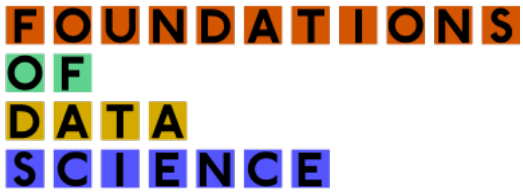

Inf2 – Foundations of Data Science 2021
Topic: Randomness, sampling and simulation

David C. Sterratt

20th November 2021

Recommended reading:

- *Computational and Inferential Thinking*, Chapter 9
- *Computational and Inferential Thinking*, Chapter 10
- *Computational and Inferential Thinking*, Chapter 14
- *Modern Mathematical Statistics with Applications*, Sections 6.1 and 6.2

1 Video: Introduction to statistical inference

Statistics The German word *Statistik* arose in the 18th century and originally referred to "data about the *state*" (country). The first use of "statistical" in the English language was in 1791 in the *Statistical Account of Scotland*. Sir John Sinclair, an elder in the Church of Scotland, sent a questionnaire to ministers in every parish (church district) in Scotland. The questionnaire asked many questions about agriculture, industry, economics, employment, poverty and education, as well as "The state of the manners, the morals, and the religious principles of the people". In fact empires and dynasties have been collecting data about population and trade for much longer than this, going back to the Han dynasty in China and the Roman Empire.

Inferential statistics This use of the word statistics above relates to the whole population. In contrast, back in the topic on Statistical Preliminaries, we looked at the difference between a sample and a population. We also considered a number of statistics that could apply to the population and to the sample: the mean, variance, standard deviation and median. **Inferential statistics** is the process of drawing conclusions about quantities that are not observed (Gelman et al., 2004).

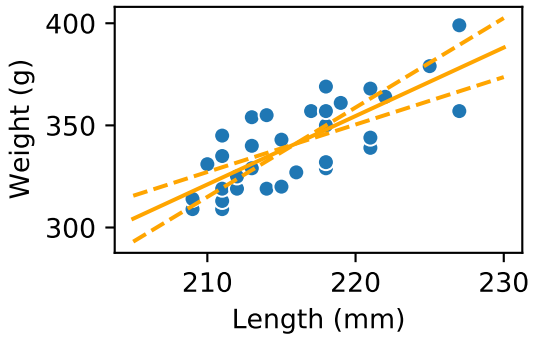

Figure 1: Uncertainty in regression line. The best estimate regression line (solid) and the lines at the edge of the 95% confidence interval (dashed lines). Note this plot is simplified, since the uncertainty in the intercept is not represented.

One example of an inferential statistics task is **estimation** of a population statistic from a sample of that population. For example, suppose we've weighed a sample of 10 wild cats from a population of 400. We know what the sample mean and sample standard deviation is. On the basis of this information, what is the best estimate of the population mean (**point estimation**), and how confident can we be in that estimate (**confidence interval estimation**)?

Inferential statistics has been around for much longer than the word "statistics". The 9th-century book "Manuscript on Deciphering Cryptographic Messages" written in Arabic by Al-Kindi (educated in Baghdad) shows how to decipher encrypted messages by frequency analysis, i.e. counting the frequency of particular letters.

Inferential statistics tasks Inferential statistics can seem like a toolbox full of tools with confusing names such as "standard error of the mean", t-test, χ^2 test, bootstrap, and a confusing set of rules about what to use each tool for. We're going to try to give you an idea of what task each tool is useful for, and how it works. There are three main tasks we will consider:

1. Estimation
2. Hypothesis testing
3. Comparing two samples (A/B testing)

Estimation We've already given one example of estimating an unobserved quantity (the population mean). Another example of an unobserved quantity is linear regression coefficients. We already know how to find (point) estimates of them, using the formulae we covered earlier in the course. But in part of the course we will learn how to estimate the **confidence intervals** around the point estimates, which will tell us how much uncertainty there is in our estimates. For example, we'll be able to say that we estimate the mean weight of squirrels in the population to be 320 ± 16g, with 95% confidence, i.e. in the confidence interval $[304, 336]$g. In a linear regression of the weight of a sample of squirrels on their length (Figure 1), we will be able to say that the best estimate of the slope of the regression line is 3.35 g/mm, but we are 95% confident that the slope lies in the interval [2.32, 4.38] g/mm.

Hypothesis testing In hypothesis testing, we are trying to ascertain which of two or more competing theories are the best explanation of the data. For example, in 1965 a court case was brought against the state of Alabama (Swain versus Alabama, 1965) due to there being no Black members of the jury in a trial. Part of the case concerned the fact that at one stage of the jury selection, 8 Black people were chosen for a jury panel of 100 people, but the fraction of Black people in the population was 26%. Our question is "Is the jury selection system biased against Black people?".

Comparing two samples (also known as A/B testing) Here we have two samples that have been treated differently, and we want to either test if the groups are different, or estimate how different they are. For example, to find out the effectiveness of a vaccine, we select a sample of volunteers from the population randomly, divide them randomly into two groups, give the vaccine to one group (Treatment group) and give the other group a placebo (Control Group). In the vaccine group 3 volunteers catch the disease, but in the placebo group 95 volunteers catch the disease. Is the vaccine effective? How much would we expect the vaccine to cut the risk of catching the disease if we give it to the whole population?

In the context of user testing, often in web applications, this is called A/B testing. A famous example was at Amazon, where a developer had the idea of presenting recommendations as you add items to a shopping cart (Kohavi et al., 2007). Amazon managers forbid the developer to work on the idea, but the developer disobeyed orders and ran a controlled experiment on users, by splitting them into two groups ("A" and "B"), one which had recommendations shown and one which didn't. The group which had recommendations shown bought more, and displaying recommendations quickly became a priority for Amazon.

Two approaches to statistical inference We are going to learn a number of techniques for undertaking point and interval estimation, hypothesis testing and comparing samples. We will also think carefully about the interpretation of these techniques. There are two main approaches to undertaking statistical inference tasks:

1. **Statistical simulations.** Here we use repeated random sampling to carry out the statistical inference procedures. The advantages of statistical simulation procedures are they often require fewer assumptions about the data and/or hypothesis, and they require somewhat less theory to understand. However, they can be compute-intensive, and care is still needed in their use.

2. **Statistical theory.**. Here we use the properties of various well-known theoretical distributions to draw inferences about our data. We need to check that the assumptions behind the distribution match the statistical question we are trying to answer. For example, a distribution of delays to flights is likely to be highly right-skewed, so we shouldn't assume a normal distribution when dealing with it. Typically, the process is not compute-intensive: very often it amounts to arithmetic and then reading of a quantity from a distribution table. These procedures come as standard in a number of stats packages, including R and Python's statsmodels.

A number of fundamental concepts underpin both the statistical theory and statistical simulations.

Plan for this semester The plan for the statistical inference topics in this semester will be:

1. Fundamental theory (the rest of this topic):

    - We'll learn how we can use statistical simulations to generate samples from a model and compute statistics for each of these samples to give a **sampling distribution**.

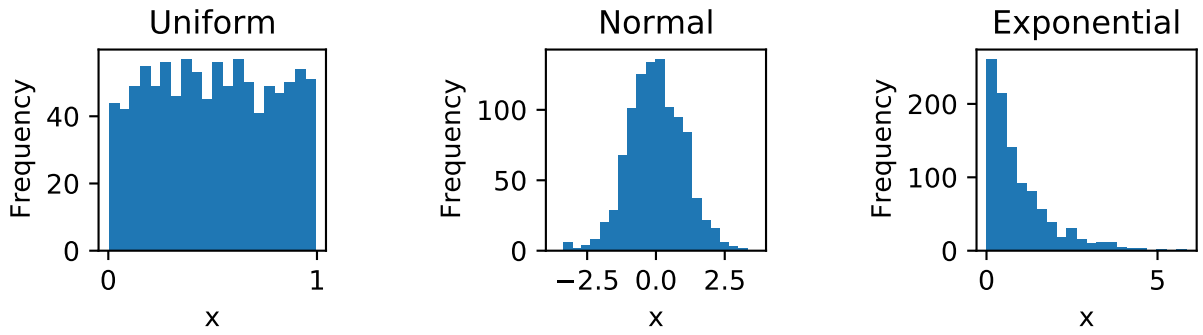

Figure 2: Histograms of 1,000 samples taken from normal, uniform and exponential distributions.

- Learn about the distribution of the mean of repeated samples from a model. This will lead us to the central limit theorem, which can help us to estimate the uncertainty in our estimate of the mean, i.e. confidence intervals, and the law of large numbers, which also helps with estimation.

2. Estimation (next lecture)
3. Hypothesis testing (third lecture)
4. An interlude - Logistic regression (fourth lecture)
5. Comparing two samples with statistical simulations (fifth lecture)

2 Video: Sampling, statistics and simulations

Sampling A prerequisite for statistical simulations is being able to sample from probability distributions and from sets of discrete items, including observed data.

Random sample In a random sample of size n from either a continuous probability distribution, or a finite population of N items the random variables $X_1, \ldots, X_n$ comprising the sample are all independent and all have the same probability distribution.

Sampling from probability distributions You should be familiar with sampling from random number generation. A standard random number generator produces numbers within an interval (e.g. [0, 1]) with uniform probability for each number, i.e. it samples from a uniform distribution. We can demonstrate the distribution of a standard random number generator by drawing many samples and plotting a histogram (Figure 2). We adapt these functions to sample from any univariate distribution, e.g. a normal distribution or an exponential distribution (Figure 2).

Sampling from a set of discrete items We can also sample from a population of discrete items. We can select n items from a set of N items at random either **without replacement** or **with replacement**. If we sample without replacement, it is as though we are pulling items of various types (e.g. coloured balls) at random out of a bag, and not replacing them. We can only sample up to N items, and also, as we remove items from the bag, the probabilities of drawing a particular type (colour) changes. If

we sample with replacement, we put the item back in the bag, before making our next choice – we can carry on doing this for ever. We could construct an algorithm for random sampling either with or without replacement from a uniform random number generator, but these functions are provided in packages such as `numpy.random.choice` in Python.

A particular application of sampling from a set of discrete items is creating a sample of a larger data set.

Non-random samples from a population We can also imagine ways of sampling that are not systematically random. For example, we might have a list of the daily takings in a restaurant. We could take the first n days. But suppose that the dataset has been sorted in terms of takings? We would then have days with low takings at the start of the list, so the statistics of the sample would not resemble the statistics of the population. We could try taking every 7th day in the list – but if the list is in date order we will always be sampling from one day of the week, e.g. Mondays. Random sampling ensures that we don't have this type of problem.

Samples of convenience When we are collecting data, it might be tempting to sample from the data that we can collect conveniently. For example, a polling company may find it easier to contact people who have more time to answer the phone, which may tend to be retired people. If we don't correct for this sort of bias, it's called a **sample of convenience**. One way of combating convenience sampling is **stratified sampling**, in which the sampling is targeted so that the proportions of attributes of the sample matches the proportions in the population.

Definition of a statistic Before going further, it's helpful to have the definition of **statistic**: "A **statistic** is any quantity whose value can be calculated from sample data."(*Modern Mathematical Statistics with Applications* 6). We probably recognise the mean, variance and median as statistics by this definition. But we've also derived other quantities from sample data, such as the correlation coefficient and regression coefficients – they are also statistics. We will follow *Modern Mathematical Statistics with Applications* and denote a statistic using an uppercase letter, to indicate that it is a random variable, since its value depends on the particular sample selected. E.g. $\overline{X}$ represents the mean and S^2 the variance.

Simulations and sampling Before considering inferential statistics proper, we will focus on running **statistical simulations**, i.e. using a computer program to make predictions from probabilistic models of real-world processes. For example, the probabilistic model of tossing a coin multiple times is that the tosses are independent and that the probability of a head is 1/2 (or perhaps another value, if with think the coin is loaded). The statistical simulation generates a sequence of heads and tails.

To do this we need to decide on:

- The statistic of interest ($\overline{X}$, S, etc.)
- The population distribution (e.g. normal with particular mean and variance) or set of discrete items
- The sample size (denoted n)
- The number of replications k

The simulation procedure is then:

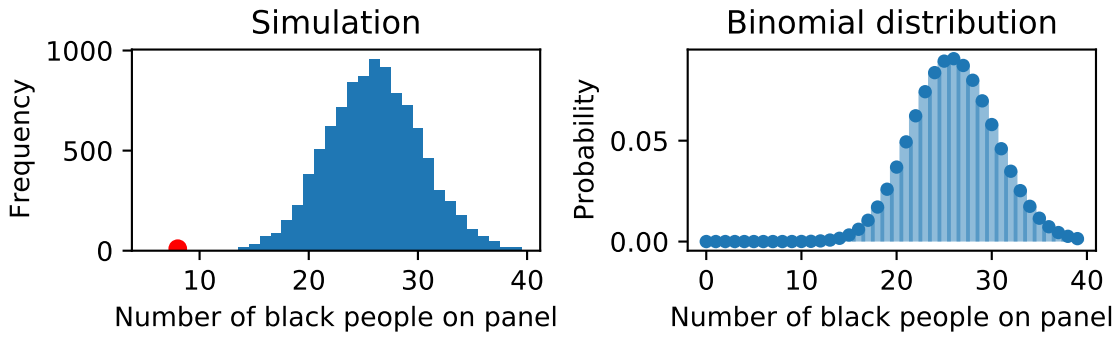

Figure 3: Results of statistical simulations of the panel size in Swain versus Alabama (1965). The blue histogram shows how many of 10 000 simulations produced jury panels of 100 with the given number of Black people on them. The red dot indicates the number of Black jurors in Swain versus Alabama (1965).

1. For i in $1, \ldots, k$
   (a) Sample n items from the population distribution or set of discrete items
   (b) Compute and store the statistic of interest for this sample
2. Generate a histogram of the k stored sample statistics

Example of hypothesis testing using a simulation experiment To demonstrate the utility of the statistical experiment we've introduced, let's look again at the example in which 26% of the population is Black and 8 Black people are selected to be on a jury panel of 100 people. The null hypothesis H_0 is "The jury panel was chosen at random from the population". We can map the null hypothesis onto the general framework above as follows:

- The statistic of interest is T_0, the number of Black people in a sample of $n = 100$ panel members
- The population distribution is a Bernoulli distribution with the sample space Black, Non-Black in which $p(\text{Black}) = 0.26$.
- The sample size is $n = 100$
- The number of replications $k = 10\,000$

We follow the procedure described in the previous section to give the results shown in Figure 3. Coding this up will be an exercise for you in the Labs. We can see that none of the 10 000 simulations of the null hypothesis produced a jury with 8 members, suggesting that we should reject the null hypothesis in favour of an alternative one. This looks like a clear-cut case; in the topic on Hypothesis testing, we'll consider in more detail how to interpret the results when the data is less distinct from the simulations.

Deriving the sampling distribution Note that in this example, we didn't have to go to the trouble of running a simulation experiment. We might have noticed that the total number of Black people will be distributed according to a binomial distribution with $n = 100$ and $p = 0.26$.

3 Video: Distributions of small samples statistics from probability distributions

Example of sampling from probability distributions In the previous example, we've sampled a total number of successes from a Bernoulli distribution. We'll now look at what happens when we sample the mean, standard deviation and median from the normal, uniform and exponential distributions by running the following simulations:

- Statistics of interest: mean $\overline{X}$, standard deviation S and median $\tilde{X}$
- Population distribution: Normal distribution with mean 0 and variance 1, Uniform distribution on [0, 1], Exponential distribution $p(x) = e^{-x}$.
- Sample size $n = 10$
- Number of replications $k = 10,000$

Figure 4. There are a number of points to notice about this plot:

Sample mean (first column), all distributions The distribution of the mean is narrower than the original distribution in every case. This is because some of the variability in the individual samples is averaged out. The standard deviation of this distribution is called the **standard error of the mean**.

Sample mean of normal distribution The distribution looks to be normal – it turns out that this is easy to prove.

Sample mean of uniform distribution The distribution is symmetric and looks to be near-normal.

Sample mean of exponential distribution The distribution is clearly skewed, but less so than the original exponential distribution.

Sample variance (second column) All these distributions are skewed, reflecting the fact that it's very unlikely to get 10 samples that are all very close together, and therefore have low variance. It turns out that there is a theoretical distribution (the χ^2 distribution) that describes the shape of sample variance from the normal distribution.

Median (third column) The main point to draw from this column is that we can use the simulation method to produce a distribution for any statistic, regardless of how easy it would be to calculate a theoretical distribution for it.

As we will see later, we could generate the sampling distribution of the mean and the variance analytically rather than by simulation. However, it is not always possible to compute sampling distributions of the desired statistics analytically, but we can always run statistical simulations.

4 Video: The distribution of the sample mean of large samples

The distribution of the sample mean A particularly common statistic of interest is the sample mean. It therefore makes sense to understand how the distribution of the sample mean depends on the distribution from which we sample and the number of samples we take.

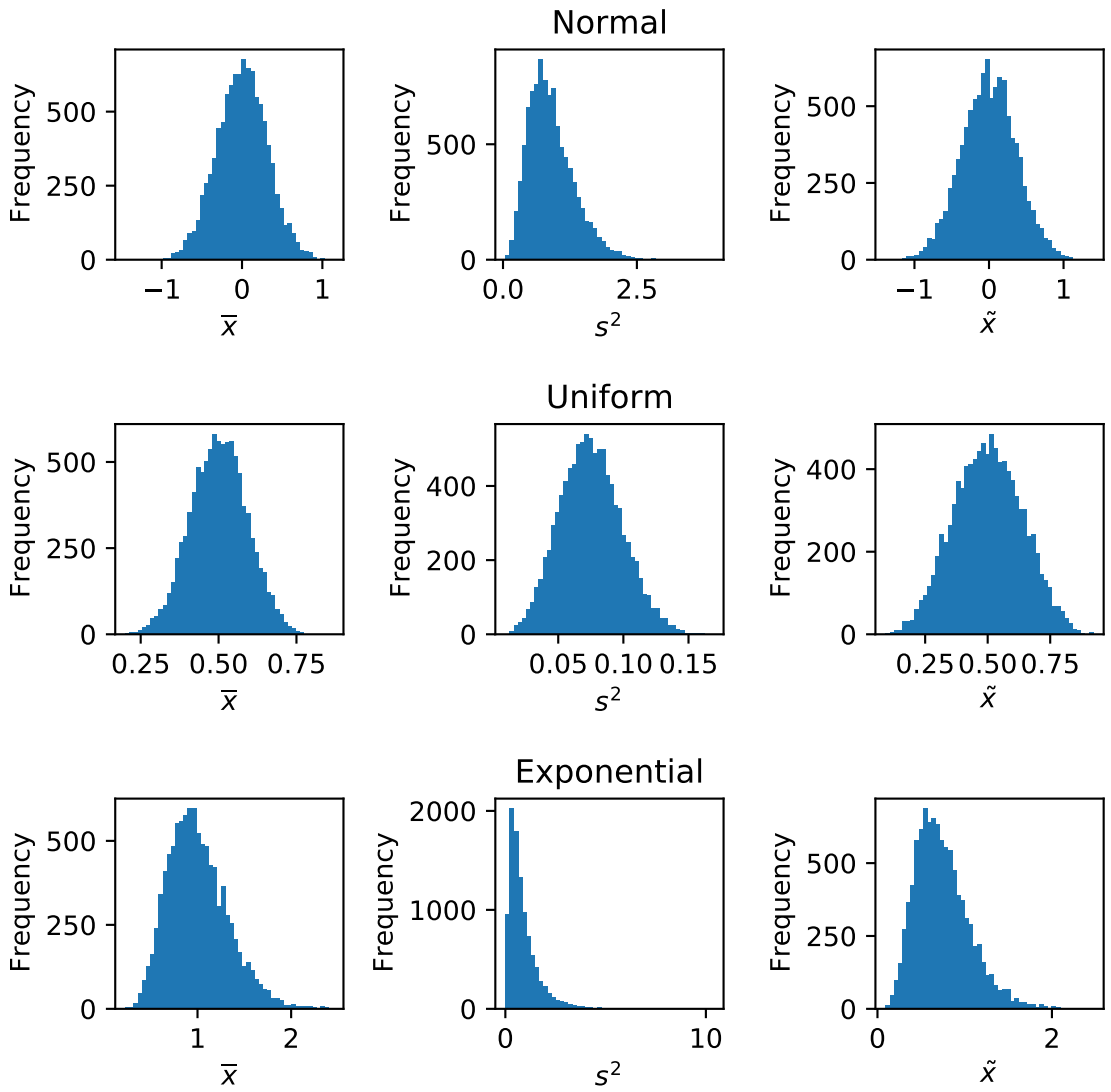

Figure 4: Sampling distribution generated by 10,000 simulations of the mean $\bar{x}$, variance s^2 and median $\tilde{x}$ of 10 samples drawn from a normal distribution (top row), uniform distribution (middle row) and exponential distribution (bottom row).

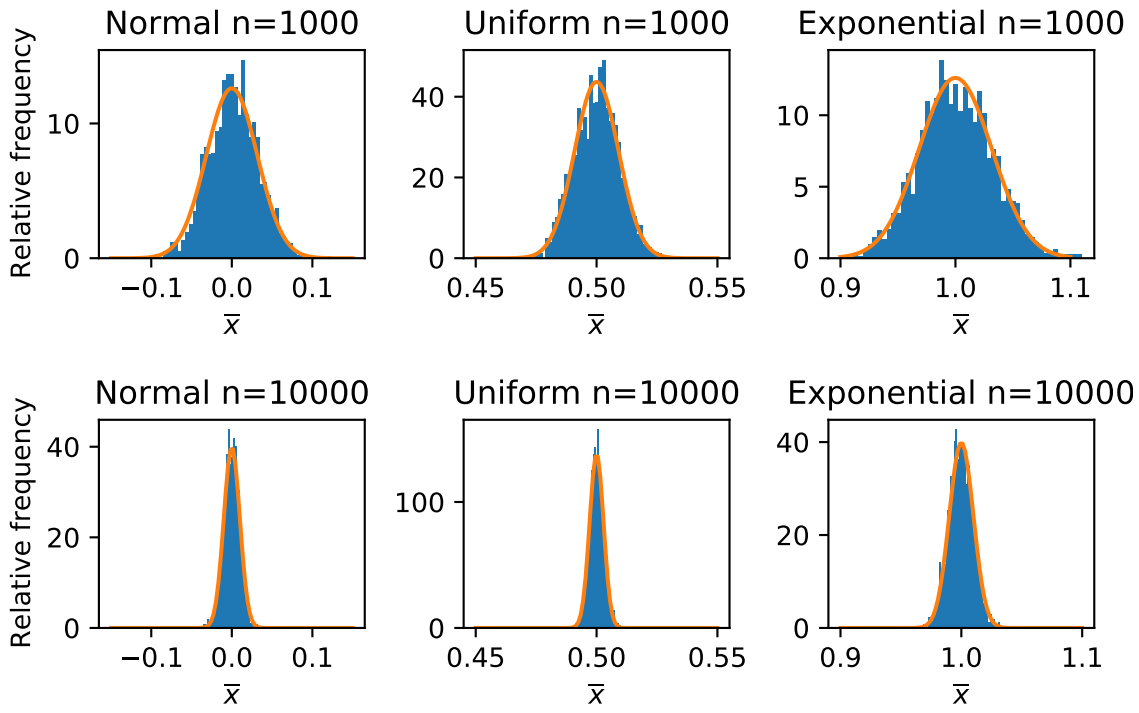

Figure 5: Distributions of means from samples of size $n = 1000$ (top row) and $n = 10000$ (bottom row) drawn from the normal, uniform and exponential distributions shown in Figure 2. The blue histograms show the histograms obtained from $k = 2000$ simulations. The orange curves are normal distributions with mean equal to the mean of the original distribution and variance $\sigma_{\overline{X}}$ equal to σ^2/n, where σ^2 is the variance of the original distribution.

We've already seen in Figure 4 that the sampling distribution of the mean of 10 items from a normal distribution is itself a normal distribution, though with smaller variance. However, the sample mean distributions for an exponential distribution in our simulation, was not normal. We can repeat the simulation experiments for the three distributions, but with larger sample sizes of $n = 1000$ and $n = 10000$ (Figure 5). What we see is remarkable: *the distributions of the sample means are all normal, regardless of whether they came from a normal, uniform or exponential distribution*. Perhaps less remarkably, we also see that as the sample size gets larger the distributions get narrower.

These simulations and observations give us the intuition for two very important statistical laws that apply to many non-normal distributions, as well as normal ones:

- The Central Limit Theorem
- The Law of Large Numbers

Central Limit Theorem Here is an informal statement of the **Central Limit Theorem** (CLT):

> The distribution of the mean [or sum] of a random sample drawn from any distribution will converge on a normal distribution. In the case of the sample mean, its expected value is the same as the mean of the population distribution, and its expected variance is a factor of n lower than the population variance. In the case of the sample sum, its expected value is the same as the product of the sample size n and the expected value of the distribution, and its expected variance is n times the variance of the population distribution.

We denote the expected variance of the mean $\sigma^2_{\overline{X}}$ and we call the standard deviation of the mean $\sigma_{\overline{X}}$, or the **standard error in the mean**, often abbreviated as SEM. It's important to note that the SEM is *not* the same as the standard deviation of the original distribution. According to the statement above, an estimate of the SEM is $\hat{\sigma}_{\overline{X}} = \sigma / \sqrt{n}$.

We can verify that this statement holds in the case of sampling a mean in Figure 5 by computing the means and SEM from the simulations and comparing with the expected values of μ (population mean) and $\sigma_{\overline{X}} = \sigma / \sqrt{n}$.

The Swain versus Alabama jury selection example demonstrates the CLT applied to a total $T_0 = \sum_{i=1}^{n} X_i$, where $X_i = 1$ indicates a Black member of the population was selected, and $X_i = 0$ indicates non-Black. The distribution is a Bernoulli distribution with population mean $\mu = p = 0.26$, the probability of picking a Black person. We can see from Figure 3 that the mean of the total is $n\mu = 100 \times 0.26 = 26$, and the variance is approximately $\sigma^2_{T_0} \approx n\sigma^2 = 19.24$, as expected for a Bernoulli distribution, giving a standard deviation of 4.38. Furthermore, the distribution is approximately normal.

The law of large numbers Here is an informal statement of the **law of large numbers**:

> In the limit of infinite n, the expected value of the sample mean $\overline{X}$ tends to the population mean μ and the variance of the sample mean $\overline{X}$ tends to 0.

Note that sometimes the law of large numbers is referred to as the "law of averages". This can lead to confusion. The law of averages is sometimes called the "Gambler's fallacy", i.e. the idea that after a run of bad luck, the chance of good luck increases. If the events that are being gambled on are independent of each other (e.g. successive tosses of the same coin), the probability of a head will be the same regardless of how many tails have preceded it.

In the second row of Figure 5 we can see that the distribution for $n = 10000$ is narrower than the distribution for $n = 1000$, and that the sample means converge on the population means. The law of large numbers says that we could, in principle, continue this process by choosing an n as large as we would like to make the variance as small as desired.

Formal statement of the central limit theorem (*Modern Mathematical Statistics with Applications* 6.2) Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit $n \to \infty$ the standardised mean $((\overline{X} - \mu)/(\sigma/\sqrt{n}))$ and standardised total $((T_0 - n\mu)/(\sqrt{n}\sigma))$ have a normal distribution. That is

$$\lim_{n\to\infty} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P\left(\frac{\overline{T_0} - n\mu}{\sqrt{n}\sigma} \leq z\right) = P(Z \leq z) = \Phi(z)$$

where $\Phi(z)$ is the cumulative distribution function (cdf) of a normal distribution with mean 0 and s.d. 1. Thus, when n is sufficiently large, $\overline{X}$ has an approximately normal distribution with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \sigma^2/n$ and the distribution of T_0 is approximately normal with mean $\mu_{T_0} = n\mu$ and variance $\sigma^2_{T_0} = n\sigma^2$. We can also say that the standardised versions of $\overline{X}$ and $\overline{T_0}$ are **asymptotically normal**.

Formal statement of the (weak) law of large numbers (*Modern Mathematical Statistics with Applications* 6.2)

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. As the number of observations n increases, the expected value of the sample mean remains $E[\overline{X}] = \mu$, but the expected variance $V[\overline{X}] = E[(\overline{X} - \mu)^2] \to 0$. We say that "$\overline{X}$ converges in mean square to μ".

More formally, the probability that the difference between the sample mean and population mean is greater than an arbitrary value ε is

$$P(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

for any value ε. Thus, as $n \to \infty$, the probability approaches 0, regardless of the value of ε.

A proof of this statement relies on **Chebyshev's inequality**, and can be found in *Modern Mathematical Statistics with Applications* 6.2.

Note that this is the statement of the weak law of large numbers. There is also a strong law, which has somewhat more stringent requirements on convergence. All distributions that obey the strong law also obey the weak law, but some distributions only obey the weak law and some obey neither law. A discussion of this topic is beyond the scope of this course; the distributions that do not obey the distribution tend to be "weird", e.g. having infinite variance.

Appendix

Frequentist versus Bayesian statistics You may have heard of the difference between Frequentist and Bayesian statistics. The two systems have different philosophical bases, but, in simpler cases, often end with similar results. Roughly speaking, the differences between the two are:

Frequentist The population is a fundamental concept. There is just one possible value of the population mean and variance, i.e. the one that exists in the population. In estimation, we are trying to estimate these quantities, and in hypothesis testing, we are trying to compare our sample with this population.

Bayesian A fundamental concept is the model of the likelihood of the data given parameters (such as the mean). The parameters themselves are uncertain. Conceptually, the population itself is generated from the model, so a number of combinations of parameters and luck may have generated the particular value of (say) the mean observed in a population. Before we have seen any data, we have an initial idea about the distribution of the parameters (the prior). The inference process involves using the data to update this prior distribution to give a distribution of the parameters given the data.

For around a century, there has been controversy about which approach is best. Broadly speaking, we will be using Frequentist approaches in this course. At the level we are working at here, it will give very similar results to Bayesian approaches. The important thing is to understand the meaning and interpretation of our inference.

References

Gelman, A., Carlin, J. B. et al. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, second ed.

Kohavi, R., Henne, R. M. et al. (2007). 'Practical guide to controlled experiments on the web: Listen to your customers not to the HiPPO'. In P. Berkhin, R. Caruana, X. Wu and S. Gaffney, eds., *KDD-2007 Proceedings of the thirteenth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 959–967. Association of Computing Machinery, New York, USA