

A guide for teachers - Years 11 and 12

Probability and statistics: Module 23

Random sampling



Education
Services
Australia



AMSI

AUSTRALIAN MATHEMATICAL
SCIENCES INSTITUTE

Random sampling - A guide for teachers (Years 11-12)

Dr Sue Finch, University of Melbourne
Professor Ian Gordon, University of Melbourne

Editor: Dr Jane Pitkethly, La Trobe University

Illustrations and web design: Catherine Tan, Michael Shaw

Full bibliographic details are available from Education Services Australia.

Published by Education Services Australia
PO Box 177
Carlton South Vic 3053
Australia

Tel: (03) 9207 9600
Fax: (03) 9910 9800
Email: info@esa.edu.au
Website: www.esa.edu.au

© 2013 Education Services Australia Ltd, except where indicated otherwise. You may copy, distribute and adapt this material free of charge for non-commercial educational purposes, provided you retain all copyright notices and acknowledgements.

This publication is funded by the Australian Government Department of Education, Employment and Workplace Relations.

Supporting Australian Mathematics Project

Australian Mathematical Sciences Institute
Building 161
The University of Melbourne
VIC 3010
Email: enquiries@amsi.org.au
Website: www.amsi.org.au

Assumed knowledge	4
Motivation	4
Content	6
Random sampling in finite populations	6
Mechanisms for generating random samples	8
Populations and sample frames	10
How biases can arise in sampling	12
Sampling from an infinite population	18
Sampling from Normal distributions	21
Sampling from exponential distributions	30
Sampling from the continuous uniform distribution	36
Sampling from the binomial distribution	40
Answers to exercises	47
References	51

Random sampling

Assumed knowledge

- A basic understanding of sampling, as covered by the series of TIMES modules *Data investigation and interpretation* (Years F–10), particularly the Year 8 module.
- The content of the modules:
 - *Probability*
 - *Discrete probability distributions*
 - *Binomial distribution*
 - *Continuous probability distributions*
 - *Exponential and normal distributions.*
- A familiarity with the idea of a random variable: A random variable X is a quantitative outcome of a random procedure. Random here refers to the inherent uncertainty of the outcome, rather than to something haphazard. Random variables can be discrete or continuous. The probability distribution of a discrete random variable X specifies the probabilities for possible values of X . The probability density function describes the probability distribution for continuous random variables.

Motivation

- What does it mean to take a ‘random’ sample?
- Can exit polls predict the outcome of an election?
- How do we know if our sample is random?
- What happens if we don’t take a random sample?
- Why is the Australian Census only conducted every five years?

In earlier years, students have seen different ways in which a ‘sample’ of data might arise or be used: surveying a sample of people, taking measurements on a sample of other objects (animals, trees, schools, companies and so on), conducting an experiment on a sample involving the random allocation of its members to different groups, and making

observations on different groups or samples. This is covered by the series of modules *Data investigation and interpretation* (Years F–10).

The context here for learning about random samples is to understand how they serve as a basis for relying on the sample data to provide quantitative information about the population from which the sample was taken. This is because, very often, the questions we ask are general in nature:

- Who do Australians prefer for Prime Minister?
- Who do Australian women prefer for Prime Minister?
- Is the proportion of Australian women who prefer candidate J for Prime Minister higher than the proportion of men with the same preference?
- What are the vitamin D levels of Australian newborns?
- How much ‘screen time’ do preschool children typically have per week?
- Does drug and alcohol use in adolescence predict success in adulthood?
- Will primary school children using ‘daily computer-assisted practice’ master arithmetic operations more quickly than those without access to the program?
- Does acupuncture have a stronger effect on the severity of migraine headaches than standard drug treatments?

In these examples, it is impractical (impossible!) to find an exact answer, that is, to find the exact value of the quantity of interest in the population (such as the proportion of women preferring candidate J as Prime Minister). Instead, we can obtain an estimate based on a sample.

There are many reasons for using samples. Most often, the cost in time and effort prohibits gathering information from the entire population of interest. It can also be easier to ensure the information is high quality in a sample. Importantly, with high-quality data collection methods and appropriate ways of selecting a sample, we can obtain accurate information about a population. In some cases, we want to infer a property of a population based on a sample from the population: the proportion of Australians who prefer candidate J . In other cases, we may wish to make a comparison of the properties of two populations: the proportions of Australian men and women preferring candidate J .

Populations are rarely static; it may not be possible to capture an entire population because it extends into the future. In asking about vitamin D levels in Australian newborns, it is likely that we want to draw a general conclusion that applies to newborns born today as well as those born tomorrow and in the future. Often we envisage that the conclusion drawn will be relevant to all humans, including humans in the future. This involves making an assumption about the stability of the world and its patterns.

Content

Random sampling in finite populations

Consider the two words in the term **random sample**. As the noun in the phrase suggests, this involves data ‘sampled’, or taken from, something else. The adjective, ‘random’, indicates that the mechanism used in obtaining the sample is based on probability, and not on conscious or unconscious preferences.

Random sampling has subtle aspects when considered formally. There are two important cases. The first is a random sample from a finite population of units.

A **unit** is a single member in a finite population we wish to study. A unit might be a person, animal, plant, school, company or other object. The **population** is the complete set of units we wish to study, and a **census** takes measurements on the entire population. A **sample** is a set of units (a subset of the population) that we take measurements on.

A **simple random sample** is a random sample selected by a method which ensures that all possible samples, of a given size, are equally likely to be chosen.

Example: Small raffle

Twelve players from a basketball club on an end-of-season trip check into a hotel. One of the twin rooms available has a balcony and a spa; the other rooms are basic. They decide to choose the two players who will get the best room by a simple raffle. Their twelve names are put into a hat and the contents are shaken well. The service manager at the reception desk is asked to draw two names out of the hat.

Assuming that the names in the hat are properly and randomly mixed, each possible pair of names is equally likely to be chosen. So to work out the chance of a particular pair being chosen, we need to find the number of distinct pairs. There are $\binom{12}{2} = 66$ possible pairs from among twelve individuals. So the chance of a particular one of these 66 pairs being chosen is equal to $\frac{1}{66} \approx 0.015$.

In this process, the order of the names is regarded as unimportant. So it does not matter, in the successful pair, which of the two names is drawn first.

The general result is that, for a simple random sample of size n chosen from a finite population of size N , the number of specific combinations is equal to the number of ways of choosing n units from N , and this is equal to

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N \times (N-1) \times (N-2) \times \cdots \times (N-n+2) \times (N-n+1)}{n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1}.$$

For a simple random sample, all of these combinations are equally likely, so the probability of a specific combination is $\frac{1}{\binom{N}{n}}$.

Example: Tattslotto

In the Saturday evening draw of Tattslotto, six winning balls are drawn at random from 45 balls numbered 1 to 45. There are 8 145 060 different ways of choosing six numbers from 45; that is, $\binom{45}{6} = 8\,145\,060$. So the probability of any specific combination of six balls being chosen is $\frac{1}{8\,145\,060}$.

An equivalent way to derive the same probability is to think of the balls being chosen in sequence. Consider a specific choice of six balls, such as $\{3, 4, 20, 37, 40, 45\}$. The probability that the first ball chosen is in this specific set is $\frac{6}{45}$. The conditional probability that the second ball chosen is one of the remaining five balls in the set, given that the first ball chosen was in the set, is equal to $\frac{5}{44}$, and so on. The conditional probability that the sixth ball chosen is in the set, given that the first five chosen were, is equal to $\frac{1}{40}$. Putting all this together using the multiplication theorem (from the module *Probability*), we obtain the probability of the specific set being chosen:

$$\text{Pr}(\text{specific set is chosen}) = \frac{6}{45} \times \frac{5}{44} \times \frac{4}{43} \times \frac{3}{42} \times \frac{2}{41} \times \frac{1}{40} = \frac{1}{8\,145\,060},$$

as before.

What kind of process is needed so that each possible combination has the same chance of selection? Each week we can watch the physical process that has been designed to try to ensure that a random sample of numbers is selected. The balls are numbered, but they need to be carefully calibrated in terms of their size, shape and weight. The balls are dropped into a transparent barrel and mixed by jets of air blowing into the barrel. After each ball is selected, jets of air mix the remaining balls again.

This process is to ensure sufficient random mixing of 45 balls so that every combination of six balls has the same probability of selection. This has the consequence that every ball in the barrel has the same chance of being selected.

The previous example illustrates an important property of random samples: what makes a sample random is how it is chosen, not what it consists of.

Exercise 1

Consider the mixing of the balls with jets of air in Tattslotto.

- a Is the selection of the first ball alone a random sample?
- b If six balls were selected with the initial mixing, but *without* the re-mixing after each ball was selected, would this be considered a random sample?
- c What may be the purpose of the re-mixing after each ball is selected?

Mechanisms for generating random samples

In some situations, generating a random sample can be straightforward. Tattslotto provides an exemplar. We have an accurate list of all the units in the population we wish to sample from (the 45 balls), and we use a random mechanism (thorough and chaotic mixing) in sampling. However, the population here is very small.

Tattslotto uses a physical randomising device. Other similar devices are referred to in the module *Probability*: for example, the use of marbles with numbers on them corresponding to dates for conscripting young men for service in Vietnam in the 1960s and 1970s.

Let's say you wish to take a random sample of 30 students from the 300 Year 11 and 12 students in your school. It is cumbersome to put 300 names in a container and shake them vigorously, and however we do it, we may have reservations about the effectiveness of the chaotic mixing, without something designed specifically for the purpose. It is more practical to use a random number generator in a computer.

As students are likely to have access to a computer with Microsoft Excel installed, we describe a method that uses Excel.

First obtain a list of all the students in a single column of an Excel worksheet. In the next column, enter the Excel formula `=RAND()` in the cell next to the first student name, and copy this formula down the column for 300 rows. This will generate, for each student, a random number between 0 and 1 from a uniform distribution. (Uniform distributions are discussed in the module *Continuous probability distributions*.)

The `RAND()` function is known (in Excel terms) as 'volatile', which means that the values will change every time an action is carried out on the worksheet. So it is important that, once a random number is generated for each student, the values (rather than the formulas) are saved. This can be done by copying the created numbers and using the 'Paste Special' function to paste the 'values'.

So now we have the list of 300 names, and alongside each name is a random number from the $U(0, 1)$ distribution, which means it is equally likely to be any number in the interval $(0, 1)$. To take a 10% sample of students we find, for example, the students with the lowest 10% of the random numbers. This can be done by sorting the name and random-number columns according to the values of the random numbers. It would be just as reasonable to find the students with the top 10% of random numbers; however, it is vital to decide which 10% will constitute the sample *before* assigning the random numbers.

It is sometimes said or thought that, if every unit in the population has the same chance of being chosen, then the sample is a simple random sample. This is not true, as the following exercise shows.

Exercise 2

Consider sampling the 300 students, in a different way, to obtain a sample of 150 students. Alongside each name, the numbers 1 and 2 are listed alternately $(1, 2, 1, 2, \dots)$. A fair coin is tossed. If the outcome is heads, the sample is taken to be the 150 students with a '1' next to their name. If the outcome is tails, the sample is the 150 students with a '2' next to their name.

- a Using this method, how many possible samples of 150 can be obtained?
- b Does every student have the same chance of selection?
- c Is this a simple random sample?

Sampling using groups

Practical sampling problems often involve large, sometimes complex, populations. Understanding the structure of the population can help in the design of a practical random sampling method. Sometimes the population can be divided into natural groups, or clusters. We may be able to get a good sample by looking at the clusters, and it can be more efficient and less costly than taking a simple random sample. For example, it is a lot more convenient to survey all students at 20 schools, rather than a simple random sample of 1000 school children. This is a random sample of clusters.

In other cases, we may wish to ensure that we sample from different groups, or strata, in our population. For example, we might wish to survey school children from different sectors (government schools, private schools, etc.). If we obtain simple random samples in each of a few strata of the population, this involves random samples within the strata.

Random selection is a vital element no matter what aspects of the structure of the population we wish to exploit in sampling. Physical randomisation devices might work well in simple situations, but most often computers are used to select random samples.

Example: Estimating the unemployment rate

Every month, the Australian Bureau of Statistics releases the national unemployment rate. It is estimated from a survey of the civilian Australian population aged 15 years or older. The survey samples dwellings, and questions are asked of the individuals in a sampled dwelling. The sampling of dwellings by the Australian Bureau of Statistics used to estimate the unemployment rate is done in several stages. Within metropolitan regions, for example, a geographic area is randomly sampled first. The geographic areas are divided into 'blocks' — groups of dwellings with boundaries like roads, parks and creeks — and a number of blocks are randomly selected. Finally, a set of dwellings is sampled within each block.

The Australian Bureau of Statistics uses random sampling at each stage of selection; a procedure like the selection of a simple random sample in Excel is used at each stage.

Populations and sample frames

In order to obtain a random sample from a defined population, we need to be able to describe the population of interest so that we can design a method to select a sample from the population. The set of units that describe the universe from which we can take a sample is called the **sample frame**. The sample frame is the 'practical' population: what we actually sample from. Even though it is often difficult to achieve this, it is important to make it match, as closely as possible, the real population of interest. In large populations this can be particularly challenging.

Consider, for example, how we could obtain a sample of Australian businesses with over 20 employees. We may be able to obtain lists from employer or business organisations. If we relied on such lists for our sample frame, we would have concerns such as:

- How do businesses with more than 20 employees get on the list? Do they have to be on it? If not, which businesses are typically omitted?
- How current is the information? How often is it updated?
- What is the quality of the information? If we are going to use the list for the purposes of contacting businesses selected in our sample, are the contact details accurately recorded?

As we will see, problems with the sample frame can seriously undermine the integrity of a sample.

Example: Estimating the unemployment rate (sample frame)

The sample frame used by the Australian Bureau of Statistics for estimating the unemployment rate describes dwellings in Australia by including three components: private dwellings, discrete indigenous communities, and non-private dwellings such as retirement homes and motels. The sample frame divides Australia into many small geographic areas. At the time of the Australian Census (every five years), a description of the dwellings within each of these small geographic areas is recorded. For the unemployment-rate survey, some of the small areas are sampled and then a subset of dwellings within each small area is sampled. Once a small area has been selected to be included, a check is made of the description of dwellings to make sure it is up-to-date.

The Australian Bureau of Statistics has a well-defined population and access to a vast sample frame. Each month the unemployment rate can be estimated from the sample selected from this frame.

Example: Jury duty

How are people selected for jury duty in the state of Victoria? In the last few years, over 60 000 people have been summoned (each year) to attend the courts for jury service.

Potential jurors are selected randomly from the Victorian electoral roll. In 2010, there were around 3.5 million voters enrolled in Victoria. Enrolment is compulsory for Australian citizens (and qualified British subjects) aged 18 years or over who have lived in Victoria for at least one month at their current address. Over the last few years, about 10% of the summoned potential jurors have been empanelled.

Exercise 3

Consider the process of obtaining a random sample of potential jurors in Victoria.

- a Assume that you want to take a sample of *eligible* voters. Would the electoral roll provide a perfect sample frame? Would there be any potential biases?
- b Assume that you had electronic access to the electoral roll and could take a simple random sample. What is the (approximate) probability of being selected as a potential juror in one year?
- c Could someone be called up for jury duty twice in one year? What would you need to know to determine this?
- d Assume that you only had access to a hard copy of the electoral roll, organised by electorate. What might be a practical way to take an approximately random sample? Would your method be a simple random sample?

How biases can arise in sampling

There is more than one way to select a sample badly, and great care is needed in both the design and the practical implementation of a sampling process to avoid biases arising. A sampling plan might be well designed, but if the plan is impractical in the field, the final sample will not be representative: it may be biased.

Convenience samples

Some methods of obtaining participants are particularly bad; relying on volunteers is one of them. This can result in strong bias. Often people who volunteer for a sample are more likely to have a vested interest in the topic of the survey than people who don't volunteer.

Recently, internet companies have begun recruiting people to join online survey panels; these are groups of people willing to answer online surveys, usually for some pay. When these people sign up for the panel, they do not have knowledge of the topics of the surveys they might be invited to participate in. However, they have agreed to answer surveys for reward, and this makes them a particular subset of the general population, and not a random selection.

Exercise 4

You will often see invitations to participate in surveys on the television and on the internet. An example is an online poll for the magazine *Cosmopolitan*, which invited readers to take part in a survey with the title 'Do you abuse prescription drugs?'

- a What do you think motivates people to participate in such polls, and who would be interested in being a part of this survey?
- b What kind of biases might arise in the results of this survey?

Problems with the sample frame

Obtaining a suitable sample frame is important for avoiding biases in the results obtained from a sample chosen from that frame. This can be tricky, as discussed in the previous section; the population of interest might not be static in space or in time. Simply selecting a sample where it is convenient may be logistically easier, but it might not give a fair representation of the population of interest.

Example: 1936 US presidential election (*The Literary Digest* poll)

In the presidential election held in the United States in 1936, the candidates were the incumbent President Franklin Roosevelt (Democrat) and Alf Landon (Republican).

The Literary Digest was a popular and widely read weekly magazine that ran a poll to predict the winner of the presidential race, and had done so correctly from 1920 to 1932. In 1936, *The Literary Digest* mailed a questionnaire to 10 million people to ask about their voting intentions. This extraordinary number of people included readers of *The Literary Digest*, registered car owners and people listed in the phone book. In one of the largest surveys ever (although not *The Literary Digest*'s largest), 2.4 million voters replied. The response rate — the percentage of people responding of those invited to participate — was $\frac{2\,400\,000}{10\,000\,000}$, or 24%.

The Literary Digest claimed 'The country will know to within a fraction of one per cent the actual popular vote of forty million.' The prediction that *The Literary Digest* made based on their survey was that Franklin Roosevelt would receive only 43% of the vote; Landon was predicted to win in a landslide. As you may know, Roosevelt won — he obtained 62% of the vote of around 40 million voters.

The failure of *The Literary Digest*'s poll was an embarrassment, and *The Literary Digest* subsequently went out of business; eventually its subscriber list was bought by *Time* magazine.

Why did *The Literary Digest* get the result so wrong? One problem was that the sample frame — the set of lists of names from which they recruited voters — was biased. Magazine readers, car club members and telephone subscribers tended to be relatively wealthy, and the wealthy at this time (during the great Depression) tended to be Republican voters. This is an example of a biased sample frame. The large number of people responding to the survey did not guarantee that the result would be accurate.

You might also wonder why *The Literary Digest* failed in 1936 when its reputation had been built on successful predictions of the results of earlier presidential elections. One reason is that economic conditions in the US were in decline, and in 1936 voting patterns related more strongly to economic circumstances than they had in the past. Biases in the sample frame used mattered less in earlier elections.

Example: 1936 US presidential election (American Institute of Public Opinion)

In 1936, the American Institute of Public Opinion also carried out a poll asking voters about their intentions in the upcoming presidential election. The institute's founder, George Gallup, understood that a very large sample would not necessarily provide an accurate result. Gallup had worked out what kinds of personal characteristics (including state, urban/rural residence, gender, age and income) related to voting patterns, and used these in the design of his sample. He set quotas for the numbers of individuals needed for each type of respondent, so that the number surveyed would reflect the population distribution. Gallup's method of filling quotas is not a random sampling method.

On 31 October 1936, Gallup reported the results of his final poll in his syndicated newspaper column 'America Speaks'; he predicted that Roosevelt would win the election with 56% of the vote. This prediction was 6% lower than the actual result, but much closer than *The Literary Digest's* prediction and, importantly, a correct prediction of a win for Roosevelt. Gallup's result was based on the responses of around 50 000 voters.

Why did the American Institute of Public Opinion's poll do better than *The Literary Digest's* poll? First, they asked about voting patterns in face-to-face interviews as well as by mail; they employed hundreds of interviewers across the country with quotas to fill, and so the survey did not suffer the same biases as *The Literary Digest's* poll. Second, the intention was to avoid bias in the sampling of individuals to be interviewed, although details about how this was carried out are sketchy.

Example: 1936 US presidential election (*The Literary Digest* versus Gallup)

George Gallup was relatively unknown before the 1936 presidential election. He had studied the polling methods used by *The Literary Digest* and thought he could do better. Gallup made a bold prediction of *The Literary Digest's* prediction of the election result before *The Literary Digest* made its own prediction! Gallup based his claim on a survey of 3000 people who were selected at random from the lists (sample frame) used by *The Literary Digest*. Gallup used the same method as *The Literary Digest* to collect information — by mail. Gallup predicted that *The Literary Digest* would call the result for Landon with Roosevelt obtaining only 44% of the vote. This was only 1% higher than the prediction *The Literary Digest* made. Gallup got enormous publicity.

Why did this American Institute of Public Opinion survey result correspond so well to *The Literary Digest's* result? The population that Gallup and his colleagues wanted to survey was the population of voters who were to be surveyed by *The Literary Digest*. For this, they had a very well-defined sample frame: the lists of magazine readers, registered car owners and telephone subscribers. Gallup sampled randomly from this list.

Problems with the response rate

When sampling human populations in particular, the willingness of invited participants to respond to the invitation and participate in the study can be an important issue. Increasingly, targeted potential respondents are *unwilling* to participate.

The **response rate** is the percentage of targeted potential respondents that actually respond. In sampling from a human population, there are many reasons why a sampled unit might not participate in the study. If the details in the sample frame are incorrect, the individual might not be able to be contacted. Contact details might be correct, but busy individuals may be hard to find. People might refuse to participate for a wide range of reasons: health, ethical issues relating to the study, lack of interest or insufficient time. This is an important issue in human surveys; it can be less problematic in surveys of other entities.

Example: 1936 US presidential election (response rates)

A common criticism of the 1936 *Literary Digest* poll is the low response rate of 24%. In Gallup's news report on 31 October 1936, where he gave the result of the final pre-election poll by the American Institute of Public Opinion, he stated that 'The number of ballots distributed in the poll was 312 551.' The final sample size for this poll is reported as 'around 50 000' (see the textbook listed in the *References* section).

There is little information available about the response rate for the American Institute of Public Opinion poll. Lawrence E. Benson, an associate of the American Institute of Public Opinion, reported a 17.3% response rate for the institute's mail surveys in 1936. It seems that the response rate to the American Institute of Public Opinion election poll must have been lower than that for the *Literary Digest* poll. Gallup would have needed about 75 000 responses for a response rate comparable to that of *The Literary Digest*.

Why is *The Literary Digest*'s response rate so often highlighted as a concern? One reason is that, when there is a poor response rate, there is a potential for non-response bias. We discuss this next.

Non-response bias

People who have been randomly selected to be part of a survey but refuse the invitation to participate can be different from the people who agree to participate.

Non-response bias occurs when the people who respond to the survey are different, on average, from those who do not. More generally, the units in a sample that cannot be contacted and the units in the sample that can be contacted may differ in important ways that relate to the purpose of the survey.

Often, unfortunately, the possibility for non-response bias is ignored. The lower the response rate, the more scope there is for non-response bias. If a large proportion of a sample fails to respond, having a large sample will not help: the results should be regarded as unreliable.

Exercise 5

Consider carrying out a long-term study of drug and alcohol use in young men, where you select a random sample of boys aged 13 to 16 and intend to follow them up several times into adulthood.

- a How might you select the sample of boys?
- b What issues might arise in following these adolescents over time?
- c Might these issues relate to the outcome we are interested in?
- d What kind of biases might this introduce?

Example

There was a state election in Victoria in late 2010. *The Age* is one of the newspapers published in Victoria, owned by Fairfax Media Limited. On 12 November 2010, *The Age* reported the outcome of an online poll they conducted. There were three main parties in Victoria at the time: Labor, the Greens and the Liberal/National coalition. In the poll, the percentage of respondents indicating that they would vote for Labor was two-thirds of those saying they would vote for the Greens, that is, considerably less.

Link

www.theage.com.au/victoria/state-election-2010/online-poll-sends-message-to-brumby-20101112-17q68.html

Exercise 6

Consider the 2010 *The Age* poll on voting intentions in the Victorian election. In the actual election, the vote for Labor was more than three times greater than that for the Greens.

- a There were over 27 000 respondents to the poll. Is that enough for it to be reliable?
- b Who was likely to have access to this poll?
- c Who was likely to respond?
- d *The Age* conceded that the poll was not scientific, but suggested that the results would disturb Labor supporters. Why was there a disclaimer?

Example: 1936 US presidential election (non-response bias)

The response rates for *The Literary Digest* and the American Institute of Public Opinion's polls of 1936 election were both relatively poor, allowing for the possibility of non-response bias. The purpose of both polls was to provide accurate estimates of the percentage of people voting for Roosevelt and for Landon. *The Literary Digest's* result was highly inaccurate; the American Institute of Public Opinion's poll was over 6% out. The voters responding to *The Literary Digest's* invitation to participate, in particular, tended to be Republican (Landon) voters; those who chose not to participate tended to be Democrat voters. Biases in the same direction are likely to have been operating in Gallup's pre-election poll, as his result underestimated the Democrat vote.

Scholars still debate the extent to which the spectacular failure of the *Literary Digest* poll was due to non-response bias versus a biased sample frame. It is likely that both were contributing factors.

Exercise 7

The Literary Digest's presidential poll of 1936 obtained responses from 2 400 000 voters from 10 000 000 sampled. Recall that 43% of the respondents indicated they would vote for Roosevelt.

- a If the poll was unbiased, what percentage vote for Roosevelt is predicted?
- b Consider how biased the survey could be. If all the non-respondents were to vote for Roosevelt, what is the predicted percentage vote for Roosevelt?
- c If all the non-respondents were to vote for Landon, what is the predicted percentage vote for Roosevelt?
- d If half of the non-respondents were to vote for Roosevelt, what is the predicted percentage vote for Roosevelt?
- e Roosevelt received 62% of the actual vote. What proportion of the non-respondents to the *Literary Digest* poll would have had to vote for Roosevelt so that his percentage vote among the 10 000 000 people sampled was 62%?

Sampling from an infinite population

So far we have been considering the first of the two cases of random sampling: sampling from a finite population. We now consider the second case: sampling from an infinite population.

The whole idea of an infinite population is clearly quite abstract. One way to think of it is to consider sampling from a finite population, and increasing the size of the population: suppose that the population size N tends to infinity. Sampling from an infinite population is handled by regarding the population as represented by a distribution. A random sample from an infinite population is therefore considered as a random sample from a distribution.

This means that there is an underlying distribution governing the random sample, typically making some values more likely than others, according to the shape of the distribution. The underlying distribution can be thought of as the distribution of some random variable X .

A **random sample 'on X '** of size n is defined to be n random variables X_1, X_2, \dots, X_n that are mutually independent and have the same distribution as X .

Over the remainder of this module, we discuss samples from different distributions. The approach is strongly visual, using diagrams to convey the important ideas.

The relevant distribution is not arbitrary: it is determined from first principles, or by appeal to the pattern in applicable historical data. We may think of the distribution of X as the underlying or 'parent' distribution, producing n 'offspring' that make up the random sample.

There are some important features of a random sample defined in this way:

- Any single element of the random sample, X_i , comes from the parent distribution, defined by the distribution of X . The distribution of X_i is the same as the distribution of X . So the chance that X_i takes any particular value is determined by the shape and pattern of the distribution of X .
- There is variation between different random samples of size n from the same underlying population distribution. Appreciating the existence of this variation and understanding it is central to the process of statistical inference, which is considered in the modules *Inference for proportions* and *Inference for means*.
- If we take a very large random sample from X , and draw a histogram of the sample, the shape of the histogram will tend to resemble the shape of the distribution of X .

- If n is small, the evidence from the sample about the shape of the parent distribution will be very imprecise: the sample may be consistent with a number of different parent distributions.
- Independence between the X_i 's is a crucial feature: if the X_i 's are not independent, then the features we discuss here may not apply, and often will not apply. And because there are n random variables, it is mutual independence that is required. This means that the conditional distribution of X_j , given the values of any number of the other X_i 's ($i \neq j$), is the same as the (unconditional) distribution of X_j . No matter what we are told about the other X_i 's, the distribution of X_j is unchanged.

A simple random sample from a very large finite population is approximately the same as a random sample from an infinite population. If we draw two numbers at random, without replacement, from a population consisting of the integers 1, 2, 3, 4, 5, the second number is clearly not independent of the first number. If we define $A =$ “first number drawn is 3” and $B =$ “second number drawn is 4”, then $\Pr(B) = \frac{1}{5}$, but $\Pr(B|A) = \frac{1}{4}$, and since these two probabilities are not equal, the events are not independent.

On the other hand, if we draw two numbers at random, without replacement, from a population consisting of the integers 1, 2, 3, ..., 10 000, then the corresponding events have the following probabilities: $\Pr(B) = \frac{1}{10\,000}$, but $\Pr(B|A) = \frac{1}{9999}$. These are different, so the two events are not independent, but they are very close, so the events in this case are approximately independent: $\Pr(B) \approx \Pr(B|A)$.

This point is illustrated as follows. Consider the following population of 100 numbers; the population distribution is shown. Think of them as marbles with numbers marked on them, positioned at the point corresponding to their number.

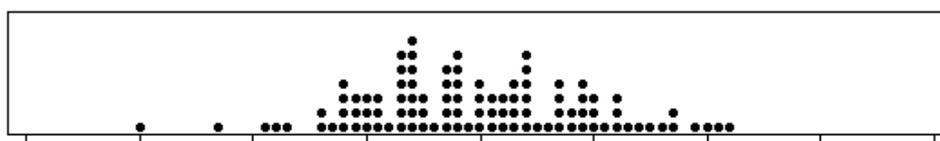


Figure 1: The distribution of a finite population of size $N = 100$.

Imagine selecting one of these marbles at random. This affects the population, perhaps noticeably: the removed marble is apparent. If a random sample of size $n = 50$ is taken from this population, it changes the population markedly: there is only half of the population left.

Now consider a much larger population of marbles with numbers on them: for example, $N = 10\,000$ (see figure 2). Removing one marble at random would not be noticeable; even taking a random sample of size $n = 50$ from this population will hardly change it all.

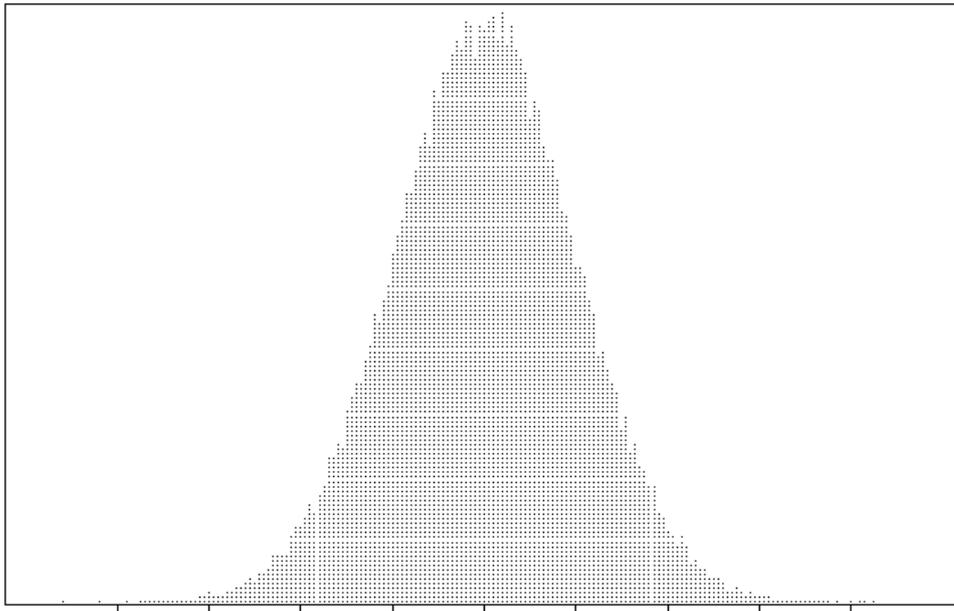


Figure 2: The distribution of a finite population of size $N = 10,000$.

Finally, imagine a huge population — think $N = 10^{100}$. Now the population is so vast that the marbles are essentially infinitesimally small (see figure 3).

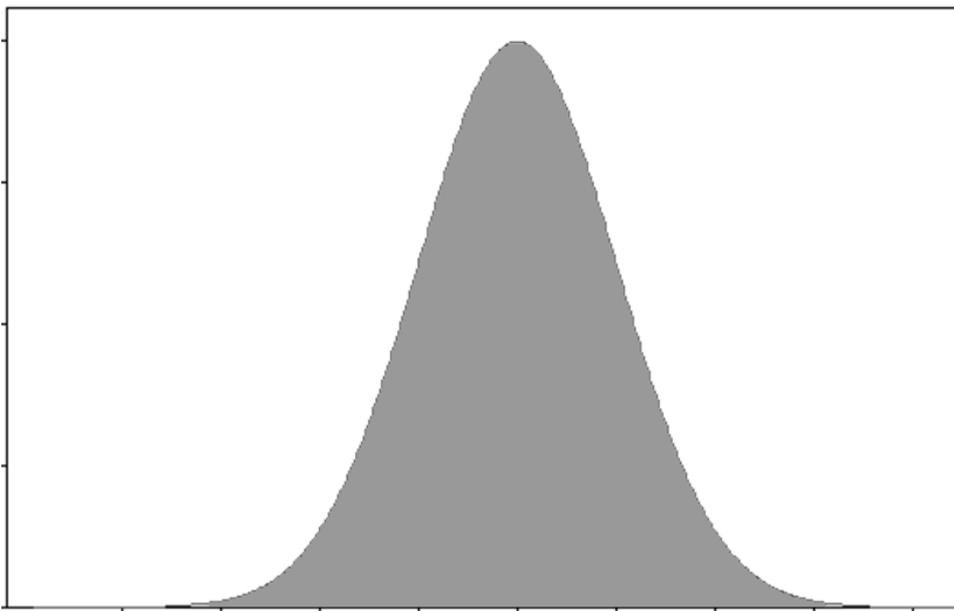


Figure 3: The distribution of a population of huge size.

Each time a marble is selected (from this ‘infinite’ pile of marbles), it will effectively be a selection from this distribution, no matter what other marbles have been selected. That is, each observation has the distribution shown in figure 3, independently of the other observations.

Sampling from Normal distributions

Normal distributions are introduced in the module *Exponential and normal distributions*. Suppose we are sampling from a Normal population with mean $\mu = 30$ and standard deviation $\sigma = 7$. For example, we could use this distribution to model the population of study scores of Year 12 students in a given subject.

This underlying distribution is shown in figure 4. Also shown is a random sample of size $n = 10$ from this distribution. The 10 observations making up the random sample are superimposed on the probability density function (pdf), to indicate that they come from this distribution. Any single one of the observations is more likely to be a value where the pdf is greater, than where it is smaller. Recall from the module *Continuous probability distributions* that, for such a random variable X , $\Pr(X \approx x) \approx f_X(x)\delta x$, where f_X is the pdf of X and we interpret $X \approx x$ as the event that X is in a small interval of width δx around x . For example, if $f_X(x_1) = 3f_X(x_2)$, then an observation near x_1 is approximately three times more probable than an observation near x_2 . This phenomenon is reflected visually in figure 4, in which the observations shown in the pdf are the dots ‘floating’ above the x -axis: there is more room for dots above x -values where the pdf is greater.

Equivalently, we can think of the sample as being obtained by considering the x - y plane and choosing n points randomly from the region under the curve: $\{(x, y) : 0 < y < f_X(x)\}$, where $f_X(x)$ is the pdf of X .

A connection with figures 1, 2 and 3 is useful. We can think of figure 4 as containing an infinite pile of marbles in the shape of the region under the curve: the black ones are the ones that happen to be selected.

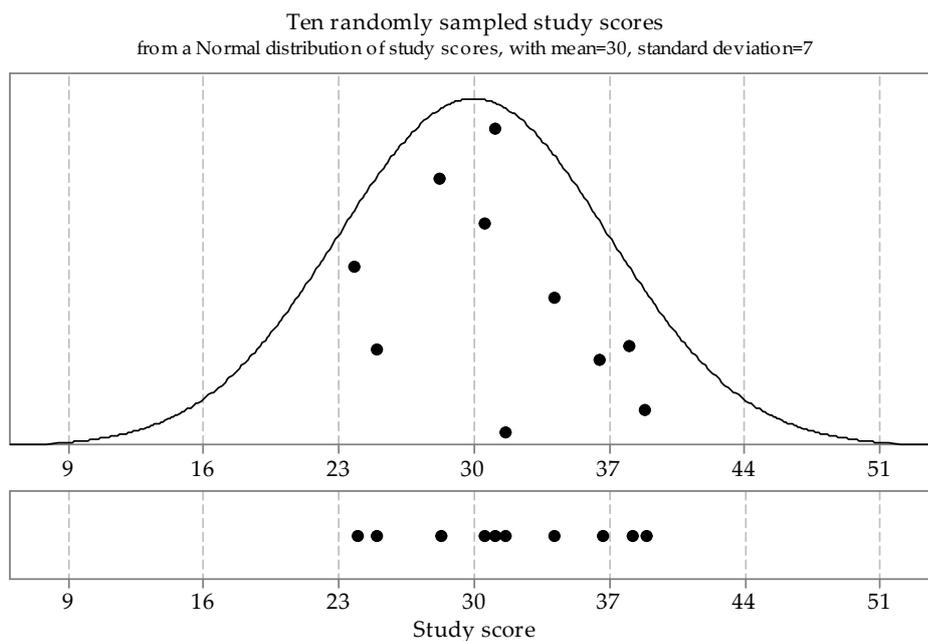


Figure 4: First random sample of size $n = 10$ from $N(30, 7^2)$.

The sample has been projected down to the x -axis in the lower part of figure 4 to give a dotplot of the data. Recall that a dotplot is a simple way to show the distribution of a small sample.

Figure 5 shows another random sample of size 10 from the same parent Normal distribution $N(30, 7^2)$. This illustrates the utterly basic but fundamentally important point that repeated samples from the same distribution are different from each other. This is obvious, since they are made up of individual observations from the population distribution, and we know that individual observations vary. Nonetheless, it is an important lesson to absorb; students who are accustomed to the deterministic reproducibility of mathematical relations may find this uncertainty disconcerting.

To make the point about this sample-to-sample variation, both the first and second samples are shown in the lower part of figure 5.

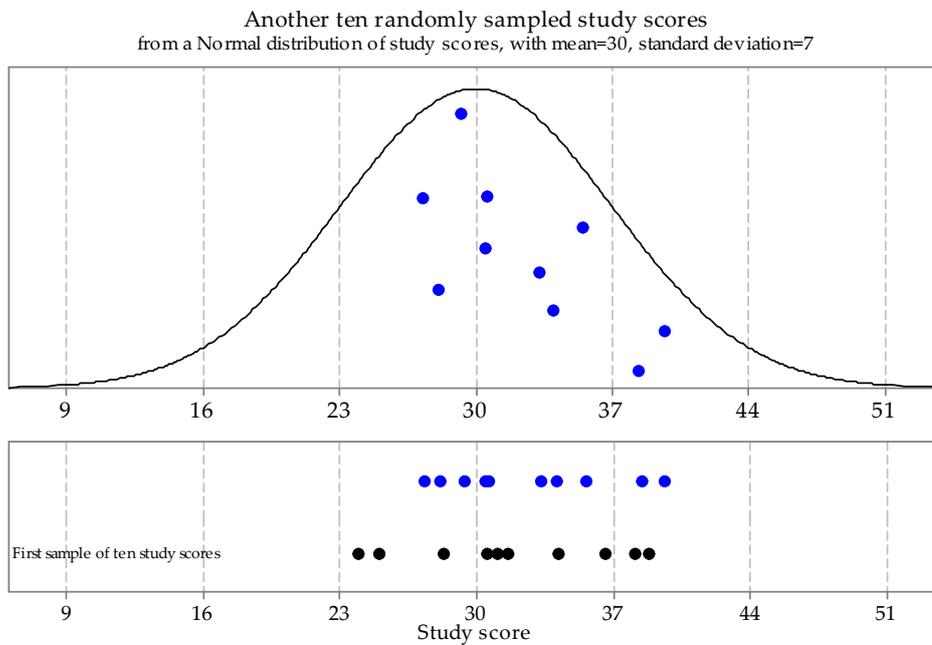


Figure 5: A second random sample of size $n = 10$ from $N(30, 7^2)$.

For good measure, figure 6 shows a third sample from the same underlying Normal distribution $N(30, 7^2)$, with lined up dotplots of the three random samples.

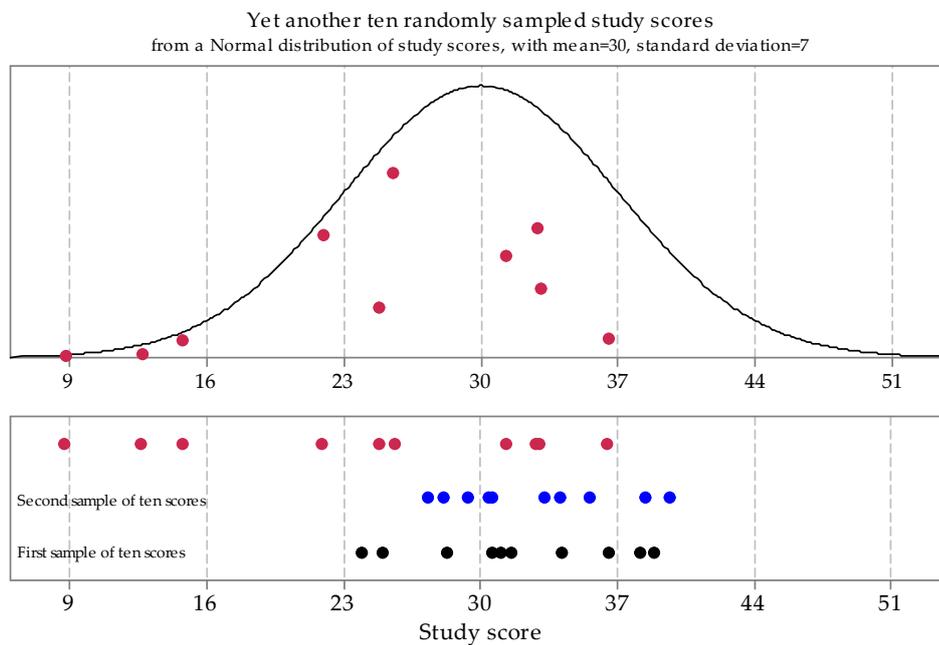


Figure 6: A third random sample of size $n = 10$ from $N(30, 7^2)$.

Now we will abandon the visual representation of the underlying Normal distribution $N(30, 7^2)$, in order to present a large number of samples of size 10 from this distribution. Figure 7 shows 100 samples of size 10 from this distribution. The choice of 100 as the number to show is arbitrary; all that is intended is to show a large enough number to get an idea of how much variation there can be among such samples. Note the extent of the differences between the samples in both location and spread.

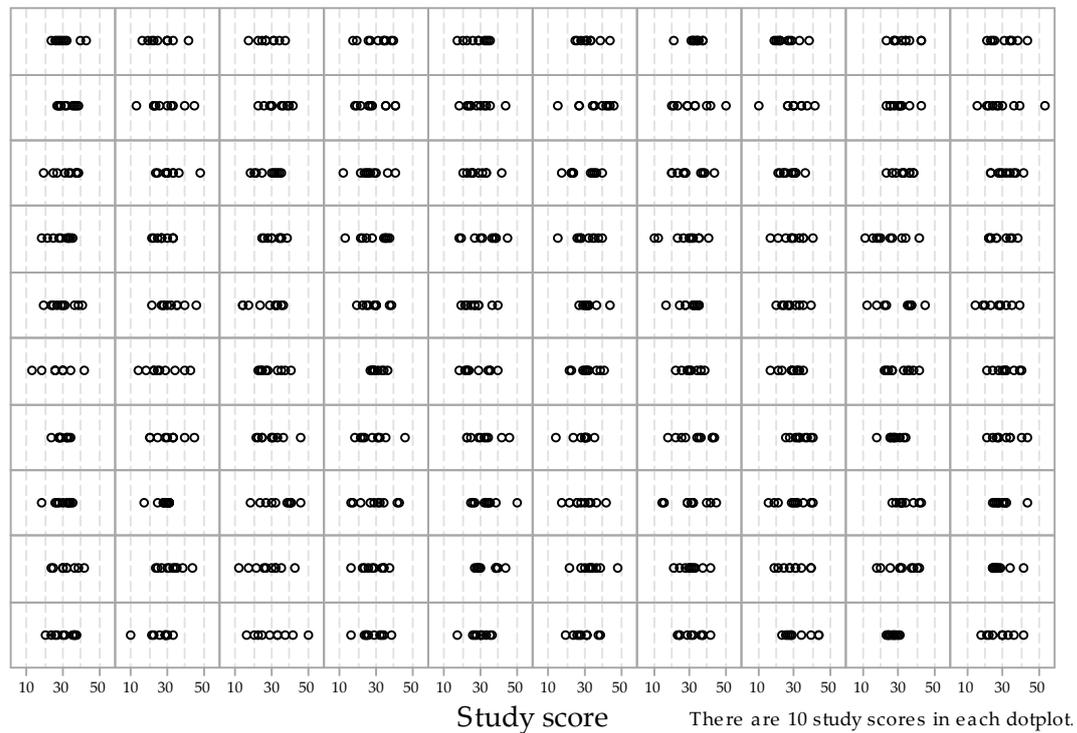


Figure 7: Dotplots of 100 random samples of size $n = 10$ from $N(30, 7^2)$.

Exercise 8

- a For a random sample of size 10 from an $N(30, 7^2)$ population, what is the probability that all 10 observations are below 30?
- b What is the probability that all 10 observations are above 30?
- c Hence, what is the probability that all 10 observations are on one side of 30?
- d Inspect figure 7 closely. Do any of the 100 samples have the feature that all 10 observations are on one side of 30?
- e Try to find the sample in figure 7 with the largest mean and the sample with the smallest mean.
- f Now consider the third random sample, represented in figure 6. Notice that there are three observations below 16 and, based on the pdf shown, these are quite low values. How unusual is such a sample?
 - i For the random variable $X \stackrel{d}{=} N(30, 7^2)$, find $\Pr(X \leq 16)$.
 - ii Suppose that a random sample of size 10 is taken from this distribution. Let Y be the number of observations in the sample that are less than or equal to 16. What is the distribution of Y ?
 - iii Hence, find $\Pr(Y \geq 3)$.
 - iv Why is it relevant to calculate $\Pr(Y \geq 3)$, and not $\Pr(Y = 3)$?

Now we begin to consider the effect of a larger sample size. Figure 8 shows histograms (not dotplots) from 100 random samples of size 20 from the same underlying distribution $N(30, 7^2)$. The same x - and y -scales are used throughout. Notice how much variation there can be: some look reasonably symmetric and unimodal, while others have marked skewness; some look relatively flat. But all are random samples on $N(30, 7^2)$. Remember that a random sample is defined by how it was chosen, and not what it consists of.

Frequency

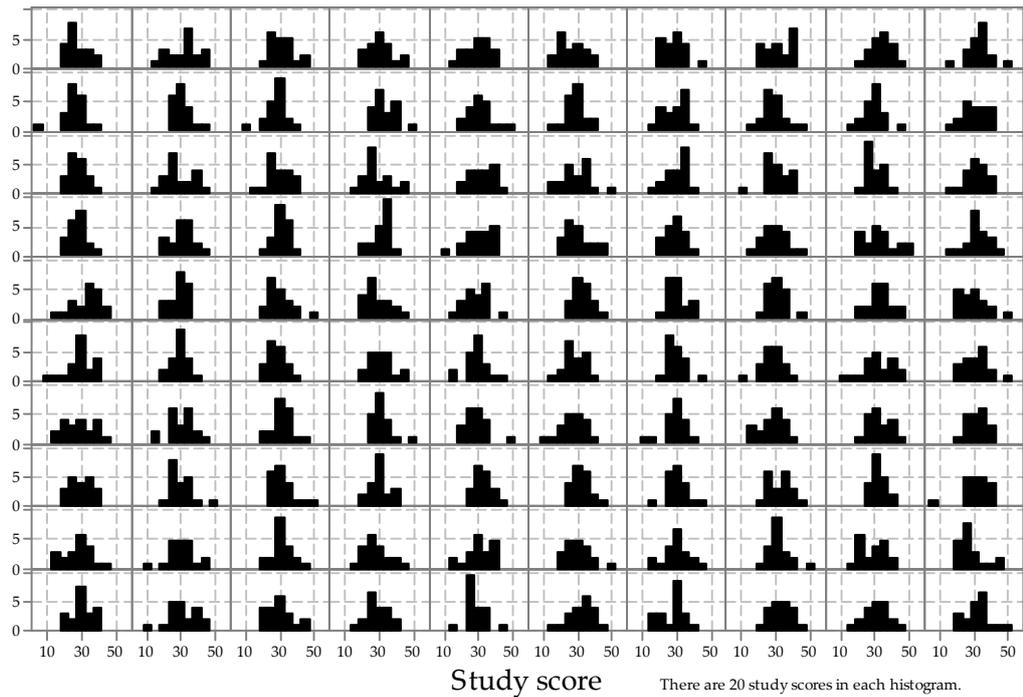


Figure 8: Histograms of 100 random samples of size $n = 20$ from $N(30, 7^2)$.

What happens if we increase the sample size to 100? This is illustrated in figure 9.

A careful comparison of figures 8 and 9 shows that the histograms in figure 9 are less variable, and closer to the shape of the underlying distribution, than in figure 8. However, even for $n = 100$, there is quite a lot of variation between the random samples from the same distribution.

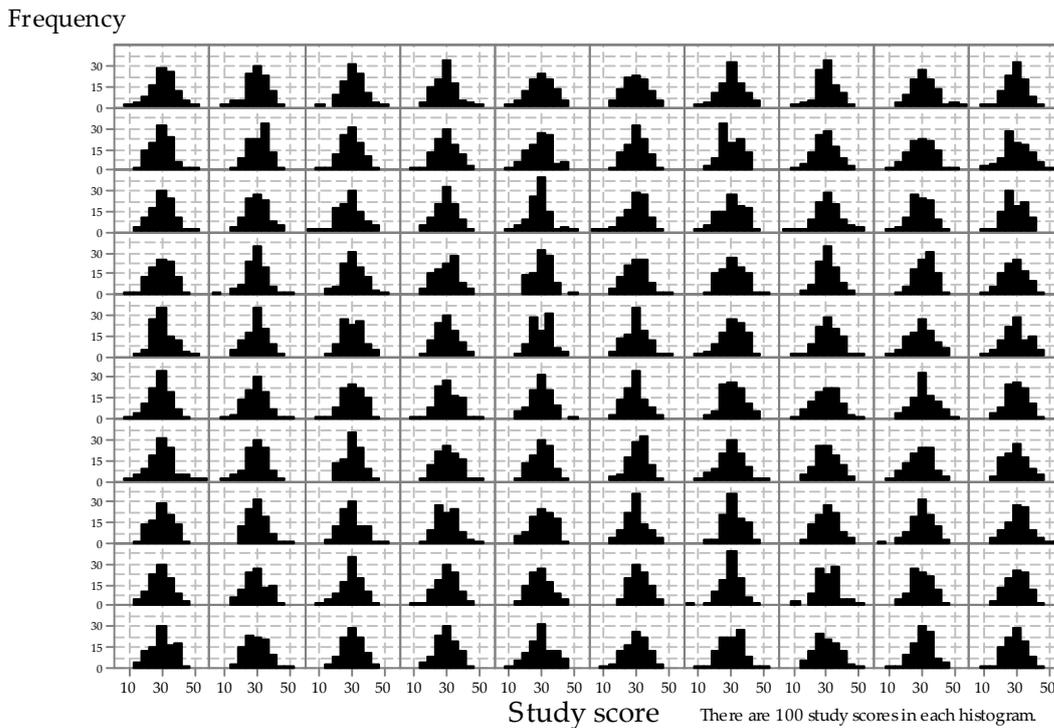


Figure 9: Histograms of 100 random samples of size $n = 100$ from $N(30, 7^2)$.

To get a clearer picture of what happens as the sample size n increases, figure 10 shows histograms from samples of very different sizes, from $n = 25$ in the first row through to $n = 10\,000$ in the last row. To facilitate comparisons, the same 'bin widths' and y -axis scale have been used throughout. Unlike the previous figures, there are 10 histograms for each sample size.

This figure shows clearly that, as the sample size increases, the histogram becomes more and more similar to the underlying distribution. Looking across the rows, the histograms are more similar to the other histograms (in the same row) for larger sample sizes.

Each histogram shows

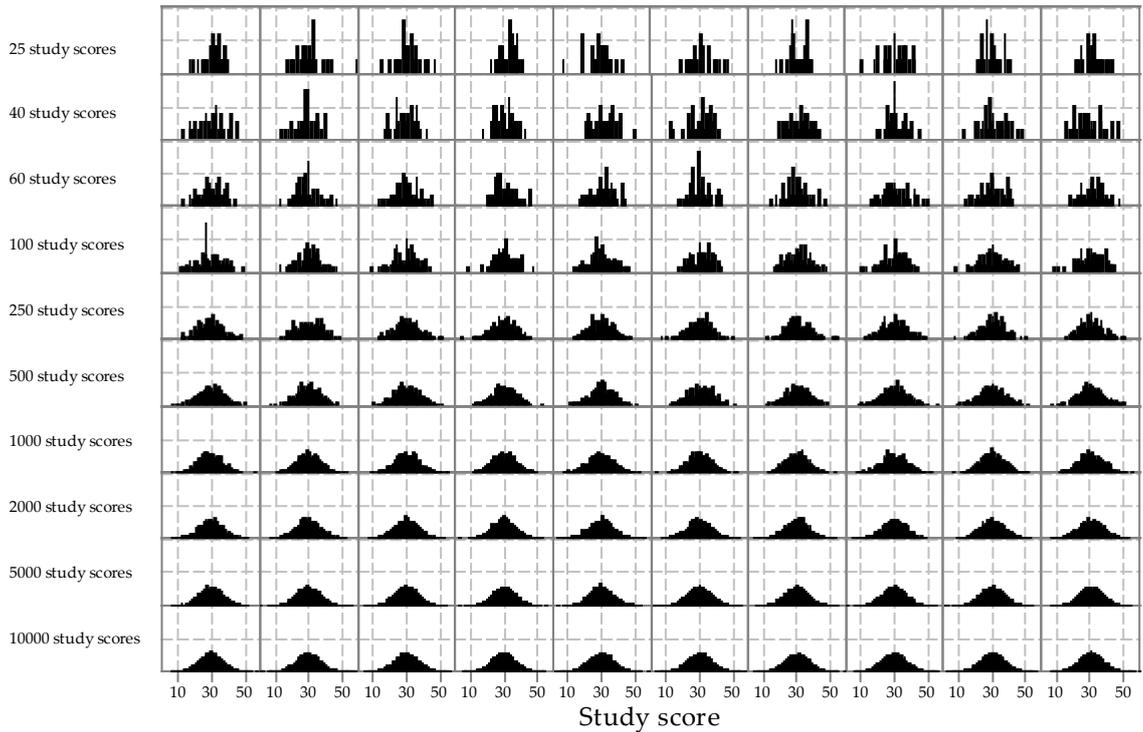


Figure 10: Histograms of random samples of varying size from $N(30, 7^2)$. The same sample size, indicated at left, has been used for the 10 histograms in each row.

The same Normal distribution has been used in all cases so far, namely $N(30, 7^2)$. Figure 11 illustrates sampling from Normal distributions with different means and standard deviations. In the upper part of the figure, the pdfs of 25 Normal distributions are shown. Each column has the same mean μ , shown by the label at the top of the column; each row has the same standard deviation σ , shown by the label at the left of the row.

In the lower part of the figure, histograms of random samples from the corresponding distributions are shown; the row and column position of the histogram corresponds to that of the distribution from which it came. The same sample size $n = 100$ has been used throughout. Note how the location and spread of the samples reflect the location and spread of the parent distributions. The locations of the histograms increases from left to right, ‘tracking’ what happens with the mean μ . Similarly, the spread of the histograms increases from top to bottom, following the standard deviation σ .

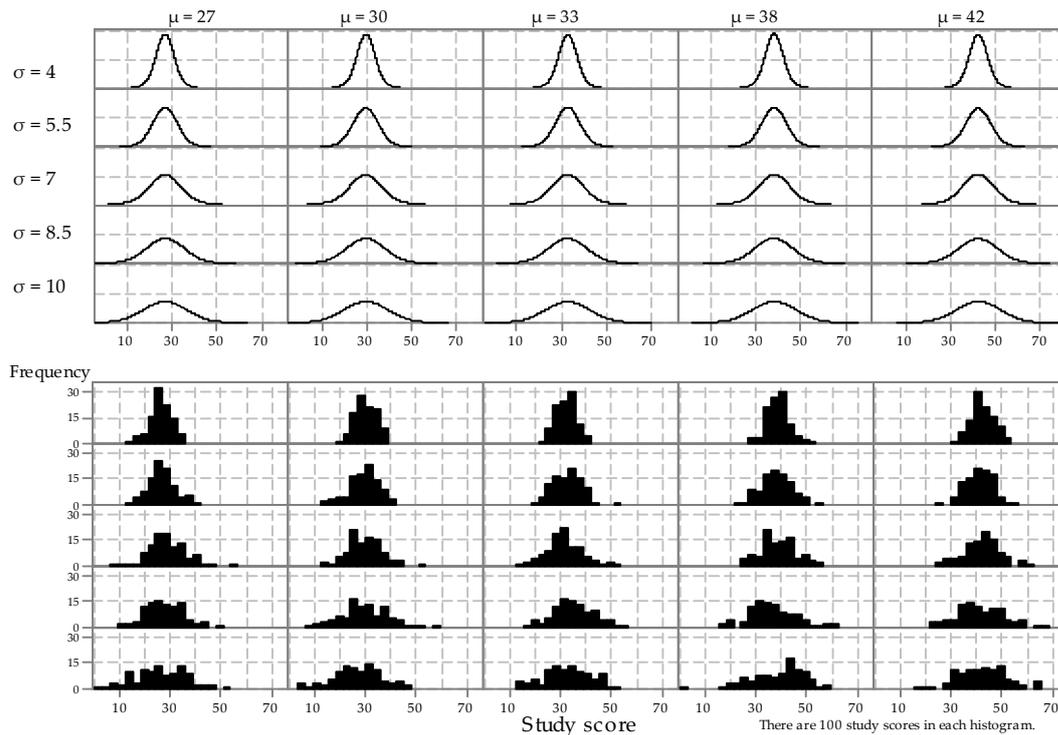


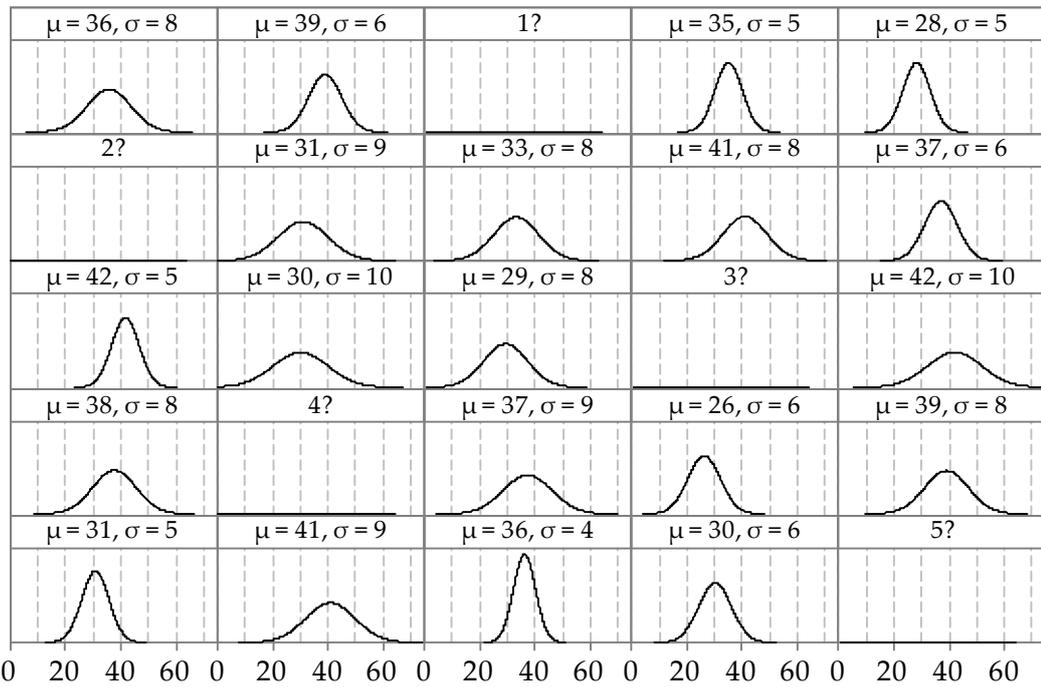
Figure 11: 25 different Normal distributions (upper part) and corresponding histograms of random samples from each of them (lower part); the same sample size $n = 100$ has been used throughout, and the same scales and bin widths have been used in all histograms.

There is a clear link between a pdf in the upper part of figure 11 and the corresponding histogram in the lower part. This suggests that we may be able to make reasonable inferences about the location and spread of a Normal distribution based on a random sample from that distribution when we are not ‘given’ the pdf. This reasoning is at the heart of statistical inference. The next exercise explores this further.

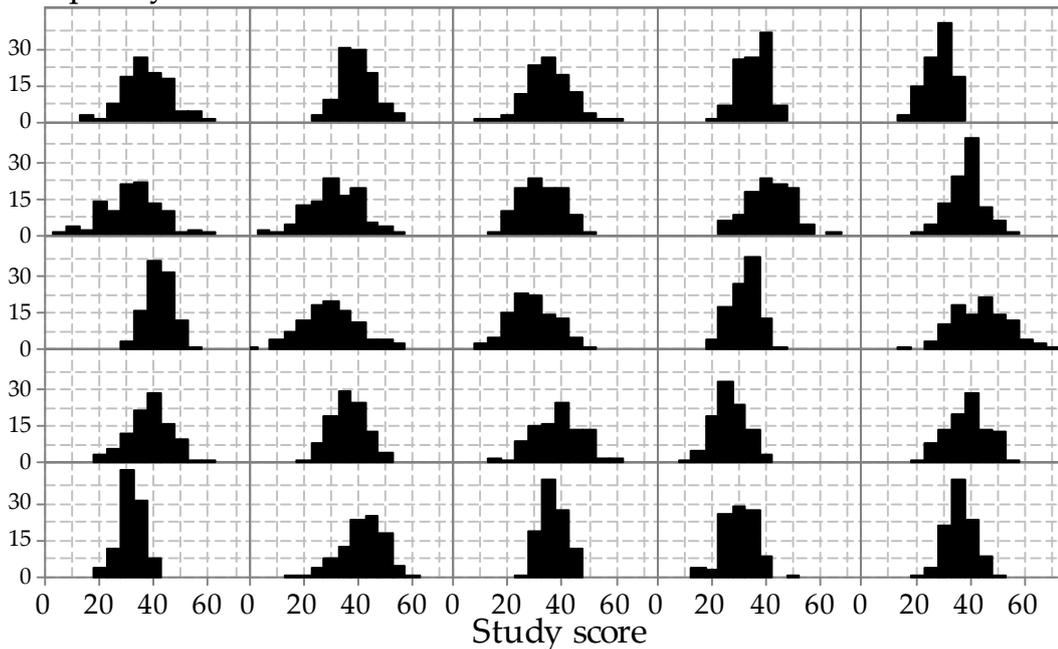
Exercise 9

Figure 12 is similar in design to figure 11. However, it uses a variety of different population means and standard deviations. The labels for the distributions are shown above them. The pdfs are shown for 20 cases; for the other five, labelled ‘1?’ to ‘5?’, the histogram is shown but the pdf from which the random sample was taken is not.

Assuming that the samples come from a Normal distribution, estimate the population mean μ and population standard deviation σ , for each of the five missing pdfs. Like the distributions shown, the five missing distributions have integer means and standard deviations.



Frequency



There are 100 study scores in each histogram.

Figure 12: 25 different Normal distributions (upper part; five are missing) and histograms of random samples from each of them (lower part); the same sample size $n = 100$ has been used throughout.

Sampling from exponential distributions

In this section, we look at random samples from exponential distributions. The sequence of diagrams is similar to that in the previous section *Sampling from Normal distributions*.

A very important message is that there are *general* features here that are the same as the Normal case. Of course, the shape of the exponential distribution is quite different from the Normal distribution, and hence the distributions of the random samples, shown in the dotplots and histograms, reflect that. However, other features apply regardless of the underlying distribution. For example, with very large sample sizes, the distributions of the samples tend to look like the parent distribution from which the samples are taken: in this case, the exponential distribution.

For this reason, much of the discussion from the previous section *Sampling from Normal distributions* applies here, and is therefore not repeated in detail.

So that the discussion has a suitable context, we use an example from the module *Exponential and normal distributions*. In the example, it is assumed that the underlying random variable represents the interval between births at a country hospital, for which the average time between births is seven days. We assume the distribution of the time between births follows an exponential distribution.

Recall from the module *Exponential and normal distributions* that, if the random variable T has an exponential distribution with rate α , which we write as $T \stackrel{d}{=} \exp(\alpha)$, then T has the following pdf:

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, $E(T) = \frac{1}{\alpha}$. So, if we are using a time unit of days and the mean is seven days, this implies that $\alpha = \frac{1}{7}$.

We take a random sample of size n from this distribution. In the context of the example, this could be considered a consecutive sequence of n births, assuming that the times between successive births are independent. (Obviously, multiple births violate this assumption, but we may deal with this by defining a ‘birth’ to be a birth event to one mother at one time, so that twins and other multiple births count as one ‘birth’ for these purposes.) It is important to keep in mind that, as usual, the model used is only intended to give some context, and it is an idealised representation with assumptions.

Figures 13 to 15 show three independent random samples from the $\text{exp}(\frac{1}{7})$ distribution. Note the variation between them.

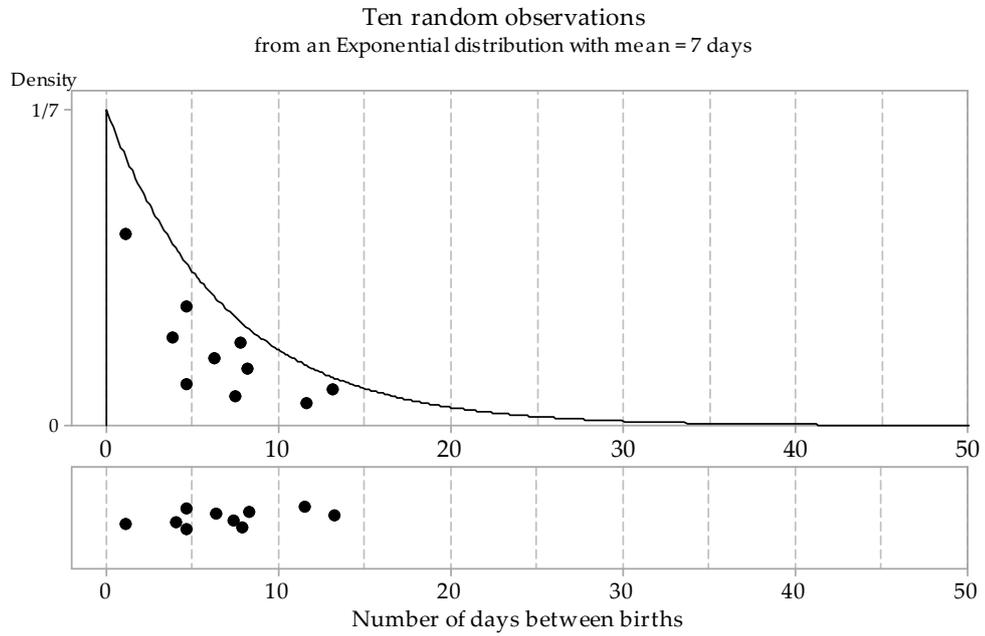


Figure 13: First random sample of size $n = 10$ from the $\text{exp}(\frac{1}{7})$ distribution.

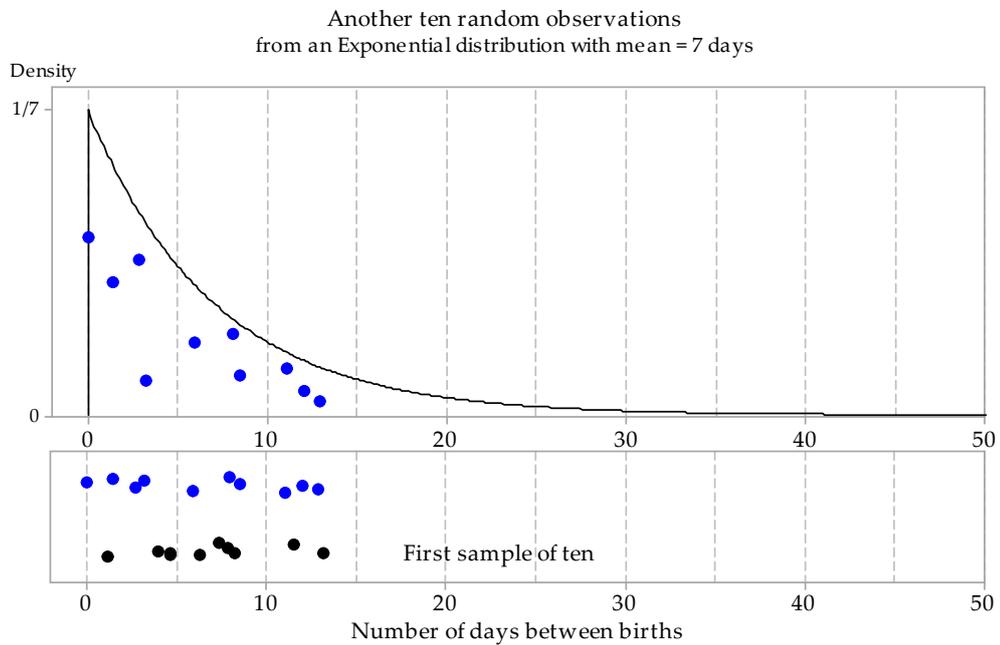


Figure 14: Second random sample of size $n = 10$ from the $\text{exp}(\frac{1}{7})$ distribution.

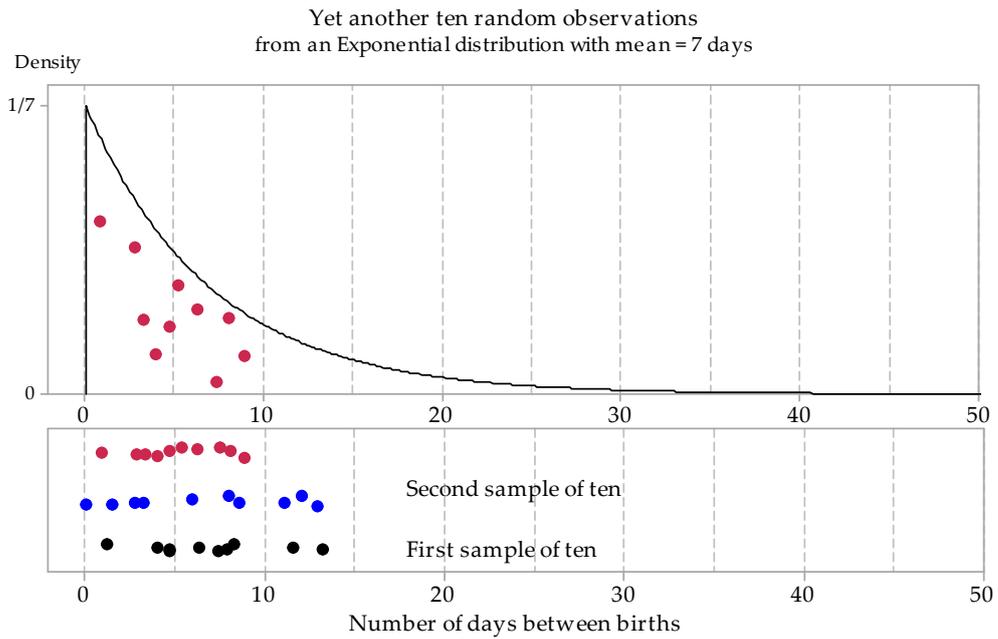


Figure 15: Third random sample of size $n = 10$ from the $\text{exp}(\frac{1}{7})$ distribution.

Figure 16 shows the distributions of 100 random samples, each of size $n = 10$, from the $\text{exp}(\frac{1}{7})$ distribution. The dotplots have been ‘jittered’ slightly and randomly in the vertical direction to assist with the detection of points that are very close to each other.

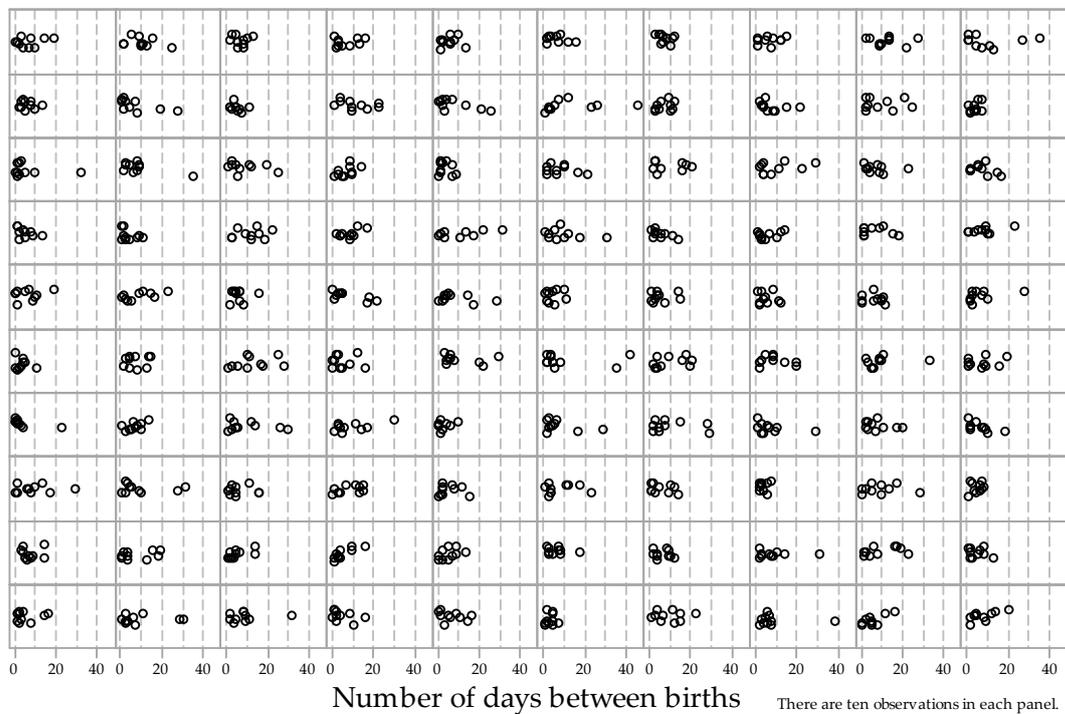


Figure 16: Dotplots of 100 random samples of size $n = 10$ from the $\text{exp}(\frac{1}{7})$ distribution.

Among the first three random samples of size 10, there was no single observation greater than 15 days. Among the 100 random samples of size 10, there are quite a few intervals of greater than 15 days, and one greater than 40 days. This gives another perspective on the simple point that samples vary. If some of the values in the parent population are unlikely, they will — correspondingly — not occur very often in random samples. But if we take enough samples, an unusual individual value will be likely to ‘turn up’ eventually.

What happens if the sample size is larger? Figures 17 and 18 show 100 histograms (in each case) for random samples of size $n = 20$ (figure 17) and $n = 100$ (figure 18) from the $\exp(\frac{1}{7})$ distribution. Note that the histograms in figure 18 tend to be closer to the shape of the parent distribution.

Frequency

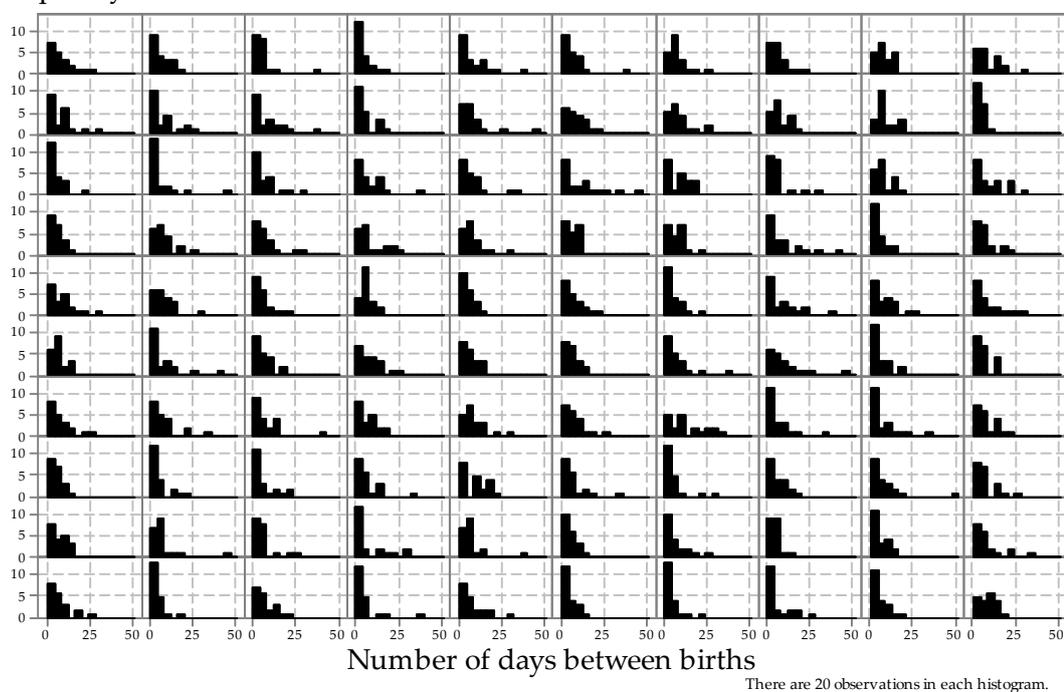
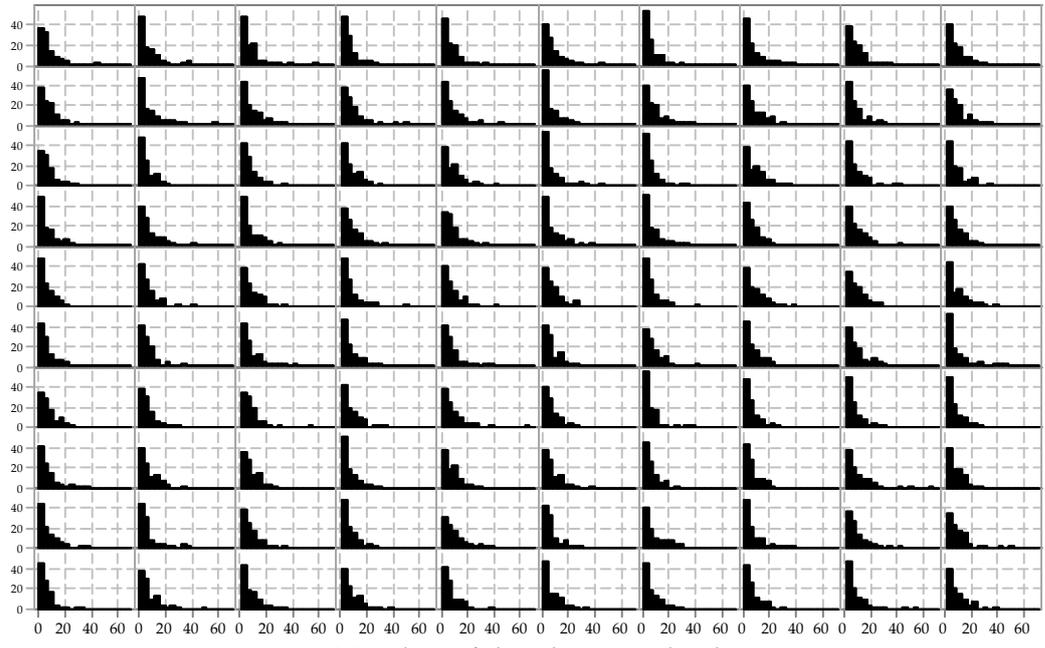


Figure 17: Histograms of 100 random samples of size $n = 20$ from the $\exp(\frac{1}{7})$ distribution.

Frequency



Number of days between births There are 100 observations in each histogram.

Figure 18: Histograms of 100 random samples of size $n = 100$ from the $\exp(\frac{1}{7})$ distribution.

Figure 19 explores more extensively the effect of varying sample sizes. As we saw for sampling from the Normal distribution, as the sample size increases the histogram becomes more and more similar to the underlying exponential distribution. Looking across the rows, the histograms are more similar to the other histograms (in the same row) for larger sample sizes.

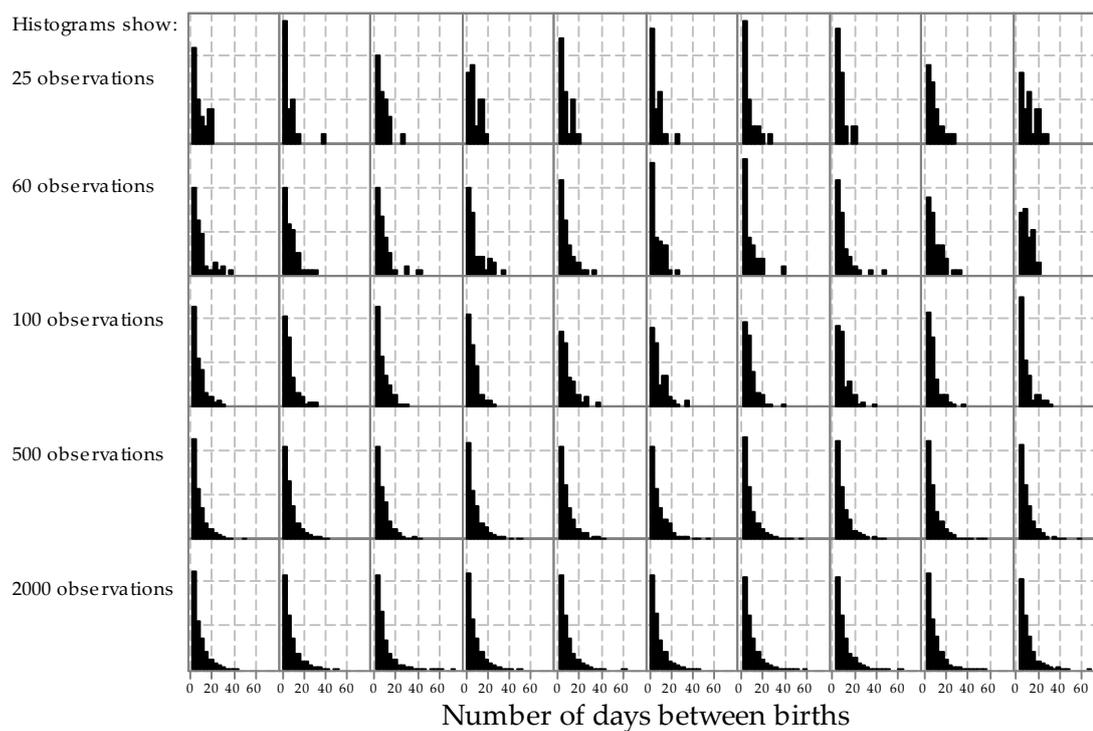


Figure 19: Histograms of random samples of varying size from the $\exp(\frac{1}{7})$ distribution. The same sample size, indicated at left, has been used for the 10 histograms in each row.

Finally, figure 20 shows samples from a number of exponential distributions with different means. Recall that the mean μ for an exponential random variable is equal to $\frac{1}{\alpha}$.

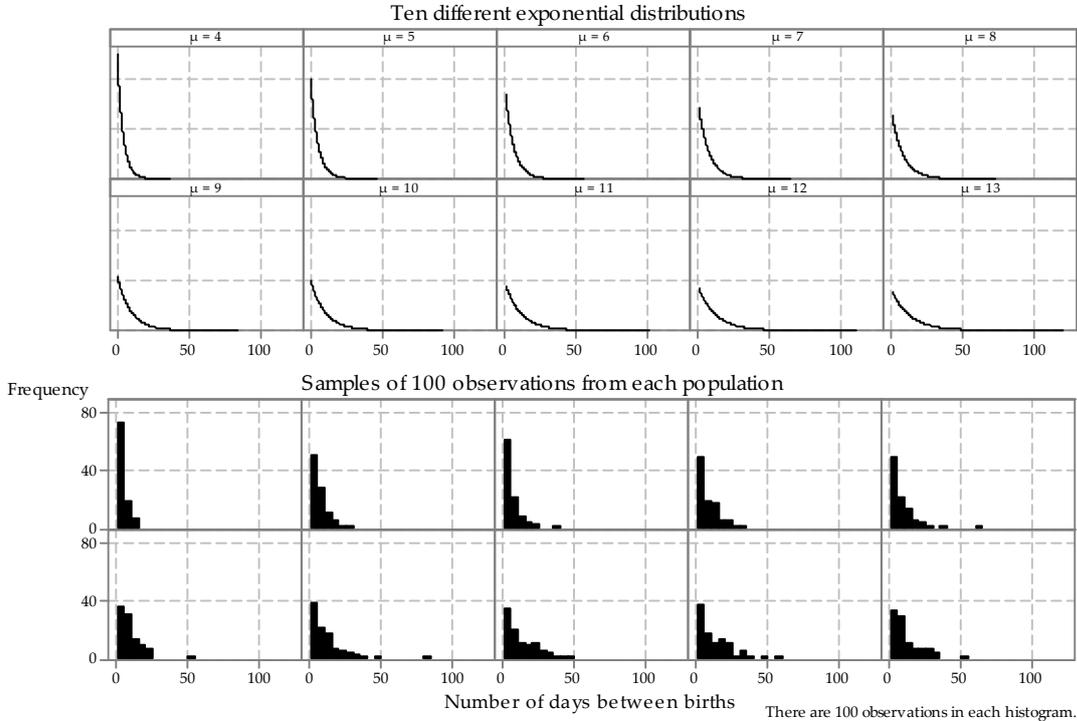


Figure 20: Random samples from different exponential distributions.

Sampling from the continuous uniform distribution

Now we consider samples from the continuous uniform distribution.

Recall from the module *Continuous probability distributions* that, if the random variable U has a uniform distribution on the interval (a, b) , which we write as $U \stackrel{d}{=} U(a, b)$, then U has the following pdf:

$$f_U(u) = \begin{cases} \frac{1}{b-a} & \text{if } a < u < b, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if $U \stackrel{d}{=} U(0, 1)$, then

$$f_U(u) = \begin{cases} 1 & \text{if } 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We saw this distribution in the section *Mechanisms for generating random samples*, where we discussed obtaining a random sample in Excel. The Excel function `RAND()` gives observations from this distribution.

Figures 21 to 23 show three independent random samples from the $U(0, 1)$ distribution, each of size $n = 10$. Note how they vary.

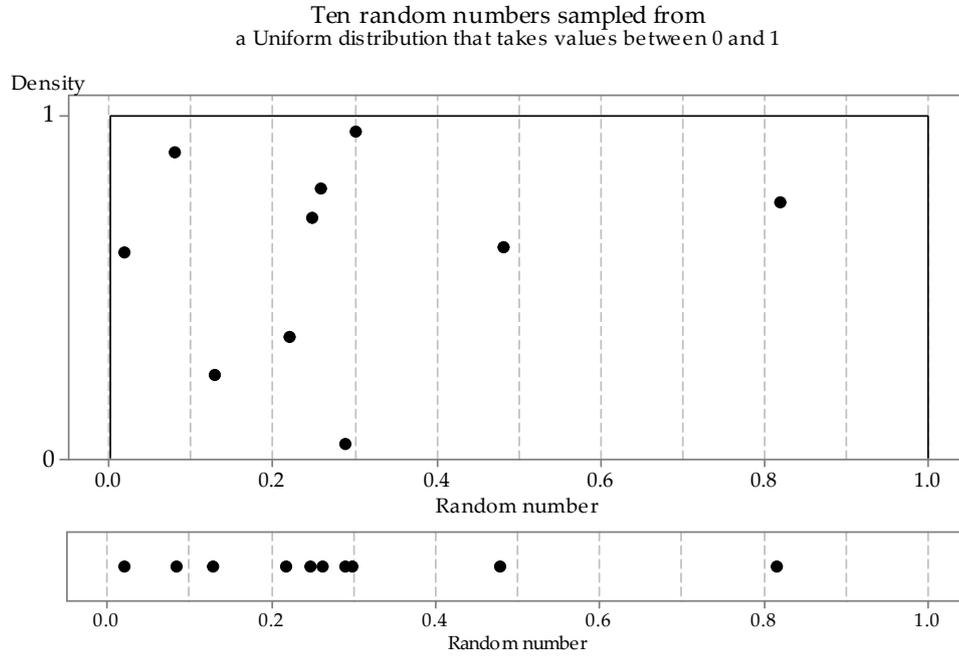


Figure 21: First random sample of size $n = 10$ from the $U(0, 1)$ distribution.

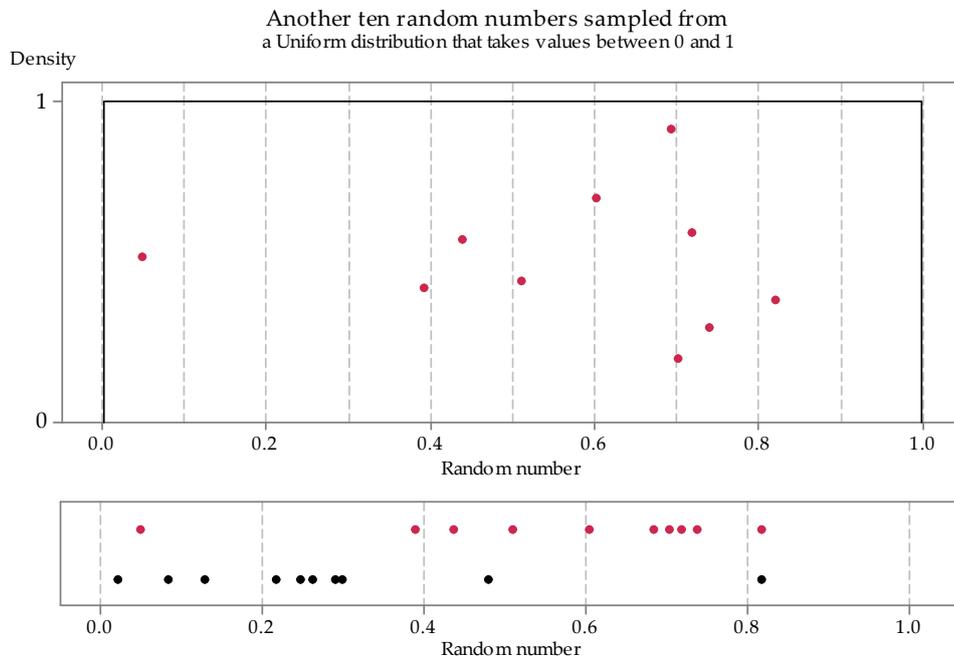


Figure 22: Second random sample of size $n = 10$ from the $U(0, 1)$ distribution.

Yet another ten random numbers sampled from a Uniform distribution that takes values between 0 and 1

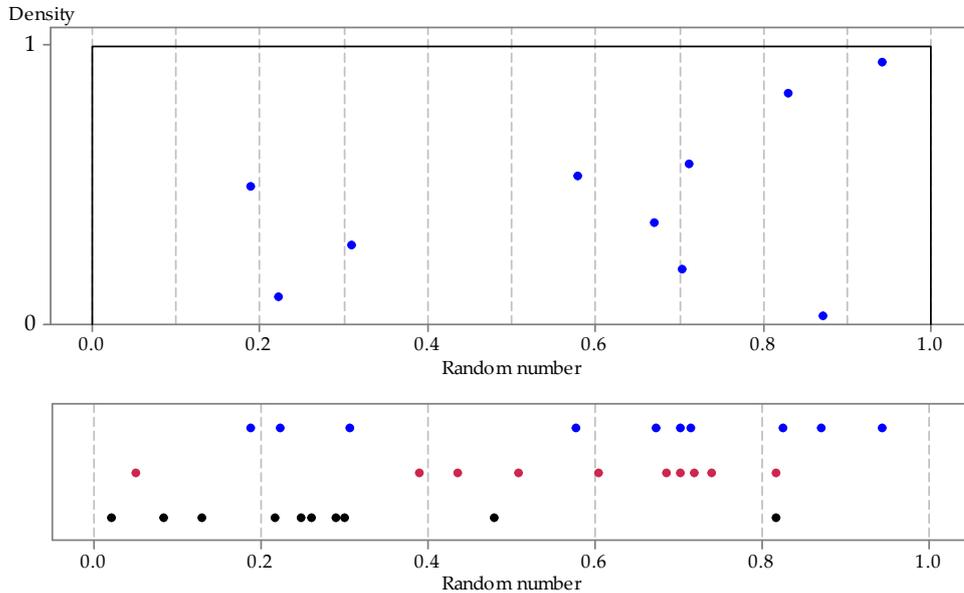


Figure 23: Third random sample of size $n = 10$ from the $U(0,1)$ distribution.

Figure 24 shows the distribution of many independent random samples of size $n = 10$ from the $U(0,1)$ distribution. Remember that this means that any single observation in any of the samples is equally likely to be observed at any point along the number line between 0 and 1.

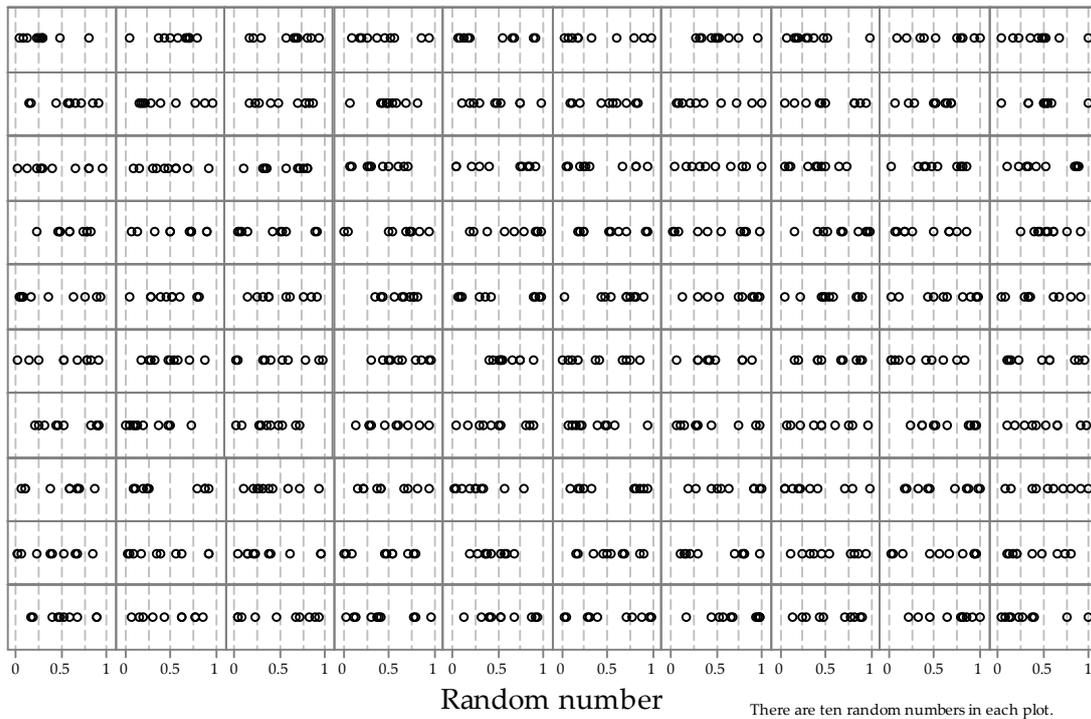


Figure 24: Dotplots of 100 random samples of size $n = 10$ from the $U(0,1)$ distribution.

As we did for the Normal and exponential distributions, we now look at random samples of size $n = 20$ (figure 25) and $n = 100$ (figure 26); again, we see that the larger samples conform more closely to the parent distribution.

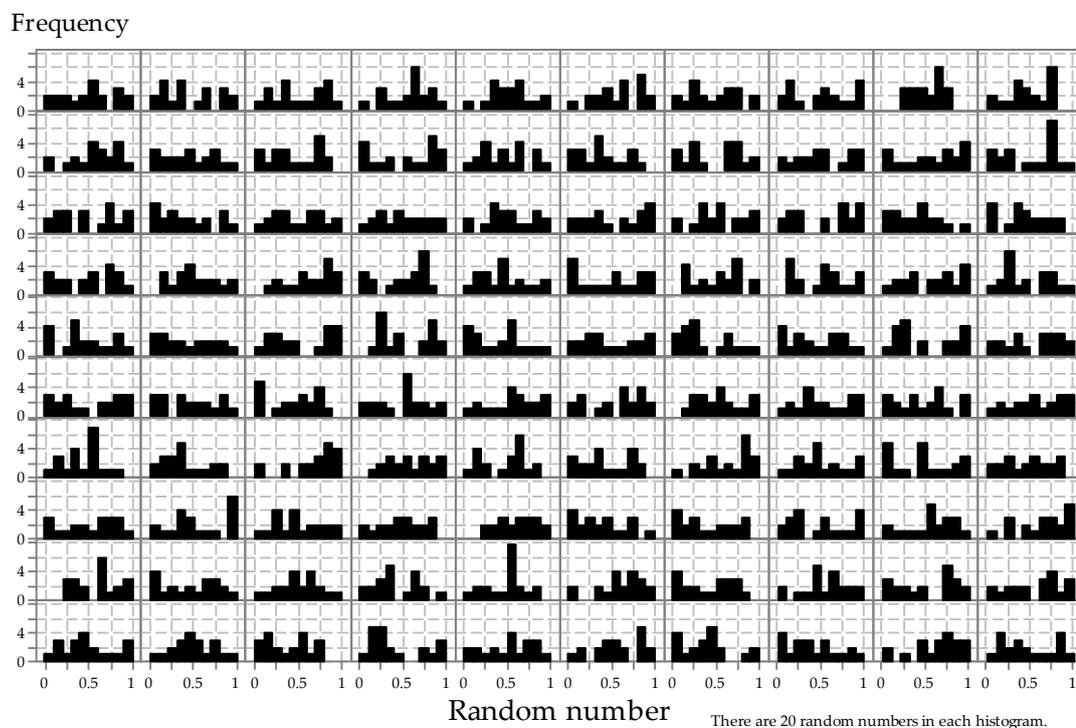


Figure 25: Histograms of 100 random samples of size $n = 20$ from the $U(0,1)$ distribution.

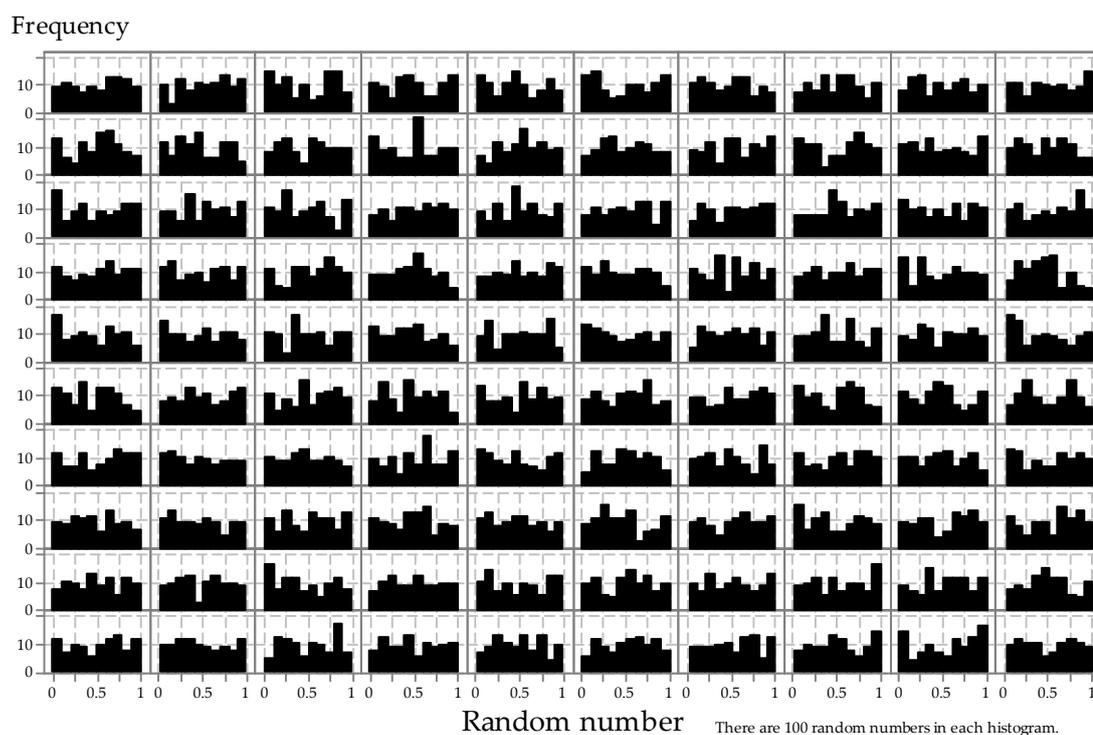


Figure 26: Histograms of 100 random samples of size $n = 100$ from the $U(0,1)$ distribution.

Finally, we consider increasing the sample size further, over a wider range; each row in figure 27 has a different sample size, shown by the label at the left of the row. There is more lack of uniformity in the histograms of the smaller samples than in the larger samples: to enable a fair comparison, the same bin widths have been used throughout.

Histograms show:

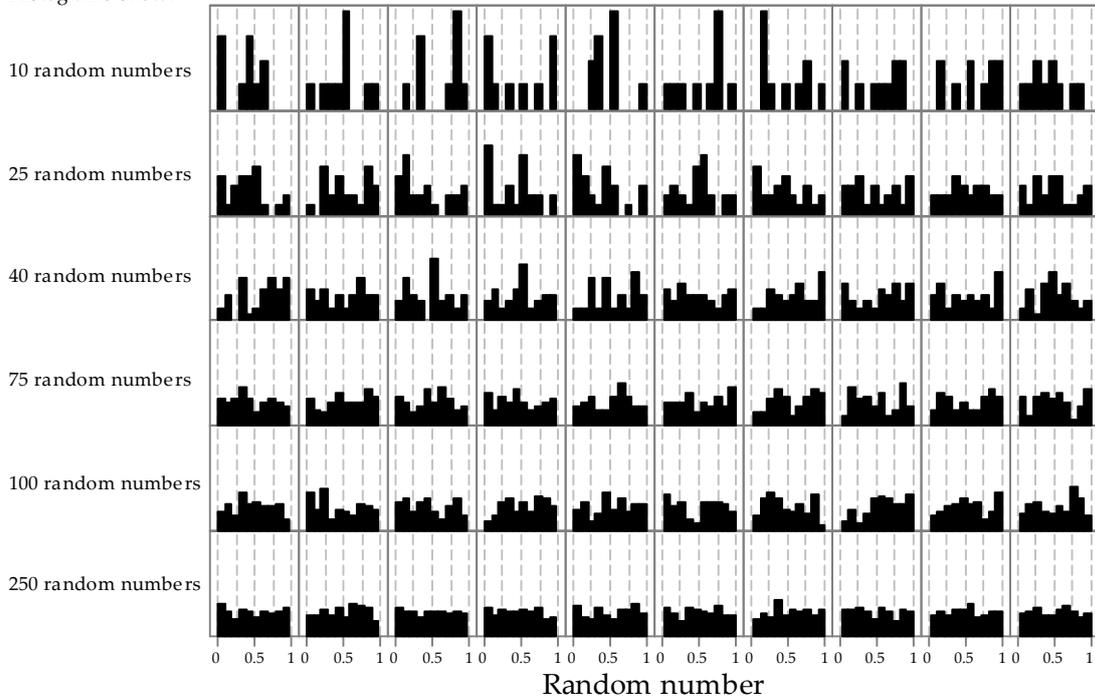


Figure 27: Histograms of random samples of varying size from the $U(0, 1)$ distribution. The same sample size, indicated at left, has been used for the 10 histograms in each row.

Sampling from the binomial distribution

In the module *Binomial distribution*, we saw that from a random sample of n observations on a Bernoulli random variable, the sum of the observations X has a binomial distribution. We revisit this theory briefly.

A Bernoulli random variable is a simple discrete random variable that takes the value 1 with probability p and the value 0 with probability $1 - p$. This is a suitable distribution to model sampling from an essentially infinite population of units in which a proportion p of the units have a particular characteristic. If we choose a unit at random from this population, the probability that it has the characteristic is equal to p , and the probability that it does not have the characteristic is equal to $1 - p$.

We may label the presence of the characteristic as '1' and its absence as '0'. So we consider $Y \stackrel{d}{=} \text{Bernoulli}(p)$, which satisfies $\Pr(Y = 1) = p$ and $\Pr(Y = 0) = 1 - p$.

We take a random sample Y_1, Y_2, \dots, Y_n from this distribution. This means that these random variables are independent and identically distributed, with $Y_i \stackrel{d}{=} \text{Bernoulli}(p)$, for $i = 1, 2, \dots, n$.

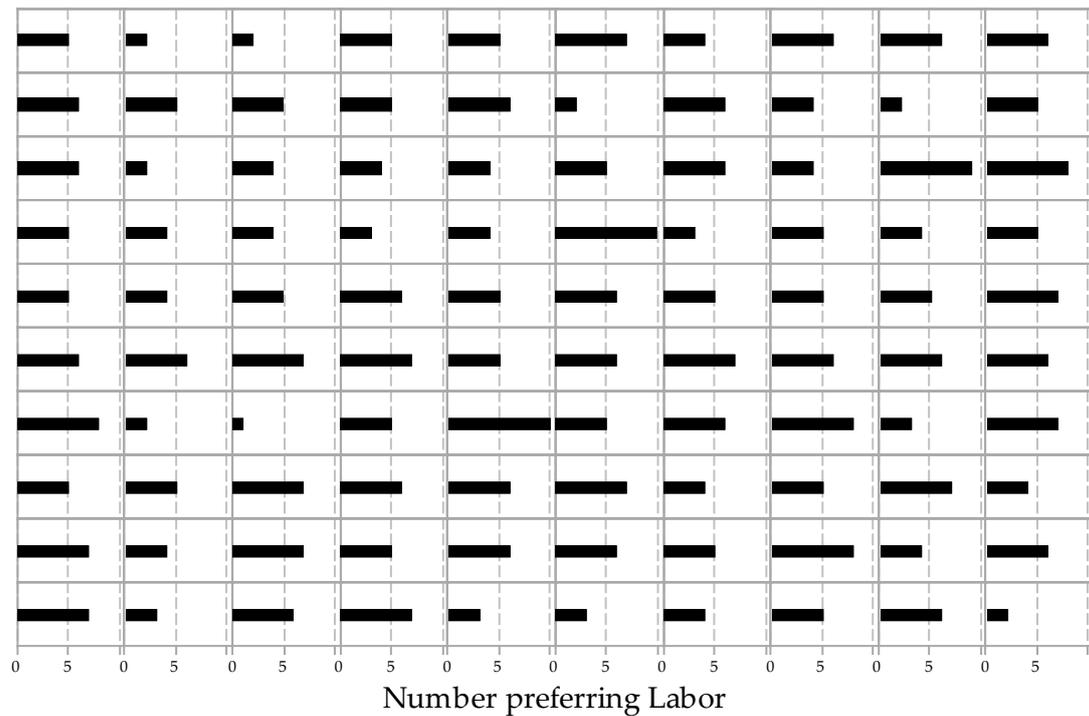
Define $X = \sum_{i=1}^n Y_i$; then $X \stackrel{d}{=} \text{Bi}(n, p)$. The sum of the Y_i 's is equal to the number of units in the sample with the characteristic of interest.

Suppose we obtain a random sample from a population of voters in Australia in which the proportion that support the Australian Labor Party is $p = 0.5$. We use this context to explore the binomial distribution. (In fact, in federal elections in Australia, the two-party-preferred vote is often quite close to 50%, corresponding to $p = 0.5$. For example, in the 2010 federal election, the two-party-preferred vote for Labor was 50.1%.) The population of voters in Australia is large enough to be regarded as effectively infinite for the purposes of this discussion.

Unlike the cases considered so far, note that each of the random samples gives a single observation from the $\text{Bi}(n, p)$ distribution (rather than many). In each case it is based on a random sample of size n from the Bernoulli distribution.

Consider figure 28. It represents 100 distinct observations on the $\text{Bi}(10, 0.5)$ distribution, coming from 100 random samples each of size $n = 10$ from the $\text{Bernoulli}(0.5)$ distribution. The details of the 10 observations in each random sample are shown by the individual zeroes and ones listed. For example, in the top left-hand sample, there are 5 ones, giving an observation $x = 5$ from the $\text{Bi}(10, 0.5)$ distribution. This observation is represented by the bar of length 5 in the top left-hand panel of the lower part of the figure.

0110101010	0100000010	0001000001	1100110010	0001101101	0001111111	0101011000	1101010110	0010111011	0101001111
0111000111	1000011110	1010001110	0011011001	1100110110	0101000000	0001111110	1000100011	0110000000	0010100111
1100110110	0100010000	0010101010	1001100100	0001011100	1010101100	1111000110	1000011010	1011111111	0110111111
1010101001	1010010010	1010100010	0110001000	1101100000	1111111111	0100110000	1000110101	1010001010	1011001010
0010100111	0001011100	1010001101	1101101010	1100010110	1001100111	0011110100	0110110100	1110100100	1011111100
0011101110	0110001111	1101110011	1110011110	1010010011	1011101001	0110111011	0010101111	1001011110	1111010001
1111001111	0000001001	0000001000	1100110010	1111111111	1011001100	0011111100	0111111011	1010000001	1010111110
1100101010	0101100101	1101101011	1110010110	1100110011	1011111010	0010100101	0011001110	0101111110	0101100100
1111111000	0010100110	1111110010	1011010001	0111010011	1101100011	1000111001	1111111001	1110001000	0001111110
1110010111	0011000010	1010101110	1101011101	0100000011	0110010000	0010011010	1101100100	0111011100	0101000000



Each panel is based on ten voters.

Figure 28: 100 random samples of size $n = 10$ from the Bernoulli(0.5) distribution (upper part) giving 100 corresponding observations from the Bi(10,0.5) distribution (lower part).

Figure 29 lists 100 random samples of size $n = 20$ from the Bernoulli(0.5) distribution. Of course, as the sample size increases, it starts to get tedious to list all those zeroes and ones! And we don't need to: given the assumptions, the information of interest is really just the total number of ones in the sample, which is exactly what is measured by the binomial random variable.

11111	01000	01001	01010	01011	11011	10111	00111	01011	00110
10101	00110	11100	00101	01011	00111	01010	10101	00001	00101
10111	10011	11101	00100	10101	00111	10101	01000	10100	11011
10111	01001	10011	00101	00010	00011	01111	10100	01011	11111
10000	10001	11001	10111	01110	10011	00000	10001	01100	10100
01000	00010	00010	01011	10111	11111	11111	10010	10110	10110
10111	10111	11111	00100	11011	00101	01100	00110	01110	00100
10101	01011	10001	00111	11000	11101	01111	11111	00110	10110
11000	10000	01010	01100	01010	10100	01110	00001	00011	00100
00111	01110	11111	01111	01101	10101	10001	01101	00111	01111
00001	10101	00011	01000	01111	10000	10100	10100	00010	10000
10011	11110	10010	01010	11110	01001	11011	01101	10100	01111
11011	11001	01110	11101	01100	01000	00000	11000	11111	00101
11111	10101	11101	10100	00111	11111	10011	11011	11010	00100
01000	00111	01101	00101	10101	11000	11100	11110	11111	10110
00110	01100	11101	01111	01001	10010	10111	11010	00010	01000
11100	00101	10110	00011	01010	10000	11011	11110	01011	00110
01100	11000	10001	10001	00010	11001	00111	10111	11000	10010
11110	01001	00101	11110	00101	00100	11010	01110	10100	00001
00101	10111	11000	10010	11101	01010	01111	00100	01001	11100
11110	00001	00100	10010	01100	11000	11001	11001	01011	00110
10010	11101	10000	01000	11100	10010	10100	10011	11100	01010
01011	11010	11111	01100	01110	10110	11100	10010	10110	11010
00111	10000	11000	11100	01001	11010	01110	11111	00011	00000
10100	11010	11100	01110	00000	01010	11100	00001	01101	11101
10101	00110	10111	11011	11010	11001	00010	11100	00011	01010
01101	01111	10100	00010	00001	11010	01001	00000	01110	10101
10001	11011	01101	10011	00111	10101	00101	10000	11110	10110
11110	10111	01010	00010	00100	10111	10110	10001	10010	10111
11111	01101	00011	01001	01001	01000	11010	11100	11000	11110
10001	01110	11100	11110	10010	11111	10001	11011	11111	11111
00101	10000	00000	10011	10111	00011	10010	01000	11001	01001
11010	10110	10011	01110	00110	11110	11000	01100	01010	10001
01000	11010	11100	00110	10101	01000	10000	01110	11100	11010
01110	10010	00111	10111	00011	10100	10111	10000	01010	00000
10111	00110	10110	01111	10001	10101	01101	00110	00101	10000
10110	00011	00011	11011	00110	10101	10001	11010	01000	10011
11100	00001	10110	01101	00001	11101	01011	10100	11000	01111
01010	10001	10001	01111	10100	10011	10010	01100	10101	11111
10011	01100	11110	00001	00100	10101	00100	10110	01001	11101

Figure 29: 100 random samples of size $n = 20$ from the Bernoulli(0.5) distribution.

The binomial observations from the Bernoulli random samples of size $n = 20$ listed in figure 29 are shown in figure 30.

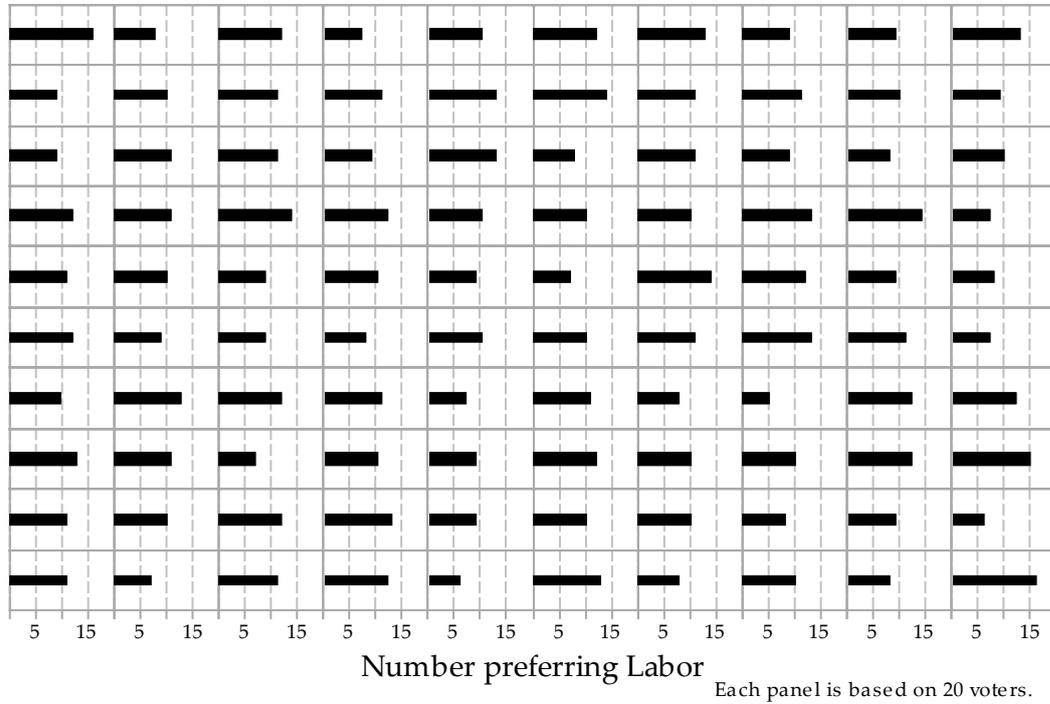
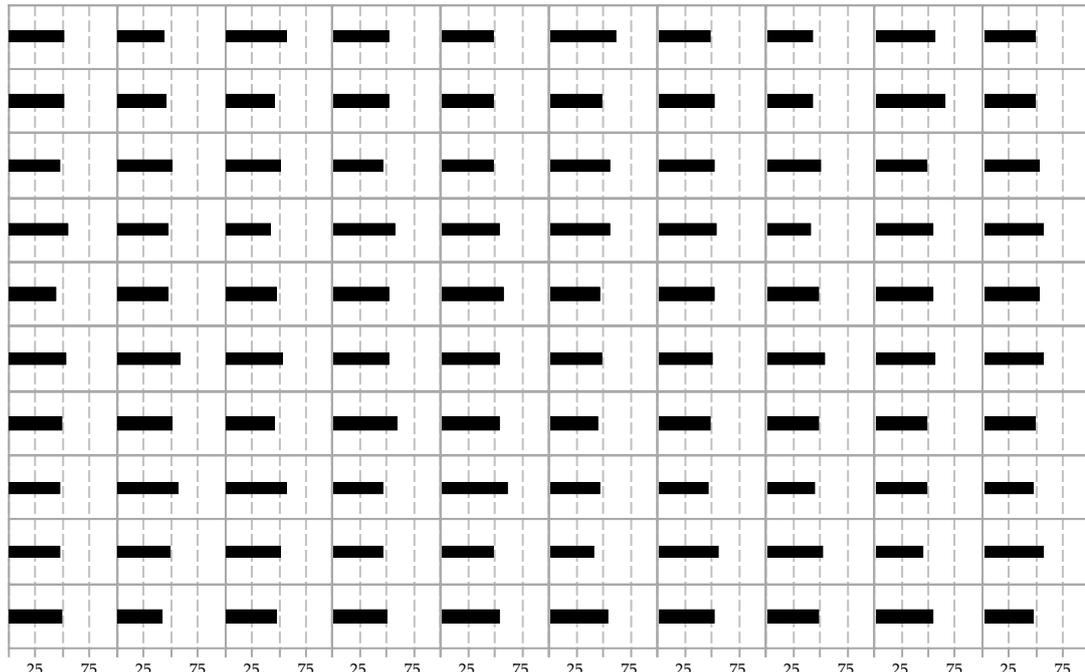


Figure 30: 100 observations from the $Bi(20,0.5)$ distribution arising from the Bernoulli random samples in figure 29.

Now we represent samples from the binomial distribution with $n = 100$ and $p = 0.5$.

0101100001011010100	01010001000001110101	00111001110110100111	0101010101010101010	0000001111000100001	1101001111101101010	1000111001000111000	111010000000101011	0101100111010100011	100010100101010110
---------------------	----------------------	----------------------	---------------------	---------------------	---------------------	---------------------	--------------------	---------------------	--------------------



Number preferring Labor
Each panel is based on 100 voters.

Figure 31: 100 random samples of size $n = 100$ from the Bernoulli(0.5) distribution (upper part) giving 100 corresponding observations from the Bi(100,0.5) distribution (lower part).

Finally, we look at what happens when we vary the value of p . Figure 32 shows independent observations from the $\text{Bi}(40, p)$ distribution for $p = 0.1, 0.2, \dots, 0.9$. Ten observations are shown for each value of p . Note how p influences the values obtained. For $p = 0.1$ (the lowest value shown), the number observed in all 10 cases was less than 10. For $p = 0.9$, all 10 observations were over 30. For $p = 0.5$, the observations were near 20.

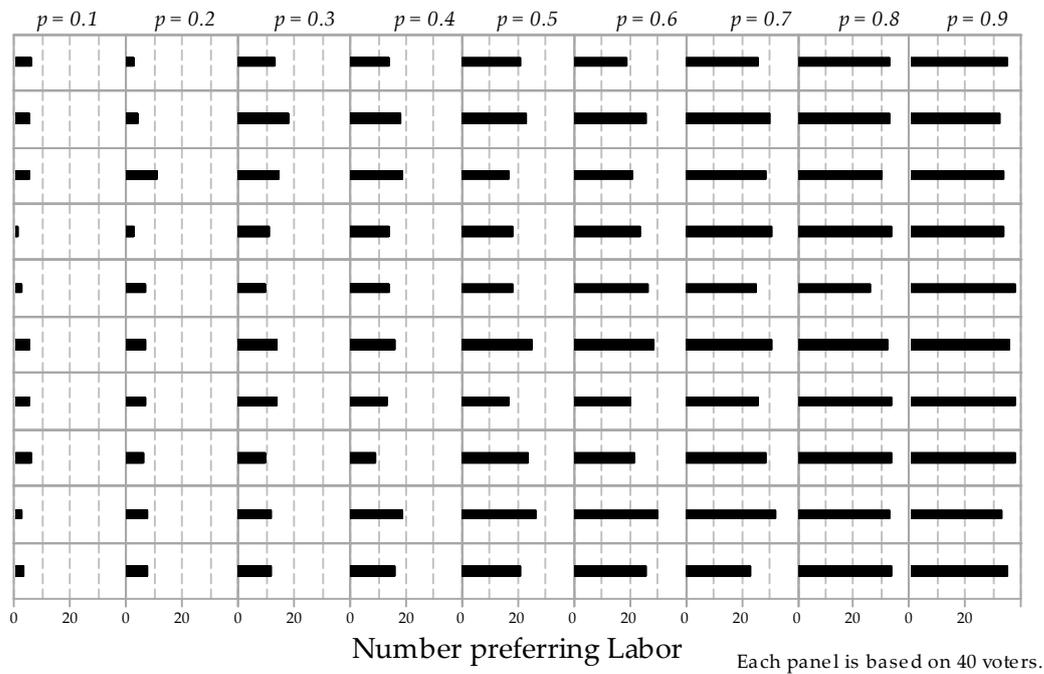


Figure 32: Observations from the $\text{Bi}(40, p)$ distribution; the value of p is constant for each column.

Answers to exercises

Exercise 1

The first stage of mixing ensures that the first ball is sampled at random. We can think of it as a random sample of size $n = 1$. In fact, any subsequent re-mixing is redundant. As the initial mixing ensures the first ball is randomly sampled, any subsequent balls sampled *without* additional mixing will also be randomly sampled. If this were not the case, the choice of the first ball could not be considered random. So the subsequent mixing is presumably done to reassure psychologically those playing the game, rather than for reasons of randomness.

Exercise 2

- a There are only two possible samples that can be chosen. Note that if a simple random sample of $n = 150$ is chosen from a population of size $N = 300$, the number of possible samples is huge: approximately 10^{89} .
- b Yes, each student has probability 0.5 of being selected.
- c This is not a simple random sample, because the process does not give every possible sample of size $n = 150$ the same chance of being selected.

Exercise 3

- a The electoral roll would be imperfect as a sample frame. For example, young people who were eligible voters but had not yet registered to vote would not be included. The extent of this potential bias might, for example, depend on the recency of an election.
- b If 60 000 people were summoned from 3.5 million voters in Victoria, the probability of selection in one year is $\frac{60\,000}{3\,500\,000} \approx 0.017$, or 1.7%.
- c There are a number of things we need to know to determine if an individual could be called up for jury duty twice in one year. If the random sample of potential jurors for a given year is taken at a single point in time, using a method like that described using Excel, then each individual will have only one chance of selection. If the sample is taken across the year, the potential to be sampled twice will depend on whether or not an individual is exempt once they have been sampled. In fact, individuals in metropolitan Melbourne who attend for jury service but do not serve as a juror on a trial are excluded from the sample frame for two years.
- d If you only had access to a hard copy of the electoral roll, organised by electorate, the sampling would need to be done in stages. You might, for example, first select a subset of electorates; you could take a simple random sample of electorates. Next

you would need to consider practical ways to sample from the book listing names in each electorate. You could, for example, find the number of pages in a selected book and take a random sample of pages. You might consider summoning all the people listed on a sampled page; this is a kind of cluster sampling. However, this is unlikely to be ideal: electoral rolls are organised alphabetically and so people on the same page are likely to be related. It would be better to randomly sample some individuals from a selected page. If random sampling is used at every stage, the resulting sample is a random sample. It would not be a simple random sample if, for example, the same number of people were sampled from each selected electorate; people in smaller electorates would have a greater chance of selection than people in larger electorates.

Exercise 4

- a The question the survey asks is quite provocative, and people might respond for a range of reasons. All will be self-selected volunteers. Some people might simply be motivated to engage with the survey because of the provoking title. Others may have strong views about prescription-drug use and want to make sure they are counted. The survey states that it is 'girls only', but there are no restrictions on who can enter the responses online. Sometimes prizes or monetary rewards are offered for participating in surveys; this can motivate participants, but was not the case for the *Cosmopolitan* survey.
- b We have no idea who has provided the answers to the questions asked in this online survey. It is open to the usual biases of self-selected samples, but also to the whims of mischievous internet users. There is no quality control, leaving great potential for bias. The actual bias is unknown, so we learn nothing, or very little, about the population of interest.

Exercise 5

- a One obvious strategy would be to access boys aged 13 to 16 through schools. The minimum school leaving age is 17, but there are options to participate in approved school-equivalent programs for 15- and 16-year-olds. Participants in these alternative programs would not be accessed through schools, and so ways of identifying and sampling boys in this group would need to be developed. This might be able to be done via registrations of participants in the approved school-equivalent programs. Sampling in stages would be needed; for example, schools and then classes might be sampled. It would be efficient to sample all boys within a given class.
- b Following these adolescents over time may become difficult as they leave school, leave home and transition into adult life.

- c It may be that the boys/men who are most difficult to track over time are those with problematic drug and alcohol use.
- d The purpose of the study is look at the effects of drug and alcohol use; if the heavier users are more difficult to track and less likely to be able to be followed up, the effects may be underestimated.

Exercise 6

- a Over 27 000 respondents sounds impressive, but if the sample is badly biased, a large sample size does not guarantee validity: think of the *Literary Digest* poll.
- b *The Age* poll was an online poll accessible by computer users. Some sectors of the community would not be able to access the poll.
- c People with an interest in the election outcome are likely to respond to this kind of poll. It is possible to vote more than once in such polls, and this may have happened.
- d The disclaimer indicated that the poll was not scientific; *The Age* polls often have a disclaimer indicating that the poll is based on views of volunteers. However, despite the caveat, *The Age* suggested that the results *will* concern Labor supporters.

Exercise 7

- a If the poll was unbiased, the percentage vote predicted for Roosevelt is the observed proportion in the poll: 43%.
- b Among the 2 400 000 respondents, there are $2\,400\,000 \times 0.43 = 1\,032\,000$ votes for Roosevelt. If all 7 600 000 non-respondents vote for Roosevelt, then the total number of votes for him is $1\,032\,000 + 7\,600\,000 = 8\,632\,000$ out of 10 000 000. Hence the predicted percentage vote for Roosevelt is 86%.
- c From the 2 400 000 respondents, there are 1 032 000 votes for Roosevelt. If all of the 7 600 000 non-respondents vote for Landon, the total number of votes for Roosevelt is $1\,032\,000 + 0 = 1\,032\,000$ out of 10 000 000. Hence the predicted percentage vote for Roosevelt is 10%.
- d From the 2 400 000 respondents, there are 1 032 000 votes for Roosevelt. If half of the 7 600 000 non-respondents vote for Roosevelt, then the total number of votes for him is $1\,032\,000 + 3\,800\,000 = 4\,832\,000$ out of 10 000 000. Hence the predicted percentage vote for Roosevelt is 48%.
- e To obtain 62% of the vote from the 10 000 000 people sampled, Roosevelt would need 6 200 000 votes. From the 2 400 000 respondents, he received 1 032 000 votes; hence, he would need to receive $6\,200\,000 - 1\,032\,000 = 5\,168\,000$ votes from the 7 600 000 non-respondents. This is $\frac{5\,168\,000}{7\,600\,000} = 0.68$, or 68%, of the non-respondents' vote. Contrast this with the 43% he received from the respondents.

Exercise 8

- a Let $X \stackrel{d}{=} N(30, 7^2)$. Then $\Pr(X < 30) = \frac{1}{2}$, because the Normal distribution is symmetric and its mean and median are equal. In a random sample, all observations are independent, so the probability that all 10 observations are below 30 is equal to the product of the individual probabilities, and is therefore equal to $(\frac{1}{2})^{10} \approx 0.001$.
- b 0.001, by symmetry.
- c 0.002; the two events are mutually exclusive, so we can add the two probabilities.
- d No. About one in every 500 samples (actually, one in every 512) will have this feature, so it is not surprising that we have observed 100 samples and none of them has the feature.
- e It is not that easy to tell this, just from inspection of the dotplots. The largest mean is 34.7, for the sample in row 2, column 6. The smallest mean is 24.1, for the sample in row 4, column 9.
- f
- i $\Pr(X \leq 16) = \Phi(-2) \approx 0.0228$.
 - ii $Y \stackrel{d}{=} \text{Bi}(10, 0.0228)$.
 - iii $\Pr(Y \geq 3) \approx 0.0013$. This is indeed quite an unusual sample.
 - iv If we regard an observation of $y = 3$ as unusual, we must regard $y \geq 4$ as even more unusual, so we count more extreme values in the calculation. Think of a very hot day in summer; suppose the maximum reached is 44.3 °C. If we wonder how extreme this is, we do not ask: 'How common are days with a maximum of 44.3 °C?' Rather, we ask: 'How common are days with a maximum of at least 44.3 °C?'

Exercise 9

Figure 33 provides the missing pdfs. You may not get the answers exactly right, based only on a visual impression. But you should be able to get within about 1 of the correct value of each mean and standard deviation, just by eye. This is a preliminary to the more formal approaches of inference, which is dealt with in the two modules *Inference for proportions* and *Inference for means*.

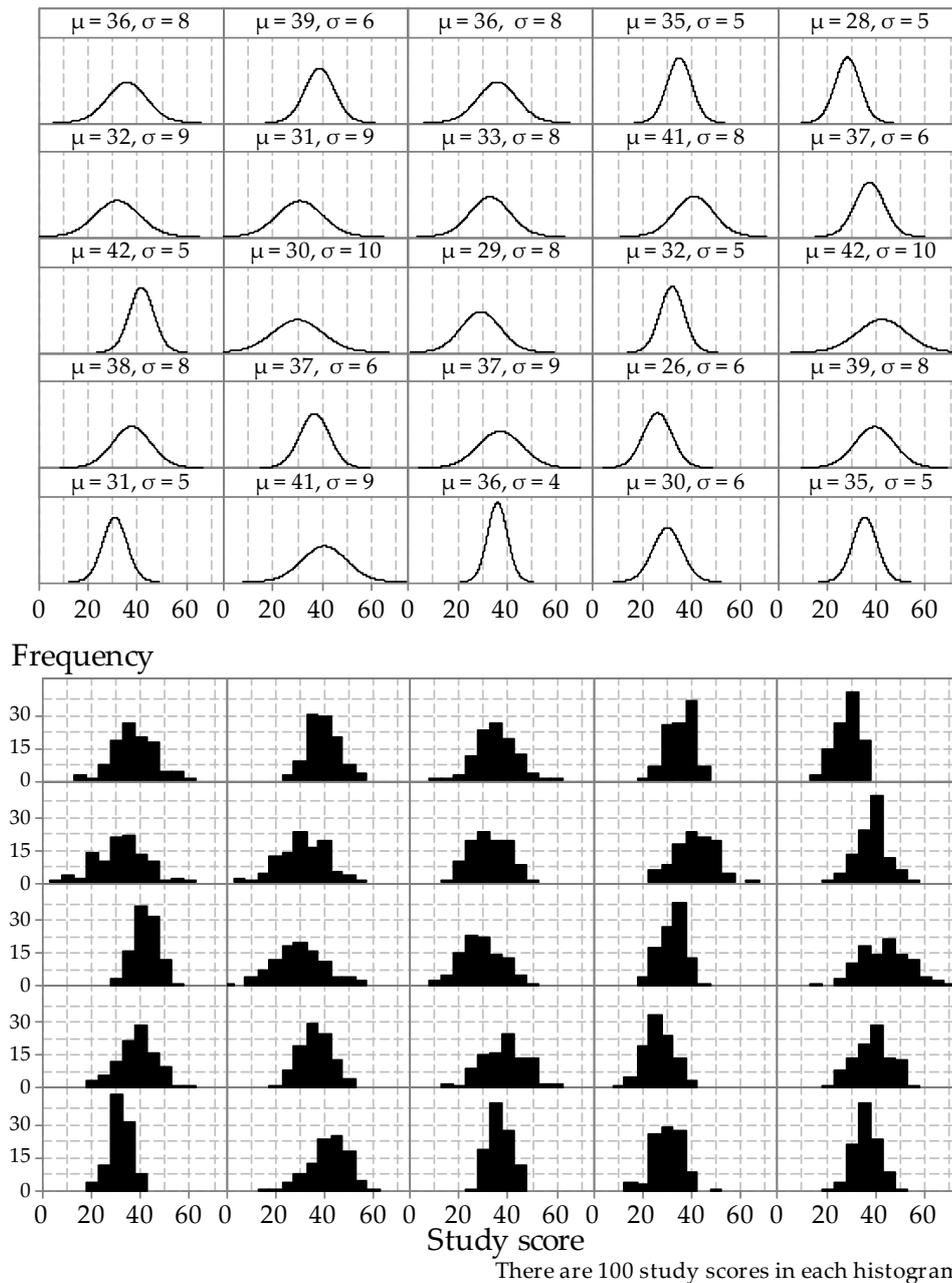


Figure 33: 25 different Normal distributions (upper part) and histograms of random samples from each of them (lower part); the same sample size $n = 100$ has been used throughout.

References

- D. A. Freedman, Robert Pisani and Roger Purves, *Statistics*, 4th edition, W. W. Norton, 2007.

0

1

2

3

4

5

6

7

8

9

10

11

12