05
Unit

Sampling and the Central Limit Theorem

Random samples, sampling distributions, the standard error, and why means pile up in a bell curve

By the end of this unit, I should be able to answer these questions:

- What is the difference between a population, a sample, and a sampling distribution?
- Why does a random sample of 1,000 people tell us more about a country than a self-selected sample of 2 million?
- How does the standard error of a sample mean depend on the sample size, and why does quadrupling the sample only halve it?
- What does the central limit theorem state, and for which kinds of data does it need large samples?
- Why do the balls in a Galton board settle into a bell shape?
- What does the law of large numbers promise a gambler, and what does it not?

An opinion poll of 1,000 people is reported with a margin of error of about 3 percentage points, whether the country has five million voters or two hundred million. A national health survey measures about 5,000 people a year and speaks for a whole country. In 1936 a magazine collected 2.4 million ballots and named the wrong winner. This unit explains how a sample is drawn so that it can stand for a population, what goes wrong when it is not, and the two theorems that say how far the mean of a sample can stray from the mean of the population it came from. It uses the mean and the standard deviation of a set of numbers and the idea of a probability distribution, each recalled in one sentence where it is needed, so that it can be read on its own.

1 Populations and samples

The **population** is the entire set of cases about which a statement is to be made, and the **sample** is the part of it that is actually observed. A number that describes the population, such as the mean height of all adult men in a country, is a **parameter**; the corresponding number computed from the sample, such as the mean height of the 2,500 men whom a health survey actually measured, is a **statistic**. Parameters are fixed and usually unknown, statistics are known and vary from sample to sample, and this unit is about the relation between the two.

Between the population and the sample stands a list, the **sampling frame**, which is the list of cases from which the sample is actually drawn, such as an electoral register, a file of postal addresses, a

directory of telephone numbers, or the enrolment list of a school. In the best case the frame and the population are the same. Often they are not, and then the sample can only represent the frame. A survey drawn from landline telephone numbers says nothing about households without a landline, however carefully the numbers are drawn, and this gap between the frame and the population is the first place where a survey can go wrong.

1.1 The simple random sample

A **simple random sample** of size n is drawn in such a way that every possible set of n cases in the frame has the same chance of being the set chosen. Two consequences follow at once: each case has the same chance n/N of being selected, where N is the number of cases in the frame, and no case is more likely than another to appear together with a given one, which means that whether one case is chosen has no effect on which other cases are chosen with it. The second property is the one that matters, because it is what rules out any connection between who is chosen and what is measured, and the formulas of this unit assume it.

In practice the cases in the frame are numbered from 1 to N, a sequence of random numbers between 1 and N is produced, and the cases whose numbers come up are the sample, with repeats skipped. The random numbers are the difficult part, because people are bad at inventing them: asked to write down random digits, they avoid repeats and runs of the same digit, which true random digits contain. In 1927 L. H. C. Tippett (1902-1985) published the first table of random digits, 41,600 of them, taken from the figures of a census report, and for decades statisticians drew their samples by reading such tables. Today a computer produces the numbers with a **pseudo-random generator**, a formula that turns a starting number, the **seed**, into a long sequence that passes the standard statistical tests of randomness. The same seed gives the same sequence, which is a virtue rather than a flaw: anyone who knows the seed can repeat the draw exactly, and every simulation in this unit states its seed for that reason.

To illustrate this, the health survey behind many of the examples below measured 2,500 men aged 20 and over in one two-year cycle. To draw a simple random sample of 25 of them, number the men from 1 to 2,500, read four-digit groups from a table or a generator, keep each group between 0001 and 2500 that has not come up before, and stop at the twenty-fifth. Nothing about the men so chosen, their height, their age, or their address, played any part in their selection.

1.2 Other ways of drawing a sample

A simple random sample is the design against which all others are measured, and three others are used because they are cheaper, more precise, or more practical.

- A **stratified sample** divides the frame into groups called strata, such as regions or age bands, and draws a simple random sample within each. Every stratum is then represented in known proportion, and a small group of special interest can be sampled more heavily than its share of the population, which is called oversampling; afterwards each stratum's result is given the weight of the stratum's true share of the population, so that the oversampled group does not distort the total.
- A **cluster sample** draws groups rather than individuals, for example schools rather than pupils or city blocks rather than households, and then observes every case in the chosen groups or a random sample of them. It saves travel and time, and it costs precision, because people in one cluster tend to resemble each other, so that a cluster of 30 neighbours tells us less than 30 people scattered across the country.

- A **systematic sample** takes every k-th case from the list after a random start. It is easy to carry out on a production line or a filing cabinet, and it is as good as a simple random sample unless the list has a cycle of length k. For example, if the list is a list of days and $k = 7$, every chosen day is the same day of the week.

Real surveys combine these. The National Health and Nutrition Examination Survey (NHANES) of the United States, whose height data appear throughout this unit, examines about 5,000 people a year in mobile examination centres. It first draws counties, then neighbourhoods within the counties, then households, then persons, and it deliberately oversamples some groups, among them Hispanic, non-Hispanic Black, and non-Hispanic Asian persons, people on low incomes, and people aged 80 and over, so that it can report on each of them with enough precision. The formulas of this unit are those of a simple random sample. The published results of such a survey are weighted for its design, which goes beyond the unit, so the heights below are used unweighted, as a population to draw from.

2 Bias and variability

A statistic can miss its parameter for two different reasons, and the whole art of sampling lies in keeping them apart. **Bias** is a systematic error of the method: the statistic is on average too high or too low, in the same direction sample after sample, because of the way the sample was drawn or the measurements were taken. **Variability** is the scatter of the statistic from one sample to the next around its own average, which exists because different samples contain different cases. Think of the two as shots at a target whose centre is the parameter: variability is the spread of the shots, and bias is the distance of their centre from the bull's eye. A large sample makes the shots cluster more tightly, but it does not move the cluster. Variability shrinks as the sample grows; bias does not shrink at all.

Four kinds of bias account for most of the surveys that have gone wrong.

1. **Selection bias** arises when the frame or the procedure favours some cases over others: telephone subscribers in the poll of 1936 described below, people who happen to walk past an interviewer outside a gym, patients of one hospital standing in for all patients.
2. **Voluntary response bias** arises when people choose themselves into the sample, as in a poll on a website, a questionnaire left on a counter, or a radio show that invites listeners to call in. Those who feel strongly, and those who are angry, answer far more often than the rest.
3. **Non-response bias** arises when cases that were properly selected cannot be reached or decline to take part, and those who do answer differ from those who do not. Response rates below 10 per cent are now usual for telephone surveys, which leaves plenty of room for this bias.
4. **Measurement bias** arises from the instrument or the question: a scale that reads one kilogram too high, a question that suggests its own answer, a request to state one's own weight, which most people understate.

The first three concern who ends up in the sample, and the fourth concerns what is recorded about them. None of the four is cured by size, and the most famous failure in the history of polling is the standard illustration of the first three.

Check Your Understanding **Question 1**

A city council puts a questionnaire about a new tram line on its website, and 12,000 residents answer; 70 per cent are in favour. A polling firm telephones a simple random sample of 600 residents drawn from the register of inhabitants; 48 per cent are in favour. Which figure deserves more trust?

(A) The 70 per cent, because a sample twenty times larger has a much smaller standard error

(B) The 48 per cent, because the residents who chose to answer online are not a random sample and their number does not repair that

(C) Neither, because the two figures differ by more than 20 percentage points and must both be wrong

(D) The 70 per cent, because people who answer a questionnaire have thought about the question more carefully

The solution is at the end of the unit.

Excursion *The Literary Digest, 1936*

The *Literary Digest* was a weekly magazine with a circulation of over a million, and from 1916 it had run a straw poll, an unofficial vote among its readers, before every presidential election, mailing ballots by the million and counting the returns. It had named the winner correctly five times in a row. In 1936 it mailed about 10 million ballots, to addresses taken from telephone directories, from lists of registered car owners, and from its own subscription list, and about 2.4 million came back. The final count, printed in the issue of 31 October 1936, gave the Republican challenger Alf Landon 1,293,669 ballots against 972,897 for the sitting president Franklin Roosevelt, which is 57 per cent for Landon, and the magazine forecast 370 electoral votes for him, which are the state-by-state votes that decide a presidential election in the United States.

On 3 November Roosevelt won 60.8 per cent of the popular vote and 46 of the 48 states. In the same season George Gallup (1901-1984), whose American Institute of Public Opinion was barely a year old, interviewed about 50,000 people chosen by quotas of age, sex, region, and income so that his sample resembled the electorate, and forecast a Roosevelt victory with 55.7 per cent. His forecast came true, and the Digest, its credibility gone, ceased publication in 1938.

Two things went wrong. In the depth of the Depression, people with a telephone and a car were richer than average and leaned towards Landon, so the frame was tilted from the start. The larger fault was non-response: only about a quarter of the ballots came back, and later analyses of a survey of 1937, which asked people whether they had received a Digest ballot and whether they had returned it, found that Roosevelt's supporters were much less likely to have sent theirs in. The people who did not answer decided the error.

Worked Example **Why 2.4 million ballots did not help**

Of the votes cast in the election for the two main candidates, Roosevelt received 62.5 per cent. The Digest's 2.4 million returns had given him 972,897/(972,897 + 1,293,669) = 42.9 per cent of the two-candidate count, which is 19.6 percentage points too low. Table 1 sets the figures side by side.

Suppose, for a moment, that the 2.4 million returns had been a simple random sample of the voters. Section 6 shows that the proportion in such a sample varies from one sample to the next by a standard deviation of about $\sqrt{0.25/n}$ when the true proportion is near one half; for now it is enough to take the formula as given. For the n = 2,266,566 returned ballots that named one of the two candidates is $\sqrt{0.25/2{,}266{,}566}$ = 0.00033, or 0.03 percentage points. Chance alone would have moved the Digest's figure by a few hundredths of a point. The actual miss of 19.6 points is about 600 times that, so it was not chance. It was bias, and bias is not reduced by collecting more ballots of the same kind: 2.4 million biased ballots missed by 19.6 percentage points, when a random sample of that size would have been within a few hundredths of a point. Gallup's 50,000 came within 6.8 points, which is also more than chance would allow for a random sample of that size, because a quota sample leaves the choice of persons to the interviewers. His method nevertheless had a much smaller bias than the Digest's, and that, not size, is what won.

Source	Number of People	Roosevelt (%)
Literary Digest returns	2,266,566	42.9
Gallup's quota sample	about 50,000	55.7
Election result	44.4 million votes	62.5

Table 1. The two polls of 1936 against the result, with Roosevelt's share of the votes for the two main candidates; the Digest's figure counts the returned ballots that named one of the two. The Digest was 19.6 percentage points off; Gallup's 50,000 were 6.8 points off. Data: Literary Digest, 31 October 1936; Gallup's final pre-election poll; official returns of the 1936 election.

The Literary Digest is not a story about the 1930s. Every online poll, every survey answered by whoever clicks, and every large administrative dataset that records some people and not others faces the same arithmetic, and the research spotlight shows what it looks like with the data of 2016.

Research Spotlight

Xiao-Li Meng Harvard University

Why 2.3 million self-selected opinions can be worth less than 400 random ones

Xiao-Li Meng, the Whipple V. N. Jones Professor of Statistics at Harvard University, asked how much a dataset that covers a large part of a population is worth when the people in it chose themselves. He proved that for any sample of n cases from a population of N, random or not, the difference between the sample mean and the population mean is the product of three factors:

1. the correlation between whether a case is recorded and the value it carries, which measures the quality of the data (a correlation of zero means that being recorded has nothing to do with the value);

2. the quantity $\sqrt{(N-n)/n}$, which depends only on the fraction of the population sampled;
3. the standard deviation of the variable.

In a simple random sample the correlation is on average about $1/\sqrt{N}$, and the product of the first two factors is then about $\sqrt{(N-n)/(nN)}$, which is close to $1/\sqrt{n}$ whenever the sample is small next to the population, so the familiar $\sigma/\sqrt{n}$ is left. When people select themselves, the correlation does not shrink with N, and for a given sample size the error grows with the square root of the population size. From the Cooperative Congressional Election Study of 2016, a survey of about 64,600 people, he estimated that correlation for the self-reported intention to vote for one candidate at about −0.005, and showed that a self-selected sample with that defect and 2.3 million responses, 1 per cent of all eligible voters, would estimate that candidate's share with the same mean squared error (the average of the squared difference between estimate and truth) as a simple random sample of about 400. The bias of Section 2 is this first factor, and his conclusion is the paradox of the title: the more the data, the surer we fool ourselves.

Read the paper: Meng (2018), Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election, The Annals of Applied Statistics 12, 685

3 Sampling distributions and the standard error

Suppose now that the sample is a proper simple random sample, so that bias is out of the way, and only variability remains. How large is it? The question can be answered exactly, and the answer is the most useful single formula of the Foundation units.

Before the sample is drawn, its mean is unknown; a different set of 25 men would have given a different mean. The sample mean is therefore a random variable, a quantity whose value is decided by chance, and like every random variable it has a **probability distribution**, the list of the values it can take together with the probabilities of taking them. The distribution of a statistic over all the samples of size n that could be drawn from the population is called its **sampling distribution**, and Figure 1 shows the three objects of this unit side by side: the population with its parameters, one sample with its statistics, and the sampling distribution that describes how the statistic varies from sample to sample.

Nobody can draw all possible samples, but a computer can draw a great many, and the worked example does exactly that. Two numbers from the data are needed for it. The **mean** of a set of values is their sum divided by their number, written μ for a population and $\bar{x}$ for a sample. The **standard deviation**, written σ for a population and s for a sample, measures how far the single values typically lie from their mean; it is the square root of the average squared distance from the mean, and for the heights below it is about 7.7 cm, so that a man 7.7 cm above or below the average is a typical rather than an unusual man.

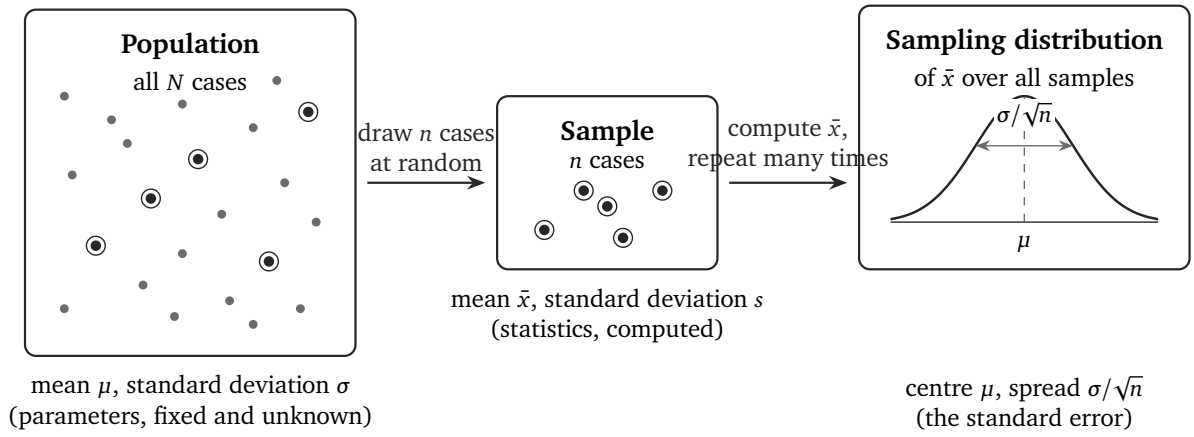


Figure 1. Population, sample, and sampling distribution. A simple random sample of n cases (ringed dots) is drawn from the population, its mean $\bar{x}$ is computed, and the process is imagined repeated over and over. The distribution of all those means is the sampling distribution of $\bar{x}$; it is centred on the population mean μ and has the standard deviation $\sigma/\sqrt{n}$, which is called the standard error.

Worked Example One thousand samples of 25 heights

The 2,500 men aged 20 and over who were measured in NHANES in 2017 and 2018 serve as a population. Their mean height is $\mu = 173.4$ cm and their standard deviation is $\sigma = 7.74$ cm. A seeded pseudo-random generator (Python's Mersenne Twister, seed 2051) drew 1,000 samples of $n = 25$ men from this population, each man drawn with replacement, which means that a man who has been drawn is put back and can be drawn again, and the mean height of each sample was recorded.

The 1,000 sample means ranged from 168.1 to 178.1 cm. Their mean was 173.3 cm, close to μ, and their standard deviation was 1.57 cm, a fifth of the standard deviation of the single heights. Of the 1,000 means, 674 lay within 1.55 cm of μ and 945 within 3.10 cm, close to the proportions, 68 and 95 per cent, that a normal distribution puts within one and two standard deviations of its centre. The same was done with samples of 4 and of 100 men, and Table 2 collects the results. In each case the standard deviation of the sample means is close to $\sigma/\sqrt{n}$, which is $7.74/\sqrt{25} = 1.548$ cm for $n = 25$. The simulation suggests, and the next box proves, that the mean of a sample of 25 heights typically misses the population mean by about 1.5 cm, a fifth of the spread of the single heights.

Sample Size n	Mean of the 1,000 Means (cm)	SD of the 1,000 Means (cm)	$\sigma/\sqrt{n}$ (cm)
4	173.54	3.98	3.87
25	173.31	1.57	1.55
100	173.44	0.79	0.77

Table 2. One thousand simulated samples of each size from the 2,500 NHANES men ($\mu = 173.42$ cm, $\sigma = 7.74$ cm), drawn with replacement with the seeds 2030, 2051, and 2126. The standard deviation of the sample means follows $\sigma/\sqrt{n}$ to within the chance variation of the simulation itself.

Key Formula **Mean and standard error of a sample mean**

$$\mathrm{E}[\bar{X}] = \mu, \qquad \mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \mathrm{SE}(\bar{x})$$

For a simple random sample of n independent draws from a population with mean μ and standard deviation σ, the sample mean $\bar{X}$ has expected value μ and standard deviation $\sigma/\sqrt{n}$. This standard deviation of the sampling distribution is called the **standard error** of the mean. When σ is unknown, which is almost always, the sample standard deviation s takes its place and the result $s/\sqrt{n}$ is the estimated standard error.

The formula has two parts. The first, $\mathrm{E}[\bar{X}] = \mu$, in which E stands for the expected value, the average of the sample mean over all the samples that could be drawn, says that the sample mean is right on average, with no systematic tendency to be high or low; a statistic with this property is called **unbiased**, and it is the mathematical form of the promise that a random sample makes and a self-selected one does not. The second says how far a single sample mean typically lands from μ, and the square root in its denominator is the most consequential square root in statistics. Capital $\bar{X}$ denotes the sample mean as a random variable, before the sample is drawn; lower case $\bar{x}$ is its value in one particular sample.

Derivation **The expected value and the variance of a sample mean**

Write the n draws as random variables $X_1, X_2, \ldots, X_n$. Each is one case drawn at random from the population, so each has expected value μ and variance σ^2 (the variance is the square of the standard deviation), and because the draws are independent, the value of one tells nothing about another. Three rules for random variables are needed; Unit 4 derives them from the definition of the expected value, and they are restated here so that this unit can be read on its own. The expected value of a sum is the sum of the expected values, whether or not the terms are independent. The variance of a sum of *independent* random variables is the sum of their variances. And multiplying a random variable by a constant a multiplies its expected value by a and its variance by a^2, because the variance is built from squared distances.

The sample mean is $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. By the first and third rules,

$$\mathrm{E}[\bar{X}] = \frac{1}{n}\left(\mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n]\right) = \frac{1}{n} \cdot n\mu = \mu.$$

By the second and third rules,

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\right) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n},$$

and taking the square root gives $\mathrm{SD}(\bar{X}) = \sigma/\sqrt{n}$. The n in the denominator of the variance arises because the summing multiplies the variance by n while the averaging divides it by n^2, and together they leave one factor $1/n$; averaging n independent quantities therefore divides the variance by n, and the standard deviation by only $\sqrt{n}$. The independence of the draws was used once, for the variance, and it is exactly what a simple random sample from a large population provides.

3.1 The square root law

The standard error falls with the square root of the sample size, and the consequence is uncomfortable for anyone who pays for data. Figure 2 draws $\mathrm{SE}(\bar{x}) = 7.74\,\mathrm{cm}/\sqrt{n}$ for the heights. Going from 25 to 100 men, four times the work, halves the standard error from 1.55 to 0.77 cm, and halving it again requires 400 men.

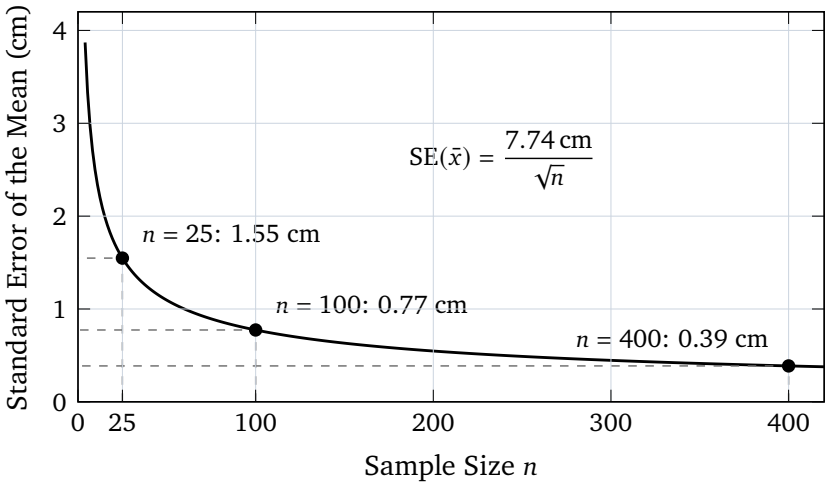


Figure 2. The standard error of the mean height against the sample size, computed from $\sigma/\sqrt{n}$ with $\sigma = 7.74$ cm. Each fourfold increase in n halves the standard error: 1.55 cm at $n = 25$, 0.77 cm at 100, and 0.39 cm at 400. Data: standard deviation of the heights of 2,500 men aged 20 and over, NHANES 2017 to 2018.

Key Formula **The sample size for a given standard error**

$$\frac{\mathrm{SE}_2}{\mathrm{SE}_1} = \sqrt{\frac{n_1}{n_2}} \qquad \text{and therefore} \qquad n_2 = n_1 \left(\frac{\mathrm{SE}_1}{\mathrm{SE}_2}\right)^2$$

If a sample of size n_1 gives the standard error SE_1, a sample of size n_2 from the same population gives SE_2. To divide the standard error by a factor k, the sample must be k^2 times as large.

Worked Example **Halving the error, and cutting it to a third**

A sample of 25 men gives a standard error of $7.74/\sqrt{25} = 1.55$ cm for the mean height. To bring it down to 0.77 cm, half as much, the Key Formula requires $n_2 = 25\times(1.55/0.77)^2 = 25\times4 = 100$ men. To bring it down to a third, 0.52 cm, it requires $25 \times 3^2 = 225$ men, and to a tenth, 0.155 cm, it requires $25\times100 = 2{,}500$ men, which happens to be the whole population of the example. So halving the standard error takes four times the sample, and cutting it to a third takes nine times the sample. The rule explains why surveys of a few thousand people are common and surveys of a few hundred thousand are rare: each further decimal place of precision costs a hundred times as many people as the one before.

Check Your Understanding **Question 2**

A simple random sample of 400 adults has a mean daily screen time of 4.2 hours with a standard deviation of 2.0 hours. What is the estimated standard error of the mean?

(A) 2.0 hours, the standard deviation of the sample

(B) 0.005 hours, the standard deviation divided by 400

(C) 0.1 hours, the standard deviation divided by 20

(D) 0.21 hours, five per cent of the mean

The solution is at the end of the unit.

4 The law of large numbers

The standard error shrinks towards zero as n grows, and this is the quantitative form of an old and famous statement. Jacob Bernoulli (1655-1705) proved in his book *Ars Conjectandi*, published in 1713 after his death, that the proportion of successes in a long series of independent trials comes as close as one likes to the probability of success, and that it does so with a probability as near to certainty as one likes, provided the series is long enough. He called it his golden theorem; it is now called the law of large numbers, and it holds for means as well as for proportions.

Theorem **The law of large numbers**

Let $X_1, X_2, X_3, \ldots$ be independent draws from one population with mean μ, and let $\bar{X}_n$ be the mean of the first n of them. Then for every margin $\varepsilon > 0$, however small,

$$\mathrm{P}\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) \longrightarrow 0 \qquad \text{as } n \to \infty.$$

In words, the probability that the sample mean differs from the population mean by more than any fixed amount can be made as small as desired by taking the sample large enough. The theorem requires independent draws and a population with a finite mean; it does not require any particular shape of the population.

When the population also has a finite standard deviation σ, the law follows from the standard error in a few lines through an inequality of Pafnuty Chebyshev (1821-1894), which says that the probability of a deviation of more than ε between $\bar{X}_n$ and μ is at most $\sigma^2/(n\varepsilon^2)$, a bound that goes to zero as n grows whatever ε is. The bound is crude, because it must hold for every shape of population; the next section gives a far sharper statement for large n.

4.1 What the law promises, and what it does not

The law of large numbers is why casinos and insurance companies exist. On a European roulette wheel with 37 pockets, a player who bets one euro on red wins one euro with probability 18/37 and loses it with probability 19/37, so the expected result is −1/37 of a euro, a loss of 2.7 cents per euro bet. A single player on a single evening may win a great deal, but over the millions of bets a casino takes in a year its average take per euro will be close to 2.7 cents, and its income is as predictable as a

factory's. An insurer who covers 100,000 houses against fire cannot say which of them will burn, but can say almost exactly how many.

What the law does not promise is a correction of the past. Suppose 1,000 tosses of a fair coin have produced 530 heads. The law says that the *proportion* of heads over the whole series will approach one half as the tossing continues. It does not say that tails are now due. The next 9,000 tosses are expected to split evenly, about 4,500 each, so the expected count after 10,000 tosses is 5,030 heads against 4,970 tails: the surplus of 30 heads is expected to persist, and the proportion nevertheless moves from 53.0 per cent to 50.3 per cent, because the same 30 are now divided by 10,000 instead of 1,000. The law works by burying the early surplus under many later tosses, never by compensating for it. The belief that a run of one outcome makes the other more likely is the gambler's fallacy, and the "law of averages" that is supposed to justify it does not exist.

Check Your Understanding **Question 3**

A fair coin has produced 530 heads in 1,000 tosses. The coin is tossed 9,000 more times. What does the law of large numbers say about the whole series of 10,000 tosses?

(A) Tails are now more likely than heads on each of the next tosses, because the law requires the two counts to even out

(B) The next 9,000 tosses are expected to split evenly, so the proportion is expected to fall from 53.0 to 50.3 per cent while the surplus of 30 heads remains

(C) The proportion is expected to stay near 53 per cent over the whole series, because the first 1,000 tosses showed that this coin favours heads

(D) After 10,000 tosses the counts are expected to be equal at 5,000 each, because later tosses cancel the surplus of 30 heads

The solution is at the end of the unit.

5 The central limit theorem

The law of large numbers says where the sample mean goes; the standard error says how far from μ it typically is. One more thing can be said, and it is the reason the normal curve appears on almost every page of statistics: the shape of the sampling distribution of the mean is known as well, and it is nearly the same whatever the shape of the population.

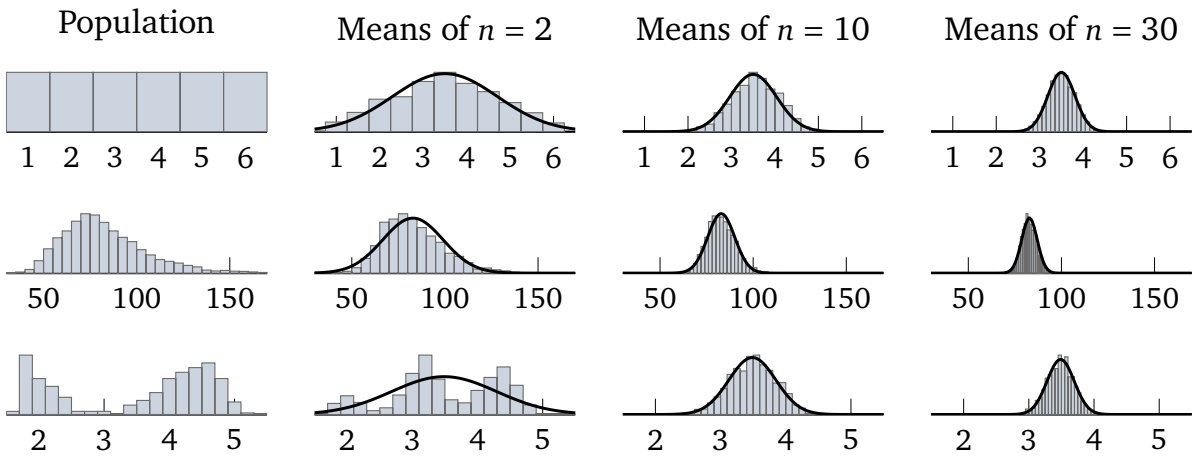


Figure 3. Three populations and the distributions of their sample means. Top row: the six faces of a fair die (uniform). Middle row: the body weights of 5,185 adults aged 20 and over, NHANES 2017 to 2018 (right-skewed; 20 adults heavier than 170 kg lie beyond the right edge). Bottom row: the durations of 272 eruptions of Old Faithful, Azzalini and Bowman (1990), from Rdatasets (bimodal). Each sampling distribution holds the means of 2,000 simulated samples drawn with replacement (seeds 5002 to 5004 for the die, 5010 to 5012 for the weights, and 5030 to 5032 for the geyser, one per sample size); bin widths are 1, 0.5, 0.2, and 0.1 for the die, 5, 5, 2, and 1 kg for the weights, and 0.2, 0.2, 0.1, and 0.05 minutes for the geyser. The curves are the normal density with mean μ and standard deviation $\sigma/\sqrt{n}$ of each population; the vertical scales differ between panels.

Figure 3 shows this for three populations that could hardly be more different. The first is the six faces of a fair die, each equally likely, a flat distribution. The second is the body weight of 5,185 adults measured in NHANES in 2017 and 2018, which is skewed to the right, with a long tail of heavy people. The third is the duration of 272 eruptions of the Old Faithful geyser in Yellowstone National Park, which has two separate humps, short eruptions of about two minutes and long ones of about four and a half. From each population 2,000 samples of size 2, 10, and 30 were drawn with a seeded generator, and the means of the samples were collected into histograms.

Three things happen in every row, and they are the three statements of this unit. The histograms of the means stay centred on the population mean. They become narrower as n grows, by the factor $\sqrt{n}$. And they take on the shape of the normal curve drawn over them, whatever the shape they started from: the flat die is almost normal at $n = 10$, the skewed weights still lean to the right at $n = 10$ and are very nearly symmetric at $n = 30$, and the two humps of the geyser, still visible as a ragged shape at $n = 2$, have merged into a single bell by $n = 10$. This is the content of the central limit theorem.

Theorem **The central limit theorem**

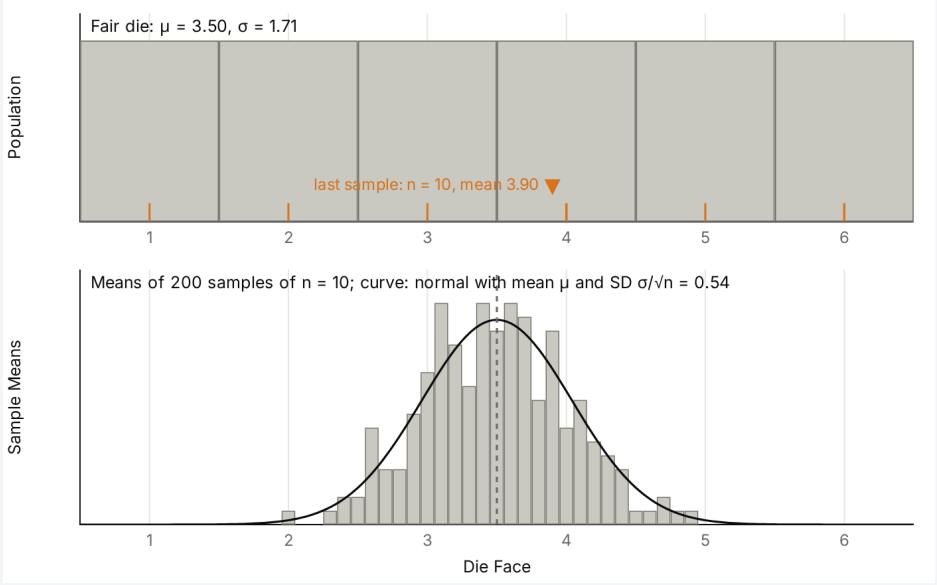
Let $X_1, X_2, \ldots, X_n$ be independent draws from one population with mean μ and finite standard deviation σ, and let $\bar{X}$ be their mean. Then, as n grows, the distribution of $\bar{X}$ approaches the

normal distribution with mean μ and variance σ^2/n,

$$\bar{X} \;\approx\; \mathrm{N}\!\left(\mu,\; \frac{\sigma^2}{n}\right) \qquad \text{for large } n,$$

whatever the shape of the population. Equivalently, the standardized mean $(\bar{X} - \mu)/(\sigma/\sqrt{n})$, which is the distance of $\bar{X}$ from μ measured in standard errors, approaches the standard normal distribution, the normal distribution with mean 0 and standard deviation 1. The same holds for the sum $X_1 + \cdots + X_n$, which is approximately $\mathrm{N}(n\mu, n\sigma^2)$. The theorem requires independent draws from the same distribution and a finite variance; how large n must be depends on the shape of the population.

Interactive **The sampling distribution of the mean, built one sample at a time**



Choose a population (a fair die, the NHANES weights and heights, or the Old Faithful eruptions), set the sample size, and draw samples: each click draws a sample with a seeded generator, shows it in the upper panel, and adds its mean to the histogram below, under the normal curve with the population mean as centre and the standard error as spread. Notice that the histogram of means narrows with the square root of the sample size and becomes a bell even for the skewed weights and the two-humped geyser. Simulated; the first 200 samples are the same for every reader.

Click the picture or this link to try it in the online version of this unit.

The last sentence of the box is where a careful reader should pause. Textbooks often give $n \geq 30$ as the point from which the normal approximation may be used, and Figure 3 shows why the number is only a habit. For a symmetric population such as the die, or the heights of Section 3, samples of 5 or 10 are already enough. For a moderately skewed population such as the weights, 30 is about right. For a strongly skewed population, such as personal incomes or insurance claims, where a few enormous values dominate the mean, several hundred observations may be needed, and for populations whose tail is so heavy that the variance is not finite at all, and such populations do exist, the theorem makes

no promise. The theorem also says nothing about the single observations. A large sample of incomes is just as skewed as a small one; it is the *mean* of the sample whose distribution becomes normal, and confusing the two is one of the most common misreadings in the subject.

The normal shape is what lets the standard error be read as a probability. A standard normal variable lies within one standard deviation of its centre with probability 68 per cent and within two with probability 95 per cent, so a sample mean lies within one standard error of μ in about 68 per cent of samples and within two in about 95 per cent, which is what the simulation of Section 3 found with 674 and 945 of its 1,000 means. The following Key Formula is the bridge to the next units, which turn it into confidence intervals and tests.

Key Formula **Standardizing a sample mean**

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

The number of standard errors by which an observed sample mean $\bar{x}$ lies above or below the population mean. By the central limit theorem, z follows approximately the standard normal distribution when n is large enough for the shape of the population, so the probability of a z beyond ± 2 is about 5 per cent and beyond ± 3 about 0.3 per cent.

5.1 The Galton board

The oldest demonstration of the theorem is a wooden board with rows of pins, and it is also the clearest, because the mechanism that produces the bell shape can be watched one ball at a time. A ball dropped onto the top pin bounces left or right, falls onto the next row, bounces again, and after passing every row drops into one of a row of slots at the bottom. Each bounce is a coin toss, and the slot the ball lands in is decided by how many of its bounces went to the right. Poured through by the hundred, the balls pile up in the slots in a bell shape, as the photograph at the head of this unit shows.

Worked Example **The Galton board as a sum of twelve bounces**

Consider a board with 12 rows of pins, so that every ball bounces 12 times, and number the slots from 0 to 12 by the number of bounces to the right that lead to them. Write $B_1, B_2, \dots, B_{12}$ for the bounces, each equal to 1 if the ball went right and 0 if it went left, each with probability one half, and independent of the others. A single bounce has expected value $1/2$ and variance $1/2 \times 1/2 = 1/4$, because the average of 0 and 1 is one half and both values lie exactly one half away from it. The slot is the sum $S = B_1 + \cdots + B_{12}$, and the rules of the derivation box in Section 3 give

$$\mathrm{E}[S] = 12 \times \tfrac{1}{2} = 6, \qquad \mathrm{Var}(S) = 12 \times \tfrac{1}{4} = 3, \qquad \mathrm{SD}(S) = \sqrt{3} = 1.73.$$

The exact distribution of S is the binomial distribution $\mathrm{Bin}(12, 1/2)$: the probability of slot k is the number of paths with k right turns, $\binom{12}{k}$, divided by the $2^{12} = 4{,}096$ equally likely paths. The middle slot, $k = 6$, has $\binom{12}{6} = 924$ paths and therefore the probability $924/4{,}096 = 22.6$ per cent; the outer slots have one path each, a probability of $1/4{,}096$. So the balls settle around slot 6 with a standard deviation of 1.73 slots, and a normal curve with that mean and standard deviation passes almost exactly through the tops of the twelve bars. The balls form a bell because each ball's position is a sum of many small, independent, equally likely steps, which is

exactly the situation of the central limit theorem, with the sum in place of the mean. The case of coin tosses is where the theorem was first discovered, by Abraham de Moivre (1667-1754) in 1733.

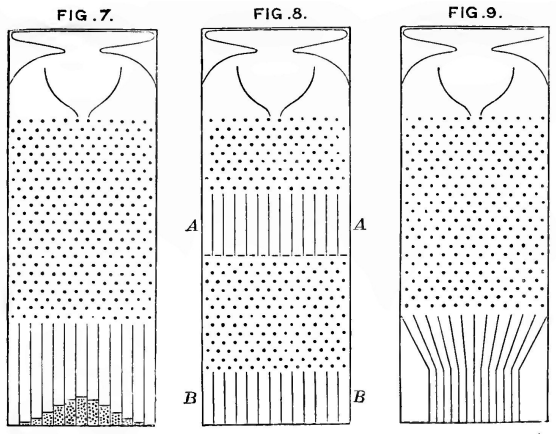


Figure 4. Galton's drawings of the apparatus in *Natural Inheritance* (1889). In Fig. 7 the shot poured through the funnel passes the rows of pins and settles in the compartments in a bell-shaped heap. Fig. 8 is the two-stage version: the shot is caught halfway down at A, where it already forms a bell, and each compartment is then released separately through a second bank of pins into the compartments at B. Fig. 9 slopes the lower compartments so that the heaps can be compared. Credit: Francis Galton, Natural Inheritance, Macmillan, 1889, page 63; Internet Archive scan, public domain.

Excursion *Galton and the quincunx*

Francis Galton (1822-1911), a cousin of Charles Darwin, an explorer and meteorologist, and a pioneer of the weather map, of correlation, and of fingerprint identification, had the first board built in the 1870s and described it in his book *Natural Inheritance* of 1889, from which Figure 4 is taken. He called it the quincunx, after the pattern of five dots on a die, which is how the pins are arranged. His interest was not in gambling. He wanted to understand why the heights of people, the sizes of seeds, and a hundred other measurements of living things follow the bell curve, and the board gave him a mechanical answer: any quantity that is the sum of many small independent influences will do so.

The two-stage board of his Fig. 8 made a subtler point. If the shot is caught halfway down, each compartment at A holds a heap that is itself a small bell, and when the compartments are released one after another, each heap spreads into a bell of its own around a different centre. Yet the final heap at B is a single, wider bell, exactly as if the shot had fallen straight through. Galton used this to argue that a population could stay normally distributed from one generation to the next even though every family scatters its children around its own average, and this argument, which he first made in a lecture of 1877, led him over the following decade to the discovery of regression, which Unit 8 tells. He wrote that he knew of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of error, his name for the normal curve, and that the law would have been personified by the Greeks and deified, had they known of it.

Check Your Understanding **Question 4**

The mean of a simple random sample is to be treated as normally distributed. For which of the following populations does that require the largest sample size?

(A) The heights of adult women, which form a symmetric bell-shaped distribution

(B) The results of rolling a fair die, which are uniform from 1 to 6

(C) The annual incomes of households, where a few very large values stretch the right tail

(D) The outcomes of a fair coin, coded 0 and 1

The solution is at the end of the unit.

The theorem has a long history. De Moivre's result of 1733 covered only coin tosses; Pierre-Simon Laplace (1749-1827) extended it in 1810 to sums of arbitrary independent quantities; the Russian mathematician Aleksandr Lyapunov (1857-1918) gave the first rigorous proof of a general version in 1901; and the name, *zentraler Grenzwertsatz*, was coined by George Pólya (1887-1985) in the title of a paper of 1920, because of the central place the theorem holds in probability. Versions for draws that are not identically distributed exist and go beyond this unit; they are the reason the bell curve appears in measurement errors and in the heights of people, none of which is a coin toss.

6 Proportions and the size of a sample

Everything said about the mean applies to a proportion, because a proportion is a mean in disguise. Give every case the value 1 if it has the property in question and 0 if it does not; then the sum of the values is the count of cases with the property, and their mean is the proportion. In the population the mean of these 0 and 1 values is the proportion p of cases with the property, and their variance is $p(1-p)$, because a value of 1 lies $1-p$ above the mean and a value of 0 lies p below it, and the average of the squared distances, $p(1-p)^2 + (1-p)p^2$, comes to $p(1-p)$; it is $1/4$ at $p = 1/2$ and smaller elsewhere. The sample proportion $\hat{p}$, read "p hat", is the sample mean of the same values, and the Key Formula of Section 3 gives its standard error at once.

Key Formula **The sampling distribution of a proportion**

$$\mathrm{E}[\hat{p}] = p, \qquad \mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}, \qquad \hat{p} \;\approx\; \mathrm{N}\left(p,\; \frac{p(1-p)}{n}\right)$$

For a simple random sample of size n from a population in which a proportion p has the property, the sample proportion $\hat{p}$ is unbiased, has the standard error $\sqrt{p(1-p)/n}$, and is approximately normal when n is large enough; the usual condition is that both np and $n(1-p)$ are at least 10. For p near one half the standard error is close to $0.5/\sqrt{n}$.

Worked Example **The proportion of heads in 100, 1,000, and 10,000 tosses**

For a fair coin, $p = 1/2$ and the standard error of the proportion of heads is $\sqrt{0.25/n} = 0.5/\sqrt{n}$. In 100 tosses it is $0.5/10 = 0.05$, or 5 percentage points, so proportions anywhere from 40 to 60 per cent are unremarkable. In 1,000 tosses it is $0.5/31.6 = 0.0158$, about 1.6 percentage points, and in 10,000 tosses $0.5/100 = 0.005$, half a percentage point.

In 2023 a group of 48 people led by František Bartoš recorded 350,757 coin flips to test a prediction made in 2007 by Persi Diaconis and two colleagues, that a coin caught in the hand lands the way it started slightly more often than half the time, about 51 per cent of the time, because it wobbles as it spins. The coins landed on their starting side in 50.8 per cent of the flips. Is 50.8 per cent distinguishable from one half? The standard error for $n = 350{,}757$ is $0.5/\sqrt{350{,}757} = 0.00084$, or 0.08 percentage points, so the observed 50.8 per cent lies $0.8/0.084 = 9.5$ standard errors above 50 per cent. A difference of nine and a half standard errors does not happen by chance under the central limit theorem, and so the 350,757 flips establish a real bias of 0.8 percentage points, which 10,000 flips, with a standard error of 0.5 points, could not have told from luck. The same bias in 100 flips would have been about one sixth of a standard error, invisible.

6.1 The poll of 1,000 people

The number in the first sentence of this unit now has a source. An opinion poll that interviews a simple random sample of $n = 1{,}000$ people, in which a candidate's true share is near one half, measures that share with a standard error of $\sqrt{0.25/1{,}000} = 0.0158$. Two standard errors, more exactly 1.96 of them, come to 0.031, and that is the figure of "about 3 percentage points" printed under every such poll; Unit 6 explains exactly what it means and how it becomes a confidence interval. Four times as many interviews, 4,000, bring it to 1.5 points, and 16,000 interviews bring it to 0.8 points. The margin also assumes a simple random sample and a complete response, which no real poll has, and the published margins of serious pollsters are corrected for their actual designs by methods beyond the unit.

One thing is absent from the formula, and its absence surprises most people: the size of the population. The standard error of a sample of 1,000 is the same whether the 1,000 were drawn from a town of 100,000 or from a country of 60 million. The exact formula for a sample drawn without replacement from a finite population of N cases multiplies $\sigma/\sqrt{n}$ by the factor $\sqrt{(N-n)/(N-1)}$, the **finite population correction**, which is 0.995 for 1,000 out of 100,000 and 0.99999 for 1,000 out of 60 million. Only when the sample is a large fraction of the population, more than 5 or 10 per cent of it, does the correction matter, and then it helps. A spoonful of soup tells the cook whether the pot is salted, however large the pot, provided the pot was stirred. The stirring is the random draw, and the Literary Digest's 2.4 million unstirred spoonfuls are the lesson that no quantity of data makes up for its absence.

Key Points

- A simple random sample gives every set of n cases the same chance of selection, which removes any link between who is chosen and what is measured; stratified, cluster, and systematic samples are practical variants, and the sampling frame decides whom a sample

can represent at all.

- Bias is a systematic error of the method and variability is the scatter from sample to sample; a larger sample reduces variability and leaves bias untouched, which is why 2.4 million self-selected ballots missed the 1936 election by 20 percentage points while a smaller sample came much closer.
- The mean of a simple random sample is unbiased, $\mathrm{E}[\bar{X}] = \mu$, and its standard error is $\mathrm{SE}(\bar{x}) = \sigma/\sqrt{n}$, so that quadrupling the sample halves the error and cutting the error to a third takes nine times the sample.
- The law of large numbers says that the sample mean comes as close to μ as desired when n is large enough; it works by burying early deviations under many later draws and promises no compensation for a run.
- The central limit theorem says that the sampling distribution of the mean is approximately $\mathrm{N}(\mu, \sigma^2/n)$ whatever the shape of the population, provided the draws are independent and the variance is finite; the n needed grows with the skewness of the population, and the single observations do not become normal.
- A proportion is the mean of values 0 and 1, so $\hat{p}$ has the standard error $\sqrt{p(1-p)/n}$, about $0.5/\sqrt{n}$ near one half, which gives a poll of 1,000 its margin of about 3 percentage points regardless of the size of the population.

Solutions to the questions in the text: 1-B, 2-C, 3-B, 4-C. The online version of this unit explains each answer.