Chapter 2

Sample Statistics

In this chapter, we explore further the nature of random variables. In particular, we are interested in the process of gaining knowledge of a particular random variable through the use of *sample statistics*. To kick off the chapter, two important concepts are described: the notion of a *population* and subsets of that population, called *samples*. Next, the utility of *sample statistics* in describing random variables is introduced, followed by a discussion of the properties of two important samples statistics: the sample mean and the sample variance. The chapter closes with a discourse on the process of *data pooling*, a process whereby multiple samples are used to obtain a sample statistic.

Chapter topics:

1. Populations and samples
2. Sample statistics
3. Data pooling

2.1 Populations and Observations

2.1.1 A Further Look at Random Variables

As we discovered in the last chapter, a random variable is a variable that contains some uncertainty, and many experiments in natural science involve random variables. These random variables must be described in terms of a probability distribution. We will now extend this concept further.

When an experiment involving random variables is performed, data are generated. These data points are often measurements of some kind. In statistics, all instances of a random variable as called *observations* of that variable. Furthermore, the set of data points actually collected is termed a *sample* of a given *population* of observations.

For example, imagine that we are interested in determining the 'typical' height of a student who attends the University of Richmond. We might choose a number of students at random, measure their heights, and then average the values. The *population* for this experiment would consist of the heights of all students attending UR, while the *sample* would be the heights of the students actually selected in the study.

To generalize: the population consists of the entire set of *all possible* observations, while the sample is a subset of the population. Figure 2.1 demonstrates the difference between sample and population.

A population can be infinitely large. For example, a single observation might consist of rolling a pair of dice and adding the values. The act of throwing the dice can continue indefinitely, and so the population would consist of an infinite number of observations. Another important exam-

Many experiments in science result in outcomes that must be described by probability distributions — i.e., they yield *random variables*.

The values collected in an experiment are *observations* of the random variable.

The *population* is the collection of all possible observations of a random variable. A *sample* is a subset of the population.

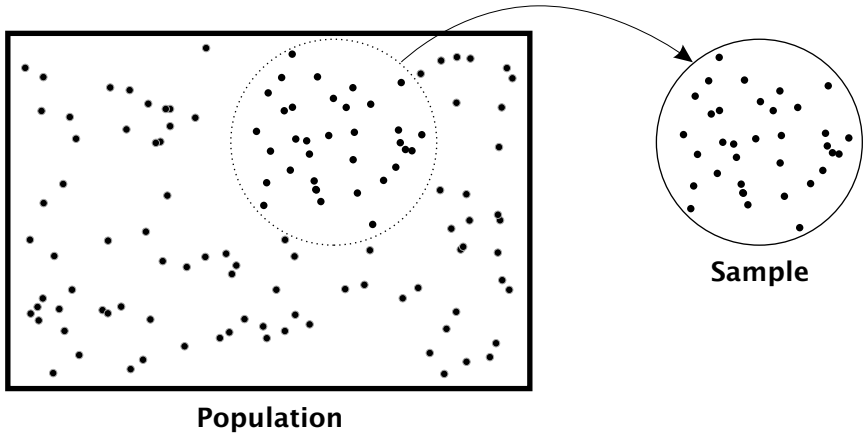

Figure 2.1: Illustration of population and sample. The *population* is the box that contains all possible outcomes of an experiment, while the *sample* contains the outcomes that are actually observed in an experiment.

ple occurs for experiments in which the measurement process introduces a random component into the observation. Since measurements can theoretically be repeated indefinitely, the population of measurements would also be infinite. The concept of a population can be somewhat abstract in such cases.

2.1.2 Population Parameters

Population parameters describe characteristics of the entire population.

A population is a collection of all possible observations from an experiment. These observations are generated according to some probability distribution. The probability distribution also determines the *frequency distribution* of the values in the population, as shown in figure 2.2.

Population parameters are values that are calculated from all the values in the population. Population parameters describe characteristics — such as location and dispersion — of the population. As such, the population parameters are characteristics of a particular experiment and the conditions under which the experiment is performed. Most scientific experiments are intended to draw conclusions about populations; thus, they are largely concerned with population parameters.

The best indicators of the location and dispersion of the observations in a population are the mean and variance. Since the probability distribution is also the frequency distribution of the population, *the mean and variance of the probability distribution of a random variable are also the mean and variance of the observations in the population*. If one had access to all the values in the population, then the mean and variance of the values in population could be calculated as follows:

The two most important population parameters are the population mean, μ_x, and the population variance, σ_x^2.

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu_x)^2 \tag{2.2}$$

where N is the number of observations in the population and x_i are the values of individual observations of the random variable x. The values μ_x

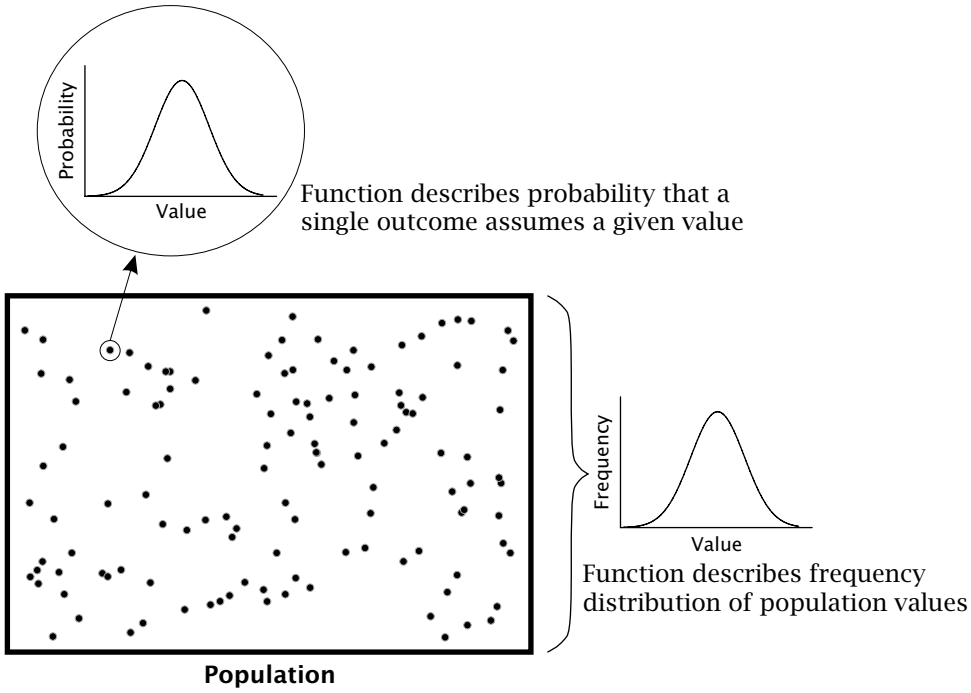

Figure 2.2: Role of probability distribution in populations. The same distribution function dictates the probability distribution of the random variable (i.e., a single observation in the population) and the frequency distribution of the entire population.

and σ_x^2 are the *population mean* and the *population variance*, respectively. The population standard deviation, σ_x, is commonly used to describe dispersion. As always, it is simply the positive root of the variance.

From eqns. 2.1 and 2.2 we can see that

- The population mean μ_x is the average of all the values in the population — the "expected value," $\mathrm{E}(x)$, of the random variable, x;
- The population variance σ_x^2 is the average of the quantity $(x - \mu_x)^2$ for all the observations in the population — the "expected value" of the squared deviation from the population mean.

Other measures of location and dispersion, such as those discussed in section 1.3, can also be calculated for the population.

2.1.3 The Sampling Process

The method used to obtain samples is very important in most experiments. Any experiment has a population of outcomes associated with it, and usually our purpose in performing the experiment is to gain insight about one (or more) properties of the population. In order to draw valid conclusions about the population when using samples, it is important that *representative samples* are used, in which the important characteristics of the population are reflected by the members of the sample. For example, if we are interested in the height of students attending UR, we would not want to choose only basketball players in our sampling procedure, since this would not be a good representation of the characteristics of the student body.

A representative sample is one whose properties mirror those of the population. One way to obtain a representative sample is to collect a random sampling of the population.

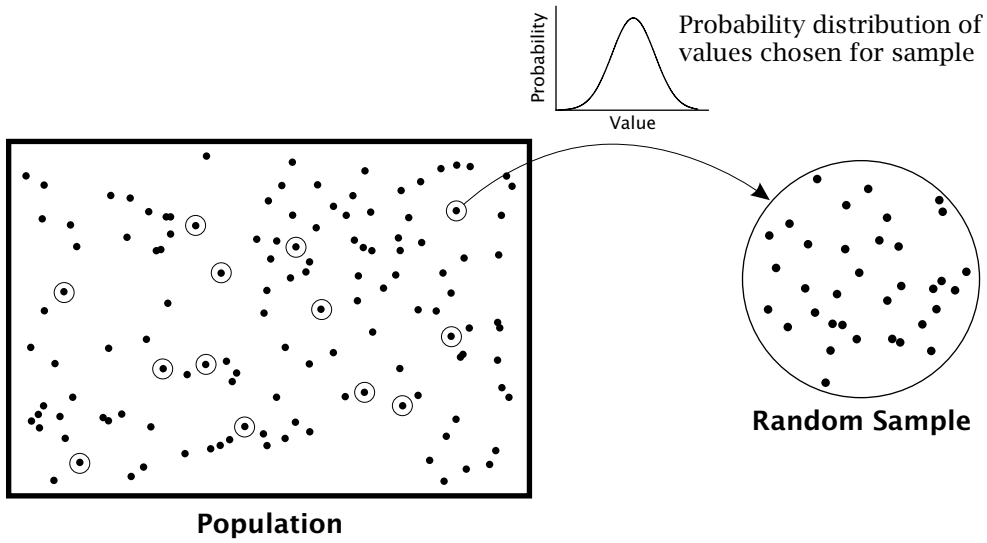

Figure 2.3: In a *random sample* objects are chosen randomly (circled points) from the population. Random sampling ensures a *representative sample*, where the probability distribution of the sampled observation is the same as the frequency distribution for the entire population.

For the same reason, we would not want to choose a sample that contains either all males or all females.

It is often difficult to ensure that a sample is a good representative of the populations. One common way to do so is a sampling procedure called *random sampling*. A truly random sampling procedure would mean that each observation in the population (e.g., each student in the university) has an equal probability of being included in the sample.

So the point of a sample is to accurately reflect the properties of the population. These properties — such as the population mean and variance — are completely described by the variable's probability distribution. *A representative sample is one in which the same distribution function also describes the probabilities of the values chosen for the sample*, as shown in figure 2.3.

Now examine figure 2.4 carefully. One thousand observations were chosen randomly[1] from a population described by a normal distribution function with $\mu_x = 50$ and $\sigma_x = 10$. If the sample is representative of the population, the same probability distribution function should determine the values of the observations in the sample. On the right side of the figure, the observed frequency distribution of values in the sample (the bar chart) is compared with a normal distribution function (with $\mu_x = 50$ and $\sigma_x = 10$). As you can see, the two match closely, implying that the sample is indeed representative. Since the same distribution function describes both population and sample, they share the same characteristics (such as location and dispersion). This is importance, since we are usually using characteristics of the sample to draw conclusions about the nature of the population from which it was obtained.

A statistician would say we 'sampled a normal distribution' (with $\mu_x = 50$, $\sigma_x = 10$).

[1] Actually, the values were generated according to an algorithm for random number generation (sometimes called a *pseudo-random number generator*). Such generators are standard fare in many computer applications — such as MS Excel — and are important components of most computer simulations.

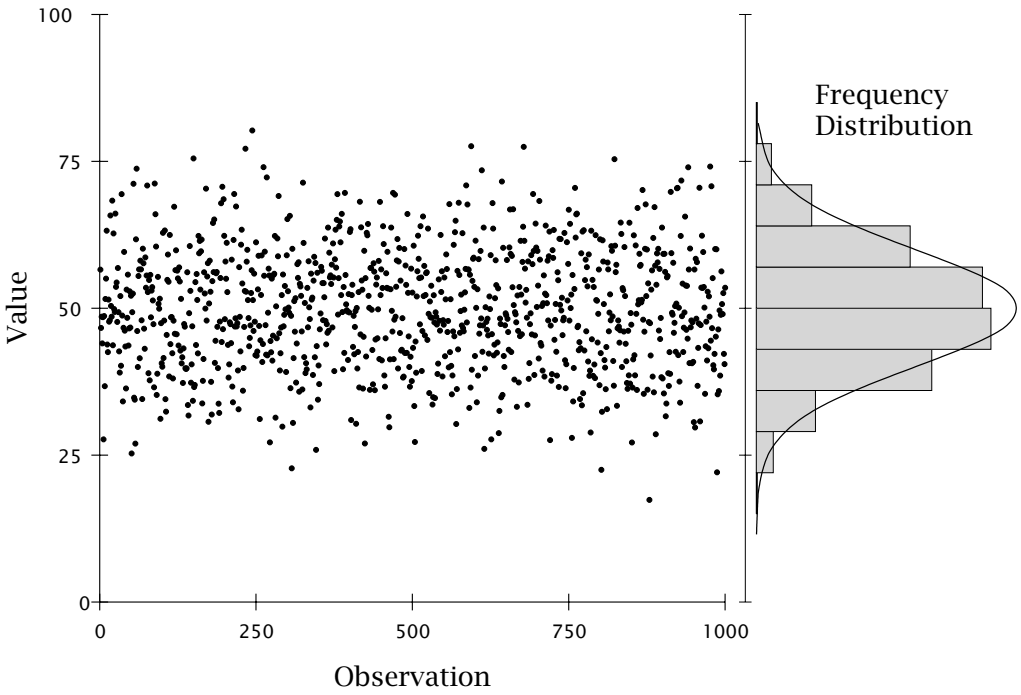

Figure 2.4: One thousand observations were chosen at random from the population of a normally-distributed variable ($\mu_x = 50$, $\sigma_x = 10$). On the right, the frequency distribution of the sample (horizontal bars) is well described by a normal distribution (solid line).

2.2 Introduction to Sample Statistics

2.2.1 The Purpose of Sample Statistics

Sample statistics describe characteristics of the sample. They are usually used to estimate population parameters.

Let's return to our example of the heights of the students at UR Suppose we were interested in comparing the mean height of the students currently attending UR with the mean height of those attending some other school. Our first task would be to determine the mean height of the students at both schools. The "brute force" approach would be to measure the height of every student at each school, average each separately, and then compare them. However, this approach is tedious, and is not usually necessary. Instead, we can measure the heights of a sample of each population, and compare the averages of these measurements.

Population parameters are values calculated using all the values in the population. These values are generally not available, so *sample statistics* are used instead. ***A sample statistic is a value calculated from the values in a sample.*** There are two primary purposes for sample statistics:

1. Sample statistics *summarize* the characteristics of the sample — just as population parameters do for populations.

2. Sample statistics *estimate* population parameters. Thus, properties of the population are inferred from properties of the sample.

The field of Statistics is largely concerned with the properties and uses of sample statistics for data analysis — indeed, that is the origin of its name. Two major branches in the field of Statistics are *Descriptive Statistics* and *Inferential Statistics*. The first of these deals with ways to summarize the

results of experiments through informative statistics and visual aids. The second — and generally more important — branch is concerned with using statistics to draw conclusions about the population from which the sample was obtained. It is with this second branch of Statistics that we will be most concerned.

2.2.2 Sample Mean and Variance

The two most important sample statistics are the sample mean, $\overline{x}$, and the sample variance, s_x^2. If there are n observations in the sample, then the sample mean is calculated using the following formula:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.3}$$

Sometimes the symbol $\overline{x}_n$ will be used to emphasize that the sample mean is calculated from n observations.

Likewise, the sample variance is calculated using the observations in the sample. If the population mean if known, then the sample variance is calculated as

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_x)^2 \tag{2.4a}$$

Equations 2.4a and 2.4b provide different estimates of σ_x^2. If μ_x is known, eqn. 2.4a gives a better estimate. Usually μ_x is not known, so eqn. 2.4b must be used to estimate σ_x^2.

Generally, however, the value of μ_x is not known, so that we must use $\overline{x}$ instead:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.4b}$$

As before, the sample standard deviation, s_x, is the positive root of the sample variance. If the value of μ_x is known, then either formula can be used to calculate sample variance — and they will give different values! Why are there two different formulas for the sample variance? Notice that when the true mean μ_x is replaced by the sample mean, $\overline{x}$, then the denominator in the formula changes from n to $n-1$. The value of the denominator is the number of *degrees of freedom*, ν, of the sample variance (or the sample standard deviation calculated from the variance). When s_x^2 (or s_x) is calculated by eqn. 2.4a, when $\nu = n$, while if eqn. 2.4b must be used — the usual scenario — $\nu = n-1$. When the sample must be used to determine s_x^2, then one degree of freedom is lost in the calculation, and the denominator must reflect the decrease. [At this point, don't worry too much about what a "degree of freedom" is; it is sufficient for now to know that more degrees of freedom is better].

Sample statistics provide estimates of population parameters:
$\overline{x} \longrightarrow \mu_x$
$s_x \longrightarrow \sigma_x$

The main purpose of the sample mean $\overline{x}$ is to estimate the population mean μ_x, just as the sample variance s_x^2 estimates the population variance σ_x^2. The equations used to calculate the sample statistics (eqns. 2.3 and 2.4a)and the corresponding population parameters (eqns. 2.1 and 2.2), look very similar; the main difference is that the population parameters are calculated using all N observations in the population set, while the sample statistics are calculated using the n observations in the sample, which is a *subset* of the population (i.e., $n > N$). Despite the similarity of the equations, there is one critical difference between population parameters and sample statistics:

A sample statistic is a *random variable* while a population parameter is a *fixed value*.

Go back and inspect fig. 2.4, which shows 1000 observations drawn from a population that follows a normal distribution. The true mean and standard deviation of the population are $\mu_x = 50$ and $\sigma_x = 10$. However, the corresponding sample statistics of the 1000 observations are $\overline{x} = 49.811$ and $s_x = 10.165$. Another sample with 1000 different observations would give different values for $\overline{x}$ and s_x; the population parameters would remain the same, obviously, since the same population is being sampled. Since $\overline{x}$ and s_x vary in a manner that is not completely predictable, they must be random variables, each with its own probability distribution.

We illustrate these concepts in the next example.

Example 2.1

The heights of students currently enrolled at UR and VCU are compared by choosing ten students (by randomly choosing student ID numbers) and measuring their heights. The collected data, along with the true means and standard deviations, are shown in the following table:

school	population parameters	sample
UR	$\mu_x = 68.07$ in $\sigma_x = 4.11$ in	73.03, 70.63, 65.32, 69.51, 71.81, 65.91, 72.41, 65.21, 65.78, 67.51
VCU	$\mu_x = 68.55$ in $\sigma_x = 4.88$ in	63.65, 68.68, 75.64, 62.45, 73.63, 70.44, 63.17, 68.01, 63.22, 65.51

Calculate and compare the means and standard deviations of these two samples.

Most calculators provide a way to calculate the mean, variance and standard deviation of a number. Compare your calculator's answer to the following results, calculated from the definitions of the sample mean and sample standard deviation.

For the two groups, the sample mean is easily calculated.

$$\begin{aligned}\overline{x}_{ur} &= \frac{1}{10}(73.03 + \cdots + 67.51)\\ &= 68.71\,\text{in}\\ \overline{x}_{vcu} &= \frac{1}{10}(63.65 + \cdots + 65.51)\\ &= 67.44\,\text{in}\end{aligned}$$

Since we know the population mean, μ_x, we may use eqn. 2.4a to calculate the sample standard deviation, s_x.

$$\begin{aligned}s_{ur} &= \sqrt{\frac{1}{10}\left[(73.03 - 68.07)^2 + \cdots + (67.51 - 68.07)^2\right]}\\ &= 3.03\,\text{in}\\ s_{vcu} &= \sqrt{\frac{1}{10}\left[(63.65 - 68.55)^2 + \cdots + (65.51 - 68.55)^2\right]}\\ &= 4.56\,\text{in}\end{aligned}$$

In most cases, the population mean μ_x will not be known, and eqn. 2.4b must be used.

$$\begin{aligned} s_{ur} &= \sqrt{\frac{1}{9}\left[(73.03-68.71)^2+\cdots+(67.51-68.71)^2\right]} \\ &= 3.13\,\text{in} \\ s_{vcu} &= \sqrt{\frac{1}{9}\left[(63.65-67.44)^2+\cdots+(65.51-67.44)^2\right]} \\ &= 4.66\,\text{in} \end{aligned}$$

Almost all calculators will use eqn. 2.4b to calculate sample standard deviation.

Let's summarize the results in a table:

	μ_x	$\overline{x}$	σ_x	s_x (by eqn. 2.4a)	s_x (by eqn. 2.4b)
UR	68.07 in	68.71 in	4.11 in	3.03 in	3.13 in
VCU	68.55 in	67.44 in	4.88 in	4.56 in	4.66 in

Notice that *in all cases the sample statistics do not equal the true values* (i.e., the population parameters). However, they appear to be reasonably good *estimates* of these values.

Since the sample means are random variables, *there will be an element of chance in the value of the sample mean*, just as for any random variable. If more samples are drawn from the population of UR or VCU students, then the sample mean will likely not be the same. This seems intuitive: if we choose two groups of ten students at UR, it is not likely that the mean height of the two groups will be the same. However, in both cases the sample mean is still an estimate of the true mean of all the students enrolled at UR.

Now, in this example, we happen to "know" that the mean height of the entire population of VCU students is more than mean height of the UR students. However, the mean height in the *sample* of UR students is greater than the mean of the sample of VCU students! Based on the sample alone, we might be tempted to conclude from this study that UR students are taller (since $\overline{x}_{ur} > \overline{x}_{vcu}$), if we didn't know already that the opposite is true (i.e., that $\mu_{ur} < \mu_{vcu}$).

Obviously, it would be wrong to conclude that UR students are taller based on our data. The seeming contradiction is due to the fact that sample statistics are only *estimates* of population parameters. There is an inherent uncertainty in this estimate, because the sample statistic is a random variable. The difference between the value of a sample statistic and the corresponding population parameter is called the *sampling error*. Sampling error is an unavoidable consequence of the fact that the sample contains only a subset of the population. Later, we will find out how to protect ourselves somewhat from jumping to incorrect conclusions due to the presence of sampling error; this is a very important topic in science.

Aside: Calculation of Sample Variance

In example 2.1, two different equations (see eqn. 2.4) are used to calculate the sample standard deviation. Which is "correct?" Neither — both are simply different estimates of the true standard deviation. A better question is, are both estimates equally "good?" The answer is no; the sample variance

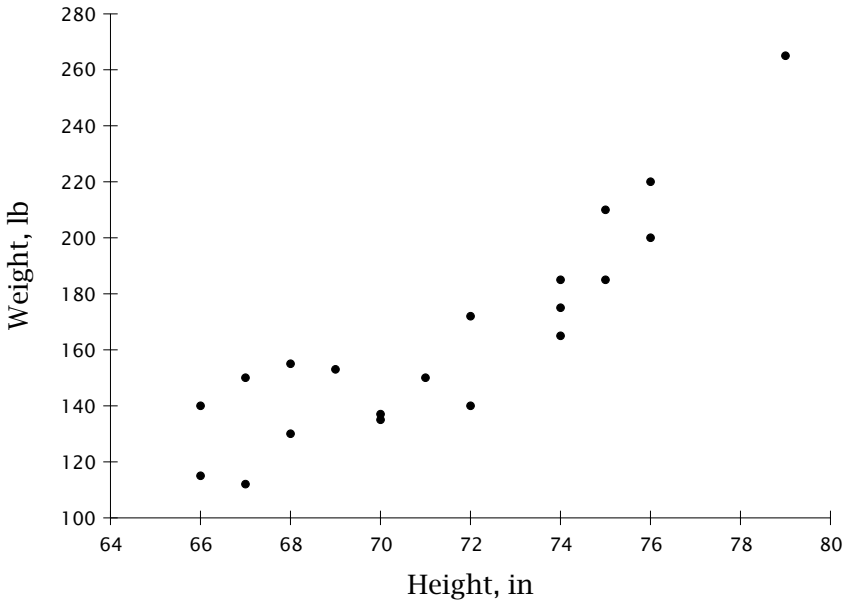

Figure 2.5: Twenty college students were chosen at random and both their height and weight were recorded. The two variables are positively correlated.

calculated using the true mean value (eqn. 2.4a), is generally the better estimate of the true variance, because it has an extra degree of freedom.

If the true mean is known, either estimate can be used as long as you use the correct number of degrees of freedom of your statistic. You will need to know the degrees of freedom in order to use the variance/standard deviation you calculate from your measurements. If we use eqn. 2.4a for the estimate, then we are essentially making use of an extra bit of information (knowledge of the value of μ_x); this is the extra degree of freedom, and results in a slightly better estimate.

Even though eqn. 2.4a gives a better estimate of the population variance, eqn. 2.4b is used much more commonly, since the value of μ_x is generally not known.

2.2.3 Covariance and Correlation (Variable Independence)

Covariance and correlation indicate the strength of the relationship between two variables

Up to this point, we have only been concerned with a single random variable at a time. However, there are times when we are interested in more than one property of a single object. In other words, we might observe the value of two or more random variables in a single experiment. These variables might or might not be related to one another; the strength of the relationship is indicated by a population parameter called the *covariance* of the two variables.

Let's imagine that we record the heights and weights of a representative sample of 20 male students at UR; figure 2.5 shows the data that might be recorded. From the figure, we see that there is a fairly strong linear association between the two variables: students with a greater height tend to weigh more. We would say that there is a *linear covariance* or *linear correlation* between the variables.

For this experiment, the population would consist of the height and weight of all students attending UR. If all N values in the population were known, we could calculate the population covariance, σ_{xy}:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{2.5}$$

where μ_x and μ_y are the population means for the height and weight, respectively.

Normally, of course, a sample is obtained from the population. The sample covariance s_{xy} of two random variables x and y is defined by one of the following two expressions:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \tag{2.6a}$$

if μ_x and μ_y are known, and

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{2.6b}$$

if μ_x and μ_y are unknown (the usual situation)

Note the similarity between the above expressions and those for the sample variance (eqn. 2.4). Essentially, the variance is the covariance of a random variable with itself: in other words, $\sigma_x^2 = \sigma_{xx}$.

As mentioned earlier, the covariance measures the association between two random variables x and y. There are two problems associated with the use of the covariance to describe the relationship between variables:

1. The covariance value is sensitive to changes in units. For example, the sample covariance of the data in fig. 2.5 is $s_{xy} = 128.9\,\text{lb}^2\text{in}^2$. If we specified the heights of the students in cm and the weight in kg, then we would obtain a different value for the sample covariance: $s_{xy} = 148.8\,\text{kg}^2\text{cm}^2$.

2. There is no obvious relationship between the magnitude of the covariance and the strength of the association between the variables. For example, we mentioned that $s_{xy} = 128.9\,\text{lb}^2\text{in}^2$ for the student data. Is this covariance value "large," indicating a high degree of linear association, or not?

A solution to these problems is to use to calculate the *correlation coefficient* (sometimes simply called the correlation) between the two variables. The population parameter ρ_{xy} is given by

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{2.7}$$

The population correlation coefficient is estimated by the sample correlation coefficient: $r_{xy} \longrightarrow \rho_{xy}$

where σ_{xy} is the population covariance, and σ_x and σ_y are the population standard deviations of x and y, respectively. The sample correlation coefficient, r_{xy}, is calculated using the sample covariance, s_{xy}, and standard deviation values, s_x and s_y:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} \tag{2.8}$$

$-1 < \rho_{xy} < +1$

The correlation coefficient is not affected by changes in units, and is always a value between -1 and 1. Values near -1 or 1 indicate strong linear associations between the two variables; for the data in fig. 2.5, $r = 0.8892$. Positive values of covariance or correlation coefficient indicate that the variables increase or decrease "together": an increase in x will tend to be accompanied by an increase in y. If the correlation coefficient is negative, then an increase in x tends to be accompanied by a *decrease* in y. If either the covariance or correlation coefficient is zero, then there is no linear association between the two variables.

Independent variables are not correlated: $\rho_{xy} = \sigma_{xy} = 0$

Two variables are said to be *correlated* when a change in one variable is generally reflected (either linearly or nonlinearly) by a change in the other. When there is no such association between variables, they are said to be ***independent***. For example, while we might expect a relationship between the variables height and weight, we probably would not anticipate any such association between height and, say, intelligence. These variables are independent of one another: the value of one (the height) does not affect the value of the other (intelligence). A necessary condition for variables to be independent is that $\rho_{xy} = 0$.

Aside: Notation Conventions in Statistics

Sample statistics, usually given as Latin characters, estimate population parameters, usually represented by Greek characters:
$\overline{x} \longrightarrow \mu_x$
$s_x \longrightarrow \sigma_x$
$r_{xy} \longrightarrow \rho_{xy}$

But sometimes a 'hat' is used to represent sample statistics:
$\hat{\mu}_x \longrightarrow \mu_x$
$\hat{\sigma}_x \longrightarrow \sigma_x$
$\hat{\rho}_{xy} \longrightarrow \rho_{xy}$

If you have been observant, you might have noticed a trend in the symbols chosen for population parameters and sample statistics: population parameters have mostly been Greek characters (e.g., μ_x, and ρ_{xy}) while the corresponding sample statistics have used Latin characters. This is a convention that is commonly followed in statistics: population parameters are Greek characters such as α, β and γ, while the sample statistics used to estimate these parameters as the corresponding Latin characters: a is a sample statistic to estimate α, b estimates β, and g estimates γ. Note that the symbol for the mean, $\overline{x}$, is something of an anomaly from this point of view.

This convention has one drawback: there are some Greek characters (such as omicron, o, or kappa, κ) that are difficult to distinguish from their Latin counterparts; other Greek characters have no counterparts in the English alphabet (e.g., χ or ω). Partly due to these facts, there is another convention: sample statistics are indicated by a 'hat' above the appropriate population parameter. Thus, the sample variance might be identified by either s_x^2 or $\hat{\sigma}_x^2$, and the sample mean may be either $\overline{x}$ or $\hat{\mu}_x$.

Being familiar with these conventions makes it easier to classify the many symbols we will come across in our study of statistics.

2.3 Properties of Sample Statistics

2.3.1 Introduction

A sample statistic reveals characteristics of the sample, just as a population parameter does the same for the entire population. One must always remember, however, the two most important facts about sample statistics:

- A sample statistic is usually meant to estimate a population parameter.

- A sample statistic is a random variable.

Since a sample statistic is a random variable, it has an associated probability distribution, just like any other random variable. The probability distribution of a sample statistic is sometimes called its ***sampling distribution***. The nature of the sampling distribution is key in determining the properties of the sample statistic. These properties are important, because they tell us how well a sample statistic does its job — in other words, how well the statistic can estimate the desired population parameter.

Let's imagine (again) that we are interested in the 'typical' height of a student who attends the University of Richmond. Imagine further that we collect three samples and calculate a sample mean for each.

1. Sample **A** contains the measured heights of 10 students chosen randomly from the population; the mean of this sample is $\overline{x}_A$.
2. Sample **B**, with mean $\overline{x}_B$, contains the measured heights of 50 male students all chosen from the same dorm.
3. Sample **C**, mean $\overline{x}_C$, contains the measured heights of 50 students chosen randomly from the population.

In each case, the idea is to use the sample mean to estimate the mean μ_x of the entire population of students. Since we collected three samples, we have three different estimates ($\overline{x}_A$, $\overline{x}_B$ and $\overline{x}_C$) of the same parameter (μ_x). So now ask yourself: which of these three estimates is "best?" Most people would intuitively choose $\overline{x}_C$ as the best estimate of μ_x; the question is, why is that estimate is better than the other two?

Actually, the best estimate of μ_x would probably be obtained by *pooling* the data from samples **A** and **C**. See section 2.4 on page 46 for more details.

The answer to this question — which sample statistic is best? — can be answered by considering the characteristics of the sampling distribution of each statistic. There are generally two key parameters of the sampling distribution: location and dispersion.

1. The location of the sampling distribution, as described by the mean (or expected value) of the statistic, reveals how *accurately* the sample statistic estimates a population parameter.
2. The dispersion of a sampling distribution is described by the standard deviation of the distribution, which is often called the *standard error* of the statistic. The standard error indicates the *uncertainty* associated with the estimate of the population parameter.

Ideally, we would want the mean of the sampling distribution to be equal to the population parameter we wish to estimate. In our example, it turns out that samples **A** and **C** both give *unbiased* estimates of the true mean μ_x of the entire student body:

$$\mathrm{E}(\overline{x}_A) = \mathrm{E}(\overline{x}_C) = \mu_x$$

However, $\overline{x}_B$ gives a *biased* estimate of μ_x:

$$\mathrm{E}(\overline{x}_B) \neq \mu_x$$

Sample **B**'s estimate of μ_x is biased because the sample is not representative of the population, since only males from a single dorm were chosen.

So both $\overline{x}_A$ and $\overline{x}_C$ are unbiased estimates of the same parameter μ_x: in other words, they are each equally accurate indicators of student height,

since they are each calculated from representative samples. So why is $\overline{x}_C$ intuitively preferred to $\overline{x}_A$? The reason, as we will discuss later in more detail, is due to the relative values of the standard deviation (i.e., the standard error) of the two estimates of μ_x. The standard error of $\overline{x}_A$ is larger than that of $\overline{x}_C$:

$\overline{x}_C$ is the preferred estimate of μ_x because it is unbiased, unlike $\overline{x}_B$, and has less uncertainty than $\overline{x}_A$.

$$\sigma(\overline{x}_A) > \sigma(\overline{x}_C)$$

A larger standard error means greater uncertainty in the estimate of the population parameter. Thus, it is more likely that $\overline{x}_C$, rather than $\overline{x}_A$, will be closer to μ_x.

This is not an isolated example. There are often situations more than one statistic available to estimate a given population parameter. Recall that both eqn. 2.4a and eqn. 2.4b are available to estimate the population variance σ_x^2. Both equations give unbiased estimates of the population variance: in other words,

Equations 2.4a and 2.4b:
$s_x^2 = \frac{1}{n}\sum(x_i - \mu_x)^2$
$s_x^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$

$$\mathrm{E}(s_x^2) = \sigma_x^2$$

Producing an unbiased estimate of σ_x^2 is the reason that $n-1$ (instead of n) is used in the denominator of eqn. 2.4b.

To summarize, sample statistics are random variables with their own probability distributions, called sampling distributions. The important characteristics of a sample statistic are bias and standard error. Bias reflects the location of the sampling distribution, and whether the mean is equal to the desired population parameter. Standard error indicates the uncertainty associated with a sample statistic: the larger the standard error, the greater the uncertainty in the statistic.

Important qualities of sample statistics as estimators of population parameters:

1. Bias
2. Standard error

2.3.2 Sampling Distribution of the Mean

The Sample Mean as a Random Variable

All sample statistics are random variables, and the sample mean is certainly no exception. In fact, we have just noted that the sample mean $\overline{x}$ becomes a more certain estimate of the population mean μ_x as the number of observations in the sample increases. To understand the properties of the sampling distribution of $\overline{x}_n$, imagine the following dice-throwing experiment:

Remember that $\overline{x}_n$ denotes the mean of n measurements.

- 200 people each have a single die
- Each person tosses the die n times and then reports the mean, $\overline{x}_n$, of the values observed.
- All 200 sample means — one from each person — are collected. The mean and standard deviation of these 200 values are calculated.

This experiment was performed, using a random number generator, for $n = 1$, 5, and 20. The results are shown in table 2.1.

Examine the table carefully, for it provides a good illustration of the typical behavior of the sample mean as a random variable. In all cases, the sample mean is an unbiased estimate of the population mean: in other words, $\mathrm{E}(\overline{x}_n) = \mu_x$ for all values of n. However, the observed standard deviation of the 200 sample means decreased markedly as the value of n increased. In fact, the central limit theorem (see next section) predicts that

Table 2.1: Results of dice-tossing experiment described in the text. The table compares the mean and standard deviation of 200 values of the sample mean, $\overline{x}_n$, of n observations. In all cases, $\overline{x}_n$ provides an unbiased estimate of μ_x, but the standard error of the mean, $s(\overline{x}_n)$, decreases as n increases.

	$n = 1$ toss	$n = 5$ tosses	$n = 20$ tosses
mean, $\overline{x}$	3.32	3.42	3.49
std dev, $s(\overline{x}_n)$	1.76	0.82	0.35

the standard deviation of the sample mean of n observations is given by the following expression.

$$\sigma(\overline{x}_n) = \frac{\sigma_x}{\sqrt{n}} \tag{2.9}$$

where $\sigma(\overline{x}_n)$ is the standard deviation of the mean of n observations (i.e., the standard error of the mean) and σ_x is the standard deviation *of the individual observations* of the random variable x.

Recall that we intuitively "trust" the mean of a large number of measurements. Now we have observed that this is because the standard deviation of the sample mean — the uncertainty of our estimate of the population mean μ_x — decreases as n gets larger. In other words, ***the larger the sample, the more likely that the sample mean is close to the population mean***. Equation 2.9 provides a quantitative measure of just how much better $\overline{x}_n$ becomes as n increases.

We're not finished with the sample mean yet. Figure 2.6 shows the distribution of the values of the 200 sample means collected in the dice-tossing experiment. For a single throw of the die (i.e., $n = 1$) we would expect that any of the integers 1–6 would be equally likely, and this is more or less what was actually observed (see plot on the top of the figure). As n increases, equation 2.9 predicts that the sample means $\overline{x}_n$ cluster more closely to the population mean $\mu_x = 3.5$, and this was indeed observed (see bottom two figures). However, something interesting is happening: the distribution of values begins to resemble a normal distribution (solid line in the bottom figures) even though the probability distribution describing a single throw of the die definitely does *not* follow a normal distribution.

This behavior points to a second important aspect of the sample mean as a random variable. Not only does $s(\overline{x}_n)$ decrease as n increases (as described in eqn. 2.9) but the probability distribution of $\overline{x}$ also begins to resemble a normal probability distribution as n increases, even if the distribution of the individual observations, x, does not follow a normal distribution. These two aspects are combined in the ***central limit theorem***.

The Central Limit Theorem

The ***central limit theorem*** is concerned with the probability distribution (i.e., the sampling distribution) of the sample mean, $\overline{x}_n$, of n observations. Let x be a random variable that follows probability distribution $p(x)$, with mean μ_x and standard deviation σ_x. The sample mean of n measurements, $\overline{x}_n$, follows a probability distribution $p(\overline{x}_n)$ with mean $\mu(\overline{x}_n)$ and standard deviation $\sigma(\overline{x}_n)$.

The central limit theorem can be stated as follows:

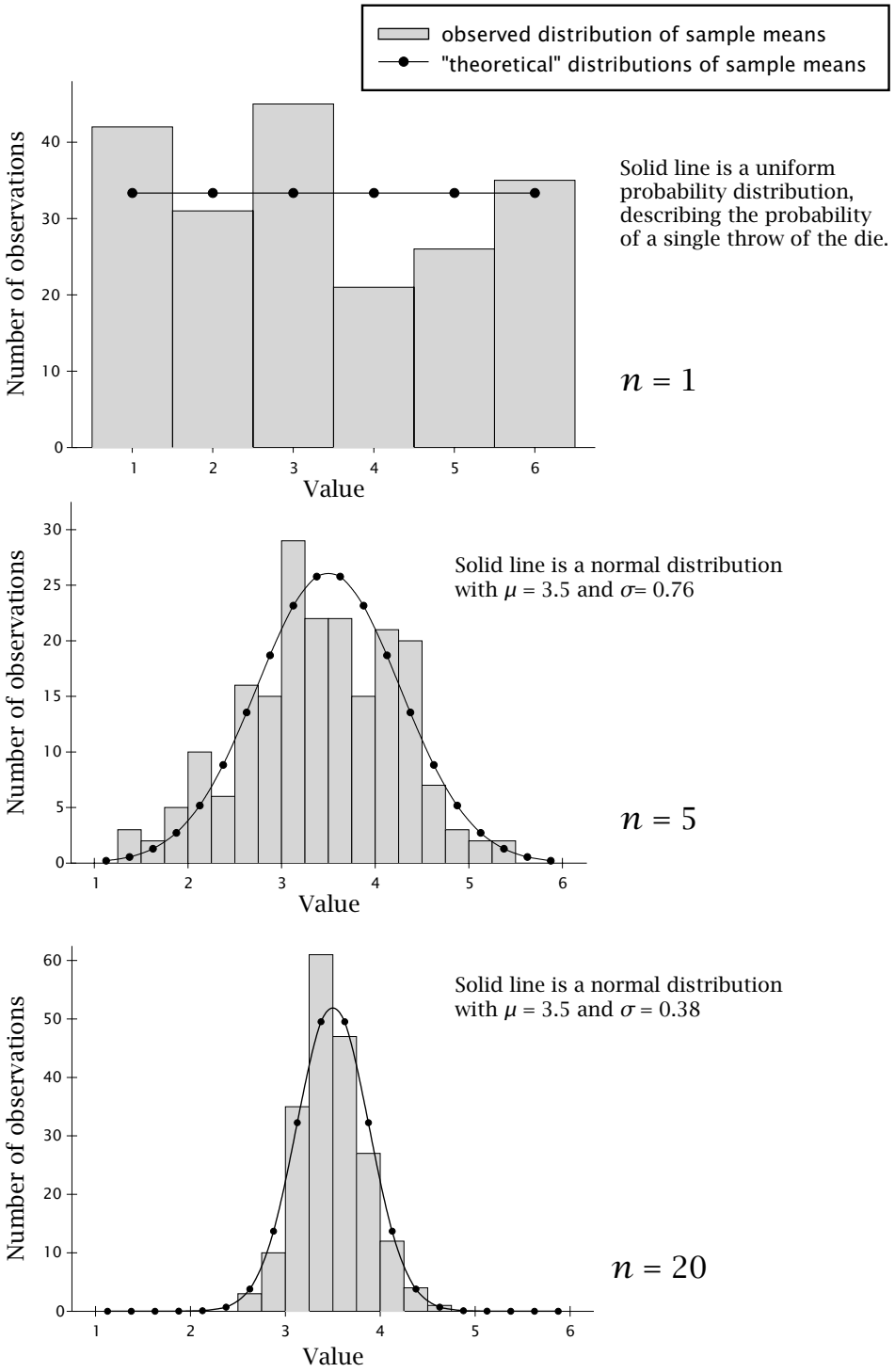

Figure 2.6: Results of dice-tossing experiment described on page 41. The bars show the distribution of 200 sample means of n measurements. As n increases, the distribution of sample means begins to follow a normal probability distribution even though the *original* distribution of the random variable (top figure) is distinctly non-normal.

The probability distribution, $p(\overline{x}_n)$, of the sample mean of n observations of a random variable x will tend toward a normal distribution with

$$\mu(\overline{x}_n) = \mu_x$$
$$\sigma(\overline{x}_n) = \frac{\sigma_x}{\sqrt{n}}$$

The central limit theorem has three very important aspects:

1. The means of the probability distributions of x and $\overline{x}_n$ are the same for any value of n. In other words, the sample mean is an *unbiased* estimate of the population mean.
2. The standard error of the mean of n observations, $\sigma(\overline{x})$ can be calculated from the standard deviation, σ_x, of the *individual* observations.
3. The probability distribution of the sample mean, $p(\overline{x}_n)$, will tend to be a normal distribution *even if the* $p(x)$ *is not a normal distribution.*

This last aspect of the central limit theorem is one of the reasons that we usually assume a normal probability distribution for measurements. Even if the original observation are not exactly normally-distributed, *their means are*, if the sample is large enough. The distribution of the mean will more closely approximate a normal distribution as n increases and as the probability distribution of the original variable x more closely resembles a normal distribution. Note that if $p(x)$ is a normal distribution, then $p(\overline{x})$ is also normal, for *any* value of n.

Figure 2.7 compares the probability distribution of x (a normally distributed variable) with the probability distributions of the sample mean $\overline{x}_n$ for various values of n. As the number of observations in the sample mean increases, the sampling distribution narrows.

Example 2.2

The levels of cadmium, which is about ten times as toxic as lead, in soil is determined by atomic absorption. In five soil samples, the following levels were measured:

8.6, 3.6, 6.1, 5.5, 8.4 ppb

Report the mean concentration of cadmium in these soil samples, as well as the standard deviation of the mean.

Remember that the standard deviation of a sample statistic is called the *standard error* of the statistic.

This is an important example, because when reporting the standard deviation of an analyte concentration determined in the laboratory, it is the standard deviation *of the mean* that is usually given.

The reported cadmium concentration in the soil would be the mean of the five measurements, calculated by the usual method.

$$\overline{x} = \frac{1}{5}(8.6 + \cdots + 8.4) = 6.44\,\text{ppb}$$

The standard deviation is the root of the variance calculated from eqn. 2.4b

$$s_x = \sqrt{\frac{1}{4}\left[(8.6 - 6.44)^2 + \cdots + (8.4 - 6.44)^2\right]} = 2.10\,\text{ppb}$$

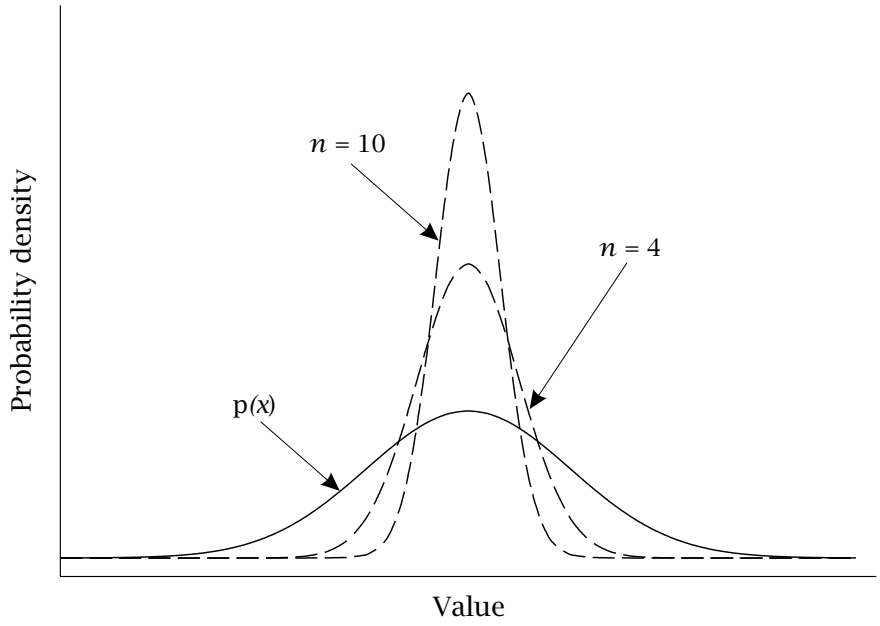

Figure 2.7: Comparison of probability distributions of individual observations (x, solid line) and sample means ($\overline{x}$, dashed lines). As the number of observations n in the sample increases, the sample mean $\overline{x}$ tends to be closer to the population mean μ_x.

But we need the standard deviation *of the mean of the 5 measurements*:

$$s(\overline{x}) = \frac{s_x}{\sqrt{5}} = 0.94\,\text{ppb}$$

Thus, the analyst would estimate the cadmium level in the soil at 6.44 ppb. The standard deviation of this estimate is 0.94 ppb.

It is very important to understand why the estimated error of this value is 0.94 ppb, and not 2.10 ppb. The latter value represents the spread in the five *individual* measurements. The standard error of 0.94 ppb, however, represents our best guess of the spread of the *mean* of five measurements. In other words, if another five soil samples were collected and the cadmium levels were measured again and averaged, and this entire process were repeated many times, the standard deviation of the mean concentrations would be close to 0.94 ppb.

Standard Error and Significant Figures

In example 2.2, the sample mean was 6.44 ppb, and the standard deviation (i.e., the standard error) of the mean was 0.94 ppb. The standard error indicates the magnitude of the uncertainty in the calculated mean.

Before continuing, we will now address the following question: how many significant figures should be used for a measurement mean and its standard error? The concept of significant figures is a crude attempt to indicate the precision of a measurement: the last significant digit of a measurement is the first "uncertain" digit of the number.

The use of standard deviation supersedes the concept of significant figures as an indicator of measurement precision. The standard deviation is much superior in this respect. Likewise, the rules governing the effect of

calculations on the significant figures are actually a crude attempt at error propagation calculations (covered in section 3.1.4).

Imagine that we calculate a measurement mean of 12.11254 cm and a standard error of 1.22564 cm. In this case, the size of the standard error makes reporting the mean to five decimal places look a little foolish. So, how many digits to keep? An argument can be made that the standard error should always have only one digit, and that the mean should be reported to the same precision. With this guideline, we would state the mean as 12 cm, with a standard error of 1 cm. The argument for this approach is that the standard error determines the final significant digit in the sample mean.

My own approach is to keep two digits in the standard error, so that we would report a mean and error of 12.1 cm and 1.2 cm, respectively. The reasoning behind this practice is that the standard error and the mean are often used to construct intervals, in hypothesis testing, in error propagation calculations, or for other purposes. Providing an extra digit avoids rounding error in such manipulations.

In the next chapter we discuss a very useful way to report measurements (as confidence intervals), and we will revisit the question of how many significant figures to retain in reported measurements. At this point it is sufficient for you to realize that the concept of significant figures is simply a crude attempt to indicate the precision of measured values. The use of standard errors and confidence intervals, however, is a much more refined indication of precision.

2.4 Combining Samples: Data Pooling

2.4.1 Introduction

Data pooling — combining observations from more than one sample — yields a pooled statistic, which is often the best estimate of a population parameter.

Experimental measurements may be quite expensive and labor-intensive to obtain. However, we generally want to have as many measurements as possible in order to obtain a reliable estimate of the property we are measuring. For this reason, measurements from a variety of different sources are sometimes combined to obtain a superior estimate, called a ***pooled statistic***. For example, we may collect a number of lake samples and measure the lead concentration in each sample; we might then combine these results to obtain a better estimate of the concentration of lead in the lake. This process of combining measurements that have been obtained under different conditions is known as ***data pooling***.

The process is best illustrated by an example.

Example 2.3

You wish to estimate the mean height of the students attending the University of Richmond. You divide the work between two assistants. The first chooses eight students randomly and measures their height. The other assistant is a little more lazy: he only manages to measure the heights of three students, (again chosen randomly from the population).

Sample 1: 67.97, 67.63, 74.67, 64.58, 63.93, 66.67, 68.95, 68.66 in.
Sample 2: 71.70, 65.01, 63.46 in.

Obtain the best possible estimate of the mean height of the entire population of students attending UR.

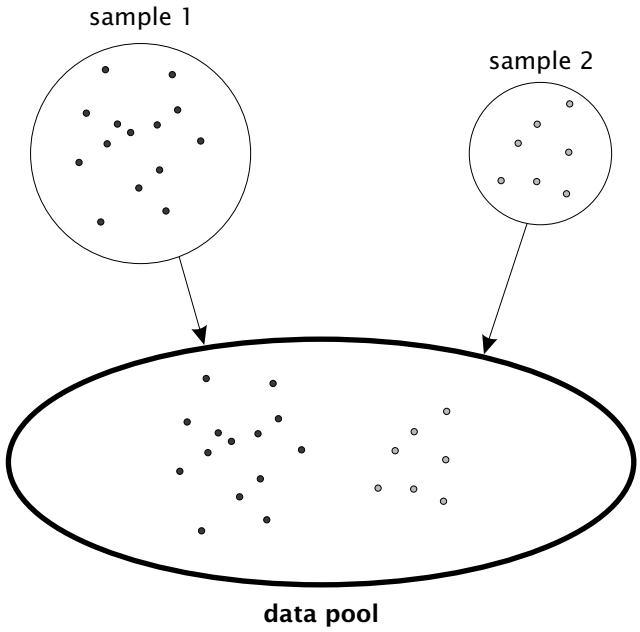

Figure 2.8: Data pooling. In this process, the observations of two or more samples are used to calculate a pooled sample statistic.

It is easy enough to calculate the means for the two samples. Let's also calculate the standard errors of these sample means.

$$\text{Sample 1, } n = 8: \quad \overline{x} = 67.88\,\text{in}, \quad s_x = 3.29\,\text{in}, \quad s(\overline{x}) = \frac{s_x}{\sqrt{8}} = 1.16\,\text{in}$$

$$\text{Sample 2, } n = 3: \quad \overline{x} = 66.73\,\text{in}, \quad s_x = 4.38\,\text{in}, \quad s(\overline{x}) = \frac{s_x}{\sqrt{3}} = 2.53\,\text{in}$$

So we have two estimates of the population mean μ_x: 67.88 in and 66.73 in. If we were forced to choose between these two estimates, then we would choose the first, since it is the mean of a larger sample and consequently has the smallest standard error. However, it is possible to get an even better estimate of μ_x by data pooling.

The problem with using the first sample to estimate μ_x is that we are essentially throwing away valuable information — the three measurements taken by the second assistant. It is better to use the measurements from both samples; this is data pooling, and in this case it will yield a *pooled mean*, $\overline{x}_{pool}$ as the estimate of the population mean, μ_x. The concept of data pooling is illustrated in figure 2.8.

For this example, the easiest way to pool the data is to simply to combine all the observations and then calculate the mean of all 11 measurements.

$$\begin{aligned}\overline{x}_{pool} &= \frac{1}{11}\sum[(67.97 + \cdots + 68.66) + (71.70 + \cdots + 63.46)] \\ &= 67.57\,\text{in}\end{aligned}$$

The best estimate of the mean height of the population is thus 67.57 in, the pooled mean. This is a better estimate than either of the sample means because more observations are used in its calculation (i.e., the standard error of the pooled mean is less than the standard error of either of the unpooled estimates).

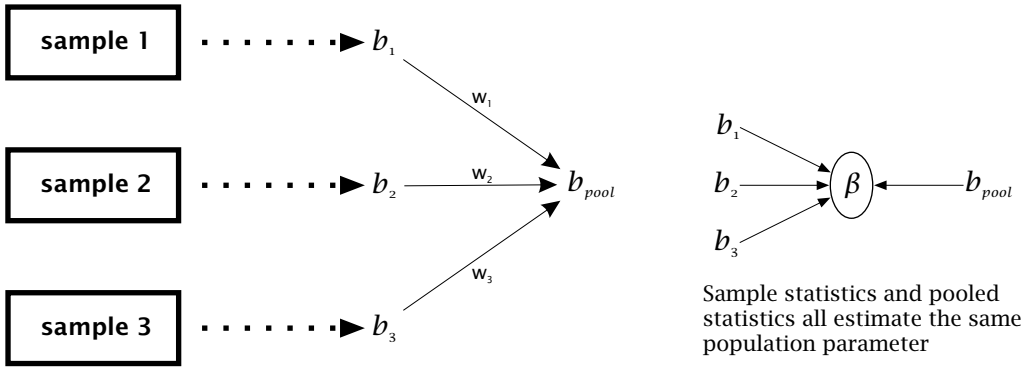

Figure 2.9: Data pooling: calculating a pooled statistic from sample statistics. Sample statistics are combined in a weighted average (eqn. 2.10) to produce a pooled statistic. The pooled statistic estimates the same population parameter as the sample statistics.

2.4.2 Pooled Statistics

A ***pooled statistic*** is an estimate of a population parameter that is calculated using the observations in more than one sample. The philosophy of a pooled statistic is that it provides the best estimate of the population parameter because it combines the information from the largest pool of observations.

Let's say that we need to obtain an estimate of the population parameter β. We collect k random samples and calculate a sample statistic b for each sample; each of these statistics must provide an unbiased estimate of the same population parameter β. We can calculate a pooled estimate of the population parameter by calculating a weighted average of the individual sample statistics:

One way to obtain a pooled statistic is calculate the weighted average of multiple sample statistics.

$$\beta_{pool} = \sum_{i=1}^{k} w_i b_i \tag{2.10}$$

where $\sum_{i=1}^{k} w_i = 1$. The weights w_i are determined by the standard error in the individual sample statistics b_i.

> **Important**: combining sample statistics in this manner to calculate a pooled estimate only makes sense if the sample statistics all estimate the same population parameter.

If the individual sample statistics b_i each provide an unbiased estimate of the shared population parameter β, then the pooled statistic b_{pool} also provides an unbiased estimate of β. See figure 2.9.

The population parameters we are most often interested in are the mean μ_x and variance σ_x^2. The general expression for calculating a pooled mean

from k sample means is

$$\overline{x}_{pool} = \sum_{i=1}^{k} \frac{n_i}{n_{tot}} \overline{x}_i \qquad (2.11)$$
$$= \frac{\sum n_i \overline{x}_i}{\sum n_i}$$

where $\overline{x}_i$ is the mean of the i^{th} sample, which contains n_i observations, and n_{tot} is the total number of observations in all k samples.

In example 2.11 we could have calculated the pooled mean by using eqn. 2.11:

$$\overline{x}_{pool} = \frac{8 \cdot 67.88 + 3 \cdot 66.73}{8 + 3}$$
$$= 67.57 \text{ in}$$

As you can see, we obtain the same result as in example 2.3. This method of calculating the pooled mean is useful in situations where only the sample means are available (and not the individual estimates).

The *pooled variance* can be calculated by using the following equation

$$s^2_{pool} = \sum_{i=1}^{k} \frac{\nu_i}{\nu_{tot}} s_i^2 \qquad (2.12)$$
$$= \frac{\sum \nu_i s_i^2}{\sum \nu_i}$$

where s_i^2 is the variance of the i^{th} sample, which has $\nu_i = n - 1$ is the degrees of freedom, and ν_{tot} is the degrees of freedom of the pooled sample variance, calculated by summing the individual ν_i values:

$$\nu_{tot} = \sum_{i=1}^{k} \nu_i$$

For small sample sizes, the standard error of the sample variance is notoriously large. Thus, it is beneficial to pool data when possible, as demonstrated in the following example.

Example 2.4

The following results show the percentage of the total available interstitial water recovered by centrifuging samples taken at different depths in sandstone:

Depth, m	Water recovered, %	Mean	Variance
7	33.3, 35.7, 31.1	33.37	5.29
8	43.6, 45.2	44.4	1.28
16	73.2, 68.7, 73.6, 70.9	71.6	5.15
23	72.5, 70.4, 65.2	69.37	14.12

Assuming that the variance for this method of water measurement is constant for all the data, calculate the pooled variance. How many degrees of freedom are in this statistic?

Four samples were obtained, and it is clearly stated that we are to assume that the variance of the observations are the same for all the samples; this is an important assumption, because it means that it is appropriate to calculate a pooled variance. For each of the samples, the variance was calculated. Each of these sample variance estimates the same population variance, σ_x^2. However, we may use eqn. 2.12 to obtain a better estimate.

Presumably, eqn. 2.4b was used to calculate the sample variances (since the population mean for each sample is unknown) and so the degrees of freedom, ν, for each sample variance, s_i^2, is given by $\nu_i = n_i - 1$. With this in mind,

The pooled variance is a weighted average of the individual sample variances.

$$\begin{aligned} s_{pool}^2 &= \frac{2 \cdot 5.29 + 1 \cdot 1.28 + 3 \cdot 5.15 + 2 \cdot 14.12}{2 + 1 + 3 + 2} \\ &= \left(\frac{2}{8}\right) 5.29 + \left(\frac{1}{8}\right) 1.28 + \left(\frac{3}{8}\right) 5.15 + \left(\frac{2}{8}\right) 14.12 \\ &= 6.95\,(\%)^2 \end{aligned}$$

The pooled variance has eight degrees of freedom ($\nu = 8$). The larger the degrees of freedom for variance estimates, the smaller the standard error of the statistic; thus, the pooled variance is superior to any of the individual sample variance values.

A pooled estimate, s_{pool}, of the population standard deviation (σ_x) can be obtained by calculating the positive root of the pooled variance.

Example 2.5

Assume the following measurements, all made using the same analytical method, have the same standard deviation, σ. Provide the best estimate of the standard deviation of the method. How many degrees of freedom are in this estimate?

32.5; 36.6; 35.7 24.9; 21.7 98.2; 102.1

Certainly we may calculate the standard deviation of each set of measurements, yielding $s_1 = 2.15$, $s_2 = 2.26$ and $s_3 = 2.76$. These sample statistics all provide estimates of the shared standard deviation, σ, of the measurements. Of these statistics, s_1 gives the best estimate, since it has two degrees of freedom (the other estimates only have one).

However, if we use s_1 as our estimate, we are essentially wasting four measurements. *Since all the measurements share the same standard deviation*, we may pool the data to provide a superior estimate.

$$\begin{aligned} s_{pool} &= \sqrt{\frac{2 \cdot s_1^2 + 1 \cdot s_2^2 + 1 \cdot s_3^2}{2 + 1 + 1}} \\ &= \sqrt{\frac{2 \cdot 4.64 + 1 \cdot 5.12 + 1 \cdot 7.61}{4}} \\ &= 2.35 \end{aligned}$$

There are four degrees of freedom in this estimate of σ, providing the best estimate of the shared population parameter.

You should learn to spot opportunities for data pooling; in real life, data does not come with a sign that reads “pool me.” Remember: when we are pooling data, there is an implicit assumption that the data share the same

value for the population parameter; indeed, it is this shared population parameter that we are trying to estimate. In the last two examples, it was stated that all measurements shared the same variance, so it is appropriate to pool all the measurements to calculate s^2_{pool} using eqn. 2.12. However, in both cases it would have been *inappropriate* to combine the measurements to calculate a pooled *mean*, since the individual samples did not share a common population mean, μ_x.

2.5 Summary and Skills

Many experiments in science produce random variables. The outcomes of these experiments, usually measurements of some kind, are called *observations* of the random variable. The collection of all possible observations is an experiment's *population*. Normally only a small fraction of the population, the *sample*, is actually produced by an experiment.

Characteristics of a population, such as location and dispersion, are described by *population parameters*. The goal of most experiments is to draw conclusions about one or more population parameters. These parameters — such as the population mean, μ_x, and population standard deviation, σ_x — could be calculated if the values of all the observations in the population were known. Since that is not normally the case, *sample statistics* must be used to estimate population parameters. The two most important sample statistics are the sample mean, $\overline{x}$, and the sample standard deviation, s_x. If several samples produce sample statistics that estimate the same population parameter, these sample statistics can be combined to produce a *pooled statistic* that is often the best possible estimate of the population parameter.

Sample statistics are themselves random variables. Like any other random variable, the sample statistics have an associated uncertainty, characterized by the *standard error* (i.e., the standard deviation) of the statistic. Also like any other random variable, sample statistics each have their own probability distribution, called the *sampling distribution* of the statistic. Perhaps the most important is the distribution of the sample mean of n measurements, $\overline{x}_n$. The *central limit theorem* describes the characteristics of this distribution. The central limit theorem states that the distribution of $\overline{x}_n$ tends toward a normal distribution with a mean of μ_x and a standard deviation of $\frac{\sigma_x}{\sqrt{n}}$.

Important skills developed in this chapter:

1. Ability to calculate $\overline{x}$ and s_x with a calculator.
2. Ability to calculate $s(\overline{x})$, and to use the proper number of significant figures to report both $\overline{x}$ and $s(\overline{x})$
3. Ability to calculated pooled statistics (especially s_{pool}) and to recognize when it is appropriate to do so.