Chapter 5

Sampling Distributions of Statistics

Chapter Objectives

- Explain what the sampling distribution of a statistic is.
- Identify the mean and standard error of the sampling distribution of the sample mean.
- State the two situations in which the sample mean will follow (at least approximately) a normal distribution.
- Obtain probabilities from the sampling distribution of the sample mean.
- Obtain percentiles from the sampling distribution of the sample mean.
- Identify the mean and standard error of the sampling distribution of the sample proportion.
- State the situation in which the sample proportion will follow (at least approximately) a normal distribution.

Key Takeaways

- The value of a statistic will vary from one random sample to the next.
- The sampling error of a statistic is the difference between its observed value and the corresponding population parameter value.
- We call the probability distribution of a statistic its sampling distribution.
- The standard error of a statistic is the standard deviation of the statistic's sampling distribution, and represents a typical amount by which the value of the statistic will differ from the corresponding population parameter value.
- The mean and standard error of the sampling distribution of the sample mean are determined by the mean and standard deviation of the population as well as the size of the sample.
- The sampling distribution of the sample mean will be (at least approximately) normal if either the sample is from a normal population or the sample size is large.
- The mean and standard error of the sampling distribution of the sample proportion are determined by the population proportion as well as the size of the sample.
- The sampling distribution of the sample proportion will be (approximately) normal if the sample size is large.

5.1 Introduction

Recall that a ***statistic*** is a numerical quantity calculated from random sample data. Examples include the sample mean, sample median, and sample standard deviation defined in Chapter 3. Because a statistic is computed from a *random* sample, its value is determined by chance. Statistics, therefore, are random variables. We call the random sample-to-sample variation in the value of a statistic ***sampling variation***,

and if the statistic is to be used to draw inferences about the population, sampling variation leads to uncertainty that should be acknowledged, or better yet, quantified.

To quantify uncertainty when a statistic is used to make inferences about a population, we'll need to identify the ***sampling distribution*** of that statistic, which is just another name for its probability distribution. The sampling distribution characterizes sampling variation in the statistic by specifying the values the statistic can take and the probabilities with which it takes those values.

The most important statistic for making inferences about a population is the sample mean $\bar{\boldsymbol{X}}$, so in this chapter we'll focus mainly on the sampling distribution of the sample mean. We'll assume that we have observations (measurements) $\boldsymbol{X_1, X_2, \ldots X_n}$ made on a random sample from a population whose mean and standard deviation are $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. We use capital letters, $\bar{X}$ and $X_1, X_2, \ldots X_n$, as a reminder that we're thinking of these values as random variables.

5.2 The Sampling Distribution of $\bar{X}$

5.2.1 Introduction

The sampling distribution of the sample mean $\bar{X}$ describes the sampling variation in the values that $\bar{X}$ takes. Its shape, center, and spread will depend on the shape, center (μ), and spread (σ) of the population from which the sample is drawn, and also on the sample size $\boldsymbol{n}$. Fig. 5.1 shows the sample means of 10 random samples, each of size $n = 4$, from a normally distributed population. We can think of each sample mean as the observed value of a random variable whose probability distribution, the sampling distribution of $\bar{X}$, is the one shown in the lower half of the figure.

Throughout this section, it's assumed that the observations $X_1, X_2, \ldots X_n$ from which $\bar{X}$ is computed are *independent* of each other. Recall from Section 2.2.2 of Chapter 2 that observations are ***independent*** if their values are unrelated and, therefore, not influenced by each other. This assumption is usually reasonable as long as the observations aren't made too close together in time or space (which, recall from Chapter 2, would lead to *pseudoreplication*).

5.2.2 Mean and Standard Error of the Sampling Distribution of $\bar{X}$

Notice in Fig. 5.1 that the sampling distribution of $\bar{X}$ is centered on μ, the mean of the population, but is less spread out than the population. The following fact tells us that this will always be the case.

Fact 5.1 Suppose $X_1, X_2, \ldots X_n$ are a random sample from *any* population (discrete or continuous) whose mean is μ and whose standard deviation is σ. Then the sampling distribution of $\bar{X}$ has mean $\boldsymbol{\mu_{\bar{X}}}$ and standard deviation $\boldsymbol{\sigma_{\bar{X}}}$ that are related to μ and σ by

$$\mu_{\bar{X}} = \mu \tag{5.1}$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \tag{5.2}$$

The mean $\mu_{\bar{X}}$ of the $\bar{X}$ distribution is the *long-run average* value that you'd get for $\bar{X}$ if you repeatedly took samples from the population. Thus (5.1) says that, *on average*, the sample mean will equal the population mean, even though a *particular* $\bar{X}$ almost certainly won't equal μ exactly. In other words, when $\bar{X}$ is used as an *estimator* of μ, it neither systematically overestimates nor systematically underestimates μ. On average, it's right on target.

In practice, though, we usually only take one sample of size n from the population, so we only have one value of $\bar{X}$ to serve as our *estimate* of μ. The discrepancy between a *particular* estimate $\bar{X}$ and the true value μ is called the ***sampling error***.

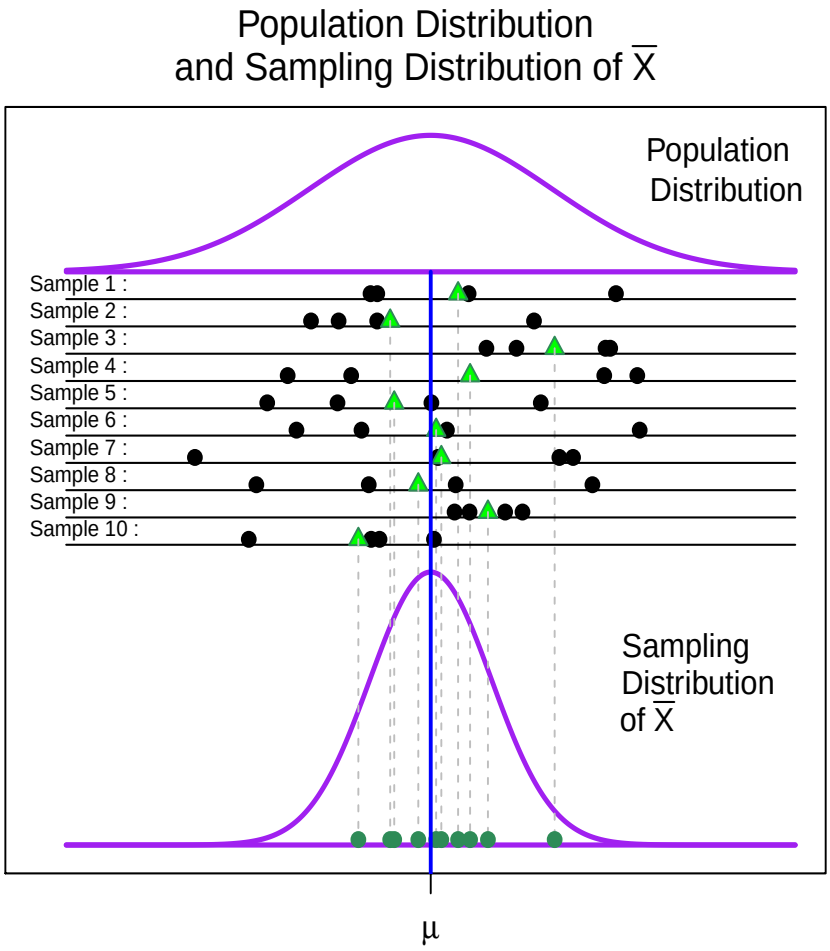


Figure 5.1: Ten samples of size $n = 4$ from a normal distribution representing a population (top). For each sample, the mean $\bar{X}$ is shown as a green triangle. The normal distribution at the bottom is the sampling distribution of $\bar{X}$. The vertical solid blue line represents the mean of both the population and the $\bar{X}$ distribution.

Sampling Error of $\bar{X}$:

$$\text{Sampling Error} \;=\; \bar{X} - \mu.$$

The sampling error indicates how far off the mark a *particular* estimate of the population mean is.

The standard deviation $\sigma_{\bar{X}}$ of the $\bar{X}$ distribution is interpreted as the size of a *typical* sampling error (for a given sample size n). Because of this, $\boldsymbol{\sigma/\sqrt{n}}$ is called the ***standard error*** of $\bar{X}$, and it serves as a measure of how precise $\bar{X}$ is as an estimator of μ. A smaller standard error means a more precise estimator. The standard error will be small if either:

- The population standard deviation σ is small, or
- The sample size n is large.

In Fig. 5.1, since $n = 4$, the standard error of $\bar{X}$ is exactly half as large as the population standard deviation ($\sigma/\sqrt{4} = \sigma/2$), and consequently the $\bar{X}$ distribution doesn't have as much spread as the population distribution. A larger sample size would've resulted in an even smaller. Fig. 5.2 shows that the sampling distribution of $\bar{X}$ becomes more concentrated about μ as n increases.

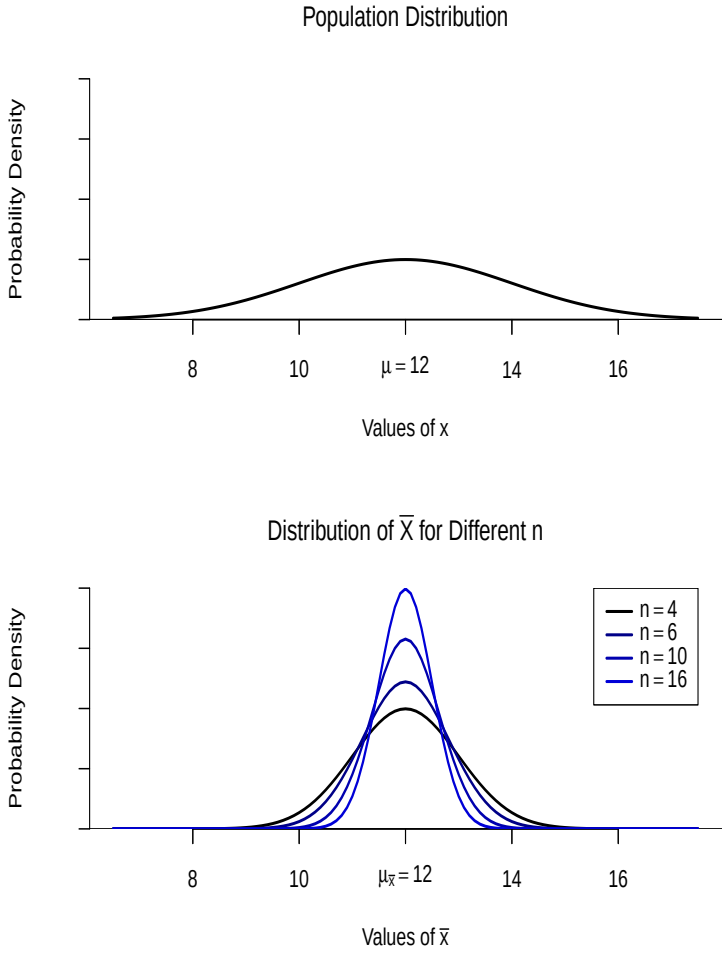


Figure 5.2: Sampling distribution of $\bar{X}$ for various sample sizes (bottom) when the sample is from a normal population (top) whose mean and standard deviation are $\mu = 12$ and $\sigma = 2$.

5.2.3 The Estimated Standard Error and Graphs with Error Bars

It's useful, when graphing sample means, to include in the graph an indication of the degree of uncertainty in each $\bar{X}$ as an estimator of its corresponding population mean μ. One way to do this is to include ***error bars*** in the graph that extend one or two standard errors above and below the sample means. In practice, we usually have to use the ***estimated standard error***, $\boldsymbol{S_{\bar{X}}}$, defined as

$$S_{\bar{X}} \;=\; S/\sqrt{n}\,,$$

where $\boldsymbol{S}$ is the sample standard deviation, because we don't know the population standard deviation σ. The next example illustrates.

Example 5.1: Graphs with Error Bars

Recall that a *bioassay* is a study to determine the toxicity of a substance. Bioassays usually involve exposing biological organisms to different concentrations of the substance and measuring their responses. In one bioassay, giant kelp (*Mactocystis pyrifera*) were exposed to copper (Cu) at five concentrations, and the lengths of their embryonic gametophyte germination tubes were measured [2], [5]. Smaller tube lengths indicate more severe toxic responses. Such damage to kelp is an indicator of potential toxicity to other aquatic life.

Five replicate observations were made at each of the five exposure concentrations. The table below shows the individual tube lengths (mm) along with the sample means, sample standard deviations, and (estimated) standard errors.

Tube Lengths for Five Cu Exposure Concentrations

	0.0 μg/L	5.6 μg/L	10.0 μg/L	18.0 μg/L	32.0 μg/L
	19.58	18.26	13.31	18.59	12.54
	18.75	16.25	18.92	12.88	10.67
	19.14	16.39	15.62	16.28	15.95
	16.50	18.70	14.30	15.38	12.54
	17.93	15.62	15.29	19.75	11.66
$\bar{X}$	18.38	17.04	15.49	16.58	12.67
S	1.21	1.35	2.12	2.71	1.99
$S_{\bar{X}} = S/\sqrt{n}$	0.54	0.60	0.95	1.21	0.89

Below, the sample means are graphed as bar heights, and error bars extend one standard error above and one standard error below the tops of the bars.

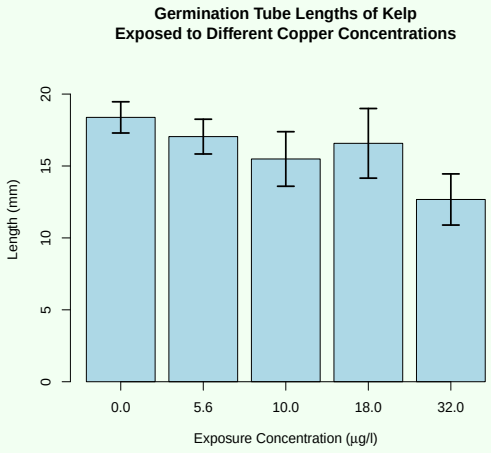


Figure 5.3: Mean tube lengths (mm) in kelp exposed to different concentrations of Cu. Error bars represent ± 1 standard error of the mean.

The error bars indicate how precise each sample mean is as an estimate of it corresponding population mean – the larger error bars indicate less precise estimates.

Comment: Graphs are sometimes made with error bars extending *two* standard errors above and below the means instead of one. Regardless of whether we use one or two standard errors, we should state it

clearly in the figure caption.

5.2.4 Normality of the Sampling Distribution of $\bar{X}$

So far, we know the mean and standard error of the $\bar{X}$ distribution, but nothing about its shape. In fact, it's not always possible to determine the shape of the $\bar{X}$ distribution, even if the shape of the population distribution is known. However, in two commonly occurring situations, the shape of the $\bar{X}$ distribution is (at least approximately) *normal*. The first is when the sample comes from a *normal* population, and the second is when the sample size is large. Knowing that $\bar{X}$ follows a normal distribution will allow us to compute probabilities of the various values that $\bar{X}$ might take, which in turn (in Chapter 6) will help us quantify uncertainty in the inferences we make about the population.

Normality of $\bar{X}$ when the Population is Normal

The following fact guarantees that if a sample is drawn from a *normal* population, the sample mean will follow a normal distribution too.

Fact 5.2 Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $\mathrm{N}(\mu, \sigma)$ population. Then

$$\bar{X} \sim \mathrm{N}\left(\mu_{\bar{X}}, \sigma_{\bar{X}}\right), \tag{5.3}$$

where

$$\mu_{\bar{X}} = \mu \qquad \text{and} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

As a consequence, if we standardize $\bar{X}$, the resulting random variable Z will follow a standard normal distribution, which is to say,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \sim \mathrm{N}(0, 1).$$

Figs. 5.1 and 5.2 both showed normal populations and $\bar{X}$ distributions that were also normal. The next example shows how we can take advantage of the normality of $\bar{X}$ to find probabilities associated with its value.

Example 5.2: Probabilities Involving $\bar{X}$

The so-called rare earth elements (REEs) are a group of 17 chemically similar elements consisting of scandium, yttrium, and the lanthanides. They're gray and silvery metals, typically soft and malleable, and are increasingly being used for industrial purposes such as magnets, batteries, and lasers as well as for agricultural purposes such as in fertilizers.

A study of REEs in river sediments in China suggests that the total REE concentration X (μg/g) in a randomly selected sediment specimen follows a normal distribution with mean $\mu = 177$ and standard deviation $\sigma = 57$ [7].

A random sample of $n = 9$ sediment specimens is to be taken and the REE concentration measured in each specimen. The sampling distribution of the mean $\bar{X}$ of these nine observations is normal with mean and standard error

$$\begin{aligned} \mu_{\bar{X}} &= 177 \\ \sigma_{\bar{X}} &= \tfrac{57}{\sqrt{9}} = 19. \end{aligned}$$

The probability that $\bar{X}$ will be within 19 μg/g of the true population mean, 177, is

$$\begin{aligned} P(158 < \bar{X} < 196) &= P\left(\frac{158-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{196-\mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(\frac{158-177}{19} < Z < \frac{196-177}{19}\right) \\ &= P(-1.00 < Z < 1.00) \\ &= P(Z < 1.00) - P(Z < -1.00) \\ &= 0.6826. \end{aligned}$$

(from a standard normal table).

The next example shows how to find *percentiles* of the $\bar{X}$ distribution,and in particular, the 2.5th and 97.5the percentiles, between which $\bar{X}$ will fall with probability 0.95, or 95%.

Example 5.3: Percentiles of the $\bar{X}$ Distribution

Continuing with the last example, we have

$$\bar{X} \sim \mathrm{N}(177, 19),$$

and the middle 95% of this distribution lies between its 2.5th and 97.5th percentiles. To find these, as explained in Chapter 4, we first obtain the corresponding percentiles of the N(0, 1) distribution from the standard normal table. From the table, they turn out to be $z = -1.96$ and $z = 1.96$. Next we "unstandardize" these using

$$x = \mu_{\bar{X}} + z\,\sigma_{\bar{X}},$$

which gives 2.5th percentile

$$x = 177 + (-1.96)(19) = 139.8$$

and the the 97.5th percentile

$$x = 177 + 1.96(19) = 214.2.$$

Thus the sample mean REE concentration $\bar{X}$ will fall between 139.8 and 214.2 with probability 0.95, or 95%.

Because the distribution of $\bar{X}$ becomes more and more concentrated about the true population mean μ as n increases (see Fig. 5.2), by using a bigger sample size, we can make it more likely that $\bar{X}$ will fall close to μ. The next example illustrates.

Example 5.4: Probabilities Involving $\bar{X}$

In Example 5.2, we found that when $n = 9$, there's a 68.26% chance that sample mean REE concentration will fall within 19 μg/g of the true population mean.

Now suppose instead that our sample size is $n = 49$. In this case, the $\bar{X}$ distribution has the same mean as before,

$$\mu_{\bar{X}} = 177,$$

but now its standard error is much smaller,

$$\sigma_{\bar{X}} = \frac{57}{\sqrt{49}} = 8.14.$$

In this case, the chance that $\bar{X}$ will fall within 19 μg/g of the population mean is

$$\begin{aligned} P(158 < \bar{X} < 196) &= P\left(\frac{158-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{196-\mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(\frac{158-177}{8.14} < Z < \frac{196-177}{8.14}\right) \\ &= P(-2.33 < Z < 2.33) \\ &= P(Z < 2.33) - P(Z < -2.33) \\ &= 0.9802, \end{aligned}$$

or 98.02%, much higher than it was with the smaller sample size.

The important point is that by using a bigger sample size, the *estimate* $\bar{X}$ of the true value μ can be made *more precise* in the sense that it's more likely to fall close to the true value.

Normality of $\bar{X}$ when the Sample Size is Large

The following remarkable fact, known as the ***Central Limit Theorem***, tells us that no matter what shape the population distribution has, as long as the sample size is large, the sampling distribution of $\bar{X}$ will be (at least approximately) normal. What this means is that when n is large, we don't need to be concerned with whether the population distribution is normal – regardless, we can compute probabilities involving $\bar{X}$ using a normal distribution. This will be important when we begin quantifying uncertainties in inferences about the population, starting in Chapter 6.

The Central Limit Theorem: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from *any* population (not necessarily normal) whose mean and standard deviation are μ and σ (with $\sigma < \infty$). Then if n is large,

$$\bar{X} \sim \mathrm{N}(\mu_{\bar{X}}, \sigma_{\bar{X}}),$$

at least approximately, where

$$\mu_{\bar{X}} = \mu \qquad \text{and} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

The larger n is, the more closely the $\bar{X}$ distribution resembles the normal distribution.

As a consequence, if we standardize $\bar{X}$, the resulting random variable Z follows (at least approximately) a standard normal distribution, that is,

$$Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} \sim \mathrm{N}(0, 1),$$

at least approximately.

Fig. 5.4 shows the Central Limit Theorem in action. The population shown at the top has a very right skewed distribution. The sampling distribution of $\bar{X}$ for various sample sizes is shown at the bottom. Notice

that the skewness is evident in the $\bar{X}$ distribution when n is small, but quickly disappears as n increases. Notice also that regardless of the sample size n, the $\bar{X}$ distribution is centered on μ, the population mean, and it becomes less spread away from μ as n increases.

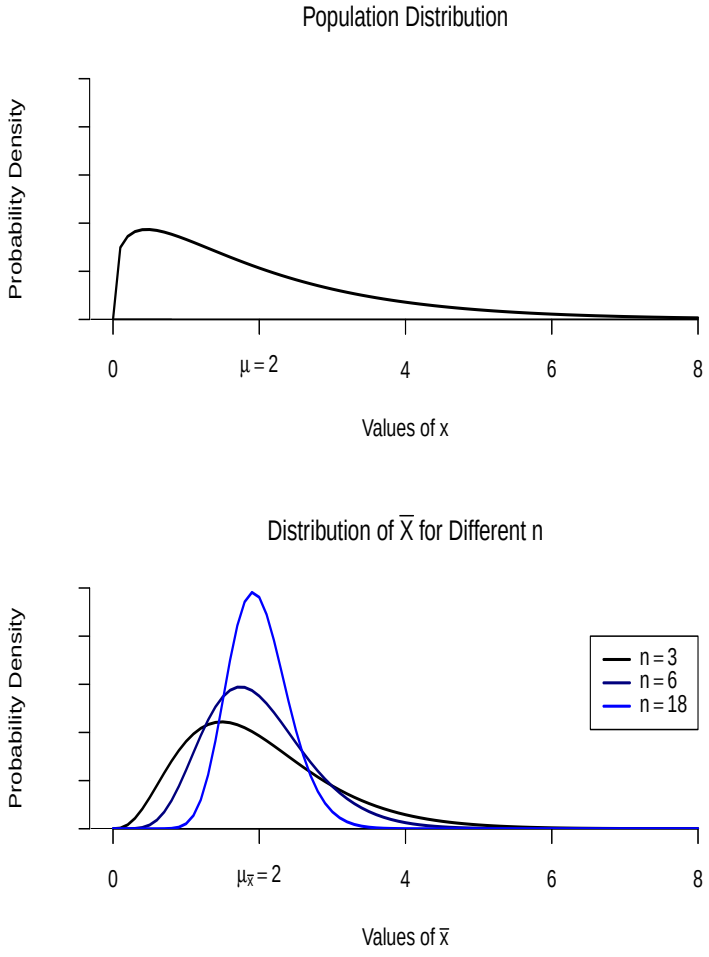


Figure 5.4: Sampling distribution of the sample mean $\bar{X}$ for various sample sizes (bottom) when the sample is from a right skewed population (top) whose mean and standard deviation are $\mu = 2$ and $\sigma = 1.8$.

In the next example, we'll see how the Central Limit Theorem can be applied to find probabilities involving $\bar{X}$ when the sample is from a non-normal population.

Example 5.5: Central Limit Theorem

Vinyl chloride is a gas used in the manufacture of building materials, automotive products, pvc piping, and electrical wire insulation. It can cause health problems when present in drinking water, and can enter the water supply via discharge from plastics factories.

Studies have found that vinyl chloride measurements in groundwater follow a right skewed distribution (see [3], for example). Suppose that in a certain region, groundwater vinyl chloride concentrations can be modeled by a right skewed distribution whose mean is 1.88 μg/L and whose standard deviation is 1.82 μg/L.

Suppose also that a random sample of $n = 100$ groundwater vinyl chloride measurements is to be taken. According to the Central Limit Theorem, because n is large, the sample mean $\bar{X}$ will follow (approximately) a normal distribution with mean and standard error

$$\begin{aligned} \mu_{\bar{X}} &= 1.88 \\ \sigma_{\bar{X}} &= \frac{1.82}{\sqrt{100}} = 0.182. \end{aligned}$$

As a consequence, the (approximate) probability that $\bar{X}$ will fall between, say, 1.70 and 2.06 is

$$\begin{aligned} P(1.70 < \bar{X} < 2.06) &= P\left(\frac{1.70 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{2.06 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &= P\left(\frac{1.70 - 1.88}{0.182} < Z < \frac{2.06 - 1.88}{0.182}\right) \\ &= P(-0.99 < Z < 0.99) \\ &= P(Z < 0.99) - P(Z < -0.99) \\ &= 0.6778 \end{aligned}$$

(from the standard normal table).

Note: Many textbooks state that $n \geq 30$ is large enough for the Central Limit Theorem to apply. However, if the population is *very* skewed, a substantially larger sample size may be needed before the $\bar{X}$ distribution becomes approximately normal. On the other hand, if the population distribution is fairly symmetric, and therefore closer to a normal distribution, a sample size of $n = 10$ or 15 may suffice.

Comment: The requirement that $\sigma < \infty$ in the Central Limit Theorem is met by almost all of the commonly used theoretical probability distributions, and indeed by all of the ones covered in Chapter 4.

Comment: The Central Limit Theorem also applies to *sums* of sample values. More precisely, if $X_1, X_2, \ldots, X_n$ is a random sample from *any* population whose mean and standard deviation are μ and σ, and we define the ***sample total***, $\boldsymbol{T}$, to be their sum,

$$T = X_1 + X_2 + \cdots + X_n,$$

then if n is large,

$$T \sim \mathrm{N}\left(n\mu, \sqrt{n}\sigma\right),$$

at least approximately. This is verified by noting that $T = n\bar{X}$ is a linear function of $\bar{X}$, which by the Central Limit Theorem is (approximately) normal, and linear functions of normal variables are themselves normal (see Subsection 4.5.1 of Chapter 4).

5.2.5 The Law of Large Numbers

We've seen that the sampling distribution of $\bar{X}$ becomes more and more concentrated about μ as the sample size n increases (see Fig. 5.2). If n is *very* large, there will be almost no variation of $\bar{X}$ away from μ, and so $\bar{X}$ is guaranteed to be essentially equal to μ. This result is known as the ***Law of Large Numbers***.

The Law of Large Numbers: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from *any* distribution whose mean is μ and whose standard deviation is σ (with $\sigma < \infty$). Let $\bar{X}$ be the sample mean. Now consider repeatedly drawing observations one at a time from the same distribution, each time incorporating the new observation into the sample and recomputing $\bar{X}$. Then the value of $\bar{X}$ will

tend to get closer and closer to the value of μ. We write this as

$$\bar{X} \to \mu \qquad \text{as} \quad n \to \infty.$$

It's the Law of Large Numbers that lets us interpret the mean μ of a probability distribution as the *long-run average* value of the random variable X. Even if a population is infinite, and can never be exhausted, the law is applicable. For example, if a radon detector is exposed to 100 pCi/L of radon, the population consists of infinitely many potential exposures of the detector to the radon (see Example 1.3 of Chapter 1). If the device is properly calibrated, then in the long-run, repeated readings will average to the true value, 100 pCi/L.

5.3 Some Theory Underlying the Sampling Distribution of $\bar{X}$

In this subsection, we look at some theoretical facts regarding sums of random variables that, together with the facts regarding linear functions of random variables given in Chapter 4, help explain expressions (5.1), (5.2), and (5.3) for the mean, standard error, and normality of the $\bar{X}$ distribution. Each fact will be preceded by an example illustrating a situation to which it may be applied.

5.3.1 Sums of Normally Distributed Random Variables

Example 5.6: Sums of Random Variables

Atmospheric sulfur dioxide (SO_2) is a gas that produces acid rain. A major source of atmospheric SO_2 is fossil fuel combustion in industrial processes. Suppose a town has two industrial plants, a cement production plant and a steel mill. The daily total SO_2 emissions X from the cement plant is a random variable that varies from day to day. The mean of its probability distribution is $\mu_X = 8,800$ lb/day. Likewise, the daily emissions Y from the steel mill is a random variable whose mean is $\mu_Y = 410$ lb/day.

Thus, on a typical day, the cement plant will emit 8,800 lb and the steel mill 410 lb, so the total *combined* emissions from these two sources on a typical day will be $8,800 + 410 = 9,210$ lb. More formally, the total SO_2 emissions on a given day is a new random variable, $X + Y$, and on average the value of $X + Y$ is $\mu_X + \mu_Y$.

The result given in the previous example is stated formally in the following fact.

Fact 5.3 If X and Y are *any* two random variables whose distributions have means μ_X and μ_Y, respectively, then the new random variable $X + Y$ follows a distribution whose mean $\boldsymbol{\mu_{X+Y}}$ is

$$\mu_{X+Y} \;=\; \mu_X + \mu_Y.$$

In words, *the mean of the sum is the sum of the means.* An analogous result can be stated about the variance and standard deviation of $X + Y$, but in this case X and Y have to be *independent* of each other. Here's an example.

Example 5.7: Sums of Random Variables

In the SO_2 emissions example, because both the cement plant emissions X and steel mill emissions Y will vary from day to day, so too will the total emissions $X + Y$. If the standard deviation of the X distribution is σ_X and that of the Y distribution is σ_Y, can we determine the standard deviation of the distribution of the total emissions $X + Y$? The answer is yes, using the following fact.

Fact 5.4 Suppose X and Y are two *independent* random variables whose standard deviations are σ_X and σ_Y, and whose variances, therefore, are σ_X^2 and σ_Y^2. Then the variance of the new random variable $X + Y$, denoted $\boldsymbol{\sigma^2_{X+Y}}$, is

$$\sigma^2_{X+Y} \;=\; \sigma^2_X + \sigma^2_Y$$

and the standard deviation $\boldsymbol{\sigma_{X+Y}}$ is

$$\sigma_{X+Y} \;=\; \sqrt{\sigma^2_X + \sigma^2_Y}.$$

Example 5.8: Sums of Random Variables

Continuing with the last example, suppose we know that the standard deviation of the concrete plant's daily emissions is $\sigma_X = 340$ lb, and the standard deviation of the steel mill's daily emissions is $\sigma_Y = 75$ lb.

On any given day, the emissions X and Y from the two plants could reasonably be assumed to be independent of each other because the emissions from one plant don't differ depending on how much is emitted from the other plant.

By the previous fact, the variance of the total emissions $X + Y$ is

$$\sigma^2_{X+Y} \;=\; \sigma^2_X + \sigma^2_Y \;=\; 340^2 + 75^2 \;=\; 121,225,$$

and so the standard deviation is

$$\sigma_{X+Y} \;=\; \sqrt{\sigma^2_X + \sigma^2_Y} \;=\; \sqrt{340^2 + 75^2} \;=\; \sqrt{121,225} \;=\; 348.2.$$

Together, the mean (from Example 5.6) and standard deviation of $X + Y$ indicate that on a typical day, the total combined emissions from the two plants will be 9,210 lb, plus or minus about 348.2 lb.

The previous two facts show how to determine the mean and standard deviation of the distribution of the sum of two random variables. We now investigate the *shape* of the distribution when the two random variables are both *normal*.

Example 5.9: Sums of Random Variables

Suppose now that we know that X, the amount of SO_2 emitted from the cement plant on a given day, is a normally distributed random variable, and the amount Y emitted from the steel mill is also normally distributed. What can be said about the shape of the distribution of the total combined emissions $X + Y$? It turns out that it too is normal, according to the following fact.

Fact 5.5 Suppose X and Y are two *independent* random variables, with $X \sim \mathrm{N}(\mu_X, \sigma_X)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$. Then

$$X + Y \sim \mathrm{N}\left(\mu_X + \mu_Y,\ \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

In words, *the sum of two independent normal random variables is itself normal.*

Example 5.10: Sums of Random Variables

Using the previous fact and the information given in Examples 5.6 and 5.8, the distribution of the total emissions $X + Y$ is

$$X + Y \sim \mathrm{N}\left(8800 + 410,\ \sqrt{340^2 + 75^2}\right),$$

that is, the total combined emissions from the two plants on a given day is a random variable that follows a $\mathrm{N}(9210, 348.2)$ distribution.

5.3.2 An Explanation of the Sampling Distribution of $\bar{X}$

We now return to the main topic of this chapter, the sampling distribution of $\bar{X}$. A justification of the properties of the $\bar{X}$ distribution given in Subsections 5.2.2 and 5.2.4 rests on the following comment.

Comment: Each of the previous three facts about sums of *two* random variables can be extended to more than two random variables. For example, the sum $X + Y + Z$ of three independent normally distributed random variables follows a normal distribution whose mean and standard deviation are $\mu_X + \mu_Y + \mu_Z$ and $\sqrt{\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2}$.

Now, if $X_1, X_2, \ldots, X_n$ is a random sample from a $\mathrm{N}(\mu, \sigma)$ population, then by the previous comment their sum follows a normal distribution. More precisely,

$$X_1 + X_2 + \ldots + X_n \sim \mathrm{N}\left(n\mu,\ \sqrt{n}\sigma\right). \tag{5.4}$$

Therefore, since

$$\bar{X} = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$

is a linear function of the normal random variable in (5.4), we know from Fact 4.3 in Chapter 4 (with $a = 1/n$) that

$$\bar{X} \sim \mathrm{N}\left(\mu, \sigma_{\bar{X}}\right),$$

where $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, as stated in (5.3) of Subsection 5.2.4.

5.4 Sampling Distribution of the Sample Proportion

Consider now data on a categorical variable that takes just *two* values, which we'll call *success* and *failure*. A categorical variable takes only *two* values is said to be ***dichotomous***. We usually summarize such data using the ***sample proportion*** of successes, denoted $\hat{\boldsymbol{P}}$.

Sample Proportion: For a data set of n observations of a dichotomous categorical variable, with categories *success* and *failure*,

$$\hat{P} = \frac{\text{Number of successes in the sample}}{\text{Sample size } n}.$$

In a population, the proportion of successes will be denoted by $\boldsymbol{p}$. Alternatively, p could be interpreted as the probability that a randomly selected individual will be a success

If our data are a random sample from the population, we can use the sample proportion $\hat{P}$ as an *estimate* of the population proportion p. As was the case when using a sample mean to estimate a population mean, we'll need to be mindful of *sampling variation* in the value of $\hat{P}$.

Example 5.11: Sample Proportion

The Annapolis River Guardians Volunteer Monitoring Program is a program for monitoring water quality in the Annapolis River, Nova Scotia, Canada [4]. Volunteers and scientists measure several water quality variables each year at various locations along the river.

One of the variables is whether or not the *E. Coli* level exceeds a safety threshold for recreational use. For the Annapolis River project, a threshold of 200 cfu/100 ml (colony forming units per 100 ml of water) was specified by the Canadian Council of Ministers of the Environment/Health Canada.

Of the 106 water specimens sampled in 2008, 29 had unsafe *E. Coli* levels and the other 77 had safe levels. The sample proportion that were unsafe was

$$\hat{P} = \frac{29}{106} = 0.27.$$

Mean and Standard Error of the Sampling Distribution of $\hat{P}$

The sampling distribution of the sample proportion $\hat{P}$ describes the sampling variation in the values that $\hat{P}$ takes. The mean and standard error of the sampling distribution are given by the following fact.

Fact 5.6 Suppose we have a random sample of size n from a dichotomous population. Let p denote the proportion of *successes* in the population. Then the sampling distribution of the statistic $\hat{P}$ has mean $\boldsymbol{\mu_{\hat{p}}}$ and standard deviation $\boldsymbol{\sigma_{\hat{p}}}$ that are related to p by

$$\mu_{\hat{p}} = p \tag{5.5}$$

and

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}. \tag{5.6}$$

The mean $\mu_{\hat{p}}$ of the distribution of $\hat{P}$ is the *long-run average* value that you'd get for $\hat{P}$ if you repeatedly took samples of size n from the population. Thus (5.5) says that, *on average*, the sample proportion will equal the population proportion, even though a *particular* sample proportion $\hat{P}$ almost certainly won't equal p exactly. In other words, when $\hat{P}$ is used as an *estimator* of p, on average the estimate is right on target.

The discrepancy between a *particular* estimate $\hat{P}$ and the true value p is called the ***sampling error*** of the estimate.

Sampling Error of $\hat{P}$:

$$\text{Sampling Error} \;=\; \hat{P} - p\,.$$

The sampling error measures how far off the mark a *particular* estimate of the population proportion is.

The standard deviation $\sigma_{\hat{p}}$ of the distribution of $\hat{P}$ is called the ***standard error*** of $\hat{P}$ because it represents the size of a *typical* error of $\hat{P}$ away from p (for a given sample size n). Thus (5.6) says that a typical sampling error is $\sqrt{p(1-p)/n}$, which serves as a measure of how precise $\hat{P}$ is as an estimator of p. A smaller standard error means a more precise estimator. The standard error will be small if either:

- The population proportion p is close to zero or one, or
- The sample size n is large.

Normality of the Sampling Distribution of $\hat{P}$

The next fact tells us that the statistic $\hat{P}$ follows a normal distribution when n is large.

Fact 5.7 Suppose we have a random sample of size n from a dichotomous population. Let p denote the proportion of *successes* in the population. Then if n is large,

$$\hat{P} \;\sim\; \mathrm{N}\left(\mu_{\hat{P}},\; \sigma_{\hat{P}}\right)$$

(approximately), where

$$\mu_{\hat{P}} \;=\; p \qquad \text{and} \qquad \sigma_{\hat{P}} \;=\; \sqrt{\frac{p(1-p)}{n}}.$$

The larger n is, the more closely the $\hat{P}$ distribution resembles the normal distribution.

As a consequence, if we standardize $\hat{P}$, the resulting random variable Z will follow a standard normal distribution, which is to say,

$$Z \;=\; \frac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}} \;\sim\; \mathrm{N}(0,\, 1)$$

(approximately).

The sample size n can be considered large enough for $\hat{P}$ to follow (approximately) a normal distribution if both

$$n\hat{P} \;\geq\; 10 \qquad \text{and} \qquad n(1-\hat{P}) \;\geq\; 10\,,$$

in other words, when there are at least ten *successes* and at least ten *failures* in the sample.

5.5 Sampling Distributions and Standard Errors of Other Statistics

We've seen that the sample mean and sample proportion both follow normal distributions (under certain circumstances, such as when the sample size n is large). Every statistic (the sample median $\tilde{X}$, sample standard deviation S, etc.), follows *some* probability distribution, which we refer to as the statistic's *sampling distribution*, but often the sampling distribution of these other statistics will be non-normal. Regardless of the shape of a statistic's sampling distribution, we refer to the standard deviation of that

distribution as the *standard error* of the statistic. These notions will be important in later chapters when we look at various statistical inference procedures.

5.6 Problems

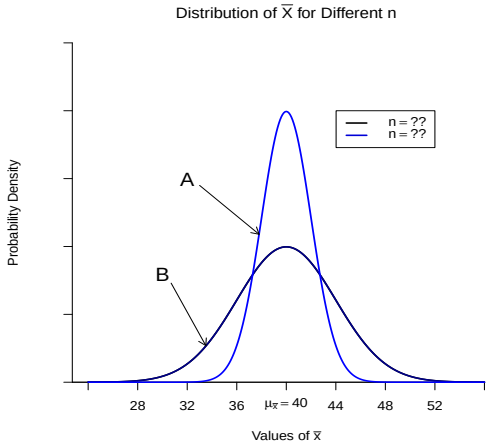
5.1 In the Southern ocean food chain, the krill species *Euphausia supeba* is the most important prey for many marine predators, from seabirds to the largest whales. According to one study, krill body lengths of are normally distributed with a mean of $\mu = 40$ mm and a standard deviation of $\sigma = 12$ mm [6].

a) A random sample of $n = 9$ krill is to be taken and the sample mean length $\bar{X}$ calculated. What are the values of the mean $\mu_{\bar{X}}$ and standard error $\sigma_{\bar{X}}$ of the sampling distribution of $\bar{X}$?

b) How can we be sure that the sampling distribution of $\bar{X}$ is a normal distribution? **Hint**: The $\bar{X}$ distribution will be normal under either of these conditions: 1) the sample came from a normal population, or 2) the sample size is large.

c) If instead the sample size was $n = 36$, what would the values of the mean $\mu_{\bar{X}}$ and standard error $\sigma_{\bar{X}}$ of the sampling distribution of $\bar{X}$ be?

d) Below are graphs of the sampling distribution of $\bar{X}$ for the two sample sizes, $n = 9$ and $n = 36$. Which sampling distribution corresponds to each sample size?



e) Calculate $P(37 < \bar{X} < 43)$, the probability that $\bar{X}$ will fall between 37 and 43 mm when $n = 9$.

f) Now calculate the probability from part e when $n = 36$ and compare your answer to the one you got for part e.

5.2 The Mexican bean beetle is a pest of dry beans, causing damage to bean crops during their vulnerable pod-filling stage in late July and early August. Female Mexican bean beetles lay eggs in clusters, typically with 40-60 eggs per cluster.

A study was carried out to investigate the Mexican bean beetle population on a bean field in Nebraska [1]. The field was divided into 30 cm stretches of crop rows. Let X denote the number of egg clusters in a randomly selected stretch. The cited study found that the probability distribution of X (that is, the population distribution) is discrete and severely right skewed, with the vast majority of the stretches having zero or one egg cluster, but a few having as many as five or six. The mean and standard deviation of the distribution are $\mu = 0.194$ and $\sigma = 0.532$.

a) If a random sample of $n = 50$ stretches is selected and the number of egg clusters counted on each stretch, what are the values of the mean $\mu_{\bar{X}}$ and standard error $\sigma_{\bar{X}}$ of the sampling distribution of the sample mean number of egg clusters $\bar{X}$?

b) Even though the distribution of egg cluster counts in the population is right skewed, how can we be sure that the sampling distribution of $\bar{X}$ is (at least approximately) normal? **Hint**: The $\bar{X}$ distribution will be normal under either of these conditions: 1) the sample came from a normal population, or 2) the sample size is large.

c) Find $P(\bar{X} > 0.31)$, the probability that $\bar{X}$ will be greater than 0.31.

5.3 Blood glucose levels in fish have been suggested as a biological indicator of environmental pollution. Results of the study cited in Problem 4.13 in Chapter 4 suggest that glucose levels in johnny darter fish (*Etheostoma nigrum* Rafinesque) in White Clay Creek, Pennsylvania follow a normal distribution with mean $\mu = 37.5$ mg/100ml and standard deviation $\sigma = 15.3$ mg/100ml.

Suppose that a random sample of $n = 100$ fish is to be taken and the sample mean glucose level $\bar{X}$ calculated.

a) What are the values of the mean $\mu_{\bar{X}}$ and standard error $\sigma_{\bar{X}}$ of the sampling distribution of $\bar{X}$?

b) How could we be sure that the sampling distribution of $\bar{X}$ is normal? **Hint**: The $\bar{X}$ distribution will be normal under either of these conditions: 1) the sample came from a normal population, or 2) the sample size is large.

c) Find $P(34.5 < \bar{X} < 40.5)$, the probability that $\bar{X}$ will fall between 34.5 and 40.5 mg/100ml when $n = 100$.

d) Find the 2.5th and 97.5th percentiles of the sampling distribution of $\bar{X}$. These are the two values that capture the middle 95% of the $\bar{X}$ distribution, and between which $\bar{X}$ has a 95% chance of falling.

5.4 A random sample of $n = 25$ air quality measurements is to be taken near a chemical warehouse to decide if there are leaks. Assume that the pollutant concentration follows a normal distribution in the vicinity of the warehouse.

a) If the true mean concentration in the vicinity of the warehouse is $\mu = 5.0$ ppb and the true standard deviation is $\sigma = 2.3$ ppb, find the probability that the sample mean $\bar{X}$ will lie within 0.9 ppb of μ (i.e. between 4.1 and 5.9).

b) Suppose instead that the true mean is $\mu = 8.0$ (but $\sigma = 2.3$ still). Now what's the probability that $\bar{X}$ will lie within 0.9 ppb of μ (i.e. between 7.1 and 8.9).?

c) In general, if the true mean concentration μ is unknown (but $\sigma = 2.3$ still), what's the probability that $\bar{X}$ will lie within 0.9 ppb of μ?

d) In parts *a* - *c*, the value 0.9 happens to be equal to $1.96\sigma/\sqrt{n}$. In general, what's the probability that a sample mean $\bar{X}$ will lie within 1.96 standard errors of the population mean μ?

e) In general, what's the probability that a sample mean $\bar{X}$ will lie within 1.64 standard errors of the population mean μ? How about within 2.58 standard errors?

5.5 To assess whether a cleanup standard has been met at a contaminated site, the contaminant will be measured in soil specimens taken at a random sample of n locations near the site, and the decision based on the value of the sample mean $\bar{X}$. To reduce the variability in the $\bar{X}$ values that might result, and

thereby to get a more precise determination of whether the cleanup standard has been met, it is decided that n should be large enough so that the standard error of $\bar{X}$ is no greater than 3.0 ppm.

If it is known (based on other studies at similar sites) that the population standard deviation of soil contaminant concentrations is $\sigma = 20.5$ ppm, how large should n be?

5.6 Consider the SO_2 emissions X and Y (in pounds) on a given day from a cement plant and steel mill, respectively, described in Examples 5.6 - 5.10. Suppose again that $X \sim \text{N}(8800, 340)$ and $Y \sim \text{N}(410, 75)$, and that X and Y are independent.

The combined total SO_2 emissions from the two plants on a given day, $X + Y$, is a new random variable. Find the probability that $X + Y$ will be greater than 9,600 pounds.

Bibliography

[1] José A. F. Barrigossi, Linda J. Young, Carol A. Gotway Crawford, Gary. L Hein, and Leon G. Higley. Spatial and probability distribution of Mexican bean beetle (Coleoptera: Coccinellidae) egg mass populations in dry bean. *Environmental Entomology*, 30(2):244–253, 2001.

[2] G. A. Chapman, D. L. Denton, and J. M. Lazorchak. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to west coast marine and estuarine organisms. Technical Report EPA/600/R-95-136, United States Environmental Protection Agency, Aug 1995.

[3] Robert D. Gibbons and Dulal K. Bhaumik. Simultaneous gamma prediction limits for ground water monitoring applications. *Ground Water Monitoring and Remediation*, 26(3):105–116, 2006.

[4] Jeffrey Glenen and Andy Sharpe. Annapolis River 2008 annual water quality monitoring report. Technical report, Clean Annapolis River Project, 2009. Report from the Annapolis River Guardians Volunteer Water Quality Monitoring Program.

[5] Walter W. Piegorsch and R.Webster West. Low dose risk estimation via simultaneous statistical inferences. *Applied Statistics*, 54(1):245 – 258, 2005.

[6] Keith Reid, Jon L. Watkins, John P. Croxall, and Eugene J. Murphy. Krill population dynamics at south Georgia 1991-1997 based on data from predators and nets. *Marine Ecology Progress Series*, 177:103–114, 1999.

[7] W. Zhu, M. Kennedy, E. W. B. de Leer, H. Zhou, and G. J. F. R. Alaerts. Distribution and modelling of rare earth elements in Chinese river sediments. *The Science of the Total Environment*, 204:233–243, 1997.