Single Maths B: Introduction to Probability

Overview

Lecturer	Prof Frank Coolen
Email	frank.coolen@durham.ac.uk
Office	CM206
Homework	None!
Webpage	On DUO

1 Introduction to Probability

1.1 Introduction

What is probability?

- Probability is the mathematical study of uncertainty.
- Probability is a useful concept like *mass* or *energy*. It is attached to events and satisfies some very simple rules.
- Some events can be said to be uncertain – we do not know their outcomes before they occur and we observe what happened.
- Standard mathematics deals only with the certain, so we need some new tools which will allow us to capture, manipulate and reason with this uncertainty.
- We begin by quantifying this uncertainty by assigning numbers to each of the possible outcomes to give a measure of "what is likely to happen."
- Larger values will indicate a particular outcome is more likely. Lower numbers will indicate an outcome is less likely.

$$\begin{aligned} \mathrm{P}\left[\text{fair coin lands heads}\right] &= \frac{1}{2}, \\ \mathrm{P}\left[\text{climate change}\right] &= ? \end{aligned}$$

Why is it useful?

- *Probability can be fundamental to our understanding of the world.* Quantum mechanics, statistical mechanics, Ising model of magnetism, genetics
- *Probability can be used to build models of complex systems or phenomena.* Epidemics, population growth, chemical interactions, financial markets, routing within networks
- The application of probability theory leads to the discipline of *statistics.*
- *Statistics can be used to analyse data gathered from experiments, and drawing conclusions under uncertainty.* Important to all the experimental sciences!

1.2 Events

- Probability theory is used to describe any process whose outcome is not known in advance with certainty. In general, we call these situations *experiments* or *trials.*

- The set of all possible outcomes of an experiment is the *sample space* S.

- An *event* is a subset of the outcomes in a sample space.

- We treat events as *sets*, and so have three basic operations to combine and manipulate them.

  Event operations

  - Let A, B be some events.
  - The event *not* A is A^c (the *complement*), which is the set of all outcomes in $\mathcal{S}$ and not in A.
  - The event A *or* B is $A \cup B$ (the *union*), which the set of all outcomes in A, or in B or in both.
  - The event A *and* B is $A \cap B$ (the *intersection*), which is the set of all outcomes that are both in A and in B.

Disjoint Events

- Two (or more) events are called *disjoint* (or *incompatible*, or *mutually exclusive*) if they *cannot occur at the same time.*

- The event which contains no outcomes is written $\emptyset$, and is called the *empty set.*

- So if A and B are disjoint, then we must have $A \cap B = \emptyset$.

Example: Cluedo

Dr. Black has been murdered! There are four possible suspects: Colonel **M**ustard, Professor **P**lum, Miss **S**carlet, Reverend **G**reen. There are three possible murder weapons: **C**andlestick, **L**ead Piping, **R**ope. There can be only one murderer and one murder weapon.

Working with Events

The following basic set of rules will be useful when working with events:

Event Rules

Commutivity:

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A$$

Associativity:

$$(A \cup B) \cup C = A \cup (B \cup C), \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

Distributivity:

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C), \qquad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

DeMorgan's laws:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

1.3 Probability

Axioms of Probability

- We associate a probability with every outcome (and hence every event) in the sample space $\mathcal{S}$.
- For any event A (i.e. any subset of $\mathcal{S}$) we define a number $\mathrm{P}[A]$ which we call the *probability* of A.
- $\mathrm{P}[A]$ is the quantification of our uncertainty about the occurrence of the event A.
- **Note:** A is an event which is a set, $\mathrm{P}[A]$ is a probability which is a number, and $\mathrm{P}[\cdot]$ is a function which maps events to numbers.

  The Axioms of Probability (Komolgorov)

  1. $0 \leq P[A] \leq 1$. Probability is a number in the interval $[0, 1]$.
  2. $P[\mathcal{S}] = 1$. Some outcome from the sample space *must* happen; certain events have probability 1.
  3. If A and B are disjoint events, then $P[A \cup B] = P[A] + P[B]$.

- We can think of these three axioms as the "Laws of Probability"

Consequences of the Axioms

- These axioms imply some additional useful properties of probabilities:

  Consequences to the Axioms of Probability

  1. $P[A^c] = 1 - P[A]$.
  2. $P[\emptyset] = 0$. Impossible events have probability zero.
  3. In general for any events A and B, $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
  4. If A and B are events, and A contains all of the outcomes in B and more, then we say that B is a *subset* of A, $B \subset A$ and $P[B] < P[A]$.

Probability Interpretations

- There are three different interpretations of probability:

  1. **Classical probability**: considers only sample spaces where every outcome is *equally likely.* If we have m outcomes in our sample space ($\#\mathcal{S} = m$), then for every outcome $s \in \mathcal{S}$ and event $A \subseteq \mathcal{S}$ we have

  $$\mathrm{P}[\{s\}] = \frac{1}{n}, \qquad \mathrm{P}[A] = \frac{\#A}{m} = \frac{\text{number of ways A can occur}}{\text{total no. outcomes}}.$$

  2. **Frequentist probability**: Suppose we repeat the trial n times, and count the number of trials where the event A occurred. The frequentist approach claims that the probability of the event A occurring is the *limit of its relative frequency* in a large number of trials:

  $$\mathrm{P}[A] = \lim_{n \to \infty} \frac{n_A}{n}$$

3. **Subjective probability** views the probability of an event as a measure of an individual's degree of belief that that event will occur.

- Regardless of which interpretation of probability we use, all probabilities must follow the same laws and axioms to be coherent.

Examples

- The probability that student A will fail a certain examination is 0.5, for student B the probability is 0.2, and the probability that both A and B will fail the examination is 0.1. What is the probability that at least one of A and B will fail the examination?
- In the Cluedo example, suppose the probabilities of the suspects being guilty are as follows:

Guilty suspect	M	P	S	G
Probability	0.5	0.25	0.1	p

1. Deduce the value of the missing probability p.
2. Find the probability that both Col. Mustard and Rev. Green are innocent.

Suggested Exercises: Q1–9.

1.4 Counting Principles for Classical Probability

Counting Principles

- Recall that in classical probability one has

$$\mathrm{P}\left[A\right] = \frac{\#A}{m} = \frac{\text{number of ways A can occur}}{\text{total no. outcomes}},$$

where A is an event in a sample space $\mathcal{S}$ in which each outcome is equally likely.

- In order to find 'the number of ways A can occur' in more complex situations, we require some clever counting skills. These skills will be based on certain *counting principles.*

Multiplication principle

If there are r experiments and the first has m_1 equally-likely outcomes, the second has m_2 equally-likely outcomes, and so on, then if all combinations are possible there are:

$$m_1 \times m_2 \times \cdots \times m_r = \prod_{i=1}^{r} m_i$$

equally-likely outcomes for the r experiments.

Examples

1. Suppose there are four different routes from Newcastle to Durham and three different routes from Durham to York. How many different routes are there from Newcastle to York that pass through Durham?

2. There are 17 boys and 13 girls in a class. The teacher needs to pick one girl and one boy. How many different possible pairs can he pick?

3. Six coins are tossed. What is the probability of getting one head and five tails?

Permutations

- Selecting r objects from a group of size n *in order* and *without replacement* gives a *permutation of size r from n.*

- The number of permulations of size r from n is

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

  The resulting arrangement is called a *permutation of size r from n.*

- Special case ($r = n$): the number of ways that r objects can be arranged in order is

$$r \times (r-1) \times (r-2) \times \cdots \times 2 \times 1 = r!$$

- (Note: If you had to choose r objects *with replacement*, then there would be n^r ways to select r objects in order, which is a special case of the multiplication principle).

Examples

1. Six different books are to be arranged on a shelf. How many possible arrangments are there?

2. A club consists of 20 members. A president and secretary have to be chosen from the membership. Determine the total possible number of ways in which these positions can be filled.

3. Six dice are rolled. What is the probability that each of the six different numbers will appear exactly once?

Combinations

- Selecting a group of size r from a group of n gives a *combination of size r from n.*

- The number of (different) combinations is

$$\frac{n!}{(n-r)!r!} = \binom{n}{r} = C_r^n$$

- Relationship between combinations and permutations:

$$C^n_r = \left(\frac{\# \text{ permutation size } r \text{ from } n}{\# \text{ permutations of a group size } r} \right)$$

- The key difference between permutations and combinations is that for permutations order is important, whereas for combinations it is not. For every combination, there are $r!$ permutation of the group members.

Examples

1. Consider the letters A, E, T. How many permutations and combinations can you form of these three letters?
2. Suppose that a subcommittee of 8 people is to be formed from a committee of size 20. How many different groups of people might be on the subcommitte?
3. Find the number of triangles which can be drawn out of n given points on a circle.

Multinomial coefficients

- How many ways are there to partition a finite set of size n into $K \geq 2$ disjoint subsets of size n_k, $k = 1, \ldots, K$?
- Count combinations repeatedly...
  - There are $\binom{n}{n_1}$ combinations for choosing the first group.
  - There are $\binom{n-n_1}{n_2}$ combinations for choosing the second group.
  - ...
  - There are $\binom{n_{K-1}+n_K}{n_{K-1}}$ combinations for choosing the K$-$1th group.
  - The Kth group is then determined.
- ... and apply multiplication principle: There are

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_{K-1}+n_K}{n_{K-1}}$$

  possible ways.

- One can easily show that the latter product can be written as

$$\frac{n!}{n_1!n_2!\ldots n_K!} \equiv \binom{n}{n_1, n_2, \ldots, n_K}.$$

- This is called the *multinomial coefficient.*
- The *multinomial coefficient* is a generalization of the binomial coefficient $\binom{n}{r} = \frac{n!}{r!(n-r)!}$, which counts the number of ways in which one group of size r (and, hence, two complementary groups!) can be chosen from a set of size n.

Examples

1. A committee of 20 is to be divided into three subcommittes of sizes 8, 8, and 4. In how many ways can this be done?
2. In how many different (that is, distinguishable) ways can you arrange the letters in STATISTICS?
3. If you arrange the letters S,S,S,T,T,T,I,I,A,C in random order, what is the probability that they spell 'Statistics'?

Suggested Exercises: Q10–16.

1.5 Conditional Probability

Conditional Probability

- For any two events A, B, the notation $P[A|B]$ means the *conditional* probability that event A occurs, given that the event B has already occurred.
- Conditional probabilities are obtained either directly or by using the *conditional probability rule*:

  The conditional probability rule

  $$\mathrm{P}[A|B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}, \quad \text{for } \mathrm{P}[B] > 0.$$

- Rearranging this equation gives the *multiplication rule*, useful in simplifying probabilities: for any two events A, B,

  The multiplication rule

  $$\mathrm{P}[A \cap B] = \mathrm{P}[A|B]\,\mathrm{P}[B].$$

1.6 Independence

Independence

- Two events are said to be *independent* when the occurrence of one has no bearing on the occurrence of the other.
- In terms of probability, if A, B are independent then

  $$\mathrm{P}[A|B] = \mathrm{P}[A]$$

  as the knowledge that B occurred is irrelevant.
- For independent events A,B, the multiplication rule can then be simplified,

  Multiplication rule for independent events

  $$\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B].$$

- **Note:** Beware of confusing independent events with disjoint events. Independent events do not affect each other in any way, whereas disjoint events cannot occur together – disjoint events are very much dependent on each other.

Example: Two Dice

Two fair dice are rolled, what is the probability that the sum of the two numbers that appear is even?

Example: Nuclear Power Station

Suppose that a nuclear power station has three separate (and independent) devices for detecting a problem and shutting down the reactor. Suppose that each device has a probability of 0.9 of working correctly. In the event of a problem, what is the probability that the reactor will be shut down?

1.7 Partitions and Bayes Theorem

Partitions

- Suppose that n events $E_1, \ldots, E_n$ are *disjoint*, and suppose that exactly one must happen. Such a collection of events is called a *partition.*
- Now we can write any other event A in combination with this partition: in general,

$$\mathrm{P}[A] = \mathrm{P}[A \cap E_1] + \mathrm{P}[A \cap E_2] + \ldots + \mathrm{P}[A \cap E_n],$$

- Using the multiplication rule, we can simplify this to get

  The partition theorem (or theorem of total probability)

$$\mathrm{P}[A] = \ \mathrm{P}[A|E_1]\,\mathrm{P}[E_1] + \mathrm{P}[A|E_2]\,\mathrm{P}[E_2] + \ldots + \mathrm{P}[A|E_n]\,\mathrm{P}[E_n].$$

- Often, this is the most convenient way of getting at certain hard-to-think-about events: to associate them with a suitable partition, and then use conditional probability to simplify matters.

Bayes Theorem

- For any two events A, B, the multiplication rule gives the formula

$$\mathrm{P}[A \cap B] = \mathrm{P}[A|B]\,\mathrm{P}[B].$$

- Another equivalent formula is obviously

$$\mathrm{P}[A \cap B] = \mathrm{P}[B \cap A] = \mathrm{P}[B|A]\,\mathrm{P}[A].$$

- By equating these two formulae and rearranging, we obtain the formula known as

  Bayes theorem

$$\mathrm{P}[A|B] = \frac{\mathrm{P}[B|A]\,\mathrm{P}[A]}{\mathrm{P}[B]}.$$

- It is useful mainly as a way of "inverting" probabilities. Often, the probability in the denominator must be calculated using the simplifying method shown in the last section; i.e. via a *partition.*

Example: Diagnosing Diseases

A clinic offers a test for a very rare and unpleasant disease which affects 1/10000 people. The test itself is 90% reliable, i.e. test results are positive 90% of the time *given you have the disease.* If you don't have the disease the test reports a false positive only 1% of the time. You decide to take the test. What is the probability that the test is positive? Your test returns a positive result. What is the probability you have the disease now?

Suggested Exercises: Q17–24.

2 Random Variables

- A *random variable* (rv) is a variable which takes different numerical values, according to the different possible outcomes of an experiment or random phenomenon.
- Random variables are *discrete* if they only take a finite number of values (e.g. outcome of a coin flip).
- The opposite is a *continuous* random variable with an infinite sample space (e.g. a real-valued measurement).

2.1 Discrete Random Variables

Discrete Random Variables and Probability Distributions

- A discrete random variable X is defined by a pair of two lists

Possible values:	x_1	x_2	x_3	$\dots$
Attached probabilities:	$\mathrm{P}[X = x_1]$	$\mathrm{P}[X = x_2]$	$\mathrm{P}[X = x_3]$	$\dots$

- This collection of all possible values with their probabilities is called the *probability distribution* of X.
- The probabilities in a probability distribution must
  1. be non-negative: $\mathrm{P}[X = x_i] \geq 0, \ \forall i$
  2. add to one: $\sum_i \mathrm{P}[X = x_i] = 1$
- The probability that X lies in an interval $[a, b]$ is then

$$\mathrm{P}[a \leq X \leq b] = \sum_{a \leq x_i \leq b} \mathrm{P}[X = x_i]$$

Joint and Marginal Distributions

- When we have two (or more) random variables X and Y, the *joint probability distribution* is the table of every possible (x, y) value for X and Y, with the associated probabilities $\mathrm{P}[X = x, Y = y]$:

	x_1	$\dots$	x_n
y_1	$\mathrm{P}\left[X = x_1, Y = y_1\right]$	$\dots$	$\mathrm{P}\left[X = x_n, Y = y_1\right]$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_m	$\mathrm{P}\left[X = x_1, Y = y_m\right]$	$\dots$	$\mathrm{P}\left[X = x_n, Y = y_m\right]$

- Given the joint distribution for the random variables (X, Y), we can obtain the distribution of X (or Y) alone – the *marginal probability distribution* for X (or Y) – by summing across the rows or columns:

$$\mathrm{P}\left[X = x\right] = \sum_{\text{all } y} \mathrm{P}\left[X = x, Y = y\right].$$

Example: Discrete Random Variables

Let X be the random variable which takes value 3 when a fair coin lands heads up, and takes value 0 otherwise. Let Y be the value shown after rolling a fair dice. Write down the distributions of X, and Y, and the joint distribution of (X, Y). You may assume that X and Y are independent. Thus find the probability that $X > Y$

2.2 Continuous Random Variables

Continuous random variables

- Discrete random variables only make sense when our sample space is finite.
- When our experimental outcome is a measurement of some quantity, then our sample space is actually part of the real line and so is infinite.
- A random variable X which can assume every real value in an interval (bounded or unbounded) is called a *continuous random variable.*
- Since our sample space is now infinite we cannot write down a table of probabilities for every possible outcome to describe the distribution of X.
- Instead, the probability distribution for X is described by a *probability density function* (pdf), $f(x)$, which is a function that *describes a curve over the range of possible values* taken by the random variable.

Continuous random variables

- A valid probability density function, $f(x)$, must
  1. be non-negative everywhere: $f(x) \geq 0, \forall x$,
  2. *integrate* to 1: $\int_{-\infty}^{\infty} f(x)\, dx = 1$,
- The probability for a range of values is given by *the area under the curve.*

$$\mathrm{P}\left[a \leq X \leq b\right] = \int_a^b f(x)\ dx$$

Note: $f(x) \neq \mathrm{P}\left[X = x\right]$. Probability densities *are not* probabilities!

- We can describe the probability by the function
$$F(x) \equiv \int_{-\infty}^{x} f(y)\; dy = \mathrm{P}\left[X \leq x\right]$$
which is called the *cumulative distribution function* (cdf) of X.
- We also have the result that $f(x) = F'(x)$.

Joint and Marginal Distributions

- When we have two (or more) continuous random variables, we describe them via their joint probability density function $f_{xy}(x, y)$, which satisfies the usual conditions for pdfs.
- The probability that X and Y fall into some region A of the xy-plane is then
$$\mathrm{P}\left[(X, Y) \in A\right] = \int_A \int f_{xy}(x, y)\, dxdy$$
- Given the joint pdf $f_{xy}(x, y)$, we can obtain the *marginal* pdf of x or y by integrating out the other variable
$$f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y)\, dy$$
- When two continuous random variables x and y are *independent*, their joint pdf can be expressed as the *product* of the marginal pdfs
$$f_{xy}(x, y) = f_x(x)\, f_y(y)$$

Example: The Exponential Distribution

Let X be a continuous random variable with probability density function
$$f(x) = \begin{cases} \beta e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0. \end{cases}$$

Show that $f(x)$ is a valid probability density function when $\beta > 0$. Find the cdf of X, and hence $\mathrm{P}\left[X > 3\right]$.

Suggested Exercises: Q25–28 (without expectation/variance part), 29–30.

---

3 Expectation and Variance

Distribution Summaries

- The distribution of a random variable X contains all of the probabilistic information about X.
- However, the entire distribution of X can often be too complex to work with.
- Summaries of the distribution, such as its average or spread can be useful for conveying information about X without trying to describe it in its entirety.
- Formally, we measure the average of the distribution by calculating its expectation, and we measure the spread by its variance.

3.1 Expectation

- Suppose that X has a discrete distribution, then the expectation of X is given by

$$\mathrm{E}\left[X\right] = \sum_{\text{all } x} x\mathrm{P}\left[X = x\right]$$

- If a random variable X has a continuous distribution with a pdf $f(\cdot)$, then the expectation of X is defined as:

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

- The value $E(X)$ is the theoretical average of the probability distribution. Because of this, it is often referred to it as the *mean* or *average* for the distribution.

Properties of Expectation

- **Expectation of a function:** If X is a random variable, then the expectation of the function $r(X)$ is given by

$$\mathrm{E}\left[r(X)\right] = \sum_{\text{all } x} r(x)\mathrm{P}\left[X = x\right], \qquad \text{or} \quad \mathrm{E}\left[r(X)\right] = \int_{-\infty}^{\infty} r(x)\, f(x)\ dx$$

- **Linearity:** If $Y = a + b\,X$ where a and b are constants, then

$$\mathrm{E}\left[Y\right] = a + b\mathrm{E}\left[X\right].$$

- **Additivity:** If X_1, X_2, ..., X_n are any random variables then

$$\mathrm{E}\left[X_1 + X_2 + \cdots + X_n\right] = \mathrm{E}\left[X_1\right] + \mathrm{E}\left[X_2\right] + \cdots + \mathrm{E}\left[X_n\right]$$

- If X_1, X_2 are any pair of *independent* random variables then

$$\mathrm{E}\left[X_1 X_2\right] = \mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right]$$

3.2 Variance

- Suppose that X is a random variable with mean $\mu = \mathrm{E}\left[X\right]$. The *variance* of X, denoted $\mathrm{Var}\left[X\right]$, is defined as follows:

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[(X - \mu)^2\right].$$

- **Note:** Since $\mathrm{Var}\left[X\right]$ is the expected value of a non-negative random variable $(X - \mu)^2$, it follows that $\mathrm{Var}\left[X\right] \geq 0$.

- We can re-write the variance formula in the following simpler form:

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2.$$

- The *standard deviation* of a random variable is defined as the square root of the variance: $\mathrm{SD}\left[X\right] = \sqrt{\mathrm{Var}\left[X\right]}$.

Properties of Variance

- For constants a and b:

$$\mathrm{Var}\left[a+bX\right]=b^2\mathrm{Var}\left[X\right], \qquad \mathrm{SD}\left[a+bX\right]=b\ \mathrm{SD}\left[X\right].$$

- If $X_1, \ldots, X_n$ are *independent* random variables, then

$$\mathrm{Var}\left[X_1+\cdots+X_n\right]=\mathrm{Var}\left[X_1\right]+\cdots+\mathrm{Var}\left[X_n\right].$$

Example: National Lottery

The National lottery has a game called ‘Thunderball’. You pick 5 numbers in the range 1-34 and one number (the Thunderball number) in the range 1-14. You win a prize if you match at least two numbers, including the Thunderball number. Let X be the amount you win in a single game. The probability distribution for X is given below. Find the expectation and variance of X.

Matches	k, Prize £	$\mathrm{P}\left[X=k\right]$
5 +Tb	250000	0.000000257
5	5000	0.000003337
4 +Tb	250	0.000037220
4	100	0.000483653
3 +Tb	20	0.001041124
3	10	0.013368984
2 +Tb	10	0.009293680
1 +Tb	5	0.029585799
Other	0	0.946185946
Sum		1.000000000

Example: Exponential Distribution (Again)

- Let X be a continuous random variable with probability density function:

$$f(x)=\begin{cases}\beta e^{-\beta x} & \text{for } x>0,\\ 0 & \text{for } x\leq 0.\end{cases}$$

- Find the expectation and variance of X.

Suggested Exercises: Q25–28, 31, 33–35.

4 The Binomial and Poisson Distributions

4.1 The Binomial distribution

- Consider the following circumstances (binomial scenario):

  1. There are n *trials*.
  2. The trials are *independent*.

3. On each trial, only *two* things can happen. We refer to these two events as *success* and *failure*.
4. The *probability of success* is the same on each trial. This probability is usually called p.
5. We count the *total number of successes*. This is a discrete random variable, which we denote by X, and which can take any value between 0 and n (inclusive).

- The random variable X is said to have a *binomial distribution* with parameters n and p; abbreviated

$$X \sim \text{Bin}(n, p)$$

- It is easy to show that if $X \sim \text{Bin}(n, p)$ then

$$\text{P}\left[X = k\right] = \binom{n}{k} p^k (1-p)^{n-k}$$

  for $k = 0, 1, \ldots, n$.

- $\binom{n}{k}$ is the *binomial coefficient* and is the number of sequences of length n containing k successes.

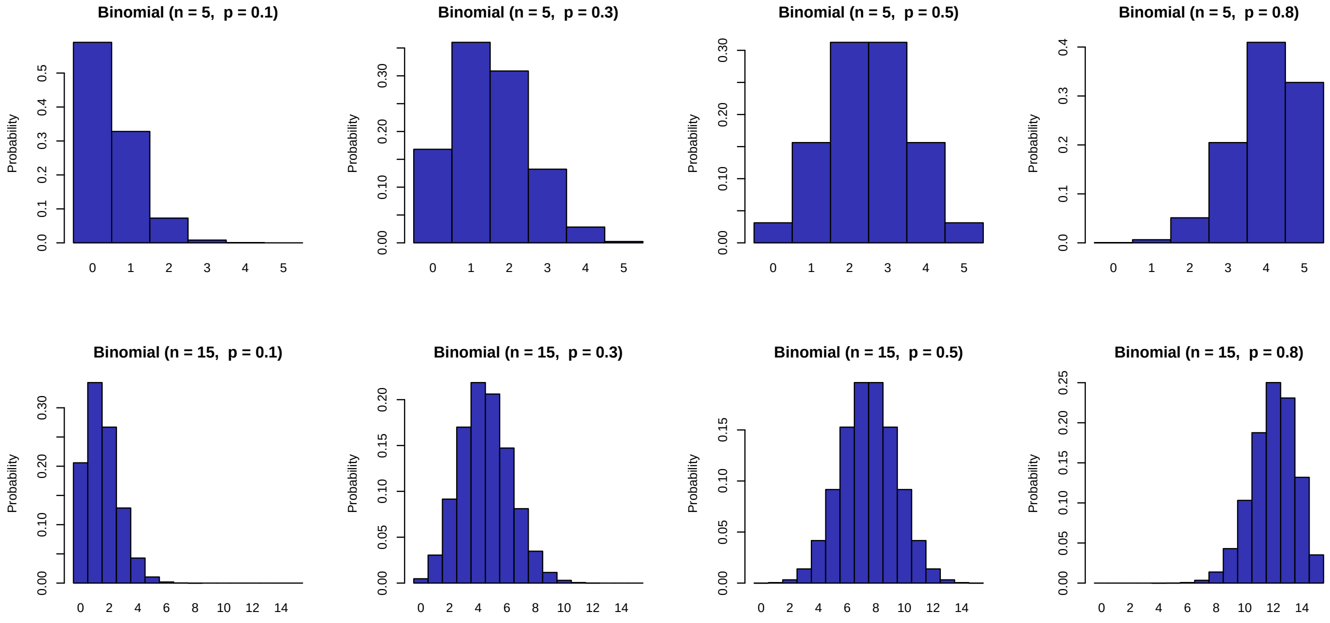
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

- The expectation and variance of X are given by

$$\begin{aligned} \text{E}\left[X\right] &= np \\ \text{Var}\left[X\right] &= np(1-p) \end{aligned}$$

The Binomial Distribution: Example

The shape of the distribution depends on n and p.



Example:

Suppose that it is known that 40% of voters support the Conservative party. We take a random sample of 6 voters. Let the random variable Y represent the number in the sample who support the Conservative party.

1. Explain why it is reasonable to model the distribution of Y by a binomial distribution.
2. Write down the probability distribution of Y as a table of probabilities.
3. Find the mean and variance of Y directly from the probability distribution.
4. Check your answers using the standard results $\mathrm{E}[Y] = np$ and $\mathrm{Var}[Y] = np(1-p)$.

4.2 The Poisson distribution

- The binomial distribution is about counting *successes* in a fixed number of well-defined trials, ie n is known and fixed.
- This can be limiting as many counts in science are open-ended counts of unknown numbers of events in time or space.
- Consider the following circumstances:
  1. Events occur randomly in time (or space) at a *fixed rate* r.
  2. Events occur *independently* of the time (or location) since the last event.
  3. We count the *total number of events* that occur in a time period s, and we let X denote the event count.
- The random variable X has a *Poisson distribution* with parameter $\lambda = rs$; abbreviated

$$X \sim \mathrm{Po}(\lambda)$$

- If $X \sim \mathrm{Po}(\lambda)$ then

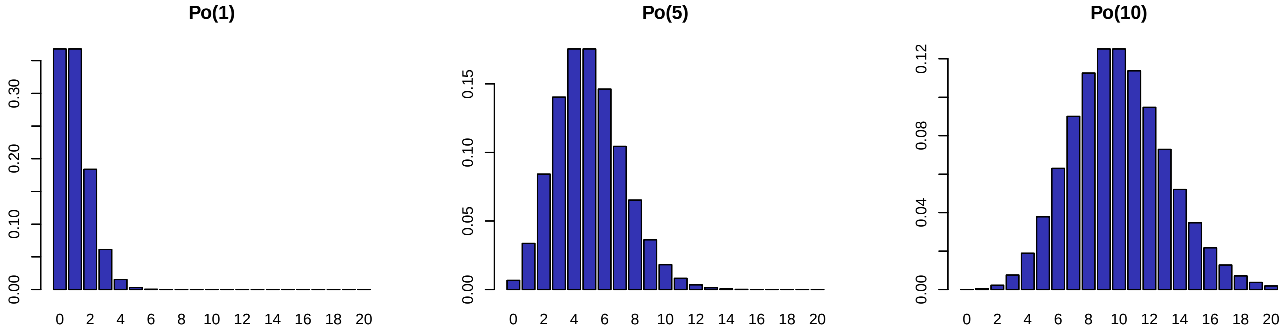
$$\mathrm{P}[X = k] = e^{-\lambda}\frac{\lambda^k}{k!}$$

  for $k = 0, 1, 2, \ldots$.
- The expectation and variance of X are given by

$$\begin{aligned} \mathrm{E}[X] &= \lambda \\ \mathrm{Var}[X] &= \lambda \end{aligned}$$

The Poisson Distribution

Like the binomial distribution, the shape of the Poisson distribution changes as we change its parameter.



Example: Yeast

Gossett, the head of quality control at Guinness brewery c. 1920 (and discoverer of the t distribution), arranged for counts of yeast cells to be made in sample vessels of fluid. He found that at a certain stage of brewing the counts were Po(0.6). Let X be the count from a sample. Find $\mathrm{P}[X \geq 4]$.

4.3 The Poisson approximation to the Binomial

The Poisson approximation to the Binomial

- Consider the Poisson scenario with events occurring randomly over a time period s at a fixed rate r.
- Now, split the time interval s into n subintervals of length s/n (very small).
- Lets consider each mini-interval as a "success" if there is an event in it.
- Now we have n independent trials with $p \approx \frac{r\,s}{n} = \frac{\lambda}{n}$
- The counts X are then binomial.
- If we assume there is no possibility of obtaining two events in the same interval, then we can say

$$\mathrm{P}[X = x] \approx \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

- It can be shown that as n increases this formula converges to

$$e^{-\lambda} \frac{\lambda^x}{x!}$$

  with $\lambda = np$.
- Hence the Binomial distribution $T \sim \mathrm{Bin}(n,p)$, can be approximated by the Poisson $T \sim \mathrm{Po}(np)$ when n is large and p is small.
- This approximation is good if $n \geq 20$ and $p \leq 0.05$, and excellent if $n \geq 100$ and $np \leq 10$.

Example: Computer Chip Failure

A manufacturer claims that a newly-designed computer chip is has a 1% chance of failure because of overheating. To test their claim, a sample of 120 chips are tested. What is the probability that at least two chips fail on testing?

Suggested Exercises: Q32–39

5 The Normal Distribution

5.1 The Normal Distribution

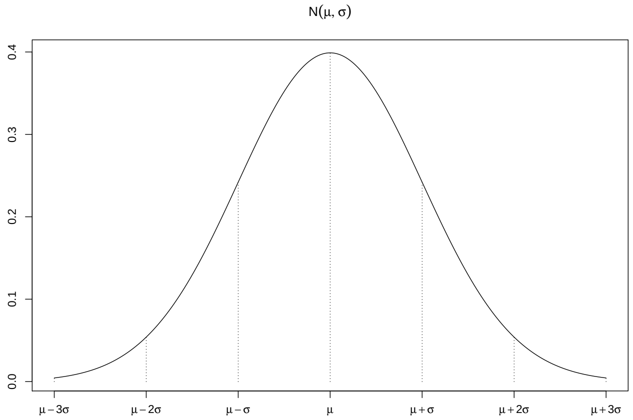
The Normal Distribution

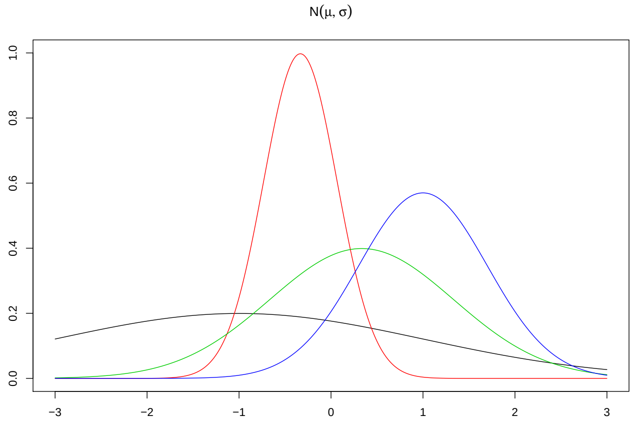
- The most widely useful continuous distribution is the *Normal* (or *Gaussian*) distribution.
- In practice, many measured variables may be assumed to be approximately normal.
- Derived quantities such as *sample means* and *totals* can also be shown to be approximately normal.
- A rv X is Normal with parameters μ and σ^2, written $X \sim \mathrm{N}(\mu, \sigma^2)$, when it has density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

for all real x, and $\sigma > 0$.

The Normal Distribution





The Standard Normal

- The *standard Normal* random variable is a normal rv with $\mu = 0$, and $\sigma^2 = 1$. It is usually denoted Z, so that $Z \sim \mathrm{N}(0, 1)$.

- The cumulative distribution function for Z is denoted $\Phi(z)$ and is

$$\mathrm{P}\left[Z < z\right] \equiv \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx.$$

- Unfortunately, there is no neat expression for $\Phi(z)$, so in practice we must rely on *tables* (or computers) to calculate probabilities.

Properties of the Standard Normal & Tables

- $\Phi(0) = 0.5$ due to the symmetry
- $\mathrm{P}\left[a \leq Z \leq b\right] = \Phi(b) - \Phi(a)$.
- $\mathrm{P}\left[Z < -a\right] = \Phi(-a) = 1 - \Phi(a) = \mathrm{P}\left[Z > a\right]$, for $a \geq 0$. Hence tables only contain probabilities for positive z.
- Φ is very close to 1 (0) for $z > 3$ ($z < -3$). Most tables stop after this point.

Example

i Find the probability that a standard Normal rv is less than -1.62.

ii Find a value c such that $P(-c \leq Z \leq c) = 0.95$.

5.2 Standardisation

- If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ is the *standardized version* of X, and $Z \sim \mathrm{N}(0, 1)$.
- Even more importantly, the distribution function for any normal rv X is given by

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

and so *the cumulative probabilities for any normal rv X can be expressed as probabilities of the standard normal Z.*

- This is why only the **standard** Normal distribution is tabulated.

Example

1. Let X be $\mathrm{N}(12, 25)$. Find $\mathrm{P}\left[X > 3\right]$
2. Let Y be $\mathrm{N}(1, 4)$. Find $\mathrm{P}\left[-1 < X < 2\right]$.

5.3 Other properties

Other properties

- Expectation and variance of Z:

$$\begin{aligned}
\mathrm{E}[Z] &= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0, \quad \text{(integrand is an odd fn)} \\
\mathrm{E}\left[Z^2\right] &= 1, \quad \text{(integrate by parts)} \\
\mathrm{Var}[Z] &= 1.
\end{aligned}$$

- Using our scaling properties it follows that for $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{gathered}
\mathrm{E}[X] = \mu, \\
\mathrm{Var}[X] = \sigma^2.
\end{gathered}$$

- If X and Y are Normally distributed then the sum $S = X + Y$ is also Normally distributed (regardless of whether X and Y are independent).

5.4 Interpolation

Interpolation

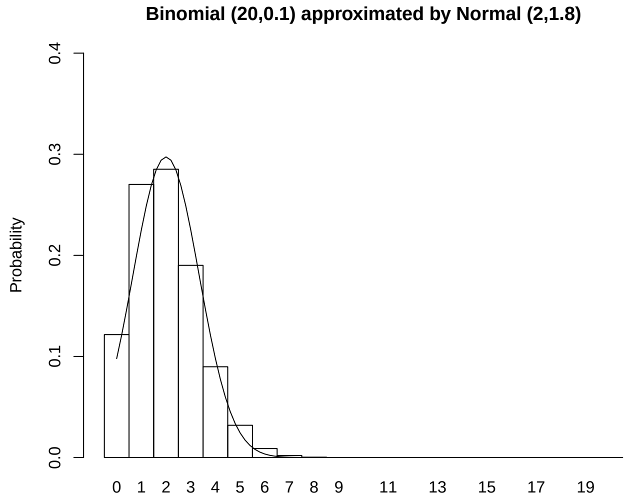
- Normal distribution tables are limited and only give us values of $\Phi(Z)$ for a fixed number of Z.
- Often, we want to know $\Phi(Z)$ for values of Z *in between* those listed in the tables.
- To do this we use *linear interpolation* - suppose we are interested in $\Phi(b)$, where $b \in [a, c]$ and we know $\Phi(a)$ and $\Phi(c)$.
- If we draw a straight line connecting $\Phi(a)$ and $\Phi(c)$ then (since Φ is smooth) we would expect $\Phi(b)$ to lie close to that line. Then

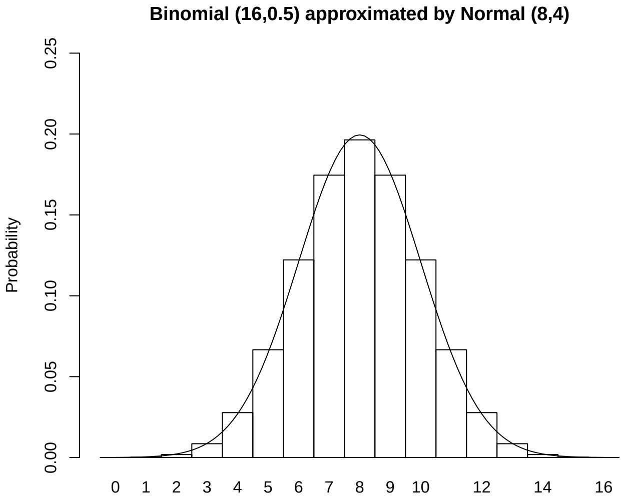
$$\Phi(b) \simeq \Phi(a) + \left(\frac{b-a}{c-a}\right)(\Phi(c) - \Phi(a))$$

Example

- Estimate the value of $\Phi(0.53)$ by interpolating between $\Phi(0.5)$ and $\Phi(0.6)$.

5.5 Normal Approximation to the Binomial





- Regardless of p, the $\mathrm{Bin}(n,p)$ histogram approaches the shape of the normal distribution as n increases. (This is actually a consequence of the *strong law of large numbers*; without going into more detail, the strong law simply says that certain distributions, under certain circumstances, converge to the normal distribution.)
- We can approximate the binomial distribution by a Normal distribution with the *same mean and variance*:

$$\mathrm{Bin}(n,p) \text{ is approximately } \mathrm{N}(np, np(1-p))$$

- The approximation is acceptable when

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

and the larger these values the better.
- Similarly to the normal approximation to the Binomial distribution, one can also justify a normal approximation to the *Poisson* distribution: In fact, for λ large (≥ 5), one has $\mathrm{Po}(\lambda) \approx \mathrm{N}(\lambda, \lambda)$.

Example: Electrical items
A certain machine produces every day $n = 1500$ electrical items. Each individual item is defect with probability $p = 0.02$. Compute the (approximative) probability that more than 40 items are defect on a given day.

Suggested Exercises: Q40–43.

---

6 The Sample Mean and the Central Limit Theorem

6.1 Experimental error

Experimental error

- We have an experiment to determine the quantity μ.
- We make a measurement X, which differs from μ by some error ϵ.
- Typically we think of each measurement as a guess (or "estimate") for μ, but with independent error.
- So we can express our measurement as
$$X = \mu + \epsilon,$$
where the error ϵ, and hence X, are random variables.
- If the experiment contains no systematic errors (bias) then $\mathrm{E}[\epsilon] = 0$.
- Let $\mathrm{Var}[\epsilon] = \sigma^2$, then
$$\begin{aligned} \mathrm{E}[X] &= \mu, \\ \mathrm{Var}[X] &= \sigma^2, \\ \mathrm{SD}[X] &= \sigma. \end{aligned}$$
- So the quality of the measurement X as a guess for μ depends on the size of the error variance, σ^2.

Improving our estimation

- **Q: Can we improve the quality of our guess at μ?**
- Yes
  - Take more measurements.
  - Combine the measurements into a summary statistic, e.g. take the average.
- Suppose we plan to make n (independent) measurements of a quantity of interest
- Each has its own independent variability (or measurement error) but has the same mean μ and sd σ.
- The first data summary that people calculate is the sample mean, $\bar{x}$, so we need to understand its theoretical behaviour.

6.2 The Sample Mean

The Setup

- There is a population of interest, with mean value μ and standard deviation σ.
- Often, **both** of these values are unknown.
- **Our goal is to say something about μ.**
- A secondary aim could be to say something about σ (more tricky).
- We want ideally to estimate μ with little or no bias and small variance, so that our guess is on target and can be expected to be close.

- We take a simple random sample of size n from this population.
- We will assume that the population is sufficiently large to make successive samples independent.
- We will label the sample values $X_1, X_2, \dots X_n$.
- **Before we see their values**, each observation is a random variable having the same mean $\mathrm{E}[X] = \mu$ and standard deviation σ.
- **After we see their values**, there is no longer any randomness in X and they are just numbers.
- We summarise this by writing that $X_1, \dots, X_n$ are independent and identically distributed (iid) with mean μ and standard deviation σ.

The Sample Mean

- We now define the sample mean to be
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
- Again, **before we see the data** the sample mean is itself a random variable $\bar{X}$.
- So it could take many possible values, and has a probability distribution, an expectation, and a variance.
- **After we see the data**, the sample mean is just a number $\bar{x}$.

Expectation and Variance

- The rules for the expected value and variance of a sum of random variables allow us to deduce that:
$$\begin{aligned} \mathrm{E}\left[\bar{X}\right] &= \mu, \\ \mathrm{Var}\left[\bar{X}\right] &= \frac{1}{n}\sigma^2, \\ \mathrm{SD}\left[\bar{X}\right] &= \frac{\sigma}{\sqrt{n}}. \end{aligned}$$
- Recall that the variance formula works only because these random variables are independent.
- Thus, the sample mean $\bar{X}$ has the same mean as the population of interest – we say $\bar{X}$ is an unbiased estimator of μ.
- Further, the sample mean has a **smaller standard deviation than a single observation**.
- By taking a sufficiently large sample size n, we could reduce its standard deviation to an arbitrarily small value.
- These two facts taken together imply that for sufficiently large n, the sample mean will turn out to give quite accurate estimates of the population mean μ, and the larger the value of n the better.

6.3 The Central Limit Theorem

The Central Limit Theorem

Theorem 1 (The Central Limit Theorem (CLT))**.** *Suppose $X_1, \dots, X_n$ are independent random variables with the same distribution (not necessarily Normal) with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Let $\bar{X}$ denote their mean rv.*
For large n, $\bar{X}$ is ***approximately Normal****.*

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

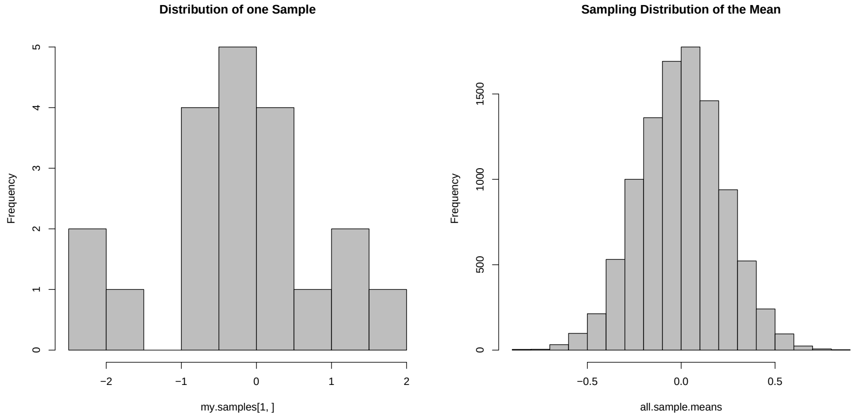
In particular,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1).$$

- The CLT is easily reformulated to a statement about the sum of random variables
- The CLT provides a plausible explanation for why the distributions of many random variables studied in physical or psychological experiments are approximately Normal.
- We now have a framework for improving the quality of our estimates for μ.
- We estimate μ by $\bar{X}$ (which is a random variable). If we take n large enough, we improve the quality of our estimate (proportionately to $\sqrt{n}$).
- We may then use the Normal distribution to make probability statements about our estimates.

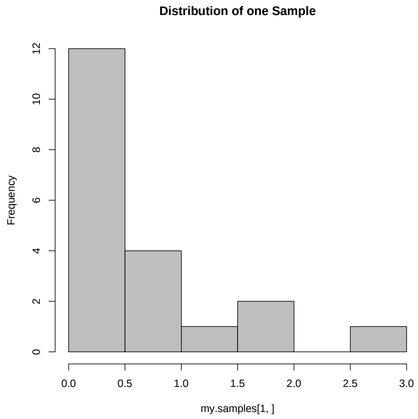
Simulation

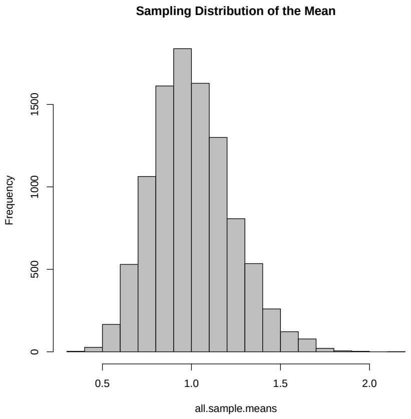
- We simulate a sample of size $n = 20$ from a $N(0, 1)$ distribution.
- Then we compute the sample mean $\bar{X}$ of the $n = 20$ observations. We do this, say, 10000 times (this number is not important) and produce a histogram of the 10000 *means*. This distribution is called the Sampling distribution of the mean.



Simulation

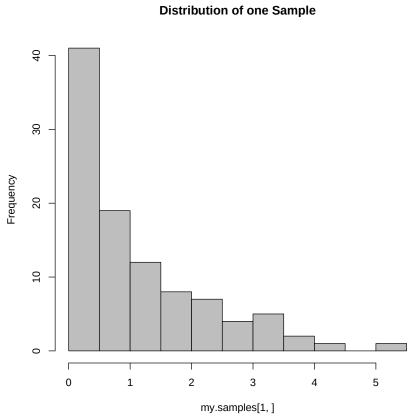
- We repeat the simulation, but now with samples of size $n = 20$ generated from the Exp(1) distribution.

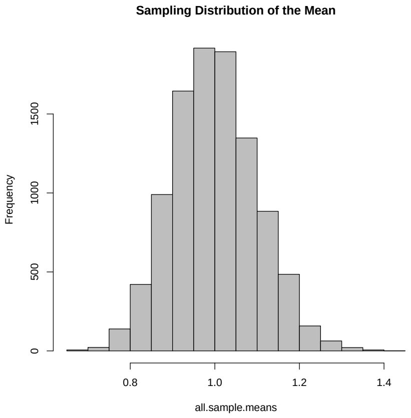




Simulation

- We repeat the simulation, but with samples of size $n = 50$ generated from the Exp(1) distribution.





Examples

- The random variables $X_1, X_2, \ldots, X_{10}$ are independent and uniformly distributed on the interval [0, 1]. Using the central limit theorem, approximate $P\left(\sum_{i=1}^{10} X_i > 7\right)$.
- A manufacturing process is designed to produce bolts with a 0.5cm diameter. Once a day, a random sample of 36 bolts is selected and the diameters recorded. If the average of the 36 values is less than 0.49cm or greater than 0.51cm, then the process is shut down for inspection and adjustment. The standard deviation for individual diameters is 0.02cm. Find approximately the probability that the line will be shut down unnecessarily (i.e. if the true process mean really is 0.5cm).

Suggested Exercises: Q44–47