Part III: Statistical methods

Chapter 10
Introduction to sampling and sampling distributions

We look at samples of data in order to learn about something, usually about something more than the sample itself. Typically we are hoping to find out about a set with many members, such that it is impossible to look at every member of the set. For example, conservation officials interested in the question of whether a license should be required to fish on a certain lake might investigate whether fishing is reducing the average size of fish by taking away relatively more of the mature fish (perhaps of a specific species). They might catch, weigh and return to the water a certain number of fish, and repeat the exercise every year for some years. (In order to judge whether any differences they see over time are genuine, as opposed to being simply the result of random differences in which fish they catch, they would want to apply methods for statistical inference described in chapters below.) In this case, it is clear that they are intending to learn about the entire population of fish in the lake from their sample of fish. A pollster who asks a sample of people how they intend to vote in the coming election is hoping to be able to predict the outcome of the election, which requires learning about the voting intentions of those who will actually vote. Note that the intentions of people who are not going to vote are irrelevant to the outcome, and so are not of concern to the pollster: the population of interest as those who are actually going to cast a vote, so that even in this case there is some subtlety to the question of what the relevant population is.

Sometimes it is less clear what population we can learn about. For example, we might survey a group of workers in a particular company in Toronto to determine whether those who undertook a training program benefited through relatively higher wages, promotions, and so on. But who are we learning about? If the results apply only to these workers and to this company, then they might not be of much interest to anyone else, and we might even be able to speak with every employee at the company if it is small enough, so that sampling would not be necessary. Would the results apply to any worker, anywhere, who undertakes training? This seems unlikely, given the diversity of the workforce and the conditions of work around the world. We might conclude, however, that the results could provide a good indicator of the likely benefits of a particular type of training program for North American workers in a certain kind of industry (so that this is the population being studied), for as long as certain general conditions remain in place. In any event, determining what population we are learning about requires some thought.

Once we are clear about what the population being studied is, we want to know how the quantities to be obtained from the sample are related to the true characteristics of the population that are of interest to us. They will not of course be identical in general, but we hope that they will be close and will tend to get closer as we take larger and larger samples. The purpose of this and the following chapters is to characterize what is known about the relation between sample quantities and population quantities: that is, what is the distribution of a sample quantity relative to a population ('true') quantity?

10.1 Sample and population

D10.1 Sample: A sample is a subset of a population that can be observed by an investigator.

The aim in sampling is to obtain a sample which is representative of the population. For example, we might be interested in how the vote will go in the upcoming (at the time of writing) referendum on Scottish independence. If we sample the population in Scotland by setting up a booth on campus at the University of Edinburgh, then almost everyone we asked will either be a university student or have a degree, will have above average (expected lifetime) income, and so on. We will not be learning the views of the poor, chronically unemployed, rural or elderly voters. Unless university students and staff happen by coincidence to have the same distribution of views as the general population, we will get a misleading view of overall voting intentions.

Normally we would prefer a 'random' sample.

D10.2 Simple random sample: A simple random sample is a sample from a population such that every member of the population is equally likely to be chosen for the sample, and successive observations in the sample are independent.

Note that this definition of 'random' is somewhat different from what might be used in other contexts in statistics; for example a random stochastic process is one which is not fully predictable, but may have some predictable part.

In some cases, it is difficult to achieve the goal of a random sample, because some members of the population are less likely to be observed than others. Researchers may therefore sometimes use a 'stratified sample'. A stratified sample is one in which the population is divided into mutually exclusive and exhaustive classes, and the final sample is designed to have the same proportion of each class as does the population are it simple random sampling may be used within each class, with the goal of obtaining an overall sample which is representative of the population. For example, we may have 8% of a particular population which is elderly (let's say, 70 or over). If we try to sample randomly from the population, however, we may find that we are getting in touch with elderly people less often, either because they are less likely to answer the phone, or come to the door, or be contacted by whatever other means we are using; perhaps we would only end up with 3% of our sample being elderly people; if their behavior patterns are different in a way that is relevant to what we are investigating, our results could therefore be misleading. We might therefore continue sampling only elderly people until we have enough to make up 8% of the overall sample. The goal remains to obtain a sample which is representative of the entire population.

When we have a sample, we will want to use it to learn about some characteristic of the population. Often, we will start by estimating the mean of some characteristic in the population, for example, the mean weight of a fish in the lake. But we know that the mean weight in our sample will not, except by outrageous coincidence, be the same to the nearest gram as in the population. So what does that mean in the sample tell us about the population mean? We can hope that they are close, but that is not very useful. In order to

answer the question well, we would like to be able to characterize the entire distribution of the sample mean, given some population mean and size of sample. If the mean weight of a trout in the lake is 746 grams, and if we catch, measure and release a sample of 100 trout, what is the distribution of possible sample means that we could find? We can answer this question, at least approximately (and with an approximation error that declines to zero as sample size increases) in a very wide variety of cases.

10.2 Sampling and distributions of samples

We can begin with a simple example that we have seen earlier, in which we can compute the exact distribution of the sample mean, to help us understand what we are trying to obtain and how to interpret it.

Consider a simple game played by two people. **A** flips a fair coin (the probability of a head = the probability a tail = 0.5) and pays $1 to **B** if the coin comes up heads, and receives $1 from **B** if the coin comes up tails. Clearly, each person is an asymmetric position, and has the same probability of being a winner, loser, or breaking even after playing N times. The population mean payoff to each player is -1(0.5) + 1(0.5) = 0, regardless of the number of times a game is played.

The sample mean – the average of what is won or lost – may of course differ. If they play three times, **A**'s possible outcomes are $\{-3, -1, 1, 3\}$ and the sample mean outcomes are $\{-1, -1/3, 1/3, 1\}$, and the same is of course true for **B**. These outcomes are not equally likely, of course, and we have seen that the probabilities can be computed in various ways. The probabilities of the four outcomes are 1/8, 3/8, 3/8, 1/8. Notice that because the number of rounds is odd, it's actually impossible to break even exactly, so it's impossible for the sample mean to equal the population mean in this case. Nonetheless, although zero is not a possible outcome, the mean of the sampling distribution is zero, just as the mean of the population distribution is zero.

This is the sampling distribution of the mean payoff after playing the game three times, and it fully describes what outcomes could emerge for the mean and what their probabilities are, given the conditions of the game.

If we were to repeat this exercise for a game of, say, 10 rounds, the distribution would be different although still centered on zero. With 10 rounds, the probabilities would be more heavily clustered near zero, and we could compute them exactly using the binomial distribution.[1] As we have seen in earlier chapters, with the distribution we can answer questions such as this: if they play the game 10 times, what is the probability that **A** has a mean loss of greater than 0.25, *i.e.* a total loss of more than $2.50? (It's the

[1] The possible outcomes are $\{-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10\}$ with corresponding means $\{-1, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 1\}$. The probabilities are

$$\left(\frac{1}{1024}\right)\{1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1\}.$$

sum of the first four probabilities, or 176/1024.) The sampling distribution allows us to make statements about the probability of the sample mean lying in different regions, given the population mean. Conversely, given an observed sample mean, it allows us to make statements about where the true population mean is likely to lie, in the more usual case where the population mean is not known.

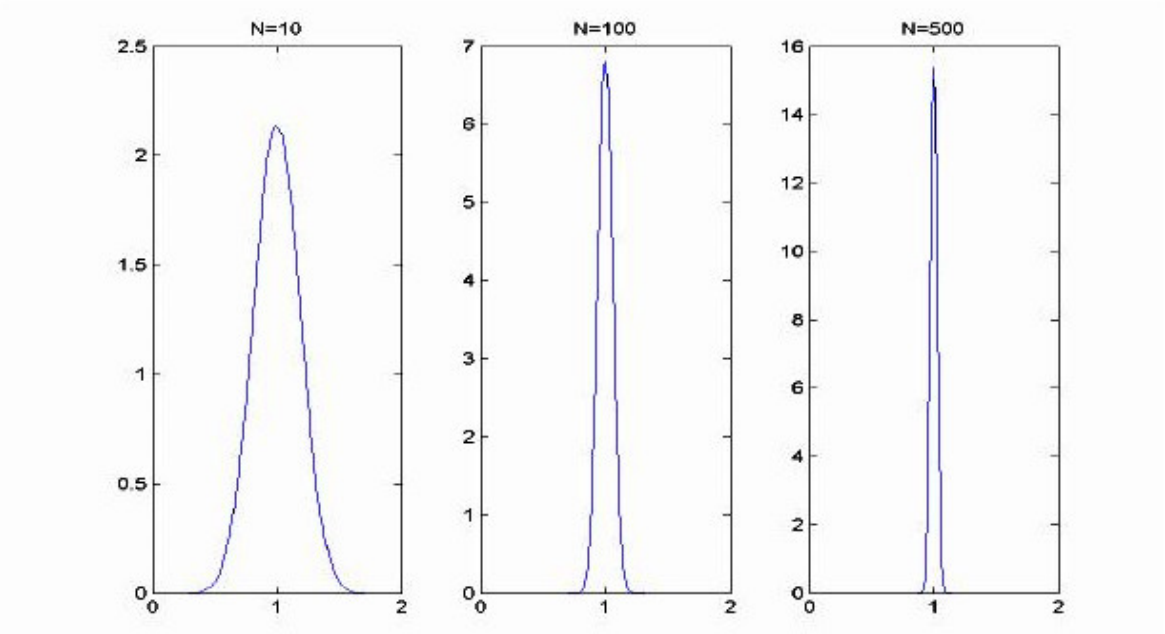
Now let's do a much larger exercise, using a computer simulation. We use a computer random-number generation algorithm to generate pseudo-random variables from either the Uniform[0,2] distribution or from the Chi-squared distribution with 1 degree of freedom. Both of these distributions have a mean of 1, so the population mean in both of these experiments is 1.

In each case we take samples of size N from the distribution, and take the mean of each sample. We do this 100,000 times for each sample size, so that we have many examples of sample means, and then we can actually estimate the density that applies to the sample mean. We do that using a kernel density estimator (which is at present not described in this book, but can be thought of for now as a development of the idea of the histogram, producing a smooth curve instead of a set of bars).[2] There are three sample sizes, so that we can observe something about the way in which the sampling distribution changes as the sample size changes.

Notice that the vertical scales are different: all of these density functions integrate to 1, so that as they become thinner they must become taller as well: that is, they become more tightly concentrated around the population mean.

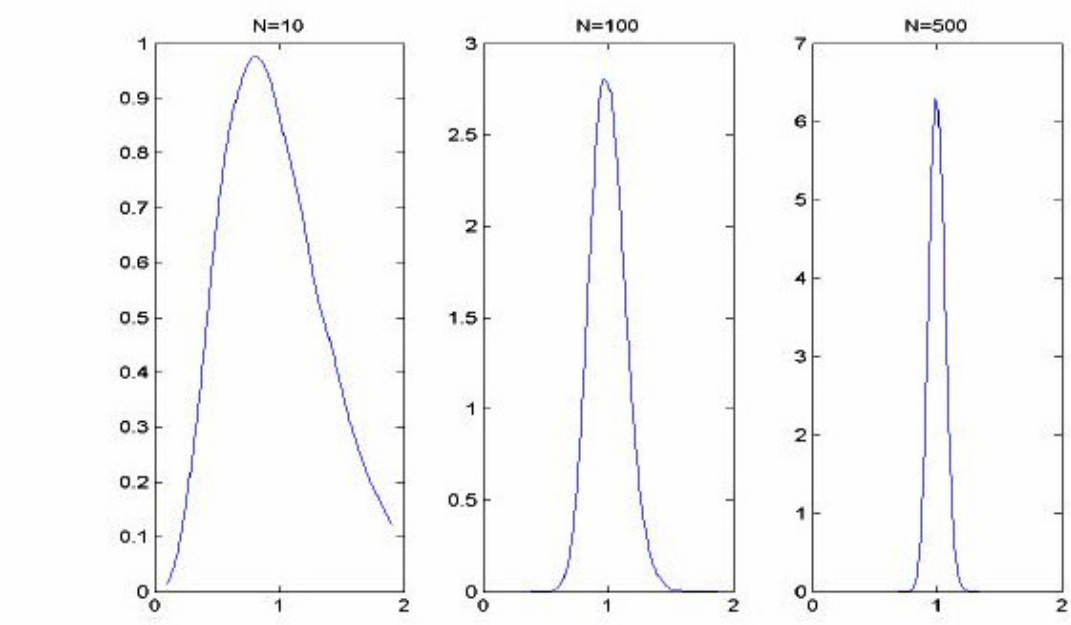
FIGURE 10.2.1

Empirical distributions of sample mean: $U[0,2]$ random variables, N=10, 100, 500



[2] See Silverman (1986) for an exposition.

FIGURE 10.2.2
Empirical distributions of sample mean:
χ^2_1 random variables, N=10, 100, 500



The top panel shows results from Uniform random variables. The Uniform distribution is symmetric, and the sampling distributions are apparently symmetric as well. In the case of the input data which are χ^2_1 , shown in the bottom panel, the sampling distribution for $N = 10$ is noticeably skewed; in fact if one looks very closely (compare the heights of the density function around 0.5 and 1.5), a tiny degree of skewing is visible at $N = 100$ as well. In the largest sample size, no skewing is visible to the naked eye.

These results suggest that the distribution of the sample mean tends to become ever more highly concentrated around the true value as the number of sample points increases, and also that the distribution of the sample values around the true value tends toward a single peaked (unimodal) and symmetric distribution as the sample size increases. Both of these results are borne out by theory, as we shall see later in Chapter 11

10.3 A simple, if unrealistic, case

A simple case in which we can work out the exact distribution of the sample mean is that in which the data actually come from a Normal distribution. Typically, however, we will observe some feature of the data that makes it impossible that the data could truly be Normal. For example, the data may be bounded on one or both sides (the unemployment rate, or the proportion of survey respondents who say they'll vote for a particular party, cannot go below zero or above 100%). Alternatively, a simple plot of the histogram of the data set may show substantial skewness. Nonetheless its useful to start by learning about this case, for several reasons:

– The sampling distribution that emerges in more realistic cases, where the data distribution is unknown, will turn out to be approximately the same as the distribution

that results in this case.

- The Normal-data case will help us to understand the reasons for the use of the t-distribution in some problems.
- We will gain some understanding, through this and results in the chapter covering Central Limit Theorems, of the distinction between exact finite-sample results and asymptotic results.
- The Normal-data case has a direct application in some circumstances, particularly in computer simulations where the input data are created to have a particular distribution.

In order to obtain the sampling distribution of the mean from a Normal population of data, we need the following result.

Theorem 10.1: Linear combinations of Normal random variables are Normal. Let $z_1, z_2, \ldots, z_N$ be independent Normal variables each of which has expectation 0 and variance σ_i^2. Then the linear combination $a_1 z_1 + a_2 z_2 + \ldots + a_N z_N$ has the distribution $\mathrm{N}(0, \sum_{i=1}^N a_i^2 \sigma_i^2)$.

Proof: See Kendall *et al.* (1991), example 11.2.

(If the expectations of the random variables are non-zero, then the expectation of the Normal distribution applying to the linear combination is simply the weighted sum of the expectations, $\sum_{i=1}^N a_i \mu_i$.)

To apply this result to obtain the distribution of the sample mean, note that the sample mean is a linear combination of the sample data, $\overline{X}_N = \sum_{i=1}^N x_i/N$, with the weight on each data point being the constant $1/N$.

Here we are treating the case in which the data are independent samples from an $\mathrm{N}(\mu, \sigma^2)$ distribution. The expectation of the sample mean is

$$\mathrm{E}\Big(\sum_{i=1}^N \frac{1}{N} x_i\Big) = \frac{1}{N} \sum_{i=1}^N \mathrm{E}(x_i) = \frac{1}{N}(N\mu) = \mu.$$

So the expectation of $\overline{X}_N$ is the same as the expectation of the sample data that are being averaged, the x_i's. This is not true for the variance; the variance of the sample mean in this independent sampling case is smaller than the variance of the data: using our earlier results on the variance of a linear combination,

$$\mathrm{Var}\Big(\sum_{i=1}^N \frac{1}{N} x_i\Big) = \mathrm{Var}\Big(\frac{1}{N} x_1 + \frac{1}{N} x_2 + \ldots + \frac{1}{N} x_N\Big)$$

$$= \Big(\sum_{i=1}^N \frac{1}{N^2} \mathrm{Var}(x_i)\Big) = \frac{1}{N^2}\Big(\sum_{i=1}^N \sigma^2\Big) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

There are several important points to note and remember about this.

- The variance declines with sample size. That is, as we get more sample points our estimator has less dispersion, and we have a better and better idea of where the true value lies. This is reflected in the graphs above, where we see the densities becoming more tightly concentrated around the true value as the sample size increases.
- The computation of the variance is very straightforward in this case because there are no co-variance terms: we have assumed that we have an independent sample. If the data were correlated, additional terms appear in the computation of the variance, and it would be larger than σ^2/N; however, as long as the correlation between subsequent observations is not perfect, the variance of the sample mean will still decline as sample information accumulates.
- Putting together the expectation and variance of the distribution of the sample mean with the fact from the theorem that it must have a normal form, we obtain the result in this case that $\overline{X}_N \sim \mathrm{N}(\mu, \sigma^2/N)$.
- We can standardize the sample mean to obtain a distribution which does not change with the sample size: subtracting the mean and dividing by the square root of the variance, we find $(\overline{X}_N - \mu)/(\sigma/\sqrt{N}) \sim \mathrm{N}(0,1)$, the standard Normal distribution.
- Multiplying numerator and denominator by the square root of sample size N in the distribution just given, the result may be rewritten as $\sqrt{N}\big((\overline{X}_N - \mu)/\sigma\big) \sim \mathrm{N}(0,1)$. This implies that scaling up the discrepancy in the estimate of the expectation by the square root of the sample size leads to a fixed, non-degenerate distribution. It follows therefore that the discrepancy itself is declining at the rate of the square root of sample size. This is an example of 'square root-N' convergence, which appears in many standard parametric problems.

So the discrepancy between the sample (estimated) and population ('true') means, divided by the standard deviation of the sample mean (we might say: the discrepancy 'measured in standard deviations') has a standard Normal distribution. Note that we write standard deviation rather than standard error, because we are referring to the population value, σ.

This is what is sometimes called an 'infeasible' or 'non-operational' statistic. Not everything on the left-hand side of the expression is observable: we dont know σ. Because we usually don't have this value, we can't actually compute this statistic.

In practice, we have to replace σ with s, that is, we replace the standard deviation with the standard error (or sample standard deviation) of the data. Does this change the sampling distribution?

Given the sampling conditions assumed, s will converge probabilistically to σ in a sense that we will define precisely in the next chapter. So in large samples,

$$\frac{\overline{X}_N - \mu}{SE(\overline{X}_N)} \text{ or } \frac{\overline{X}_N - \mu}{s/\sqrt{N}} \text{ should be very close to } \frac{\overline{X}_N - \mu}{SD(\overline{X}_N)} \text{ or } \frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}},$$

and so the former should have a distribution close to N(0,1). This turns out to be true.

But in fact, for this case, the exact distribution that applies for any given sample size N (not just the asymptotic result) has been worked out, and we do not need to use an approximation.[3]

Theorem 10.2: t-distribution. Let Z be a random variable with the standard normal distribution $\big((Z \sim \mathrm{N}(0,1)\big)$ and let W the random variable with the Chisquared distribution with r degrees of freedom ($W \sim \chi^2_r$). Then if Z and W are independent, the ratio

$$\frac{Z}{\sqrt{W/r}}$$

has the Student's t-distribution with r degrees of freedom.

Proof: See Mood *et al.* (1974), section 4.5.

We will see below that in this case of independent random sampling from Normal data, the sample variance s^2 in fact has a χ^2_{N-1} distribution, so that the feasible statistic $(\overline{X}_N - \mu)/(s/\sqrt{N})$ will be distributed as t_{N-1}. Note again that this result, which gives an exact distribution applicable to any particular sample size, has been obtained under the generally unrealistic assumption that the data that we are sampling themselves have a Normal distribution. In more general circumstances where we do not know this, we will have to rely on an asymptotic approximation to get the distribution of this feasible statistic, as described in the next chapter.

10.4 Using a sampling distribution

Consider then that we have a sample of size N from a population with expectation μ and variance σ^2. What can we deduce from this?

If we take a given population mean, we can answer questions about where the sample mean is likely to be – what is the probability that it will lie in a certain interval, for example, or the probability that it will lie more than a certain distance away from the population mean. If we determine how tightly concentrated the distribution of the sample mean is around the true expectation for a given sample size, it will be useful in determining what sample size we need to use to get a given degree of precision. In practical sampling problems where we have a sample already, we are interested in the converse: given our estimate of the mean, $\overline{X}_N$, what is the probability that the true expectation lies in a certain interval?

To answer these questions, let's manipulate the expressions above, working with the feasible or operational form of the statistic:

$$\frac{\overline{X}_N - \mu}{s/\sqrt{N}} \sim t_{N-1}.$$

[3] In 1908, by William S. Gosset, 1876-1937. Because Gossett used the pseudonym Student, the distribution is often called Student's t-distribution.

Since the t-distribution with N 1 degrees of freedom has a known form, and the quantiles and so on have been tabulated, we can compute an interval so that the expression on the left-hand side above has a given probability of lying in that interval. Using the notation q_α for the 1-α-quantile of the relevant distribution (t_{N-1}), we can define the interval such that

$$P\Big(-q_{\alpha/2} < \frac{\overline{X}_N - \mu}{s/\sqrt{N}} < q_{\alpha/2}\Big) = 1 - \alpha, \tag{10.a}$$

where we have used $\alpha/2$ in each case so that we have a probability $\alpha/2$ of the actual outcome lying outside this interval both above and below the interval, adding up to a total probability of α outside the interval (1-α inside the interval). Often, α is taken to be 5% (0.05), so that the interval spans the interior 95% of the distribution, leaving 2.5% on both the left and right tails.

When working with the Normal distribution, either to describe the infeasible case or in using the approximation from asymptotic theory that we will learn in the next chapter, it is common to use the notation $z_{\alpha/2}$ to describe the corresponding quantiles from the Normal. This notation is also sometimes used for the t-distribution.

Let us now manipulate this expression further. The expression (10.a) above contains $\overline{X}_N - \mu$ in the middle: if we know one of these, we will be able to obtain a statement about the other.

If we perform the same operation on each of the quantities in parentheses, we will not change the probability: so multiplying through by the denominator of the expression in the middle, we can obtain

$$P\Big(-q_{\alpha/2}(s/\sqrt{N}) < \overline{X}_N - \mu < q_{\alpha/2}(s/\sqrt{N})\Big) = 1 - \alpha.$$

If we now subtract $\overline{X}_N$, from each of the three terms, we obtain a statement purely about μ:

$$P\big(-\overline{X}_N - q_{\alpha/2}(s/\sqrt{N}) < -\mu < -\overline{X}_N + q_{\alpha/2}(s/\sqrt{N})\Big) = 1 - \alpha,$$

and then if we multiply through by -1 (the inequality signs must then be reversed: for example $5 > 4 > 3$ implies that $-5 < -4 < -3$),

$$P\Big(\overline{X}_N + q_{\alpha/2}(s/\sqrt{N}) > \mu > \overline{X}_N - q_{\alpha/2}(s/\sqrt{N})\Big) = 1 - \alpha.$$

This expression says that the population mean μ lies in the interval $\overline{X}_N \pm q_{\alpha/2}(s/\sqrt{N})$ with probability α. So we have succeeded in obtaining a probability statement about where the population mean lies, although we only observe the sample. Notice that as N gets larger, this interval gets narrower: more information produces a more precise statement.

For the t-distribution with a large number of degrees of freedom N-1 (or for the standard Normal distribution), $q_{\alpha/2}$ (or $z_{\alpha/2}$ in the notation commonly used for the standard Normal) is approximately 1.96: that is, about 2.5% of the distribution lies below -1.96, and

about 2.5% of the distribution lies above 1.96.[4] So the probability interval just stated is what lies behind the commonly-remembered result that there is a 95% probability that the population mean of something will lie within about ± two standard errors of the sample mean.

To take a numerical example, consider the population of fish in the lake mentioned earlier. We catch 100 fish, and find an average weight of 746 g, with a standard error of 205 g. As usual in real data, a moment's reflection tells us that these data could not literally be Normal: weight cannot be negative, so the distribution is bounded below, unlike the Normal. As we said, knowing that the data are Normal is generally unrealistic. Let's go on with this example anyway, because it will turn out below that the results that we have just stated will turn out to be a good approximation in a wide range of circumstances even though the data are not Normal.

So using the intervals given above, and using $q_{0.025} = 1.96$, we have

$$P\Big(746 + 1.96(205/\sqrt{100}) > \mu > 746 - 1.96(205/\sqrt{100})\Big) = 0.95$$

or since $1.96(20.5) = 40.18$,

$$P(786.18 > \mu > 705.82) = 0.95.$$

So, given the conditions assumed to hold in this sampling experiment, we can be 95% sure that the mean weight of the fish in the lake is between about 706 g and 787 g (rounding to three significant digits).

10.5 Simple case continued: distribution of the sample variance

The sample variance has been defined as follows:

$$s^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \overline{X}_N)^2,$$

where $\overline{X}_N = N^{-1} \sum_{i-1}^{N} x_i$. We can prove the following Theorem.

Theorem 10.3 Suppose that the data are normally distributed, the observations being mutually independent and normally distributed, each with expectation μ and variance σ^2. Then $(N-1)s^2/\sigma^2$ has a χ^2 distribution with N – 1 degrees of freedom.

[4] The value of $q_{\alpha/2}$ can be obtained from tables for particular values of the degrees of freedom, or from a computer program that computes the inverse of the cumulative distribution function: that is, given a value of the CDF such as 0.99, a program will calculate the corresponding quantile which gives $\mathrm{CDF}(q_{\alpha/2}) = 0.99$.

Proof: Let $w_i = (x_i - \mu)/\sigma,\ i = 1, \ldots, N$. Clearly each w_i has the standard Normal distribution, and they are mutually independent. If we define $\overline{W}_N$ as the average of these centered and rescaled random variables, then

$$\sum_{i=1}^{N}(w_i - \overline{W}_N)^2 = \sum_{i=1}^{N}\frac{1}{\sigma^2}\big(x_i - \mu - N^{-1}\sum_{j=1}^{N}(x_j - \mu)\big)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{N}\big(x_i - N^{-1}\sum_{j=1}^{N}x_j\big)^2$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \overline{X}_N)^2 = (N-1)s^2/\sigma^2.$$

Consider a set of coefficients $a_{ik},\ i, k = 1, \ldots, N$, that satisfy the relations

$$\sum_{k=1}^{N} a_{ik}a_{jk} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad i, j = 1, \ldots, N.$$

Let $y_i = \sum_{k=1}^{N} a_{ik}w_k,\ i = 1, \ldots, N$. By Theorem 10.1, linear combinations of Normal random variables are Normal, The expectation of y_i is obviously zero, and its variance is $\sum_{k=1}^{N} a_{ik}^2 = 1$. It follows that y_i has the standard Normal distribution for each $i = 1, \ldots, N$. If $i \neq j$, the covariance of y_i and y_j is

$$\mathrm{cov}(y_i, y_j) = \mathrm{E}(y_i y_j) = \sum_{k=1}^{N}\sum_{l=1}^{N} a_{ik}a_{jl}\mathrm{E}(w_k w_l).$$

But the covariance of the independent random variables w_k and w_l is zero if $k \neq l$, and so only the terms in the double sum above for which $k = l$ are nonzero. Therefore, if $i \neq j$,

$$\mathrm{cov}(y_i, y_j) = \sum_{k=1}^{N} a_{ik}a_{jk}\mathrm{E}(w_k^2) = \sum_{k=1}^{N} a_{ik}a_{jk} = 0.$$

Now $y_i^2 = \sum_{k=1}^{N}\sum_{l=1}^{N} a_{ik}a_{il}w_k w_l$. Consequently,

$$\sum_{i=1}^{N} y_i^2 = \sum_{k=1}^{N}\sum_{l=1}^{N} w_k w_l \sum_{i=1}^{N} a_{ik}a_{il} = \sum_{k=1}^{N} w_k^2,$$

since $\sum_{i=1}^{N} a_{ik}a_{il} = 1$ only if $k = l$, and is zero otherwise. Next, let $a_{Nk} = N^{-1/2},\ k = 1, \ldots, N$. We see immediately that $\sum_{k=1}^{N} a_{Nk}a_{Nk} = 1$, as required. Then $y_N \equiv \sum_{k=1}^{N} a_{Nk}w_k = N^{-1/2}\sum_{k=1}^{N} w_k = N^{1/2}\overline{W}_N$.

Observe next that

$$\sum_{i=1}^{N}(w_i - \overline{W}_N)^2 = \sum_{i=1}^{N}\big(w_i^2 - 2\overline{W}_N w_i + \overline{W}_N^2\big) = \sum_{i=1}^{N} w_i^2 - 2N\overline{W}_N^2 + N\overline{W}_N^2 = \sum_{i=1}^{N} w_i^2 - N\overline{W}_N^2.$$

From what we have just seen, the last expression here is $\sum_{i=1}^{N} y_i^2 - y_N^2 = \sum_{i=1}^{N-1} y_i^2$. Consequently, $(N-1)s^2/\sigma^2$ is the sum of $N-1$ squared independent standard Normal random variables, and so it has a χ^2 distribution with N – 1 degrees of freedom. This completes the proof.

It is a legitimate question to ask where the coefficients a_{ik} come from. In the proof, we gave an explicit definition only of a_{Nk}, $k = 1, \ldots N$. For the other coefficients, there is no unique definition, but here is one set of coefficients that satisfies the requirements. For $i = 1, \ldots, N-1$,

$$a_{ik} = \begin{cases} 0 & \text{for } k < i \\ [(N-k)/(N-k+1)]^{1/2} & \text{for } k = i \\ -[(N-k)(N-k+1)]^{1/2} & \text{for } k > i \end{cases}$$

The proof that these coefficients do indeed satisfy the requirements is left as a (tedious) exercise.

We have said several times that this case is unrealistic, because it assumes that the data are Normal and that they are known to be Normal. In a typical problem, we do not know the distribution from which the data come. How then will we find the distribution that results when we perform some operation such as taking the sample mean or variance, when we don't even know the distribution of the input data?

Perhaps surprisingly, it is possible to answer questions under these circumstances, using an invariance principle. An invariance principle states that, for any (input) distribution that has certain characteristics, performing some operation on the data will tend to produce, as sample size grows, a particular (output) distribution. The Central Limit Theorem, which we will discuss next, is an example of an invariance principle and states that the distribution of the standardized sample mean of data that have a few simple characteristics will converge toward the standard Normal distribution. With this result, it is not necessary to make unfounded assumptions about the nature of the data that we are analyzing.