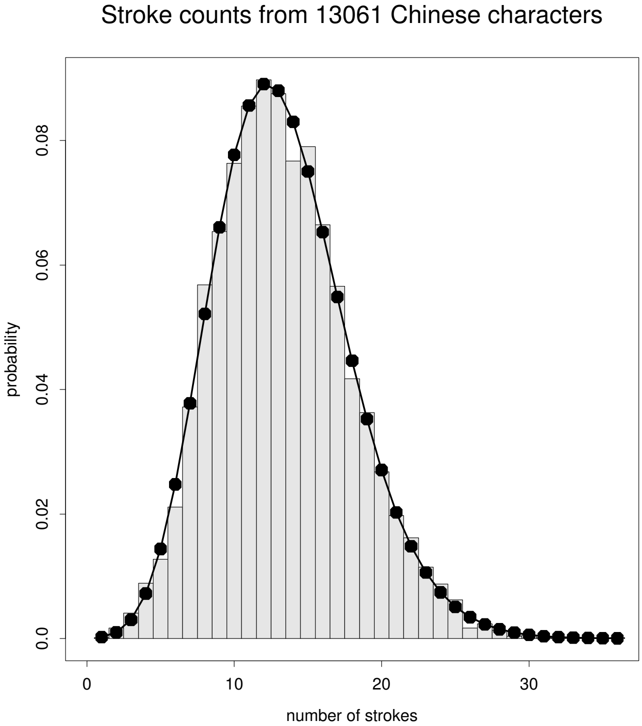
Can a Poisson distribution fit any data? The answer is *no: in fact the Poisson distribution is very inflexible.*

Here are stroke counts from 13061 Chinese characters. Y is the number of strokes in a randomly chosen character. The best-fitting Poisson distribution *(found by MLE)* is overlaid.

The fit of the Poisson distribution is *awful.*



It turns out, however, that the Chinese stroke distribution is well-described by *a Negative Binomial model.*



The best-fitting Negative Binomial distribution *(found by MLE)* is NegBin$(k = 23.7, p = 0.64)$. The fit is *very good.*

However, Y does not represent the number of failures before the k'th success: the Negative Binomial is a *subjective model.*

3.8 Statistical regression modelling

Statistical regression modelling is a fundamental technique used in data analysis in science and business. In this section we give an introduction to the idea of regression modelling, using the simplest example of modelling a straight line through the origin of a scatterplot.

In statistical regression, we explore the ***relationship between two variables.***

One variable, x, is typically *under our control.*

We select several different values of x. At each value of x, we make measurements of the other variable, Y.

The other variable, Y, is regarded as ***random.***

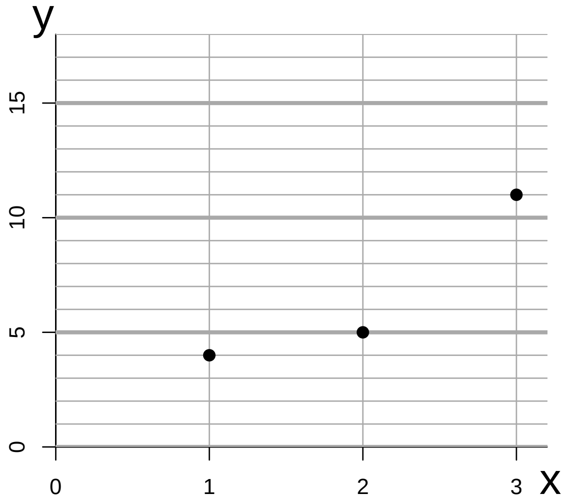
The distribution of Y ***depends upon the value of*** x at which we measure it.

We write (x_i, Y_i) for the i'th pair of measurements, where $i = 1, 2, \ldots, n$.

After the measurements are observed, we use lower-case letters and write (x_i, y_i).

Example: Where would you draw the best-fit line through the origin?

(x_1, y_1)	$= (1, 4)$
(x_2, y_2)	$= (2, 5)$
(x_3, y_3)	$= (3, 11)$



- x is called the ***predictor variable,*** because *it predicts the distribution of* Y.
- Y is called the ***response variable,*** because *it is observed in response to selecting a particular value of* x.
- You might sometimes see x called the 'independent variable' and Y called the 'dependent variable'. Although it is widely used, this terminology is confusing because x is not independent of Y in a statistical sense. Most statisticians avoid this language and use the terms ***predictor*** and ***response*** instead.

How does the distribution of Y depend upon x?

In regression modelling, we generally assume that *the MEAN of Y has some relationship with the value of x.*

The simplest regression model is a straight line through the origin. In this model, we assume that:

$$\mathbb{E}(Y) = \beta x\,,$$

where the ***slope parameter*** β is what we want to estimate.

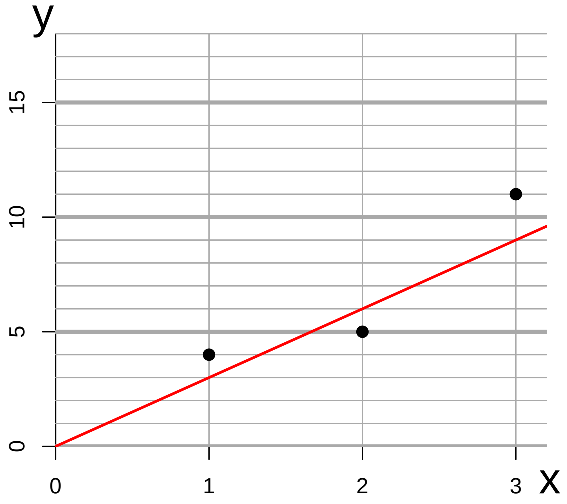
More specifically, in each of the pairs (x_i, Y_i) for $i = 1, \ldots, n$, we assume the same relationship $\mathbb{E}(Y_i) = \beta x_i$.

- The parameter β stays the same for all $i = 1, \ldots, n$. It gives the *slope of the best-fit line through the origin.*
- The mean of Y changes as x changes. *When x is large, Y has a larger mean than when x is small (assuming β is positive). The mean of Y is $\mathbb{E}(Y_i) = \beta x_i$.*

The line has equation $y = \beta x$.

(Here, $\beta = 3$.)

The line shows the MEAN of the distribution of Y at each point x.



Why do we want to fit a line to these points?

Our main interest is in the ***relationship*** between x and Y. In regression through the origin, this relationship is captured by the slope of the line, β.

Example:

- x represents some level of experience: e.g. $x = 1$ *for children in their first year of school, $x = 2$ for 2nd-years, etc.*
- Y represents some sort of achievement: e.g. Y *could be a reading score or numeracy score.*
- The slope of the line, β, tells us about *the improvement in children's scores from one year to the next.*
- The school needs to prove that β is sufficiently high, and not 0 or negative!

Statistical model for Y

So far we have only specified the relationship between x and the mean of Y: $\mathbb{E}(Y_i) = \beta x_i$.

In order to estimate the slope β, we need to specify *the whole distribution of* Y.

Example 1: *Let* $Y_i \sim$ *Poisson*(βx_i). *Then* $\mathbb{E}(Y_i) = \beta x_i$ *for each* $i = 1, \ldots, n$.

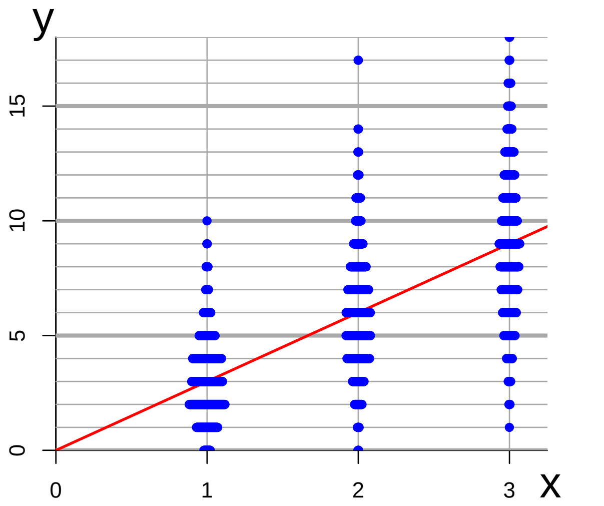
This model could be suitable if Y_i measures a ***count*** of some item that depends upon x_i and has no upper limit. In the school example, Y_i could be a number of achievements or credits accumulated over the years.

As another example, a university student could create a model in which x_i is the percentage course credit awarded for an assignment, and Y_i is the time in hours spent on the assignment. We would expect more time to be spent on assignments with higher credit, but there will be randomness (scatter) about the straight-line relationship. A student might use this model to *look for outliers, to decide whether a particular assignment takes an unreasonably long time for the amount of credit awarded!*

Properties of $Y_i \sim$ Poisson(βx_i):

- Y_i *takes values* $0, 1, 2, \ldots$ *with no upper limit.*
- *Var*$(Y_i) = \mathbb{E}(Y_i) = \beta x_i$, *so variance increases with the mean.*

Scatter of Y~Poisson(βx) about the mean



It is often appropriate to allow the variance to increase with the mean. If you estimate an assignment is going to take you 1 hour, you are unlikely to be wrong by 10 hours: there is ***low variance*** about a mean of $\mathbb{E}(Y) = 1$. Conversely, if you estimate the assignment will take 100 hours, it could easily take 10 hours more or less: there is ***higher variance*** about the mean of $\mathbb{E}(Y) = 100$.

Although the Poisson distribution allows variance to increase with the mean, it also makes a very specific assumption about the increase: *under the Poisson distribution, the variance is always EQUAL to the mean.*

This assumption is often good enough, but equally it is often too restrictive. The usual problem is that the Poisson distribution doesn't allow the variance to increase enough as the mean gets larger. If so, modellers often use a more flexible distribution such as the *Negative Binomial.*

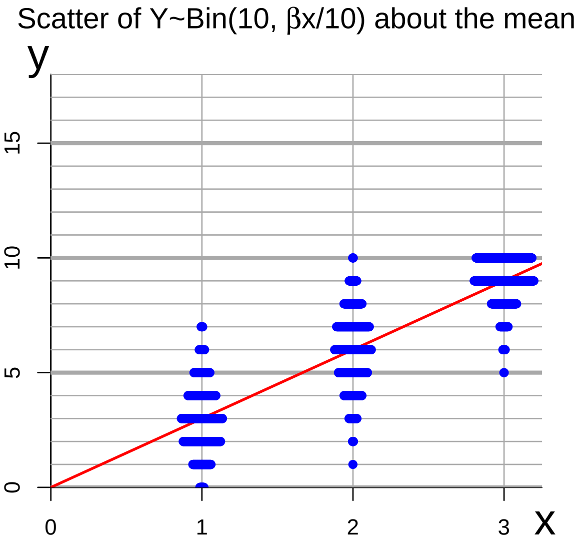
Example 2: *If* $Y_i \sim$ *Binomial*$(n = 10, p = \frac{\beta x_i}{10})$*, then* $\mathbb{E}(Y_i) = n \times p = \beta x_i$ *for each* i*.*

(Note: we could use $p = \gamma x_i$ instead of $p = \frac{\beta x_i}{10}$. We use $\frac{\beta x_i}{10}$ so that we can compare the distribution of Y_i between Examples 1, 2, and 3 with the same value of β.)

This model could be suitable if Y_i measures a ***score out of 10*** on some test. In the school example, we might expect older children (larger x_i) to achieve a higher score than younger children on the same test.

Properties of $\boldsymbol{Y_i \sim}$ Binomial $\left(\mathbf{10}, \frac{\boldsymbol{\beta x_i}}{\mathbf{10}}\right)$:

- Y_i takes values $0, 1, 2, \ldots, 10$ *with a strict upper limit at 10.*
- $\text{Var}(Y_i) = np(1-p) = \beta x_i \left(1 - \frac{\beta x_i}{10}\right)$, *which is respectively 2.1, 2.4, 0.9 when* $\beta = 3$ *and* $x = 1, 2, 3$. *The variance becomes small as* $\mathbb{E}(Y)$ *gets close to the upper limit of 10.*



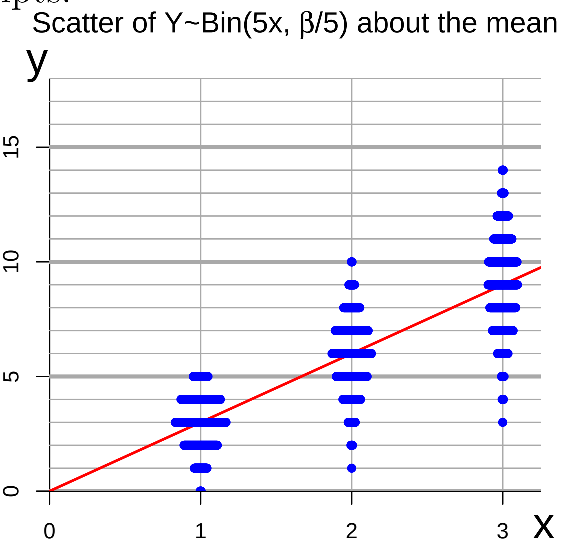
Example 3: *If* $Y_i \sim$ *Binomial*$(n = 5x_i, p = \frac{\beta}{5})$*, then* $\mathbb{E}(Y_i) = n \times p = \beta x_i$ *for each* i*.*

(We use the peculiar formulation $p = \frac{\beta}{5}$ so we can keep the same value of β to compare with Examples 1 & 2.)

For example, a person at a fairground can pay $\$x$, corresponding to \$1, \$2, or \$3, to get respectively 5, 10, or 15 chances to throw a ball through a net. Their winnings are related to Y_i, the number of times they succeed in throwing the ball through the net out of their $5x_i$ attempts.

Properties of $\boldsymbol{Y_i \sim}$ Binomial $\left(\boldsymbol{5x_i}, \frac{\boldsymbol{\beta}}{\mathbf{5}}\right)$:

- Y_i takes values $0, 1, 2, \ldots, 5x_i$ *with an upper limit at* $5x_i$*.*
- $\text{Var}(Y_i) = np(1-p) = \beta x_i \left(1 - \frac{\beta}{5}\right)$, which is respectively 1.2, 2.4, 3.6 when $\beta = 3$ and $x = 1, 2, 3$. *The variance increases with* $\mathbb{E}(Y)$ *but (unusually)* Var(Y) *is smaller than* $\mathbb{E}(Y)$*.*



Difference between statistical regression and our previous models

- In section 3.6, we had n independent random observations $Y_1, \ldots, Y_n$. These observations were ***drawn from the same distribution:*** they were ***independent, identically distributed (iid).*** In the example in section 3.6, each $Y_i \sim \text{Poisson}(\lambda)$, and we wanted to estimate the common parameter λ.
- In statistical regression, we again have n independent random variables $Y_1, \ldots, Y_n$, but this time they have ***different distributions: for example,*** $Y_i \sim$ ***Poisson***(βx_i).

  The different distributions are linked by a common parameter, β, that describes how the distribution of the response variable Y changes as the predictor variable x changes. ***Our interest is in estimating this parameter*** β.

Estimation by maximum likelihood

To estimate the parameter β, we use maximum likelihood as usual.

We assume that the response variables $Y_1, \ldots, Y_n$ are ***independent, conditional on the corresponding predictor variables*** $x_1, \ldots, x_n$.

For observations $Y_1 = y_1, \ldots, Y_n = y_n$, the likelihood is:

$$\begin{aligned} L(\beta\,; y_1, \ldots, y_n) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \,|\, x_1, \ldots, x_n\,; \beta) \\ &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \,|\, x_i\,; \beta) \qquad \textbf{\textit{by independence.}} \end{aligned}$$

The log-likelihood is:

$$\begin{aligned} \log\left\{L(\beta\,; y_1, \ldots, y_n)\right\} &= \log\left\{\prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \,|\, x_i\,; \beta)\right\} \\ &= \sum_{i=1}^{n} \log\left\{\mathbb{P}(Y_i = y_i \,|\, x_i\,; \beta)\right\}. \end{aligned}$$

We maximize the likelihood (or more often, the log-likelihood) with respect to β as usual. The only difference from typical likelihood maximization is that ***we have to remember that*** $\mathbb{P}(Y_i = y_i \,|\, x_i\,; \beta)$ ***is different for every different value of*** x_i.

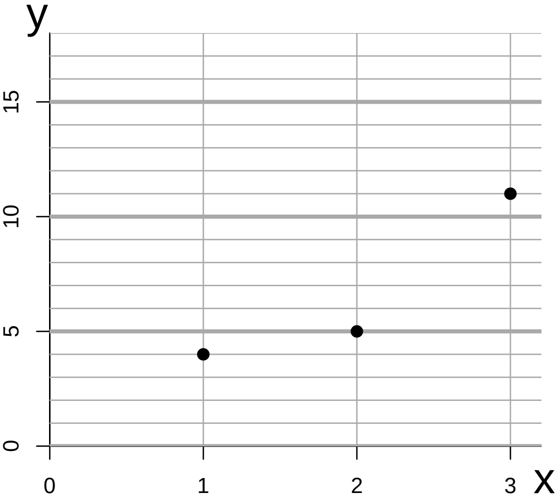
Example: Poisson regression

Recall the scenario shown at the beginning of this section:

$(x_1, y_1) = (1, 4)$

$(x_2, y_2) = (2, 5)$

$(x_3, y_3) = (3, 11)$



Consider the model $Y_i \sim \text{Poisson}(\beta x_i)$, for $i = 1, \ldots, n$.
Maximize the likelihood to find the maximum likelihood estimator, $\widehat{\beta}$.

Also find the exact variance, $\text{Var}(\widehat{\beta})$ in terms of the unknown parameter β, and suggest a suitable estimator $\widehat{\text{Var}}(\widehat{\beta})$ for the variance.

Evaluate $\widehat{\beta}$ and $\widehat{\text{Var}}(\widehat{\beta})$ for the data shown above, where $n = 3$ and $x_i = i$ for $i = 1, 2, 3$. Mark your estimated best-fit line on the graph shown.

Solution: For $Y_i \sim \text{Poisson}(\beta x_i)$, the likelihood is (from the previous page):

$$\begin{aligned}
L(\beta\,; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i \,|\, x_i\,; \beta) \\
&= \prod_{i=1}^{n} \frac{(\beta x_i)^{y_i}}{y_i\,!} e^{-\beta x_i} \\
&= \left(\prod_{i=1}^{n} \frac{x_i^{y_i}}{y_i\,!} \right) \beta^{(y_1 + \ldots + y_n)}\, e^{-\beta(x_1 + \ldots + x_n)} \\
&= K\, \beta^{(y_1 + \ldots + y_n)}\, e^{-\beta(x_1 + \ldots + x_n)}
\end{aligned}$$

where K is a constant: does not depend upon β

$$= K\, \beta^{n\overline{y}}\, e^{-n\overline{x}\beta}.$$

Differentiate and set to 0 for the MLE:

$$\begin{aligned}
0 &= \frac{d}{d\beta} L(\beta\,; y_1, \ldots, y_n) \\
&= \frac{d}{d\beta}\left\{K\,\beta^{n\overline{y}}\,e^{-n\overline{x}\beta}\right\} \\
&= K\left\{n\overline{y}\beta^{(n\overline{y}-1)}\,e^{-n\overline{x}\beta} - \beta^{n\overline{y}}\,n\overline{x}\,e^{-n\overline{x}\beta}\right\} \\
&= K\beta^{(n\overline{y}-1)}\,e^{-n\overline{x}\beta}\,\left\{n\overline{y} - \beta n\overline{x}\right\} \\
\Rightarrow\quad 0 &= n\overline{y} - \beta n\overline{x} \qquad \text{or } \beta = 0, \infty \\
\Rightarrow\quad \widehat{\beta} &= \frac{\overline{y}}{\overline{x}}, \qquad \textit{assuming a unique maximum in } 0 < \beta < \infty.
\end{aligned}$$

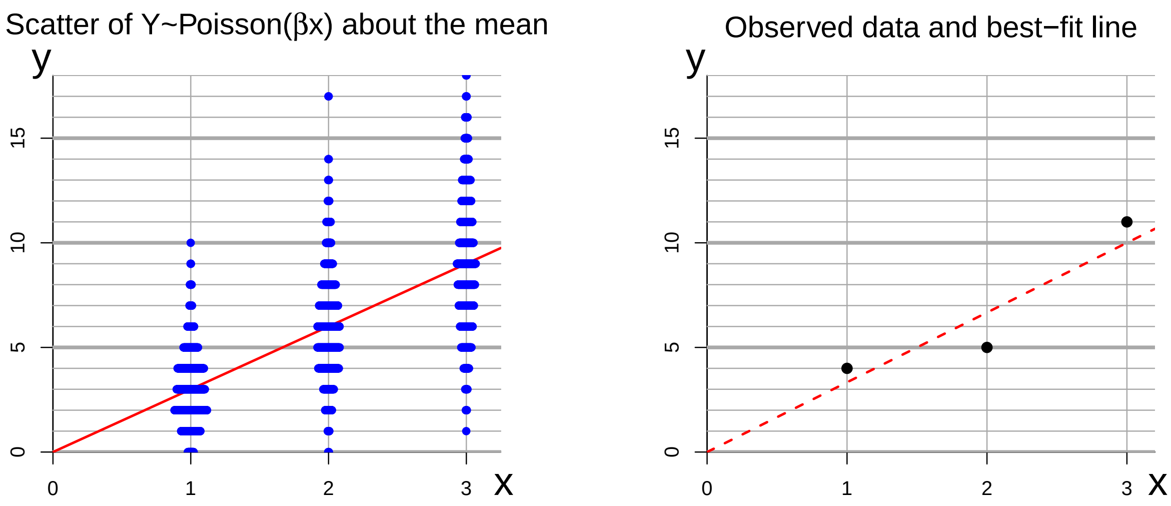
So the MLE is:

$$\widehat{\beta} = \frac{\overline{Y}}{\overline{x}} = \frac{Y_1 + \ldots + Y_n}{x_1 + \ldots + x_n}\,.$$

For the particular case $(x_1, y_1) = (1, 4)\,;\ (x_2, y_2) = (2, 5)\,;\ (x_3, y_3) = (3, 11)$, *we have:*

$$\begin{aligned}
\widehat{\beta} &= \frac{y_1 + y_2 + y_3}{x_1 + x_2 + x_3} \\
&= \frac{4 + 5 + 11}{1 + 2 + 3} \\
&= \frac{20}{6} \\
\Rightarrow\quad \widehat{\beta} &= 3.33.
\end{aligned}$$

Add the best-fit line $y = 3.33x$ *to the graph overleaf by picking two points the line must go through: e.g.* $(x, y) = (0, 0)$ *and* $(x, y) = (3, 10)$.



(a) True line with $\beta = 3$; and true distributions of $Y_i \sim \text{Poisson}(3x_i)$.

(b) Observed data and the estimated best-fit line of $y = \widehat{\beta}x$ using $\widehat{\beta} = 10/3 = 3.33$.

Find the variance, $\text{Var}(\widehat{\beta})$, using $\widehat{\beta} = \dfrac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}$:

$$\widehat{\beta} = \frac{1}{n\overline{x}}(Y_1 + Y_2 + \ldots + Y_n)$$

$$\textit{So} \quad \textit{Var}(\widehat{\beta}) = \left(\frac{1}{n\overline{x}}\right)^2 \left\{ \textit{Var}(Y_1) + \textit{Var}(Y_2) + \ldots + \textit{Var}(Y_n) \right\}$$

by independence of $Y_1, \ldots, Y_n$

$$= \frac{1}{n^2\,\overline{x}^2}(\beta x_1 + \beta x_2 + \ldots + \beta x_n)$$

because $Y_i \sim$ *Poisson*(βx_i) *so Var*$(Y_i) = \beta x_i$

$$= \frac{\beta n\overline{x}}{n^2\,\overline{x}^2}$$

$$= \frac{\beta}{n\overline{x}}\,.$$

Notice that Var$(\widehat{\beta})$ *depends upon the unknown true value of* β*, and that the variance gets smaller as* n *increases: large* n *means large sample sizes, for fixed* $\overline{x}$*.*

Suggested estimator, $\widehat{\textbf{Var}}(\widehat{\beta})$: *Use the obvious one,* $\widehat{\textit{Var}}(\widehat{\beta}) = \dfrac{\widehat{\beta}}{n\overline{x}}$.

For the data above with $\widehat{\beta} = 3.333$: $\widehat{\textit{Var}}(\widehat{\beta}) = \dfrac{3.333}{1+2+3} = \dfrac{3.333}{6} = \dfrac{5}{9} = 0.556\,.$

Chapter 4: Continuous Random Variables

4.1 Introduction

When Mozart performed his opera *Die Entführung aus dem Serail*, the Emperor Joseph II responded wryly, 'Too many notes, Mozart!'

In this chapter we meet a different problem: *too many numbers!*

We have met ***discrete*** random variables, for which we can *list all the values and their probabilities, even if the list is infinite:*

e.g. for $X \sim \text{Geometric}(p)$,

x	0	1	2	...
$f_X(x) = \mathbb{P}(X = x)$	p	pq	pq^2	...

But suppose that X takes values in a *continuous set, e.g.* $[0, \infty)$ *or* $(0, 1)$.

We can't even begin to list all the values that X can take. For example, how would you list all the numbers in the interval $[0,\ 1]$?

- the smallest number is 0, but what is the next smallest? 0.01? 0.0001? 0.0000000001? We just end up talking *nonsense.*

In fact, there are so many numbers in any continuous set that *each of them must have probability 0.*

If there was a probability > 0 for all the numbers in a continuous set, however 'small', there simply wouldn't be enough probability to go round.

> *A continuous random variable takes values in a continuous interval* $(a,\ b)$.
>
> *It describes a continuously varying quantity such as time or height.*
>
> *When X is continuous,* $\mathbb{P}(X = x) = 0$ *for ALL x.*
>
> *The probability function is meaningless.*

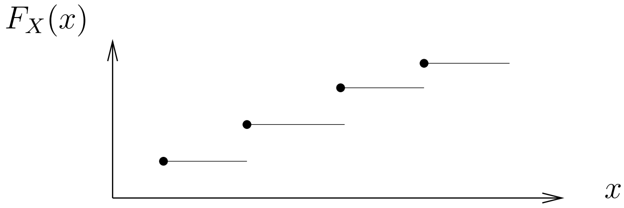
Although we cannot assign a probability to any *value* of X, we *are* able to assign probabilities to *intervals:*
eg. $\mathbb{P}(X = 1) = 0$, *but* $\mathbb{P}(0.999 \leq X \leq 1.001)$ *can be* > 0.

This means we should use *the distribution function,* $F_X(x) = \mathbb{P}(X \leq x)$.

The cumulative distribution function, $F_X(x)$

Recall that for ***discrete*** random variables:

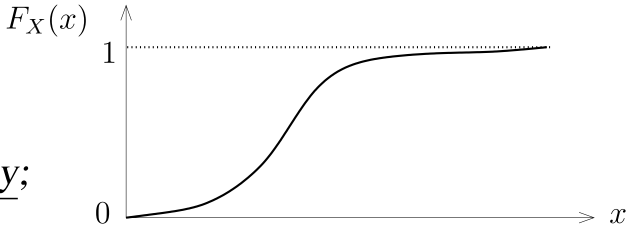
- $F_X(x) = \mathbb{P}(X \leq x)$;
- $F_X(x)$ *is a step function: probability accumulates in discrete steps;*



- $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F(b) - F(a)$.

For a ***continuous random variable:***

- $F_X(x) = \mathbb{P}(X \leq x)$;
- $F_X(x)$ *is a continuous function: probability accumulates continuously;*



- *As before,* $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F(b) - F(a)$.

However, for a *continuous* random variable,

$$\mathbb{P}(X = a) = 0.$$

So it makes *no difference* whether we say $\mathbb{P}(a < X \leq b)$ *or* $\mathbb{P}(a \leq X \leq b)$.

For a continuous random variable,

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a).$$

This is ***not*** true for a discrete random variable: in fact,

For a discrete random variable with values $0, 1, 2, \ldots$,

$$\mathbb{P}(a < X < b) = \mathbb{P}(a + 1 \leq X \leq b - 1) = F_X(b - 1) - F_X(a).$$

Endpoints are not important for continuous r.v.s.
Endpoints are very important for discrete r.v.s.

4.2 The probability density function

Although the cumulative distribution function gives us an interval-based tool for dealing with continuous random variables, it is not very good at telling us what the distribution *looks like.*
For this we use a different tool called the *probability density function.*

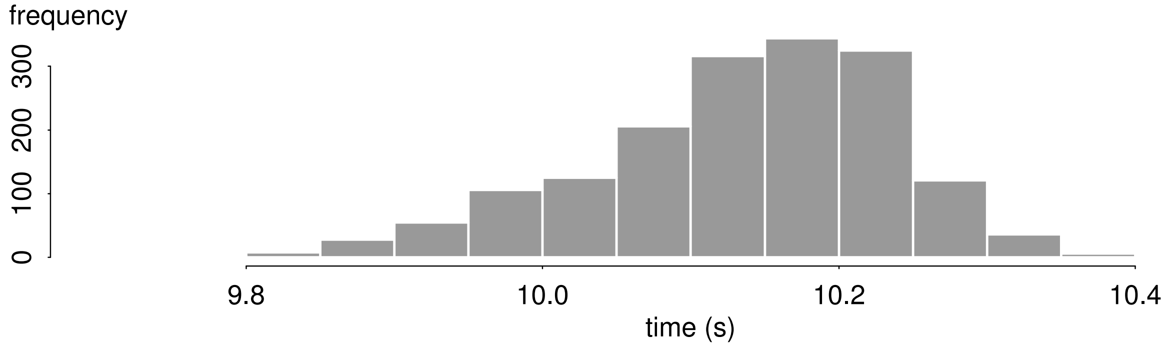
The probability density function (p.d.f.) is the best way to describe and recognise a continuous random variable. We use it all the time to calculate probabilities and to gain an intuitive feel for the shape and nature of the distribution. Using the p.d.f. is like recognising your friends by their faces. You can chat on the phone, write emails or send txts to each other all day, but you never really know a person until you've seen their face.

Just like a cell-phone for keeping in touch, the cumulative distribution function is a tool for facilitating our interactions with the continuous random variable. However, we never really understand the random variable until we've seen its 'face' — the probability density function. Surprisingly, it is quite difficult to describe exactly what the probability density function *is.* In this section we take some time to motivate and describe this fundamental idea.

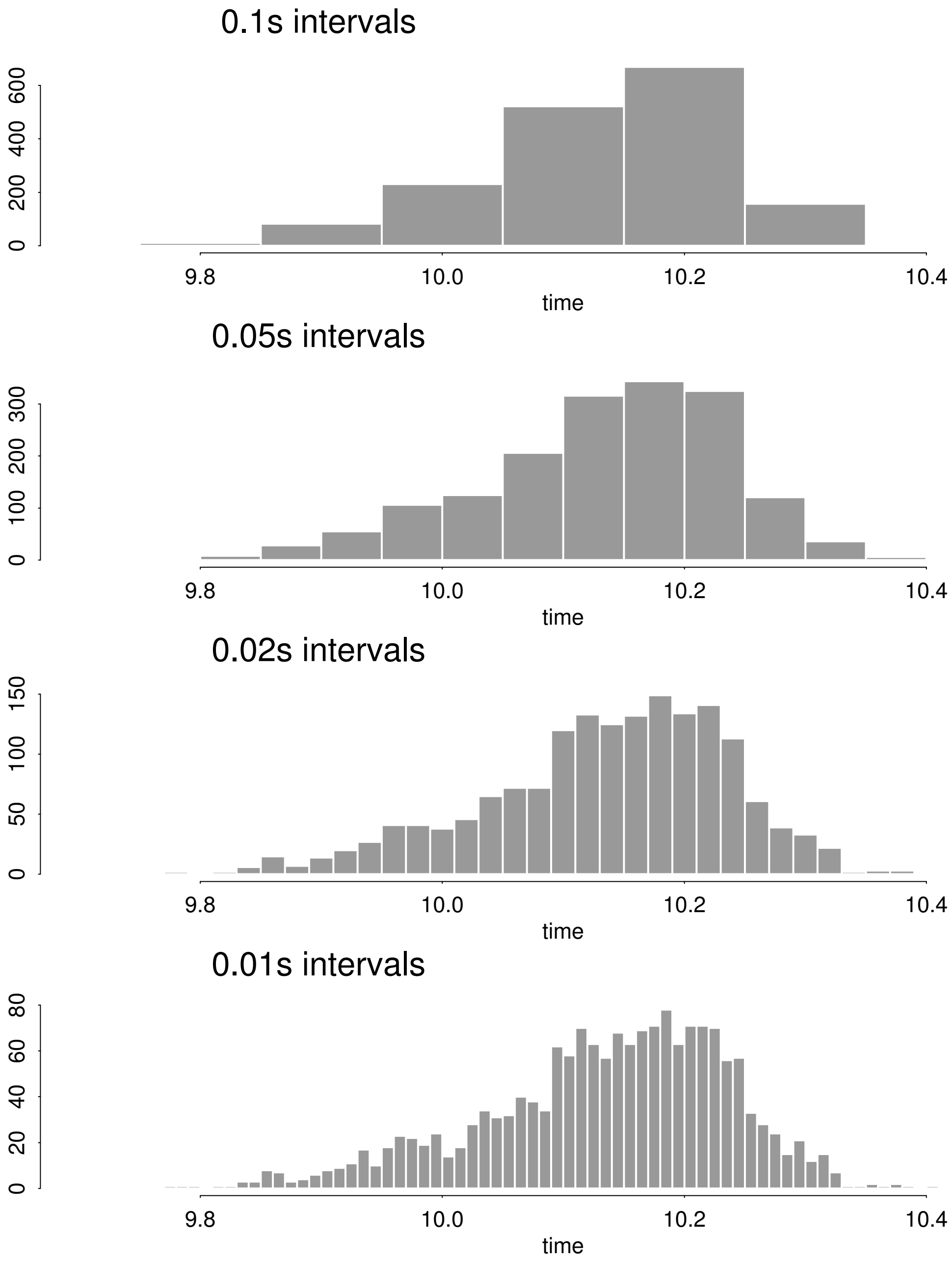
All-time top-ten 100m sprint times

The histogram below shows the best 10 sprint times from the 168 all-time top male 100m sprinters. There are 1680 times in total, representing the top 10 times up to 2002 from each of the 168 sprinters. Out of interest, here are the summary statistics:

```
Min. 1st Qu. Median  Mean 3rd Qu.  Max.
9.78 10.08   10.15  10.14 10.21   10.41
```



We could plot this histogram using different time intervals:



We see that ***each histogram has broadly the same shape, although the heights of the bars and the interval widths are different.***

The histograms tell us the most intuitive thing we wish to know about the distribution: its *shape*:

- the *most probable* times are ***close to 10.2 seconds;***
- the distribution of times has a ***long left tail (left skew);***
- times below 10.0s and above 10.3 seconds have ***low probability.***

We could fit a curve over any of these histograms to show the desired shape, but the problem is that *the histograms are not standardized:*

• every time we change the interval width, *the heights of the bars change.*

How can we derive a curve or function that captures the common shape of the histograms, but keeps a constant height? What should that height be?

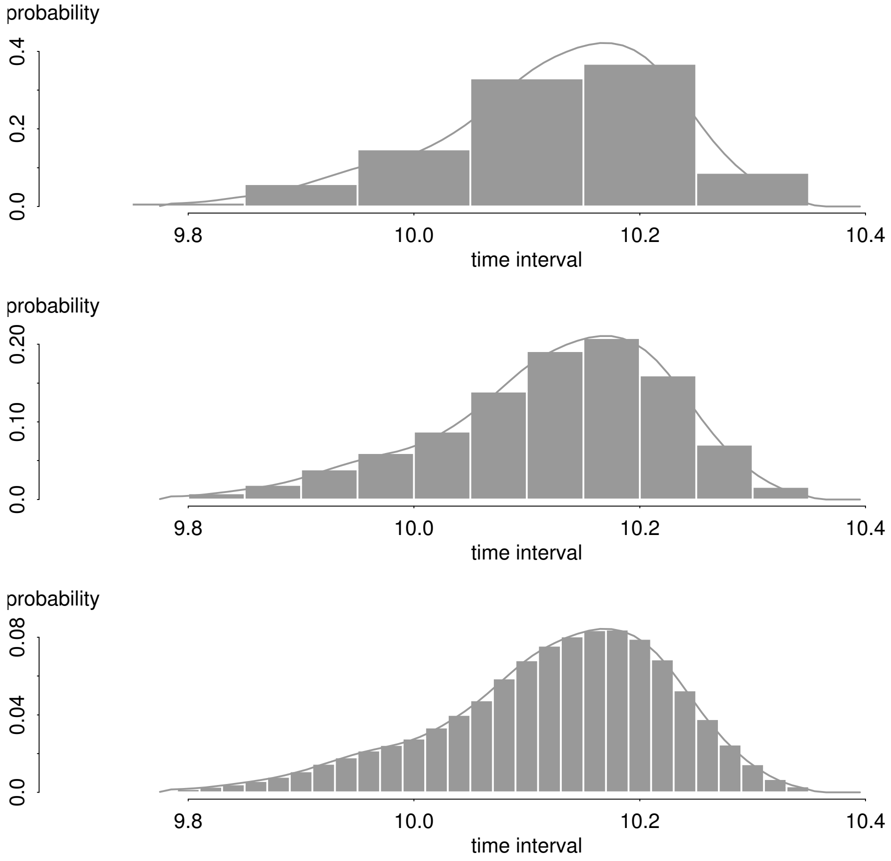
The standardized histogram

We now focus on an *idealized* (smooth) version of the sprint times distribution, rather than using the exact 1680 sprint times observed.

We are aiming to derive a curve, or function, that captures the shape of the histograms, but will keep the same height for any choice of histogram bar width.

First idea: plot the probabilities instead of the frequencies.

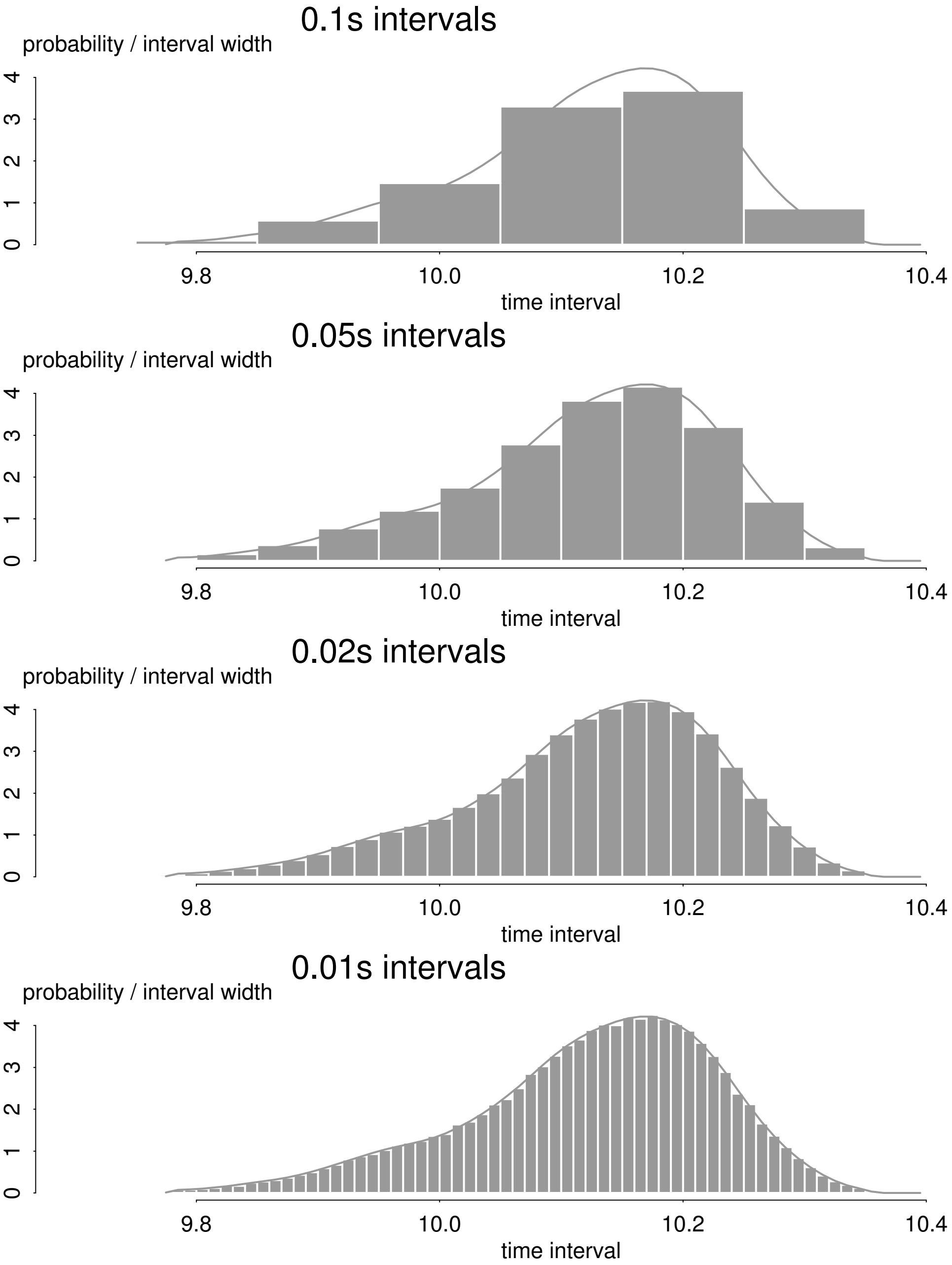
The height of each histogram bar now represents the probability of getting an observation in that bar.



This doesn't work, because *the height (probability) still depends upon the bar width. Wider bars have higher probabilities.*

Second idea: plot the probabilities divided by bar width.

The height of each histogram bar now represents the probability of getting an observation in that bar, divided by the width of the bar.



This seems to be exactly what we need! *The same curve fits nicely over all the histograms and keeps the same height regardless of the bar width.*

These histograms are called *standardized histograms.*

The nice-fitting curve is *the probability density function.*

But. . . what *is* it?!

The probability density function

We have seen that there is a single curve that fits nicely over any standardized histogram from a given distribution.

This curve is called the *probability density function (p.d.f.).*

We will write the p.d.f. of a continuous random variable X as $p.d.f. = f_X(x)$.

The p.d.f. $f_X(x)$ is *NOT the probability of x — for example, in the sprint times we can have $f_X(x) = 4$, so it is definitely NOT a probability.*

However, as the histogram bars of the standardized histogram get narrower, the bars get closer and closer to the p.d.f. curve. The p.d.f. is in fact the *limit of the standardized histogram as the bar width approaches zero.*

What is the height of the standardized histogram bar?

For an interval from x to $x+t$, the standardized histogram plots *the probability of an observation falling between x and $x+t$, divided by the width of the interval, t.*

Thus the height of the standardized histogram bar over the interval from x to $x+t$ is:

$$\frac{\textit{probability}}{\textit{interval width}} = \frac{\mathbb{P}(x \leq X \leq x+t)}{t} = \frac{F_X(x+t) - F_X(x)}{t},$$

where $F_X(x)$ is the cumulative distribution function.

Now consider the limit as the histogram bar width (t) goes to 0: *this limit is DEFINED TO BE the probability density function at x, $f_X(x)$:*

$$f_X(x) = \lim_{t \to 0} \left\{ \frac{F_X(x+t) - F_X(x)}{t} \right\} \qquad \textit{by definition.}$$

This expression should look familiar: it is *the derivative of $F_X(x)$.*

The probability density function (p.d.f.) is therefore *the function*

$$f_X(x) = F_X'(x).$$

It is defined to be a single, unchanging curve that describes the SHAPE of any histogram drawn from the distribution of X.

Formal definition of the probability density function

Definition: Let X be a continuous random variable with distribution function $F_X(x)$. The **probability density function (p.d.f.)** of X is defined as

$$\boxed{f_X(x) = \frac{dF_X}{dx} = F_X'(x).}$$

It gives:

- *the RATE at which probability is accumulating at any given point, $F_X'(x)$;*
- *the SHAPE of the distribution of X.*

Using the probability density function to calculate probabilities

As well as showing us the shape of the distribution of X, the probability density function has another major use:

- *it calculates probabilities by integration.*

Suppose we want to calculate $\mathbb{P}(a \leq X \leq b)$.

We already know that: $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$.

But we also know that:

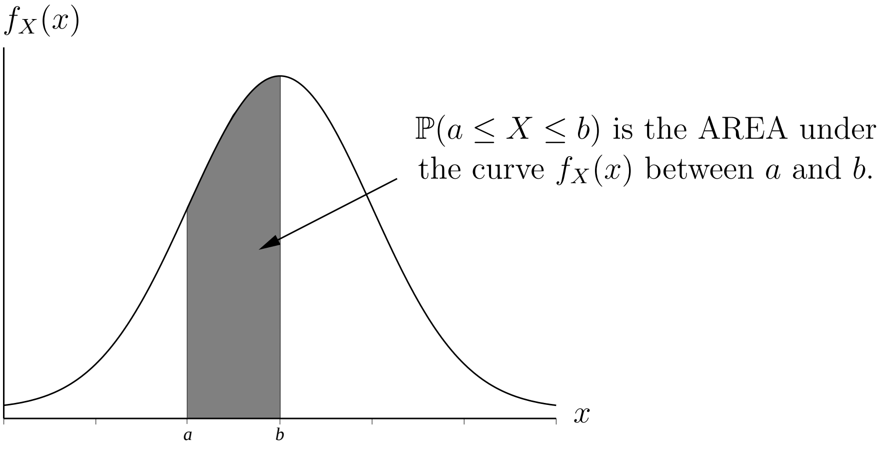
$$\begin{aligned} \frac{dF_X}{dx} &= f_X(x), \\ \text{so} \qquad F_X(x) &= \int f_X(x)\,dx \qquad \text{(without constants).} \\ \textit{In fact:} \qquad F_X(b) - F_X(a) &= \int_a^b f_X(x)\,dx. \end{aligned}$$

This is a very important result:

Let X be a continuous random variable with probability density function $f_X(x)$. Then

$$\boxed{\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \in [\,a, b\,]\,) = \int_a^b f_X(x)\,dx\,.}$$

This means that *we can calculate probabilities by integrating the p.d.f.*

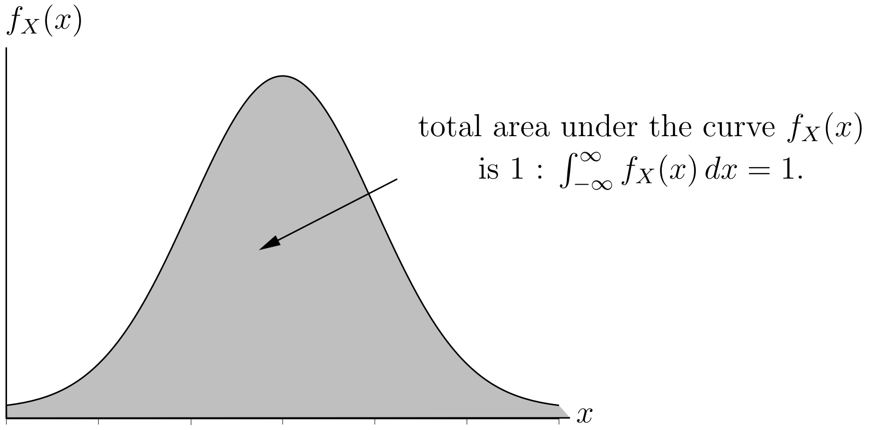


$\mathbb{P}(a \leq X \leq b)$ is the AREA under the curve $f_X(x)$ between a and b.

The **total area under the p.d.f. curve** is:

$$\textit{total area} = \int_{-\infty}^{\infty} f_X(x)\,dx = F_X(\infty) - F_X(-\infty) = 1 - 0 = 1.$$

This says that the total area under the p.d.f. curve is equal to the total probability that X takes a value between $-\infty$ and $+\infty$, which is 1.



total area under the curve $f_X(x)$ is 1 : $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Using the p.d.f. to calculate the distribution function, $F_X(x)$

Suppose we know the probability density function, $f_X(x)$, and wish to calculate the distribution function, $F_X(x)$. We use the following formula:

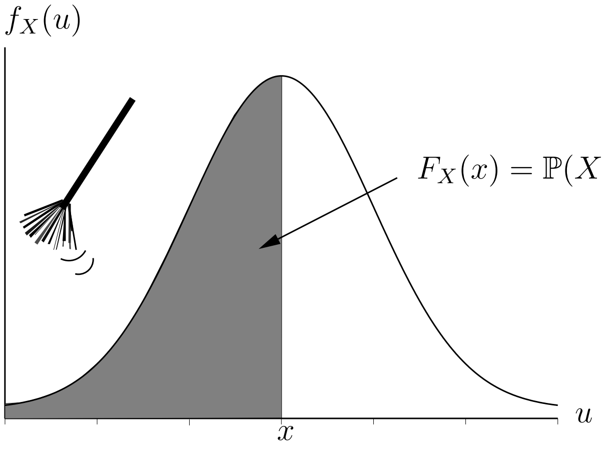
Distribution function, $\boxed{F_X(x) = \int_{-\infty}^{x} f_X(u)\,du.}$

Proof:

$$\int_{-\infty}^{x} f(u)du \quad = \quad F_X(x) - F_X(-\infty) \quad = \quad F_X(x) - 0 \quad = \quad F_X(x).$$

Using the dummy variable, u:

Writing $F_X(x) = \displaystyle\int_{-\infty}^{x} f_X(u)\,du$ means:
integrate $f_X(u)$ as u ranges from $-\infty$ to x.



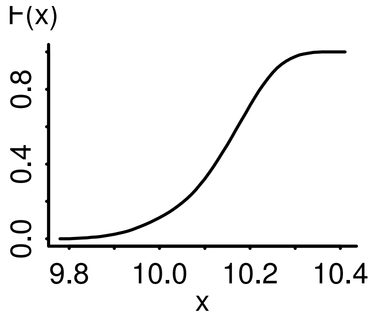
Writing $F_X(x) = \displaystyle\int_{-\infty}^{x} f_X(x)\,dx$ is *WRONG and MEANINGLESS: LOSES A MARK every time.*

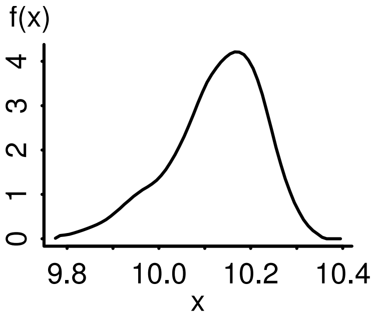
In words, $\int_{-\infty}^{x} f_X(x)\,dx$ means: *integrate $f_X(x)$ as x ranges from $-\infty$ to x. It's nonsense!*

How can x range from $-\infty$ to x?!

Why do we need $f_X(x)$? Why not stick with $F_X(x)$?

These graphs show $F_X(x)$ and $f_X(x)$ from the men's 100m sprint times (X is a random top ten 100m sprint time).



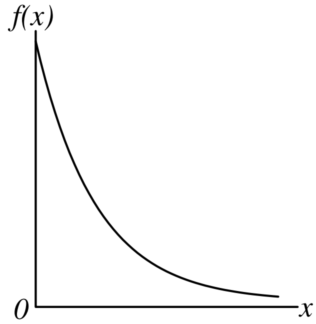


Just using $F_X(x)$ gives us very little intuition about the problem. For example, which is the region of highest probability?

Using the p.d.f., $f_X(x)$, we can see that it is about *10.1 to 10.2 seconds.*

Using the c.d.f., $F_X(x)$, we would have to *inspect the part of the curve with the steepest gradient: very difficult to see.*

Example of calculations with the p.d.f.



Let $\quad f_X(x) = \begin{cases} k\,e^{-2x} & \text{for } 0 < x < \infty, \\ 0 & \text{otherwise.} \end{cases}$

(i) Find the constant k.

(ii) Find $\mathbb{P}(1 < X \leq 3)$.

(iii) Find the cumulative distribution function, $F_X(x)$, for *all* x.

(i) *We need:*

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\,dx &= 1 \\ \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} k\,e^{-2x}\,dx &= 1 \\ k\left[\frac{e^{-2x}}{-2}\right]_0^{\infty} &= 1 \end{aligned}$$

$$\begin{aligned}
\frac{-k}{2}(e^{-\infty} - e^{0}) &= 1 \\
\frac{-k}{2}(0 - 1) &= 1 \\
k &= 2.
\end{aligned}$$

(ii)

$$\begin{aligned}
\mathbb{P}(1 < X \leq 3) &= \int_1^3 f_X(x)\,dx \\
&= \int_1^3 2\,e^{-2x}\,dx \\
&= \left[\frac{2e^{-2x}}{-2}\right]_1^3 \\
&= -e^{-2\times 3} + e^{-2\times 1} \\
&= 0.132.
\end{aligned}$$

(iii)

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(u)\,du \\
&= \int_{-\infty}^{0} 0\,du + \int_0^x 2\,e^{-2u}\,du \quad \text{for } x > 0 \\
&= 0 + \left[\frac{2e^{-2u}}{-2}\right]_0^x \\
&= -e^{-2x} + e^{0} \\
&= 1 - e^{-2x} \quad \text{for } x > 0.
\end{aligned}$$

When $x \leq 0$, $F_X(x) = \int_{-\infty}^{x} 0\,du = 0$.

So overall,

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ 1 - e^{-2x} & \text{for } x > 0. \end{cases}$$

<u>**Total area under the p.d.f. curve is 1:**</u> $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$

<u>**The p.d.f. is NOT a probability:**</u> $f_X(x) \geq 0$ *always,*
but we do NOT require $f_X(x) \leq 1$*.*

<u>**Calculating probabilities:**</u>

1. If you only need to calculate *one* probability $\mathbb{P}(a \leq X \leq b)$: *integrate the p.d.f.:*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx.$$

2. If you will need to calculate *several* probabilities, it is easiest to *find the distribution function,* $F_X(x)$*:*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du.$$

Then use: $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$ *for any* a*,* b*.*

<u>**Endpoints:**</u> *DO NOT MATTER for continuous random variables:*

$$\mathbb{P}(X \leq a) = \mathbb{P}(X < a) \quad \textit{and} \quad \mathbb{P}(X \geq a) = \mathbb{P}(X > a)\,.$$

4.3 The Exponential distribution

When will the next volcano erupt in Auckland? We never quite answered this question in Chapter 3. The Poisson distribution was used to count the *number of volcanoes that would occur in a fixed space of time.*

We have not said *how long* we have to wait for the next volcano: this is a *continuous random variable.*

Auckland Volcanoes

About 50 volcanic eruptions have occurred in Auckland over the last 100,000 years or so. The first two eruptions occurred in the Auckland Domain and Albert Park — right underneath us! The most recent, and biggest, eruption was Rangitoto, about 600 years ago. There have been about 20 eruptions in the last 20,000 years, which has led the Auckland Council to assess current volcanic risk by assuming that volcanic eruptions in Auckland follow a Poisson process with rate $\lambda = \frac{1}{1000}$ volcanoes per year. For background information, see: `www.aucklandcouncil.govt.nz` and search for 'volcanic hazard'.

Distribution of the waiting time in the Poisson process

The length of time between events in the Poisson process is called the *waiting time.*

To find the distribution of a continuous random variable, we often work with the *cumulative distribution function,* $F_X(x)$.

This is because $F_X(x) = \mathbb{P}(X \leq x)$ gives us a *probability*, unlike the p.d.f. $f_X(x)$. We are comfortable with handling and manipulating probabilities.

Suppose that $\{N_t : t > 0\}$ forms a Poisson process with rate $\lambda = \frac{1}{1000}$.

N_t is the *number of volcanoes to have occurred by time t, starting from now.*

We know that

$$N_t \sim \textit{Poisson}(\lambda t)\,; \qquad \textit{so}\ \mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

Let X be a continuous random variable giving the *number of years waited before the next volcano, starting now.* We will derive an expression for $F_X(x)$.

(i) When $x < 0$:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\textit{ less than 0 time before next volcano}) = 0.$$

(ii) When $x \geq 0$:

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \mathbb{P}(\textit{amount of time waited for next volcano is} \leq x) \\ &= \mathbb{P}(\textit{there is at least one volcano between now and time } x) \\ &= \mathbb{P}(\textit{\# volcanoes between now and time } x \textit{ is} \geq 1) \\ &= \mathbb{P}(N_x \geq 1) \\ &= 1 - \mathbb{P}(N_x = 0) \\ &= 1 - \frac{(\lambda x)^0}{0!} e^{-\lambda x} \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Overall: $$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & \textit{for } x \geq 0, \\ 0 & \textit{for } x < 0. \end{cases}$$

The distribution of the waiting time X is called the *Exponential distribution* because of the exponential formula for $F_X(x)$.

Example: What is the probability that there will be a volcanic eruption in Auckland within the next 50 years?

Put $\lambda = \frac{1}{1000}$. *We need* $\mathbb{P}(X \leq 50)$.

$$\mathbb{P}(X \leq 50) = F_X(50) = 1 - e^{-50/1000} = 0.049.$$

There is about a ***5% chance*** that there will be a volcanic eruption in Auckland over the next 50 years. This is the figure given by the Auckland Council at the above web link.

The Exponential Distribution

We have defined the Exponential(λ) distribution to be the distribution of *the waiting time (time between events) in a Poisson process with rate* λ*.*

We write $X \sim$ *Exponential*(λ)*, or* $X \sim$ *Exp*(λ).

However, just like the Poisson distribution, the Exponential distribution has many other applications: it does not always have to arise from a Poisson process.

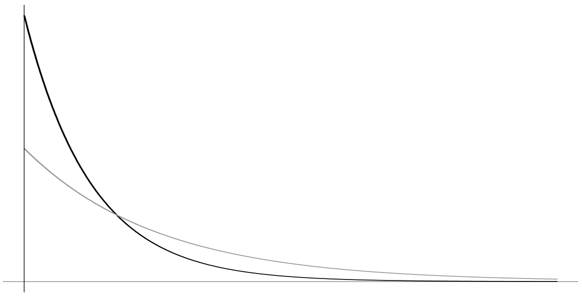
Let $X \sim \text{Exponential}(\lambda)$. *Note:* $\lambda > 0$ *always.*

Distribution function:

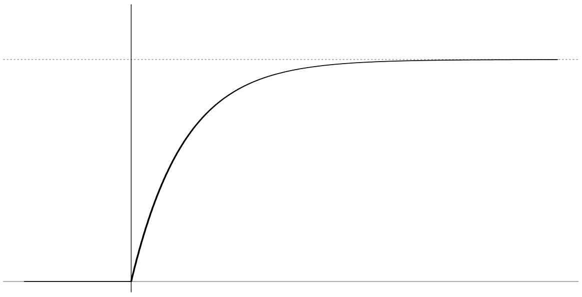
$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Probability density function:

$$f_X(x) = F'_X(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{for } x \geq 0, \\ 0 & \textit{for } x < 0. \end{cases}$$



P.d.f., $f_X(x)$



C.d.f., $F_X(x) = \mathbb{P}(X \leq x)$.

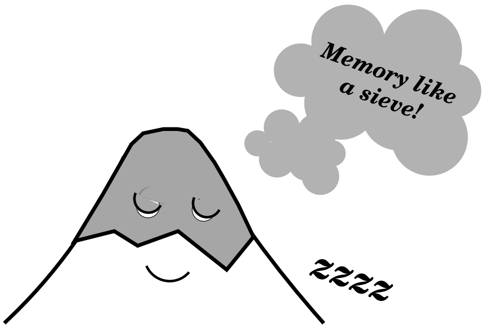
Link with the Poisson process

Let $\{N_t : t > 0\}$ be a Poisson process with rate λ. Then:

- N_t is the number of events to occur by time t;
- $N_t \sim \text{Poisson}(\lambda t)$; so $\mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$;
- Define X to be *either* the time till the first event, *or* the time from now until the next event, *or* the time between any two events.

Then $X \sim$ *Exponential*(λ).
X is called the *waiting time of the process.*

Memorylessness



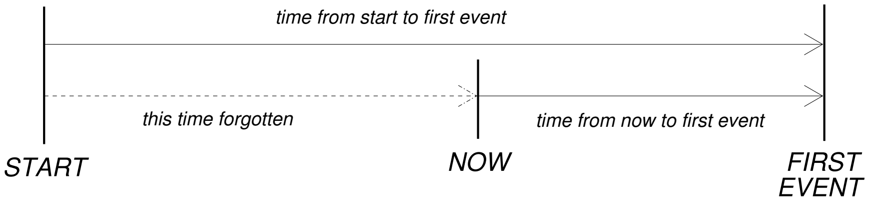
We have said that the waiting time of the Poisson process can be defined *either* as the time from the start to the first event, *or* the time from now until the next event, *or* the time between any two events.

All of these quantities have the ***same distribution:*** $X \sim$ ***Exponential***(λ).

The derivation of the Exponential distribution was valid for all of them, because events occur at a constant average rate in the Poisson process.

This property of the Exponential distribution is called ***memorylessness:***

- the distribution of the time from *now* until the first event is the same as the distribution of the time from *the start* until the first event: ***the time from the start till now has been forgotten!***



The Exponential distribution is famous for this memoryless property: it is the *only* continuous memoryless distribution.

For volcanoes, memorylessness means that ***the 600 years we have waited since Rangitoto erupted have counted for nothing.***

The chance that we still have 1000 years to wait for the next eruption is the same today as it was 600 years ago when Rangitoto erupted.

Memorylessness applies to *any* Poisson process. It is not always a desirable property: you don't want a memoryless waiting time for your bus!

The Exponential distribution is often used to model *failure times* of components: for example $X \sim \text{Exponential}(\lambda)$ is the amount of time before a light bulb fails. In this case, memorylessness means that 'old is as good as new' — or, put another way, 'new is as bad as old'! A memoryless light bulb is quite likely to fail almost immediately.

For private reading: proof of memorylessness

Let $X \sim \text{Exponential}(\lambda)$ be the *total* time waited for an event.

Let Y be the amount of *extra* time waited for the event, given that we have *already* waited time t (say).

We wish to prove that Y has the same distribution as X, i.e. that the time t already waited has been 'forgotten'. This means we need to prove that $Y \sim \text{Exponential}(\lambda)$.

Proof: We will work with $F_Y(y)$ and prove that it is equal to $1 - e^{-\lambda y}$. This proves that Y is Exponential(λ) like X.

First note that $X = t + Y$, because X is the *total* time waited, and Y is the time waited after time t. Also, we must condition on the event $\{X > t\}$, because we know that we have already waited time t. So $\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq t + y \,|\, X > t)$.

$$\begin{aligned}
F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(X \leq t + y \,|\, X > t) \\
&= \frac{\mathbb{P}(X \leq t + y \quad \text{AND} \quad X > t)}{\mathbb{P}(X > t)} \\
& \qquad\qquad\qquad\qquad \text{(definition of conditional probability)} \\
&= \frac{\mathbb{P}(t < X \leq t + y)}{1 - \mathbb{P}(X \leq t)} \\
&= \frac{F_X(t + y) - F_X(t)}{1 - F_X(t)} \\
&= \frac{(1 - e^{-\lambda(t+y)}) - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})} \\
&= \frac{e^{-\lambda t} - e^{-\lambda(t+y)}}{e^{-\lambda t}} \\
&= \frac{e^{-\lambda t}(1 - e^{-\lambda y})}{e^{-\lambda t}} \\
&= 1 - e^{-\lambda y}. \qquad \text{So } Y \sim \text{Exponential}(\lambda) \text{ as required.}
\end{aligned}$$

Thus the *conditional* probability of waiting time y *extra*, given that we have already waited time t, is the same as the probability of waiting time y in total. The time t already waited is forgotten. □

4.4 Likelihood and estimation for continuous random variables

- For discrete random variables, we found the likelihood using the *probability function,* $f_X(x) = \mathbb{P}(X = x)$.
- For continuous random variables, we find the likelihood using the *probability density function,* $f_X(x) = \frac{dF_X}{dx}$.
- Although the notation $f_X(x)$ *means something different for continuous and discrete random variables, it is used in exactly the same way for likelihood and estimation.*

Note: Both discrete and continuous r.v.s have the same definition for the cumulative distribution function: $F_X(x) = \mathbb{P}(X \leq x)$.

Example: Exponential likelihood

Suppose that:

- $X \sim$ *Exponential*(λ)*;*
- λ *is unknown;*
- *the observed value of* X *is* x.

Then the likelihood function is:

$$L(\lambda\,; x) = f_X(x) = \lambda e^{-\lambda x} \quad \textit{for } 0 < \lambda < \infty.$$

We estimate λ by *setting* $\dfrac{dL}{d\lambda} = 0$ *to find the MLE,* $\hat{\lambda}$.

Two or more independent observations

Suppose that $X_1, \ldots, X_n$ are continuous random variables such that:

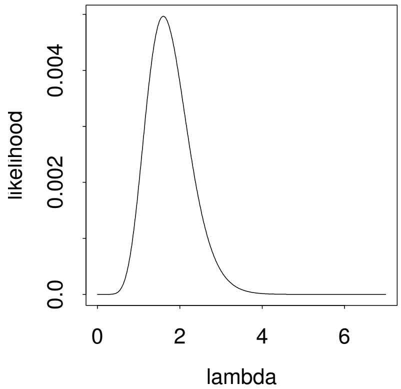
- $X_1, \ldots, X_n$ *are INDEPENDENT;*
- *all the* X_i*s have the same p.d.f.,* $f_X(x)$*;*

then the likelihood is

$$f_X(x_1) f_X(x_2) \ldots f_X(x_n).$$

Example: Suppose that $X_1, X_2, \ldots, X_n$ are independent, and $X_i \sim \text{Exponential}(\lambda)$ for all i. Find the maximum likelihood estimate of λ.

Likelihood graph shown for $\lambda = 2$ and $n = 10$. $x_1, \ldots, x_{10}$ generated by R command `rexp(10, 2)`.



Solution:

$$\begin{aligned} L(\lambda\,; x_1, \ldots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \quad \textit{for } 0 < \lambda < \infty. \end{aligned}$$

Define $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ *to be the sample mean of* $x_1, \ldots, x_n$. *So*

$$\sum_{i=1}^{n} x_i = n\overline{x}.$$

Thus

$$L(\lambda\,; x_1, \ldots, x_n) = \lambda^n e^{-\lambda n \overline{x}} \quad \textit{for } 0 < \lambda < \infty.$$

Solve $\dfrac{dL}{d\lambda} = 0$ *to find the MLE of* λ*:*

$$\begin{aligned} \frac{dL}{d\lambda} = n\lambda^{n-1}e^{-\lambda n\overline{x}} - \lambda^n \times n\overline{x} \times e^{-\lambda n \overline{x}} &= 0 \\ n\lambda^{n-1}e^{-\lambda n \overline{x}}(1 - \lambda\overline{x}) &= 0 \\ \Rightarrow \quad \lambda = 0, \quad \lambda = \infty, \qquad \lambda &= \frac{1}{\overline{x}}\,. \end{aligned}$$

The MLE of λ *is*

$$\hat{\lambda} = \frac{1}{\overline{x}}\,.$$

4.5 Hypothesis tests

Hypothesis tests for continuous random variables are just like hypothesis tests for discrete random variables. The only difference is:

- *endpoints matter for discrete random variables, but not for continuous random variables.*

Example: discrete. Suppose $H_0 : X \sim \text{Binomial}(n = 10,\ p = 0.5)$, and we have observed the value $x = 7$. Then the *upper-tail* p-value is

$$\mathbb{P}(X \geq 7) = 1 - \mathbb{P}(X \leq 6) = 1 - F_X(6).$$

Example: continuous. Suppose $H_0 : X \sim \text{Exponential}(2)$, and we have observed the value $x = 7$. Then the *upper-tail* p-value is

$$\mathbb{P}(X \geq 7) = 1 - \mathbb{P}(X \leq 7) = 1 - F_X(7).$$

Other than this trap, the procedure for hypothesis testing is the same:

- Use H_0 to specify the distribution of X completely, and offer a one-tailed or two-tailed alternative hypothesis H_1.
- Make observation x.
- Find the one-tailed or two-tailed p-value as the probability of seeing an observation *at least as weird* as what we have seen, if H_0 is true.
- That is, find the probability under the distribution specified by H_0 of seeing an observation *further out in the tails* than the value x that we have seen.

Example with the Exponential distribution

A very very old person observes that the waiting time from Rangitoto to the next volcanic eruption in Auckland is 1500 years. Test the hypothesis that $\lambda = \frac{1}{1000}$ against the one-sided alternative that $\lambda < \frac{1}{1000}$.

Note: If $\lambda < \frac{1}{1000}$, we would expect to see *BIGGER values of X, NOT smaller.* This is because X is the time between volcanoes, and λ is the rate at which volcanoes occur. A smaller value of λ means *volcanoes occur less often, so the time X between them is BIGGER.*

Hypotheses: *Let* $X \sim$ *Exponential*(λ).

$$\begin{aligned} H_0 &: \lambda = \frac{1}{1000} \\ H_1 &: \lambda < \frac{1}{1000} \quad \textit{one-tailed test} \end{aligned}$$

Observation: $x = 1500$ *years.*

Values weirder than $x = 1500$ years: *all values BIGGER than* $x = 1500$.

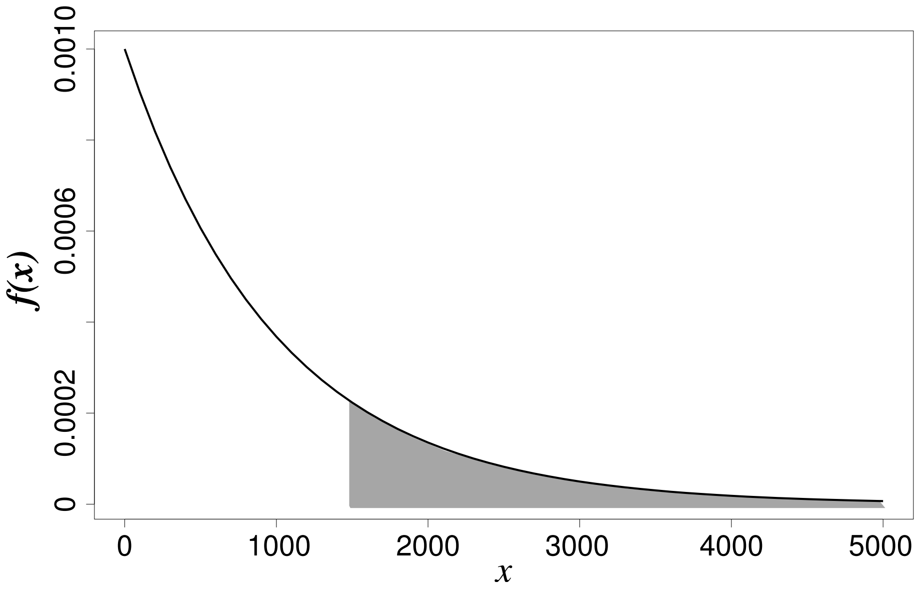
***p*-value:** $\mathbb{P}(X \geq 1500)$ *when* $X \sim$ *Exponential*$(\lambda = \frac{1}{1000})$.

So

$$\begin{aligned} p - \textit{value} &= \mathbb{P}(X \geq 1500) \\ &= 1 - \mathbb{P}(X \leq 1500) \\ &= 1 - F_X(1500) \quad \textit{when } X \sim \textit{Exponential}(\lambda = \tfrac{1}{1000}) \\ &= 1 - (1 - e^{-1500/1000}) \\ &= 0.223. \end{aligned}$$

***R* command:** `1-pexp(1500, 1/1000)`

Interpretation: *There is no evidence against* H_0. *The observation* $x = 1500$ *years is consistent with the hypothesis that* $\lambda = 1/1000$, *i.e. that volcanoes erupt once every 1000 years on average.*



4.6 Expectation and variance

Remember the expectation of a ***discrete*** random variable is *the long-term average:*

$$\mu_X = \mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x x f_X(x).$$

(For each value x, we add in the value and multiply by the proportion of times we would expect to see that value: $\mathbb{P}(X = x)$.)

For a ***continuous*** random variable, *replace the probability function with the probability density function, and replace $\sum_x$ by $\int_{-\infty}^{\infty}$:*

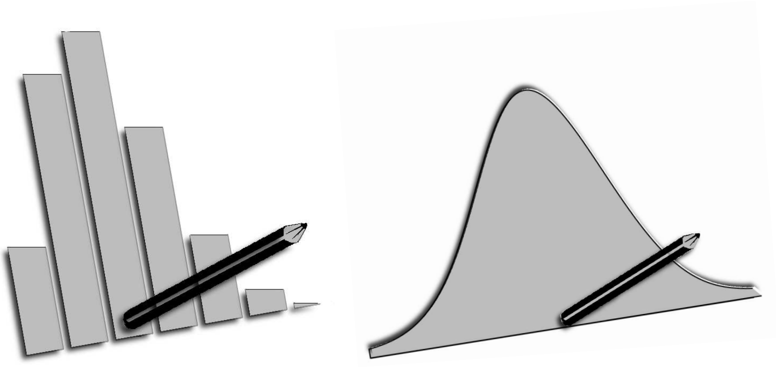
$$\mu_X = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

where $f_X(x) = F'_X(x)$ is the probability density function.

Note: There exists no concept of a 'probability function' $f_X(x) = \mathbb{P}(X = x)$ for continuous random variables. In fact, if X is continuous, then $\mathbb{P}(X = x) =$ *0 for all x.*

The idea behind expectation is the same for both discrete and continuous random variables. $\mathbb{E}(X)$ is:

- the long-term average of X;
- a 'sum' of values multiplied by how common they are: $\sum xf(x)$ *or* $\int xf(x)\,dx$.



Expectation is also the balance point of $f_X(x)$ for both continuous and discrete X. Imagine $f_X(x)$ cut out of cardboard and balanced on a pencil.

Discrete:	**Continuous:**
$\mathbb{E}(X) = \sum_x x f_X(x)$	$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$
$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x)$ Transform the values, leave the probabilities alone;	$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$ Transform the values, leave the probability density alone.
$f_X(x) = \mathbb{P}(X = x)$	$f_X(x) = F_X'(x)$ **(p.d.f.)**

Variance

If X is continuous, its variance is defined in exactly the same way as a discrete random variable:

$$\boxed{\textit{Var}(X) = \sigma_X^2 = \mathbb{E}\left((X - \mu_X)^2 \right) = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.}$$

For a continuous random variable, we can either compute the variance using

$$\textit{Var}(X) = \mathbb{E}\left((X - \mu_X)^2 \right) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx,$$

or

$$\textit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - (\mathbb{E}X)^2.$$

The second expression is usually easier (although not always).

Properties of expectation and variance

All properties of expectation and variance are *exactly the same for continuous and discrete random variables.*

For *any* random variables, X, Y, and $X_1, \ldots, X_n$, continuous or discrete, and for constants a and b:

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$
- $\mathbb{E}(ag(X) + b) = a\mathbb{E}(g(X)) + b.$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$
- $\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$
- $\text{Var}(aX + b) = a^2 \textit{Var}(X).$
- $\text{Var}(ag(X) + b) = a^2 \textit{Var}(g(X)).$

The following statements are generally true *only when X and Y are INDEPENDENT:*

- $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ *when X, Y independent.*
- $\text{Var}(X + Y) = \textit{Var}(X) + \textit{Var}(Y)$ *when X, Y independent.*

4.7 Exponential distribution mean and variance

When $X \sim \text{Exponential}(\lambda)$, then:

$$\boxed{\mathbb{E}(X) = \frac{1}{\lambda} \qquad \textit{Var}(X) = \frac{1}{\lambda^2}\,.}$$

Note: If X is the waiting time for a Poisson process with rate λ events per year (say), it makes sense that $\mathbb{E}(X) = \frac{1}{\lambda}$. For example, if $\lambda = 4$ events per hour, the average time waited between events is $\frac{1}{4}$ hour.

Proof: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx.$

Integration by parts: recall that $\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx.$

Let $u = x,$ *so* $\frac{du}{dx} = 1,$ *and let* $\frac{dv}{dx} = \lambda e^{-\lambda x},$ *so* $v = -e^{-\lambda x}.$

Then

$$\begin{aligned}
\mathbb{E}(X) = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx &= \int_0^{\infty} u\frac{dv}{dx}\,dx \\
&= \Big[uv\Big]_0^{\infty} - \int_0^{\infty} v\frac{du}{dx}\,dx \\
&= \Big[-xe^{-\lambda x}\Big]_0^{\infty} - \int_0^{\infty} (-e^{-\lambda x})\,dx \\
&= 0 + \left[\tfrac{-1}{\lambda}e^{-\lambda x}\right]_0^{\infty} \\
&= \tfrac{-1}{\lambda} \times 0 - \left(\tfrac{-1}{\lambda} \times e^0\right) \\
\therefore \quad \mathbb{E}(X) &= \tfrac{1}{\lambda}\,.
\end{aligned}$$

Variance: $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \frac{1}{\lambda^2}.$

Now $\mathbb{E}(X^2) = \displaystyle\int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^{\infty} x^2\lambda e^{-\lambda x}\,dx.$

Let $u = x^2,$ so $\frac{du}{dx} = 2x,$ and let $\frac{dv}{dx} = \lambda e^{-\lambda x},$ so $v = -e^{-\lambda x}.$

Then

$$\begin{aligned}
\mathbb{E}(X^2) = \Big[uv\Big]_0^{\infty} - \int_0^{\infty} v\frac{du}{dx}\,dx &= \Big[-x^2e^{-\lambda x}\Big]_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x}\,dx \\
&= 0 + \frac{2}{\lambda}\int_0^{\infty} \lambda xe^{-\lambda x}\,dx \\
&= \frac{2}{\lambda} \times \mathbb{E}(X) = \frac{2}{\lambda^2}\,.
\end{aligned}$$

So

$$\begin{aligned}
\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 &= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 \\
\text{Var}(X) &= \frac{1}{\lambda^2}\,. \qquad \square
\end{aligned}$$

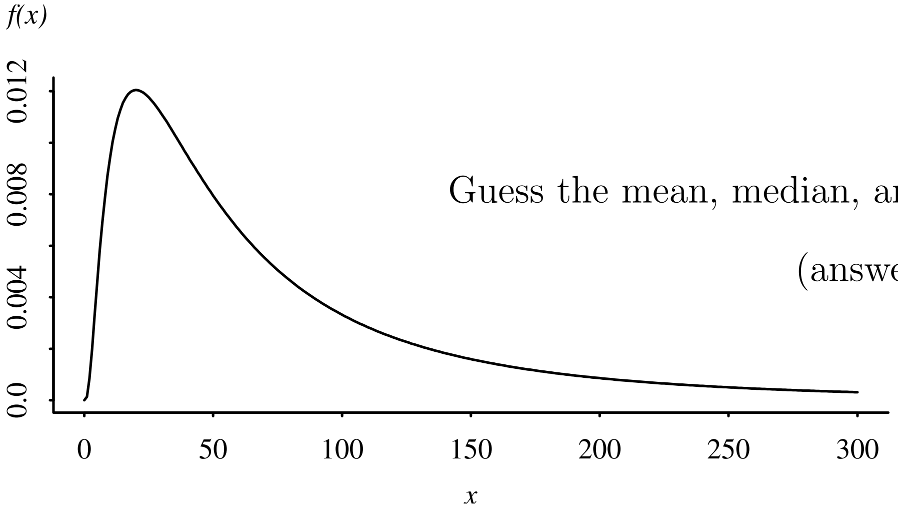
Interlude: Guess the Mean, Median, and Variance

For any distribution:

- the ***mean*** is the ***average*** that would be obtained if a large number of observations were drawn from the distribution;
- the ***median*** is the ***half-way point*** of the distribution: every observation has a 50-50 chance of being above the median or below the median;
- the ***variance*** is the ***average squared distance*** of an observation from the mean.

Given the probability density function of a distribution, we should be able to guess roughly the distribution mean, median, and variance . . . but it isn't easy! Have a go at the examples below. As a hint:

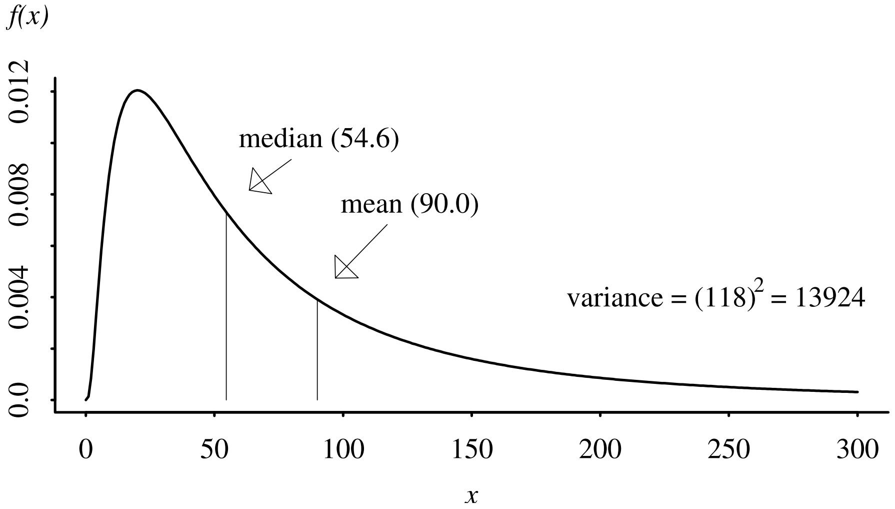
- the ***mean*** is the ***balance-point*** of the distribution. Imagine that the p.d.f. is made of cardboard and balanced on a rod. The mean is the point where the rod would have to be placed for the cardboard to balance.
- the ***median*** is the half-way point, so it divides the p.d.f. into two equal areas of 0.5 each.
- the ***variance*** is the average ***squared*** distance of observations from the mean; so to get a ***rough*** guess (not exact), it is easiest to guess an average distance from the mean and square it.



Guess the mean, median, and variance.

(answers overleaf)

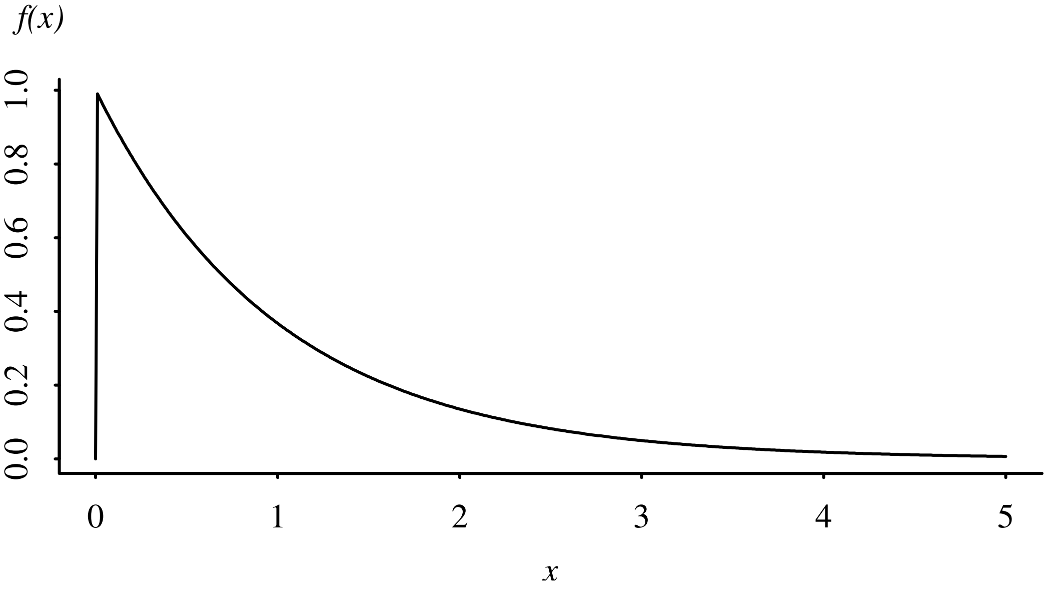
Answers:



Notes: The mean is larger than the median. This always happens when the distribution has a long right tail (positive skew) like this one.

The variance is ***huge*** . . . but when you look at the numbers along the horizontal axis, it is quite believable that the average squared distance of an observation from the mean is 118^2. Out of interest, the distribution shown is a Lognormal distribution.

Example 2: Try the same again with the example below. Answers are written below the graph.



Answers: Median = 0.693; Mean=1.0; Variance=1.0.

4.8 The Uniform distribution

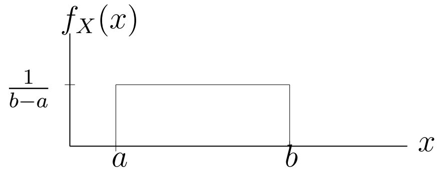
X has a **Uniform distribution on the interval $[a, b]$** if X *is equally likely to fall anywhere in the interval* $[a, b]$.

We write $X \sim$ *Uniform*$[a, b]$, *or* $X \sim$ *U*$[a, b]$.
Equivalently, $X \sim$ *Uniform*(a, b), *or* $X \sim$ *U*(a, b).

Probability density function, $f_X(x)$

If $X \sim U[a, b]$, *then*

$$\boxed{f_X(x) = \begin{cases} \dfrac{1}{b-a} & \textit{if } a \le x \le b, \\ 0 & \textit{otherwise.} \end{cases}}$$



Distribution function, $F_X(x)$

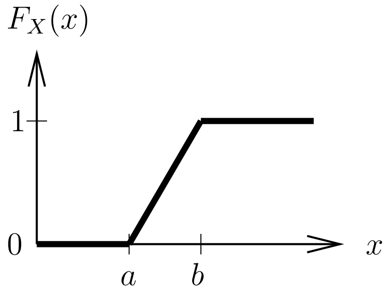
$$F_X(x) = \int_{-\infty}^{x} f_Y(y)\,dy \;=\; \int_a^x \frac{1}{b-a}\,dy \qquad \textit{if} \qquad a \le x \le b$$

$$= \left[\frac{y}{b-a}\right]_a^x$$

$$= \frac{x-a}{b-a} \qquad \textit{if} \qquad a \le x \le b.$$

Thus

$$\boxed{F_X(x) = \begin{cases} 0 & \textit{if } x < a, \\ \frac{x-a}{b-a} & \textit{if } a \le x \le b, \\ 1 & \textit{if } x > b. \end{cases}}$$



Mean and variance:

If $X \sim$ *Uniform*$[a,b]$, $\boxed{\mathbb{E}(X) = \frac{a+b}{2}, \qquad \textit{Var}(X) = \frac{(b-a)^2}{12}.}$

Proof:

$$\begin{aligned}\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_a^b x\left(\frac{1}{b-a}\right)dx &= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ &= \left(\frac{1}{b-a}\right)\cdot\frac{1}{2}(b^2-a^2) \\ &= \left(\frac{1}{b-a}\right)\frac{1}{2}(b-a)(b+a) \\ &= \frac{a+b}{2}.\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}[(X-\mu_X)^2] = \int_a^b \frac{(x-\mu_X)^2}{b-a}\,dx &= \frac{1}{b-a}\left[\frac{(x-\mu_X)^3}{3}\right]_a^b \\ &= \left(\frac{1}{b-a}\right)\left\{\frac{(b-\mu_X)^3-(a-\mu_X)^3}{3}\right\}\end{aligned}$$

But $\mu_X = \mathbb{E}X = \frac{a+b}{2}$, so $b-\mu_X = \frac{b-a}{2}$ and $a-\mu_X = \frac{a-b}{2}$.
So,

$$\begin{aligned}\mathrm{Var}(X) = \left(\frac{1}{b-a}\right)\left\{\frac{(b-a)^3-(a-b)^3}{2^3\times 3}\right\} &= \frac{(b-a)^3+(b-a)^3}{(b-a)\times 24} \\ &= \frac{(b-a)^2}{12}. \qquad \square\end{aligned}$$

Example: let $X \sim \mathrm{Uniform}[0,1]$. Then

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

$\mu_X = \mathbb{E}(X) = \frac{0+1}{2} = \frac{1}{2}$ *(half-way through interval* $[0,1]$).

$\sigma_X^2 = \textit{Var}(X) = \frac{1}{12}(1-0)^2 = \frac{1}{12}.$

4.9 The Change of Variable Technique: finding the distribution of $g(X)$

Let X be a continuous random variable. Suppose

- *the p.d.f. of X, $f_X(x)$, is known;*
- *the r.v. Y is defined as $Y = g(X)$ for some function g;*
- *we wish to find the p.d.f. of Y.*

We use the *Change of Variable technique.*

Example: Let $X \sim \text{Uniform}(0, 1)$, and let $Y = -\log(X)$.

The p.d.f. of X is $f_X(x) = 1$ *for* $0 < x < 1$.

What is the p.d.f. of Y, $f_Y(y)$?

Change of variable technique for monotone functions

Suppose that $g(X)$ is *a monotone function* $\mathbb{R} \to \mathbb{R}$.

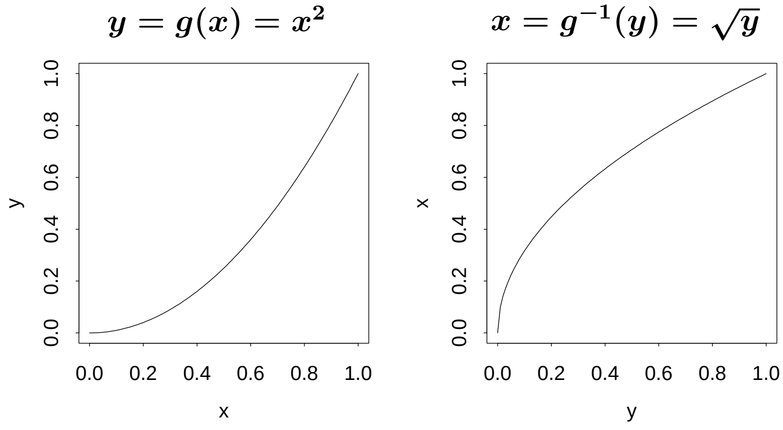
This means that g *is an increasing function, or g is a decreasing f^n.*

When g is monotone, *it is invertible, or (1–1) ('one-to-one').*

That is, *for every y there is a unique x such that $g(x) = y$.*

This means that the inverse function, $g^{-1}(y)$, is well-defined as a function for a certain range of y.

When $g : \mathbb{R} \to \mathbb{R}$, as it is here, then g can *only* be (1–1) if it is monotone.



Change of Variable formula

Let $g : \mathbb{R} \to \mathbb{R}$ be a monotone function and let $Y = g(X)$. Then *the p.d.f. of $Y = g(X)$ is*

$$\boxed{f_Y(y) = f_X(g^{-1}(y)) \left| \tfrac{d}{dy} g^{-1}(y) \right| .}$$

Easy way to remember

Write $\quad y = y(x)(= g(x))$
$\therefore \quad x = x(y)(= g^{-1}(y))$

Then $\quad \boxed{f_Y(y) = f_X\left(x(y)\right) \left| \tfrac{dx}{dy} \right| .}$

Working for change of variable questions

1) Show you have checked $g(x)$ *is monotone over the required range.*

2) *Write* $y = y(x)$ *for* x *in <range of* x*>, e.g. for* $a < x < b$.

3) *So* $x = x(y)$ *for* y *in <range of* y*>:*
for $y(a) < y(x) < y(b)$ *if* y *is increasing;*
for $y(a) > y(x) > y(b)$ *if* y *is decreasing.*

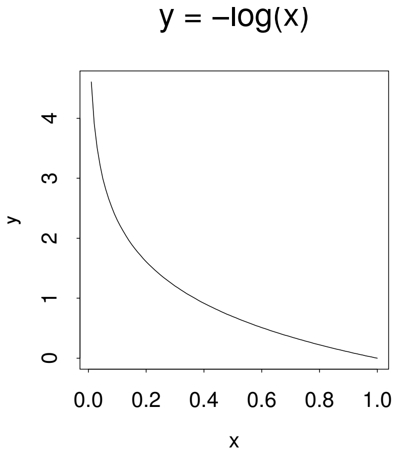
4) *Then* $\left| \dfrac{dx}{dy} \right|$ = *<expression involving* y*>.*

5) *So* $f_Y(y) = f_X(x(y)) \left| \dfrac{dx}{dy} \right|$ *by Change of Variable formula,*
$= \ldots .$
Quote range of values of y *as part of the FINAL answer.*

Refer back to the question to find $f_X(x)$*: you often have to deduce this from information like* $X \sim$ *Uniform*$(0, 1)$ *or* $X \sim$ *Exponential*(λ).
Or it may be given explicitly.

Note: *There should be no x's left in the answer!*

$x(y)$ *and* $\left|\dfrac{dx}{dy}\right|$ *are expressions involving y only.*



Example 1: Let $X \sim \text{Uniform}(0, 1)$, and let $Y = -\log(X)$. Find the p.d.f. of Y. Hence **name** the distribution of Y, with parameters.

1) $y(x) = -\log(x)$ *is monotone decreasing, so we can apply the Change of Variable formula.*

2) *Let* $y = y(x) = -\log x$ *for* $0 < x < 1.$

3) *Then* $x = x(y) = e^{-y}$ *for* $-\log(0) > y > -\log(1),$ *ie.* $0 < y < \infty.$

4) $$\left|\frac{dx}{dy}\right| = \left|\frac{d}{dy}(e^{-y})\right| = \left|-e^{-y}\right| = e^{-y} \quad \text{for} \quad 0 < y < \infty.$$

5) *So*
$$\begin{aligned} f_Y(y) &= f_X(x(y))\left|\frac{dx}{dy}\right| \quad \text{for} \quad 0 < y < \infty \\ &= f_X(e^{-y})e^{-y} \quad \text{for} \quad 0 < y < \infty. \end{aligned}$$

But $X \sim$ *Uniform*$(0, 1)$*, so* $f_X(x) = 1$ *for* $0 < x < 1$*,*
$\Rightarrow f_X(e^{-y}) = 1$ *for* $0 < y < \infty.$

Thus $f_Y(y) = f_X(e^{-y})e^{-y} = e^{-y}$ *for* $0 < y < \infty$*. So* $Y \sim$ *Exponential*(1).

Note: In change of variable questions, *you lose a mark for:*

1. *not stating $g(x)$ is monotone over the required range of x;*
2. *not giving the <u>range</u> of y for which the result holds, as part of the FINAL answer. (eg. $f_Y(y) = \ldots$ for $0 < y < \infty$).*

Example 2: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} \frac{1}{4}x^3 & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = 1/X$. Find the probability density function of Y, $f_Y(y)$.

Let $Y = 1/X$. *The function* $y(x) = 1/x$ *is monotone decreasing for* $0 < x < 2$, *so we can apply the Change of Variable formula.*

Let $y = y(x) = 1/x$ *for* $0 < x < 2$.

Then $x = x(y) = 1/y$ *for* $\frac{1}{0} > y > \frac{1}{2}$, *i.e.* $\frac{1}{2} < y < \infty$.

$$\left|\frac{dx}{dy}\right| = |-y^{-2}| = 1/y^2 \qquad \text{for } \tfrac{1}{2} < y < \infty.$$

Change of variable formula:

$$\begin{aligned} f_Y(y) &= f_X(x(y))\left|\frac{dx}{dy}\right| \\ &= \frac{1}{4}(x(y))^3\left|\frac{dx}{dy}\right| \\ &= \frac{1}{4}\times\frac{1}{y^3}\times\frac{1}{y^2} \\ &= \frac{1}{4y^5} \qquad \text{for} \quad \frac{1}{2} < y < \infty. \end{aligned}$$

Thus

$$f_Y(y) = \begin{cases} \dfrac{1}{4y^5} & \text{for } \frac{1}{2} < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

For mathematicians: proof of the change of variable formula

Separate into cases where g is increasing and where g is decreasing.

i) g increasing

g is increasing if $u < w \Leftrightarrow g(u) < g(w)$. $\quad \circledast$
Note that putting $u = g^{-1}(x)$, and $w = g^{-1}(y)$, we obtain

$$\begin{aligned} g^{-1}(x) < g^{-1}(y) &\Leftrightarrow g(g^{-1}(x)) < g(g^{-1}(y)) \\ &\Leftrightarrow x < y, \end{aligned}$$

so g^{-1} is also an increasing function.

Now

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) &= \mathbb{P}(X \leq g^{-1}(y)) \quad \text{put} \left\{ \begin{array}{l} u = X, \\ w = g^{-1}(y) \end{array} \right. \text{in } \circledast \text{ to see this.} \\ &= F_X(g^{-1}(y)). \end{aligned}$$

So the p.d.f. of Y is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= F'_X(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) \quad \text{(Chain Rule)} \\ &= f_X(g^{-1}(y)) \frac{d}{dy}(g^{-1}(y)) \end{aligned}$$

Now g is increasing, so g^{-1} is also increasing (by overleaf), so $\frac{d}{dy}(g^{-1}(y)) > 0$, and thus $f_Y(y) = f_X(g^{-1}(y)) |\frac{d}{dy}(g^{-1}(y))|$ as required.

ii) g decreasing, i.e. $\quad u > w \iff g(u) < g(w). \quad (\star)$

(Putting $u = g^{-1}(x)$ and $w = g^{-1}(y)$ gives $g^{-1}(x) > g^{-1}(y) \iff x < y$, so g^{-1} is also decreasing.)

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \geq g^{-1}(y)) \qquad \text{(put } u = X,\, w = g^{-1}(y) \text{ in } (\star)) \\ &= 1 - F_X(g^{-1}(y)). \end{aligned}$$

Thus the p.d.f. of Y is

$$f_Y(y) = \frac{d}{dy}\Big(1 - F_X(g^{-1}(y))\Big) = -f_X\Big(g^{-1}(y)\Big)\frac{d}{dy}\Big(g^{-1}(y)\Big).$$

This time, g is decreasing, so g^{-1} is also decreasing, and thus

$$-\frac{d}{dy}\Big(g^{-1}(y)\Big) = \left|\frac{d}{dy}\Big(g^{-1}(y)\Big)\right|.$$

So once again,

$$f_Y(y) = f_X\Big(g^{-1}(y)\Big)\left|\frac{d}{dy}\Big(g^{-1}(y)\Big)\right|. \qquad \square$$

4.10 Change of variable for non-monotone functions: non-examinable

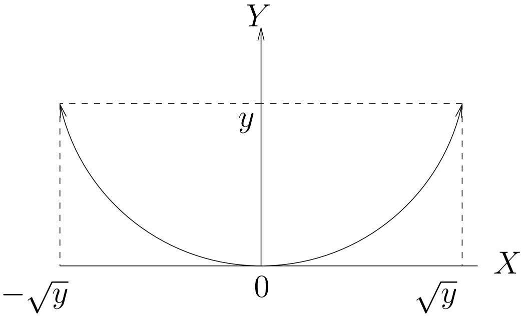
Suppose that $Y = g(X)$ and g is ***not*** monotone. We wish to find the p.d.f. of Y. We can sometimes do this *by using the distribution function directly.*

Example: Let X have ***any*** distribution, with distribution function $F_X(x)$. Let $Y = X^2$. Find the p.d.f. of Y.

Clearly, $Y \geq 0$*, so* $F_Y(y) = 0$ *if* $y < 0$.

For $y \geq 0$*:*

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$



So

$$F_Y(y) = \begin{cases} 0 & \textit{if } y < 0, \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \textit{if } y \geq 0. \end{cases}$$

So the p.d.f. of Y is

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y &= \frac{d}{dy}(F_X(\sqrt{y})) - \frac{d}{dy}(F_X(-\sqrt{y})) \\ &= \tfrac{1}{2}y^{-\frac{1}{2}}F'_X(\sqrt{y}) + \tfrac{1}{2}y^{-\frac{1}{2}}F'_X(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}}\Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big) \quad \text{for } y \geq 0. \end{aligned}$$

$$\boxed{\therefore \qquad f_Y(y) = \frac{1}{2\sqrt{y}}\Big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\Big) \text{ for } y \geq 0, \text{ whenever } Y = X^2.}$$

Example: Let $X \sim \text{Normal}(0, 1)$. This is the familiar bell-shaped distribution (see later). The p.d.f. of X is:

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

Find the p.d.f. of $Y = X^2$.

By the result above, $Y = X^2$ has p.d.f.

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \cdot \frac{1}{\sqrt{2\pi}}(e^{-y/2} + e^{-y/2}) \\ &= \frac{1}{\sqrt{2\pi}}y^{-1/2}e^{-y/2} \ \text{ for } \ y \geq 0. \end{aligned}$$

This is in fact the Chi-squared distribution with $\nu = 1$ degrees of freedom.

The Chi-squared distribution is a special case of the Gamma distribution (see next section). This example has shown that if $X \sim \text{Normal}(0, 1)$, then $Y = X^2 \sim$Chi-squared(df=1).

4.11 The Gamma distribution

The Gamma(k, λ) distribution is a very flexible family of distributions.

It is defined as the *sum of k independent Exponential r.v.s:*

if $X_1, \ldots, X_k \sim$ Exponential$(\lambda)$and $X_1, \ldots, X_k$ are independent,
then $X_1 + X_2 + \ldots + X_k \sim$ Gamma(k, λ).

Special Case: When $k = 1$, *Gamma*$(1, \lambda) =$ *Exponential*(λ)
(the sum of a single Exponential r.v.)

Probability density function, $f_X(x)$

For $X \sim \text{Gamma}(k, \lambda)$,

$$\boxed{f_X(x) = \begin{cases} \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}}$$

Here, $\Gamma(k)$, called the Gamma function of k, is a constant that ensures $f_X(x)$ integrates to 1, i.e. $\int_0^\infty f_X(x)dx = 1$. It is defined as $\Gamma(k) = \displaystyle\int_0^\infty y^{k-1} e^{-y}\, dy$. When k is an integer, $\Gamma(k) = (k-1)!$

Mean and variance of the Gamma distribution:

For $X \sim \text{Gamma}(k, \lambda)$,

$$\boxed{\mathbb{E}(X) = \frac{k}{\lambda} \;\; \textit{and} \;\; \textit{Var}(X) = \frac{k}{\lambda^2}}$$

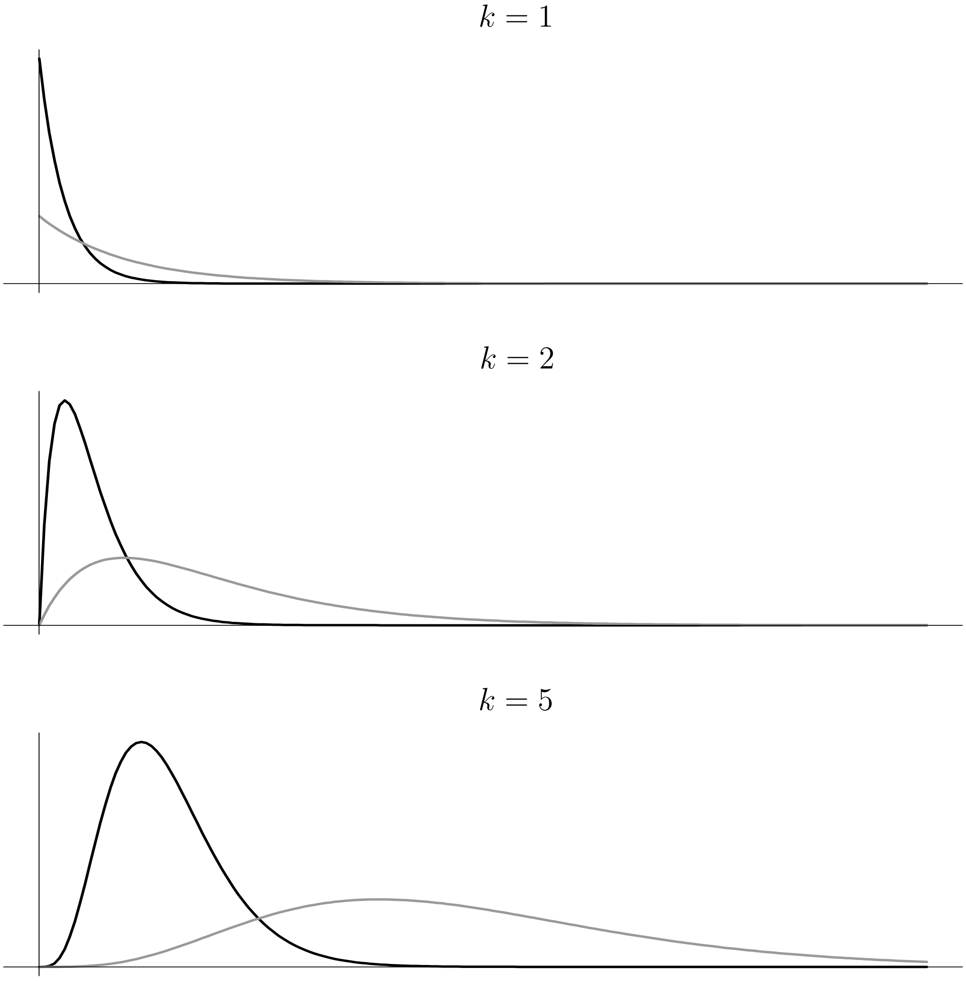
Relationship with the Chi-squared distribution

The Chi-squared distribution with ν degrees of freedom, χ^2_ν, is a special case of the Gamma distribution.

$$\chi^2_\nu = \text{Gamma}(k = \tfrac{\nu}{2}, \lambda = \tfrac{1}{2}).$$

So if $Y \sim \chi^2_\nu$, then $\mathbb{E}(Y) = \frac{k}{\lambda} = \nu$, and $\text{Var}(Y) = \frac{k}{\lambda^2} = 2\nu$.

Gamma p.d.f.s



$k = 1$

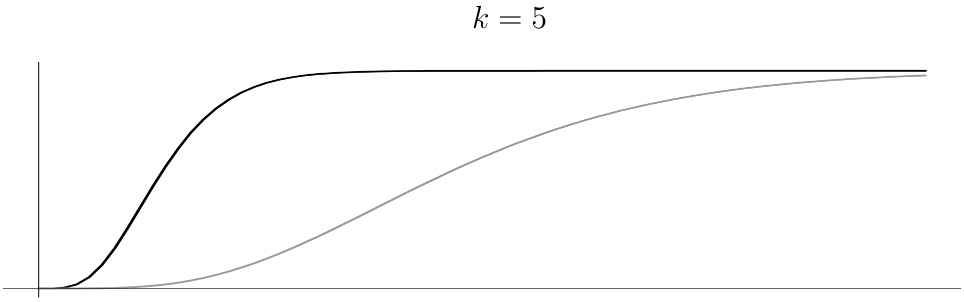
$k = 2$

$k = 5$

Notice: right skew (long right tail); flexibility in shape controlled by the 2 parameters

Distribution function, $F_X(x)$

There is no closed form for the distribution function of the Gamma distribution. If $X \sim \text{Gamma}(k, \lambda)$, then $F_X(x)$ can can only be calculated *by computer.*



$k = 5$

Proof that $\mathbb{E}(X) = \frac{k}{\lambda}$ and $\mathrm{Var}(X) = \frac{k}{\lambda^2}$ (non-examinable)

$$\begin{aligned}
\mathbb{E}X = \int_0^\infty x f_X(x)\,dx &= \int_0^\infty x \cdot \frac{\lambda^k x^{k-1}}{\Gamma(k)} e^{-\lambda x}\,dx \\
&= \frac{\int_0^\infty (\lambda x)^k e^{-\lambda x}\,dx}{\Gamma(k)} \\
&= \frac{\int_0^\infty y^k e^{-y} (\frac{1}{\lambda})\,dy}{\Gamma(k)} \qquad \text{(letting } y = \lambda x,\ \tfrac{dx}{dy} = \tfrac{1}{\lambda}\text{)} \\
&= \frac{1}{\lambda} \cdot \frac{\Gamma(k+1)}{\Gamma(k)} \\
&= \frac{1}{\lambda} \cdot \frac{k\,\Gamma(k)}{\Gamma(k)} \qquad \text{(property of the Gamma function),} \\
&= \frac{k}{\lambda}\,.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 &= \int_0^\infty x^2 f_X(x)\,dx - \frac{k^2}{\lambda^2} \\
&= \int_0^\infty \frac{x^2 \lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}\,dx - \frac{k^2}{\lambda^2} \\
&= \frac{\int_0^\infty (\frac{1}{\lambda})(\lambda x)^{k+1} e^{-\lambda x}\,dx}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{1}{\lambda^2} \cdot \frac{\int_0^\infty y^{k+1} e^{-y}\,dy}{\Gamma(k)} - \frac{k^2}{\lambda^2} \qquad \left[\text{where } y = \lambda x, \frac{dx}{dy} = \frac{1}{\lambda}\right] \\
&= \frac{1}{\lambda^2} \cdot \frac{\Gamma(k+2)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{1}{\lambda^2} \frac{(k+1)k\,\Gamma(k)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\
&= \frac{k}{\lambda^2}\,. \qquad \square
\end{aligned}$$

Gamma distribution arising from the Poisson process

Recall that the waiting time between events in a Poisson process with rate λ has the ***Exponential*(λ) distribution.**

That is, if X_i =time waited between event $i-1$ and event i, then $X_i \sim$ ***Exp*(λ).**

The time waited from time 0 to the time of the kth event is

$$X_1 + X_2 + \ldots + X_k, \text{ the sum of } k \text{ independent Exponential}(\lambda) \text{ r.v.s.}$$

Thus the time waited until the kth event in a Poisson process with rate λ has the ***Gamma*(k, λ) distribution.**

Note: There are some similarities between the Exponential(λ) distribution and the (discrete) Geometric(p) distribution. Both distributions describe the 'waiting time' before an event. In the same way, the Gamma(k, λ) distribution is similar to the (discrete) Negative Binomial(k, p) distribution, as they both describe the 'waiting time' before the kth event.

4.12 The Beta Distribution: non-examinable

The Beta distribution has two parameters, α and β. We write $X \sim \text{Beta}(\alpha, \beta)$.

P.d.f.

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function $B(\alpha, \beta)$ is the *Beta function* and is defined by the integral

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx, \quad \text{for } \alpha > 0,\ \beta > 0.$$

It can be shown that $$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Chapter 5: The Normal Distribution and the Central Limit Theorem

The Normal distribution is the familiar bell-shaped distribution. It is probably the most important distribution in statistics, mainly because of its link with the Central Limit Theorem, which states that *any large sum of independent, identically distributed random variables is approximately Normal:*

$$\boxed{\begin{array}{c} X_1 + X_2 + \ldots + X_n \sim \textit{approx Normal} \\ \textit{if } X_1, \ldots, X_n \textit{ are i.i.d. and } n \textit{ is large.} \end{array}}$$

Before studying the Central Limit Theorem, we look at the Normal distribution and some of its general properties.

5.1 The Normal Distribution

The Normal distribution has two parameters, *the mean, μ, and the variance, σ^2.*

μ and σ^2 satisfy $-\infty < \mu < \infty, \quad \sigma^2 > 0.$

We write $X \sim$ ***Normal***(μ, σ^2), *or* $X \sim$ ***N***(μ, σ^2).

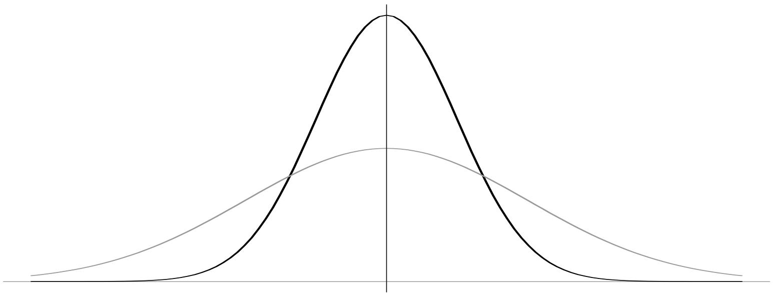
Probability density function, $f_X(x)$

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\{-(x-\mu)^2/2\sigma^2\}} \quad \textit{for } -\infty < x < \infty.}$$

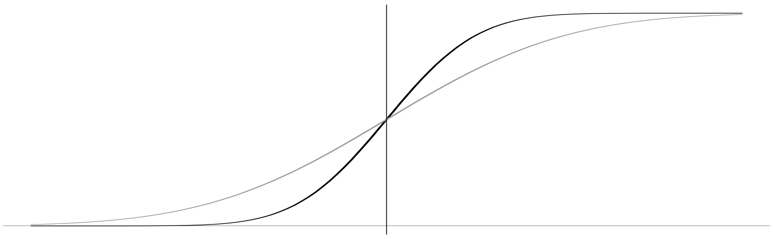
Distribution function, $F_X(x)$

There is no closed form for the distribution function of the Normal distribution. If $X \sim \text{Normal}(\mu, \sigma^2)$, then $F_X(x)$ can can only be calculated *by computer.* *R* command: $F_X(x) =$ `pnorm(x, mean=`μ`, sd=sqrt(`σ^2`))`.

Probability density function, $f_X(x)$



Distribution function, $F_X(x)$



Mean and Variance

For $X \sim \text{Normal}(\mu, \sigma^2)$, $\boxed{\mathbb{E}(X) = \mu, \quad \textit{Var}(X) = \sigma^2.}$

Linear transformations

If $X \sim \text{Normal}(\mu, \sigma^2)$, then for any constants a and b,

$$\boxed{aX + b \sim \textit{Normal}\Big(a\mu + b,\ a^2\sigma^2\Big).}$$

In particular, *put* $a = \dfrac{1}{\sigma}$ *and* $b = -\dfrac{\mu}{\sigma}$, *then*

$$\boxed{X \sim \textit{Normal}(\mu\ \ \sigma^2) \quad \Rightarrow \quad \left(\frac{X-\mu}{\sigma}\right) \sim \textit{Normal}(0,\ 1).}$$

$Z \sim \text{Normal}(0, 1)$ is called the *standard Normal random variable.*

Proof that $aX + b \sim \text{Normal}\Big(a\mu + b,\ a^2\sigma^2\Big)$:

Let $X \sim \text{Normal}(\mu, \sigma^2)$, and let $Y = aX + b$. We wish to find the distribution of Y. Use *the change of variable technique.*

1) $y(x) = ax + b$ *is monotone, so we can apply the Change of Variable technique.*

2) *Let* $y = y(x) = ax + b$ *for* $-\infty < x < \infty$.

3) *Then* $x = x(y) = \frac{y-b}{a}$ *for* $-\infty < y < \infty$.

4) $\left|\dfrac{dx}{dy}\right| = \left|\dfrac{1}{a}\right| = \dfrac{1}{|a|}.$

5) *So* $f_Y(y) = f_X(x(y))\left|\dfrac{dx}{dy}\right| = f_X\left(\dfrac{y-b}{a}\right)\dfrac{1}{|a|}.$ $(\star)$

But $X \sim$ *Normal*(μ, σ^2), *so* $f_X(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$

$$\textit{Thus } f_X\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(\frac{y-b}{a}-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(y-(a\mu+b))^2/2a^2\sigma^2}.$$

Returning to $(\star)$,

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right)\cdot\frac{1}{|a|} = \frac{1}{\sqrt{2\pi a^2\sigma^2}}e^{-(y-(a\mu+b))^2/2a^2\sigma^2} \quad \textit{for } -\infty < y < \infty.$$

But this is the p.d.f. of a Normal$(a\mu + b,\ a^2\sigma^2)$ *random variable.*

So, if $X \sim$ *Normal*(μ, σ^2), *then* $aX + b \sim$ *Normal*$\Big(a\mu + b,\ a^2\sigma^2\Big)$.

Sums of Normal random variables

If X and Y are ***independent***, and $X \sim \text{Normal}(\mu_1,\ \sigma_1^2)$, $Y \sim \text{Normal}(\mu_2,\ \sigma_2^2)$, then

$$X+Y \sim \textit{Normal}\Big(\mu_1+\mu_2,\ \ \sigma_1^2+\sigma_2^2\Big).$$

More generally, if $X_1, X_2, \ldots, X_n$ are ***independent***, and $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$, then

$$a_1X_1+a_2X_2+\ldots+a_nX_n \sim \text{Normal}\Big(\ (a_1\mu_1+\ldots+a_n\mu_n),\quad (a_1^2\sigma_1^2+\ldots+a_n^2\sigma_n^2)\ \Big).$$

For mathematicians: properties of the Normal distribution

1. Proof that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

The full proof that $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\{-(x-\mu)^2/(2\sigma^2)\}}\,dx = 1$

relies on the following result:

$$\text{FACT:}\qquad \int_{-\infty}^{\infty} e^{-y^2}\,dy \ = \ \sqrt{\pi}.$$

This result is non-trivial to prove. See Calculus courses for details.

Using this result, the proof that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ follows by using the change of variable $y = \dfrac{(x-\mu)}{\sqrt{2}\sigma}$ in the integral.

2. Proof that $\mathbb{E}(X) = \mu$.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\,dx$$

Change variable of integration: let $z = \frac{x-\mu}{\sigma}$: then $x = \sigma z + \mu$ and $\frac{dx}{dz} = \sigma$.

$$\text{Then}\quad \mathbb{E}(X) \ = \ \int_{-\infty}^{\infty} (\sigma z+\mu)\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-z^2/2}\cdot\sigma\,dz$$

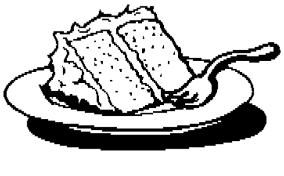
$$= \underbrace{\int_{-\infty}^{\infty} \frac{\sigma z}{\sqrt{2\pi}} \cdot e^{-z^2/2}\,dz}_{\substack{\text{this is an odd function of } z \\ \text{(i.e. } g(-z) = -g(z)\text{), so it} \\ \text{integrates to 0 over range} \\ -\infty \text{ to } \infty.}} \quad + \quad \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz}_{\text{p.d.f. of } N(0,1) \text{ integrates to 1.}}$$

$$\begin{aligned} \text{Thus} \quad \mathbb{E}(X) &= 0 + \mu \times 1 \\ &= \mu. \end{aligned}$$

3. Proof thatVar(X) = σ^2.

$$\begin{aligned}
\text{Var}(X) &= E\left\{(X-\mu)^2\right\} \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/(2\sigma^2)}\,dx \\
&= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2\, e^{-z^2/2}\,dz \qquad \left(\text{putting } z = \frac{x-\mu}{\sigma}\right) \\
&= \sigma^2 \left\{ \frac{1}{\sqrt{2\pi}} \left[-ze^{-z^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \right\} \quad \text{(integration by parts)} \\
&= \sigma^2 \{0+1\} \\
&= \sigma^2. \qquad \square
\end{aligned}$$

5.2 The Central Limit Theorem (CLT)



also known as... **the Piece of Cake Theorem**

The Central Limit Theorem (CLT) is one of the most fundamental results in statistics. In its simplest form, it states that if a large number of independent random variables are drawn from ***any*** distribution, then the distribution of their sum (or alternatively their sample average) always converges to the Normal distribution.

Theorem (The Central Limit Theorem):

Let $X_1, \ldots, X_n$ be independent r.v.s with mean μ and variance σ^2, from ANY distribution.
For example, $X_i \sim$ Binomial(n, p) for each i, so $\mu = np$ and $\sigma^2 = np(1-p)$.

> *Then the sum $S_n = X_1 + \ldots + X_n = \sum_{i=1}^{n} X_i$ has a distribution that tends to Normal as $n \to \infty$.*

The ***mean*** of the Normal distribution is $\mathbb{E}(S_n) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mu$.

The ***variance*** of the Normal distribution is

$$\begin{aligned} \textit{Var}(S_n) &= \textit{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} \textit{Var}(X_i) \quad \textit{because } X_1, \ldots, X_n \textit{ are independent} \\ &= n\sigma^2. \end{aligned}$$

So $\boxed{S_n = X_1 + X_2 + \ldots + X_n \to \textit{Normal}(n\mu,\ n\sigma^2) \ \textit{ as } n \to \infty.}$

Notes:

1. This is a remarkable theorem, because the limit holds for ***any*** distribution of $X_1, \ldots, X_n$.

2. A sufficient condition on X for the Central Limit Theorem to apply is that Var(X) is finite. Other versions of the Central Limit Theorem relax the conditions that $X_1, \ldots, X_n$ are independent and have the same distribution.

3. The ***speed*** of convergence of S_n to the Normal distribution depends upon the distribution of X. Skewed distributions converge more slowly than symmetric Normal-like distributions. It is usually safe to assume that the Central Limit Theorem applies whenever $n \geq 30$. It might apply for as little as $n = 4$.

Distribution of the sample mean, $\overline{X}$, using the CLT

Let $X_1, \ldots, X_n$ be independent, identically distributed with mean $\mathbb{E}(X_i) = \mu$ and variance *Var*$(X_i) = \sigma^2$ for all i.

The sample mean, $\overline{X}$, is defined as:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

So $\overline{X} = \frac{S_n}{n}$, where $S_n = X_1 + \ldots + X_n \sim$ *approx Normal*$(n\mu,\ n\sigma^2)$ *by the CLT.*

Because $\overline{X}$ is a scalar multiple of a Normal r.v. as n grows large, $\overline{X}$ *itself is approximately Normal for large n:*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \sim \textit{approx Normal}\left(\mu,\ \frac{\sigma^2}{n}\right) \quad \textit{as } n \to \infty.$$

The following three statements of the Central Limit Theorem are equivalent:

$$\boxed{\begin{aligned} \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} &\sim \text{approx Normal}\left(\mu,\ \tfrac{\sigma^2}{n}\right) \text{ as } n \to \infty. \\ S_n = X_1 + X_2 + \ldots + X_n &\sim \text{approx Normal}\left(n\mu,\ n\sigma^2\right) \text{ as } n \to \infty. \\ \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} &\sim \text{approx Normal}\,(0,\ 1) \text{ as } n \to \infty. \end{aligned}}$$

The essential point to remember about the Central Limit Theorem is that large sums or sample means of independent random variables converge to a Normal distribution, *whatever the distribution of the original r.v.s.*

More general version of the CLT

A more general form of CLT states that, if $X_1, \ldots, X_n$ are independent, and $\mathbb{E}(X_i) = \mu_i$, $\text{Var}(X_i) = \sigma_i^2$ (not necessarily all equal), then

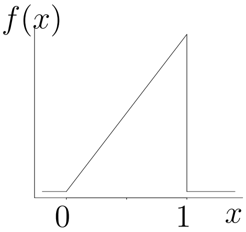
$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \quad \to \quad \text{Normal}(0, 1) \quad \text{as } n \to \infty.$$

Other versions of the CLT relax the condition that $X_1, \ldots, X_n$ are independent.

The Central Limit Theorem in action : simulation studies

The following simulation study illustrates the Central Limit Theorem, making use of several of the techniques learnt in STATS 210. We will look particularly at *how fast the distribution of S_n converges to the Normal distribution.*

Example 1: Triangular distribution: $f_X(x) = 2x$ for $0 < x < 1$.



Find $\mathbb{E}(X)$ and $\text{Var}(X)$:

$$\begin{aligned}\mu = \mathbb{E}(X) &= \int_0^1 x f_X(x)\,dx \\ &= \int_0^1 2x^2\,dx \\ &= \left[\frac{2x^3}{3}\right]_0^1 \\ &= \frac{2}{3}.\end{aligned}$$

$$\begin{aligned}\sigma^2 = \textit{Var}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\ &= \int_0^1 x^2 f_X(x)\,dx - \left(\frac{2}{3}\right)^2 \\ &= \int_0^1 2x^3\,dx - \frac{4}{9} \\ &= \left[\frac{2x^4}{4}\right]_0^1 - \frac{4}{9} \\ &= \frac{1}{18}.\end{aligned}$$

Let $S_n = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are *independent.*

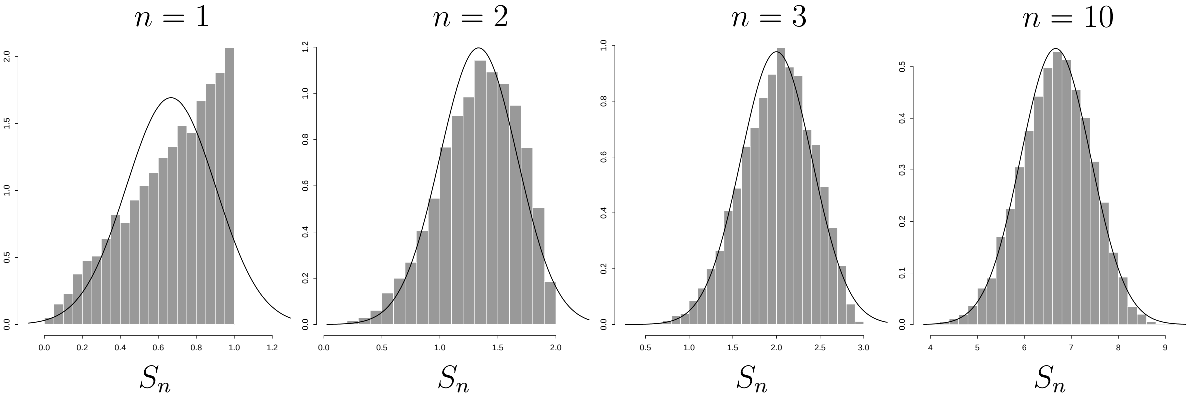
Then

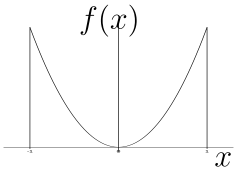
$$\mathbb{E}(S_n) = \mathbb{E}(X_1 + \ldots + X_n) = n\mu = \frac{2n}{3}$$

$$\begin{aligned}\textit{Var}(S_n) = \textit{Var}(X_1 + \ldots + X_n) &= n\sigma^2 \quad \textit{by independence} \\ \Rightarrow \quad \textit{Var}(S_n) &= \frac{n}{18}.\end{aligned}$$

So $S_n \sim$ *approx Normal* $\left(\frac{2n}{3}, \frac{n}{18}\right)$ *for large n, by the Central Limit Theorem.*

The graph shows histograms of 10 000 values of $S_n = X_1 + \ldots + X_n$ for $n = 1, 2, 3$, and 10. The Normal p.d.f. Normal$(n\mu, n\sigma^2) =$ Normal$(\frac{2n}{3}, \frac{n}{18})$ is superimposed across the top. Even for n as low as 10, the Normal curve is a very good approximation.





Example 2: U-shaped distribution: $f_X(x) = \frac{3}{2}x^2$ for $-1 < x < 1$.

We find that $\mathbb{E}(X) = \mu = 0$, *Var*$(X) = \sigma^2 = \frac{3}{5}$. *(Exercise)*

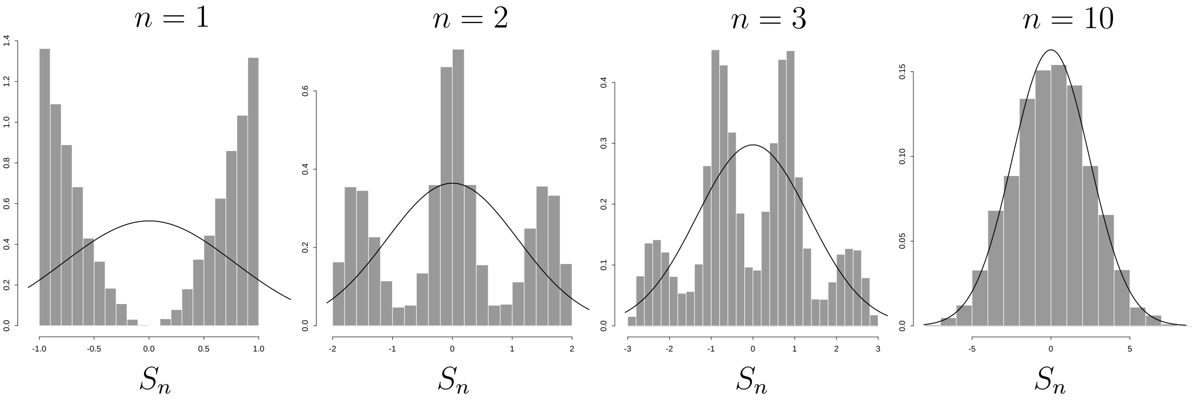
Let $S_n = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are *independent.*

Then

$$\begin{aligned} \mathbb{E}(S_n) = \mathbb{E}(X_1 + \ldots + X_n) = n\mu &= 0 \\ \textit{Var}(S_n) = \textit{Var}(X_1 + \ldots + X_n) &= n\sigma^2 \quad \textit{by independence} \\ \Rightarrow \quad \textit{Var}(S_n) &= \frac{3n}{5}. \end{aligned}$$

So $S_n \sim$ *approx Normal* $\left(0, \frac{3n}{5}\right)$ *for large n, by the CLT.*

Even with this highly non-Normal distribution for X, the Normal curve provides a good approximation to $S_n = X_1 + \ldots + X_n$ for n as small as 10.



Normal approximation to the Binomial distribution, using the CLT

Let $Y \sim \text{Binomial}(n, p)$.

We can think of Y as the *sum of n Bernoulli random variables:*

$$Y = X_1 + X_2 + \ldots + X_n, \text{ where } X_i = \begin{cases} 1 & \textit{if trial } i \textit{ is a "success" (prob} = p\textit{),} \\ 0 & \textit{otherwise (prob} = 1 - p\textit{)} \end{cases}$$

So $Y = X_1 + \ldots + X_n$ *and each* X_i *has* $\mu = \mathbb{E}(X_i) = p, \sigma^2 =$ *Var*$(X_i) = p(1-p)$.

Thus by the CLT,

$$\begin{aligned} Y = X_1 + X_2 + \ldots + X_n &\rightarrow \textit{Normal}(n\mu, n\sigma^2) \\ &= \textit{Normal}\Big(np, np(1-p)\Big). \end{aligned}$$

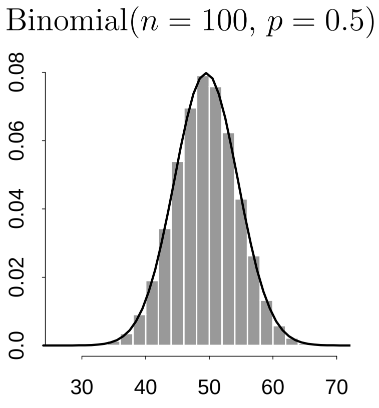
Thus,

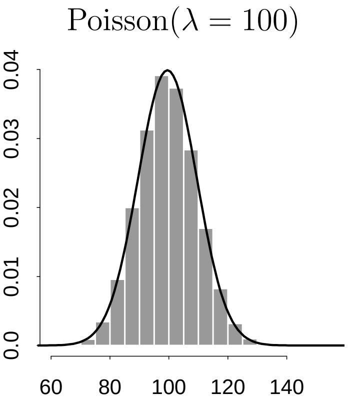
$$\boxed{\textit{Bin}(n, p) \rightarrow \textit{Normal}\Big(\underbrace{np}_{\text{mean of Bin}(n,p)} \quad , \quad \underbrace{np(1-p)}_{\text{var of Bin}(n,p)} \Big) \text{ as } n \rightarrow \infty \text{ with } p \text{ fixed.}}$$

The Binomial distribution is therefore well approximated by the Normal distribution when n is large, for any fixed value of p.

The Normal distribution is also a good approximation to the Poisson(λ) distribution when λ is large:

$$\boxed{\textit{Poisson}(\lambda) \rightarrow \textit{Normal}(\lambda, \lambda) \textit{ when } \lambda \textit{ is large.}}$$


Binomial($n = 100$, $p = 0.5$)


Poisson($\lambda = 100$)

Why the Piece of Cake Theorem? . . .

- The Central Limit Theorem makes whole realms of statistics into a *piece of cake.*
- After seeing a theorem this good, you deserve *a piece of cake!*

5.3 Confidence intervals

Example: Remember the margin of error for an opinion poll?

An opinion pollster wishes to estimate the level of support for Labour in an upcoming election. She interviews n people about their voting preferences. Let p be the true, unknown level of support for the Labour party in New Zealand. Let X be the number of of the n people interviewed by the opinion pollster who plan to vote Labour. Then $X \sim \textit{Binomial}(n, p)$.

At the end of Chapter 2, we said that the maximum likelihood estimator for p is

$$\widehat{p} = \frac{X}{n}.$$

In a large sample (large n), we now know that

$$X \sim \textit{approx Normal}(np,\ npq) \qquad \textit{where } q = 1 - p.$$

So

$$\widehat{p} = \frac{X}{n} \sim \textit{approx Normal}\left(p,\ \frac{pq}{n}\right) \qquad \textit{(linear transformation of Normal r.v.)}$$

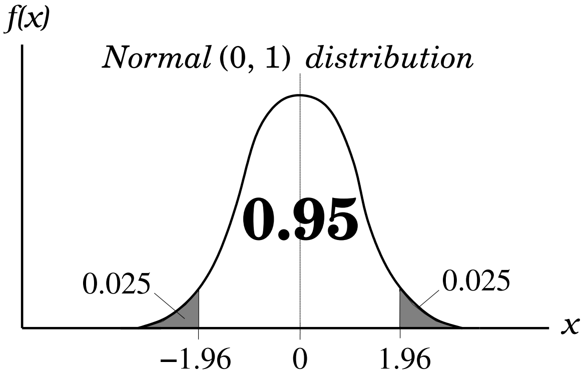
So

$$\frac{\widehat{p} - p}{\sqrt{\frac{pq}{n}}} \sim \textit{approx Normal}(0,\ 1).$$

Now if $Z \sim \text{Normal}(0, 1)$, we find (using a computer) that the 95% central probability region of Z is from -1.96 to $+1.96$:

$$\mathbb{P}(-1.96 < Z < 1.96) = 0.95.$$

Check in *R*: `pnorm(1.96, mean=0, sd=1) - pnorm(-1.96, mean=0, sd=1)`



Putting $Z = \dfrac{\widehat{p} - p}{\sqrt{\frac{pq}{n}}}$, we obtain

$$\mathbb{P}\left(-1.96 < \frac{\widehat{p} - p}{\sqrt{\frac{pq}{n}}} < 1.96\right) \simeq 0.95.$$

Rearranging *to put the unknown p in the middle:*

$$\mathbb{P}\left(\widehat{p} - 1.96\sqrt{\frac{pq}{n}} \;<\; p \;<\; \widehat{p} + 1.96\sqrt{\frac{pq}{n}}\right) \simeq 0.95.$$

This enables us to form an estimated 95% confidence interval for the unknown parameter p: *estimated 95% confidence interval is*

$$\widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \quad \textit{to} \quad \widehat{p} + 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

The 95% confidence interval *has RANDOM end-points, which depend on $\widehat{p}$. About 95% of the time, these random end-points will enclose the true unknown value, p.*

Confidence intervals are extremely important for helping us to assess *how useful our estimate is.*

A narrow confidence interval suggests *a useful estimate (low variance);*
A wide confidence interval suggests *a poor estimate (high variance).*

When you see newspapers quoting the ***margin of error*** on an opinion poll:

- Remember: *margin of error* $= 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$*;*
- Think: *Central Limit Theorem!*
- Have: *a piece of cake.*

Confidence intervals for the Poisson $\boldsymbol{\lambda}$ parameter

We saw in section 3.6 that if $X_1, \ldots, X_n$ are independent, identically distributed with $X_i \sim \text{Poisson}(\lambda)$, then the maximum likelihood estimator of λ is

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

Now $\mathbb{E}(X_i) = \mu = \lambda$, and $\text{Var}(X_i) = \sigma^2 = \lambda$, for $i = 1, \ldots, n$.

Thus, when n is large,

$$\widehat{\lambda} = \overline{X} \sim \text{approx Normal}(\mu,\ \frac{\sigma^2}{n})$$

by the Central Limit Theorem. In other words,

$$\widehat{\lambda} \sim \textit{approx Normal}\left(\lambda,\ \frac{\lambda}{n}\right) \quad \text{as } n \to \infty.$$

We use the same transformation as before to find approximate 95% confidence intervals for λ as n grows large:

Let $Z = \dfrac{\widehat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{n}}}$. We have $Z \sim \textit{approxNormal}(0, 1)$ *for large* n.

Thus:

$$\mathbb{P}\left(-1.96 < \frac{\widehat{\lambda} - \lambda}{\sqrt{\frac{\lambda}{n}}} < 1.96\right) \simeq 0.95.$$

Rearranging *to put the unknown* λ *in the middle:*

$$\mathbb{P}\left(\widehat{\lambda} - 1.96\sqrt{\frac{\lambda}{n}}\ <\ \lambda\ <\ \widehat{\lambda} + 1.96\sqrt{\frac{\lambda}{n}}\right) \simeq 0.95.$$

So our ***estimated*** 95% confidence interval for the unknown parameter λ is:

$$\widehat{\lambda} - 1.96\sqrt{\frac{\widehat{\lambda}}{n}} \quad \textit{to} \quad \widehat{\lambda} + 1.96\sqrt{\frac{\widehat{\lambda}}{n}}\ .$$

Why is this so good?

It's clear that it's important to measure precision, or reliability, of an estimate, otherwise the estimate is almost worthless. However, we have already seen various measures of precision: variance, standard error, coefficient of variation, and now confidence intervals. Why do we need so many?

- The true variance of an estimator, e.g. $\text{Var}(\widehat{\lambda})$, is the most convenient quantity to work with mathematically. However, it is on a non-intuitive scale (squared deviation from the mean), and it usually depends upon the unknown parameter, e.g. λ.

- The ***standard error*** is $\text{se}(\widehat{\lambda}) = \sqrt{\widehat{\text{Var}}\left(\widehat{\lambda}\right)}$. It is an ***estimate*** of the square root of the true variance, $\text{Var}(\widehat{\lambda})$. Because of the square root, the standard error is a direct measure of deviation from the mean, rather than squared deviation from the mean. This means it is measured in more intuitive units. However, it is still unclear how we should comprehend the information that the standard error gives us.

- The beauty of the Central Limit Theorem is that it gives us an incredibly easy way of understanding what the standard error is telling us, using ***Normal-based asymptotic confidence intervals*** as computed in the previous two examples.

  Although it is beyond the scope of this course to see why, the Central Limit Theorem guarantees that almost ***any*** maximum likelihood estimator will be Normally distributed as long as the sample size n is large enough, subject only to fairly mild conditions.

  Thus, *if we can find an estimate of the variance, e.g.* $\widehat{Var}(\widehat{\lambda})$, *we can immediately convert it to an estimated 95% confidence interval using the Normal formulation:*

$$\widehat{\lambda} - 1.96\sqrt{\widehat{Var}\left(\widehat{\lambda}\right)} \quad \textit{to} \quad \widehat{\lambda} + 1.96\sqrt{\widehat{Var}\left(\widehat{\lambda}\right)} \;,$$

  or equivalently,

$$\widehat{\lambda} - 1.96\,\text{se}(\widehat{\lambda}) \quad \text{to} \quad \widehat{\lambda} + 1.96\,\text{se}(\widehat{\lambda}) \;.$$

  The confidence interval has an easily-understood interpretation: *on 95% of occasions we conduct a random experiment and build a confidence interval, the interval will contain the true parameter.*

  So the Central Limit Theorem has given us an incredibly simple and powerful way of converting from a hard-to-understand measure of precision, $\text{se}(\widehat{\lambda})$, to a measure that is easily understood and relevant to the problem at hand. Brilliant!

Chapter 6: Wrapping Up

Probably the two major ideas of this course are:

- likelihood and estimation;
- hypothesis testing.

Most of the techniques that we have studied along the way are to help us with these two goals: expectation, variance, distributions, change of variable, and the Central Limit Theorem.

Let's see how these different ideas all come together.

6.1 Estimators — the good, the bad, and the estimator PDF

We have seen that an *estimator* is a capital letter replacing a small letter. What's the point of that?

Example: Let $X \sim \text{Binomial}(n, p)$ with known n and observed value $X = x$.

- The maximum likelihood *estimate* of p is $\widehat{p} = \frac{x}{n}$.
- The maximum likelihood *estimator* of p is $\widehat{p} = \frac{X}{n}$.

Example: Let $X \sim \text{Exponential}(\lambda)$ with observed value $X = x$.

- The maximum likelihood *estimate* of λ is $\widehat{\lambda} = \frac{1}{x}$.
- The maximum likelihood *estimator* of λ is $\widehat{\lambda} = \frac{1}{X}$.

Why are we interested in *estimators*?

The answer is that ***estimators are random variables.*** This means they have ***distributions, means,*** and ***variances*** that tell us how well we can trust our single observation, or ***estimate,*** from this distribution.

Good and bad estimators

Suppose that $X_1, X_2, \ldots, X_n$ are independent, and $X_i \sim \text{Exponential}(\lambda)$ for all i. λ is unknown, and we wish to estimate it.

In Chapter 4 we calculated the maximum likelihood estimator of λ:

$$\widehat{\lambda} = \frac{1}{\overline{X}} = \frac{n}{X_1 + X_2 + \ldots + X_n}.$$

Now $\widehat{\lambda}$ is a *random variable with a distribution.*

For a given value of n, we can calculate the p.d.f. of $\widehat{\lambda}$. How?

We know that $T = X_1 + \ldots + X_n \sim$ *Gamma*(n, λ) *when* $X_i \sim$ *i.i.d. Exponential*(λ).

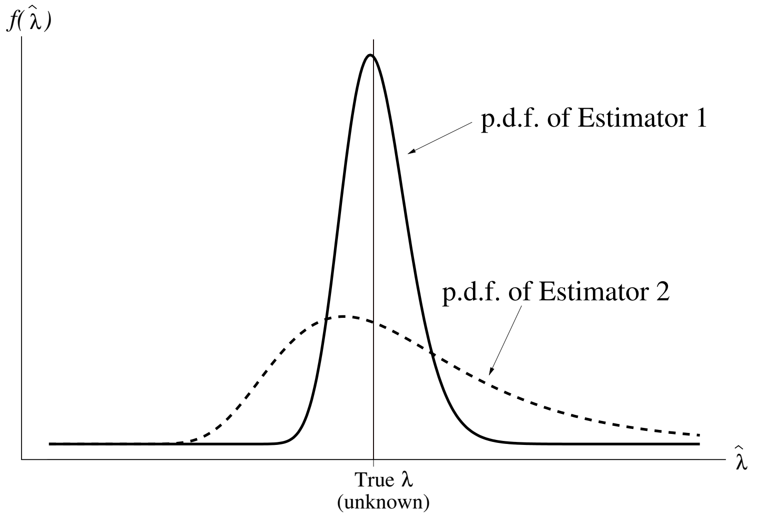
So we know *the p.d.f. of* T.

Now $\widehat{\lambda} = \frac{n}{T}$.

So we can find the p.d.f. of $\widehat{\lambda}$ using the *change of variable technique.*

Here are the p.d.f.s of $\widehat{\lambda}$ for *two different values of* n:

- Estimator 1: $n = 100$. *100 pieces of information about* λ*.*
- Estimator 2: $n = 10$. *10 pieces of information about* λ*.*



Clearly, the more information we have, the better. The p.d.f. for $n = 100$ is focused much more tightly about the true value λ (unknown) than the p.d.f. for $n = 10$.

It is important to recognise what we do and don't know in this situation:

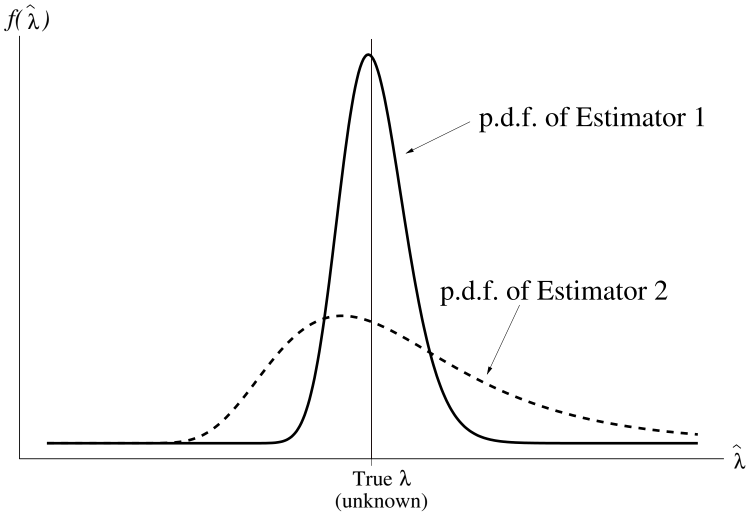
What we don't know:

- *the true λ;*
- *WHERE we are on the p.d.f. curve.*

What we do know:

- *the p.d.f. curve;*
- *we know we're SOMEWHERE on that curve.*

So we need an estimator such that *EVERYWHERE on the estimator's p.d.f. curve is good!*



This is why we are so concerned with *estimator variance.*

A ***good estimator*** has *low estimator variance:* everywhere on the estimator's p.d.f. curve is guaranteed to be good.

A ***poor estimator*** has *high estimator variance:* some places on the estimator's p.d.f. curve may be good, while others may be very bad. Because we *don't know where we are on the curve,* we can't trust *any* estimate from this poor estimator.

The estimator variance tells us how much the estimator can be trusted.

Note: We were lucky in this example to happen to know that $T = X_1 + \ldots + X_n \sim$ Gamma(n, λ) when $X_i \sim$ i.i.d. Exponential(λ), so we could find the p.d.f. of our estimator $\widehat{\lambda} = n/T$. We won't usually be so lucky: so what should we do? *Use the Central Limit Theorem!*

Example: calculating the maximum likelihood estimator

The following question is in the same style as the exam questions.

Let X be a continuous random variable with probability density function

$$f_X(x) = \begin{cases} \dfrac{2(s-x)}{s^2} & \text{for } 0 < x < s\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Here, s is a parameter to be estimated, where s is the maximum value of X and $s > 0$.

(a) Show that $\mathbb{E}(X) = \dfrac{s}{3}$.

Use $\mathbb{E}(X) = \displaystyle\int_0^s x f_X(x)\,dx = \frac{2}{s^2}\int_0^s (sx - x^2)\,dx$.

(b) Show that $\mathbb{E}(X^2) = \dfrac{s^2}{6}$.

Use $\mathbb{E}(X^2) = \displaystyle\int_0^s x^2 f_X(x)\,dx = \frac{2}{s^2}\int_0^s (sx^2 - x^3)\,dx$.

(c) Find $\text{Var}(X)$.

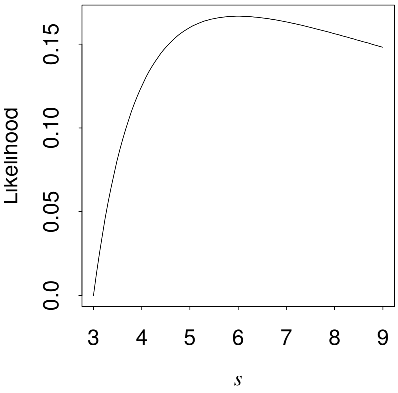
Use $\textit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$. *Answer:* $\textit{Var}(X) = \frac{s^2}{18}$.

(d) Suppose that we make a single observation $X = x$. Write down the likelihood function, $L(s\,;\,x)$, and state the range of values of s for which your answer is valid.

$L(s\,;x) = \dfrac{2(s-x)}{s^2}$ *for* $x < s < \infty$.

(e) The likelihood graph for a particular value of x is shown here.

Show that the maximum likelihood estimator of s is $\widehat{s} = 2X$. You should refer to the graph in your answer.



$$\begin{aligned} L(s\,;\,x) &= 2s^{-2}(s-x) \\ \text{So} \qquad \frac{dL}{ds} &= 2\left\{-2s^{-3}(s-x)+s^{-2}\right\} \\ &= 2s^{-3}(-2(s-x)+s) \\ &= \frac{2}{s^3}(2x-s). \end{aligned}$$

At the MLE,

$$\frac{dL}{ds}=0 \quad \Rightarrow \quad s=\infty \quad \text{or} \quad s=2x.$$

From the graph, we can see that $s=\infty$ *is not the maximum. So* $s=2x$.

Thus the maximum likelihood estimator is

$$\widehat{s}=2X.$$

(f) Find the estimator variance, $\text{Var}(\widehat{s})$, in terms of s. Hence find the estimated variance, $\widehat{\text{Var}}(\widehat{s})$, in terms of $\widehat{s}$.

$$\begin{aligned} \text{Var}(\widehat{s}) &= \text{Var}(2X) \\ &= 2^2\,\text{Var}(X) \\ &= 4\times\frac{s^2}{18} \qquad \text{by (c)} \\ \text{Var}(\widehat{s}) &= \frac{2s^2}{9}. \end{aligned}$$

$$\text{So also:} \qquad \widehat{\text{Var}}(\widehat{s}) = \frac{2\widehat{s}^2}{9}.$$

(g) Suppose we make the single observation $X = 3$. Find the maximum likelihood estimate of s, and its estimated variance and standard error.

$$\begin{aligned}
\widehat{s} &= 2X = 2 \times 3 = 6. \\
\widehat{\mathit{Var}}(\widehat{s}) &= \frac{2\widehat{s}^2}{9} = \frac{2 \times 6^2}{9} = 8 \\
se(\widehat{s}) &= \sqrt{\widehat{\mathit{Var}}(\widehat{s})} = \sqrt{8} = 2.82.
\end{aligned}$$

This means $\widehat{s}$ is a POOR estimator: the twice standard-error interval would be $6 - 2 \times 2.82$ to $6 + 2 \times 2.82$: that is, 0.36 to 11.64 !

Taking the twice standard error interval strictly applies only to the Normal distribution, but it is a useful rule of thumb to see how 'good' the estimator is.

(h) Write a sentence in plain English to explain what the maximum likelihood estimate from part (g) represents.

The value $\widehat{s} = 6$ is the value of s under which the observation $X = 3$ is more likely than it is at any other value of s.

6.2 Hypothesis tests: in search of a distribution

When we do a hypothesis test, we need a ***test statistic:*** some random variable with a ***distribution*** that we can specify exactly under H_0 and that differs under H_1.

It is ***finding the distribution*** that is the difficult part.

- **Weird coin:** is my coin fair? Let X be the number of heads out of 10 tosses. $X \sim \text{Binomial}(10, p)$. We have an easy distribution and can do a hypothesis test.
- **Too many daughters?** Do divers have more daughters than sons? Let X be the number of daughters out of 190 diver children. $X \sim \text{Binomial}(190, p)$. Easy.

- **Too long between volcanoes?** Let X be the length of time between volcanic eruptions. If we assume volcanoes occur as a Poisson process, then $X \sim$ Exponential(λ). We have a simple distribution and test statistic (X): we can test the observed length of time between eruptions and see if it this is a believable observation under a hypothesized value of λ.

More advanced tests

Most things in life are not as easy as the three examples above.

Here are some observations. Do they come from a distribution (any distribution) with mean 0?

```
3.96  2.32 -1.81 -0.14  3.22  1.07 -0.52  0.40  0.51  1.48
1.37 -0.17  1.85  0.61 -0.58  1.54 -1.42 -0.85  1.66  1.54
```

Answer: yes, they are Normal(0, 4), but how can we tell?

What about these?

```
 3.3 -30.0  -7.8   3.4  -1.3  12.6  -9.6   1.4  -6.4 -11.8
-8.1   8.1  -9.0   8.1 -13.7  -5.0  -6.6  -5.6   2.5   9.0
```

Again, yes they do (Normal(0, 100) this time), but how can we tell? The unknown variance (4 versus 100) interferes, so that the second sample does not cluster about its mean of 0 at all.

What test statistic should we use?

If we don't know that our data are Normal, and we don't know their underlying variance, ***what*** can we use as our X to test whether $\mu = 0$?

Answer: a clever person called W. S. Gossett (1876-1937) worked out an answer. He called himself only 'Student', possibly because he (or his employers) wanted it to be kept secret that he was doing his statistical research as part of his employment at Guinness Brewery. The test that 'Student' developed is the familiar Student's t-test. It was originally developed to help Guinness decide how large a sample of people should be used in its beer tastings!

Student used the following test statistic for the unknown mean, μ:

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

Under $H_0 : \mu = 0$, the distribution of T is known: T has p.d.f.

$$f_T(t) = \left(\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \right) \left(1 + \frac{t^2}{n-1} \right)^{-n/2} \qquad \text{for } -\infty < t < \infty.$$

T is the Student's t-distribution, derived as the ratio of a Normal random variable and an independent Chi-Squared random variable. If $\mu \neq 0$, observations of T will tend to lie out in the tails of this distribution.

The Student's t-test is exact when the distribution of the original data $X_1, \ldots, X_n$ is Normal. For other distributions, it is still approximately valid in large samples, by the Central Limit Theorem.

It looks difficult

It is! Most of the statistical tests in common use have deep (and sometimes quite impenetrable) theory behind them. As you can probably guess, Student did not derive the distribution above without a great deal of hard work. The result, however, is astonishing. With the help of our best friend the Central Limit Theorem, Student's T-statistic gives us a test for $\mu = 0$ (or any other value) that can be used with any large enough sample.

The Chi-squared test for testing proportions in a contingency table also has a deep theory, but once researchers had derived the ***distribution of a suitable test statistic,*** the rest was easy. In the Chi-squared goodness-of-fit test, the Pearson's chi-square test statistic is shown to have a Chi-squared distribution under H_0. It produces larger values under H_1.

One interesting point to note is the pivotal role of the Central Limit Theorem in all of this. The Central Limit Theorem produces approximate Normal distributions. Normal random variables squared produce Chi-squared random variables. Normals divided by Chi-squareds produce t-distributed random variables. A ratio of two Chi-squared distributions produces an F-distributed random variable. All these things are not coincidental: ***the Central Limit Theorem rules!***