Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 4: Estimation

(I.) Overview of Estimation

In most studies or experiments, we are interested in investigating one or more variables whose values change across units of the population. These variables are referred to as "random variables," which were defined formally in Lecture 2. However, a more intuitive and less formal definition of a random variable is 'a function that measures some characteristic of a unit and whose outcome is not known (cannot be predicted with certainty) until it is measured.' In our data, which represent a sample from the underlying population, we observe various values of the underlying random variables(s); the observed values of the variable(s) that occur in a sample are often referred to as "realisations" of the underlying random variable. In general, random variables will be denoted by capital letters (e.g., *X, Y, Z*), and lower cases letters (e.g., *x, y, z*) will denote values or realisations of a random variable. One exception is the letter *N*, which is commonly used to denote the population size, rather than a r.v., while *n* will almost always refer to the sample size. As has been stated before, in this course we will address only situations in which the population size is, for all practical purposes, infinite (i.e., $N=\infty$).

Turning to the central topic of estimation, we reiterate that estimation techniques involve using the data to make a 'best guess' at the population attribute(s) we hope to deduce. For now, we will say that this guess is 'best' in the sense that it is a good representative of the underlying population attribute(s) of interest. In this lecture, we will focus on deducing only parameters, rather than structural properties, of the underlying random variable(s) of interest. In other words, in this lecture, we will be concerned with "point estimation." We should note, however, that it is, in general, possible to estimate various structural properties of the underlying random variable(s); as an example, "kernel density estimation" refers to using the sample realisations of a continuous random variable to estimate the population density curve (i.e., the pdf) for that variable. Since we will concern ourselves with estimating population parameters in this lecture, we should note once more that a parameter is a numerical characteristic of the population of interest or, identically, a numerical function of the random variable(s) of interest. Usually, a parameter is denoted by a Greek letter (e.g., $\beta, \theta, \lambda, \sigma, \mu, \tau$, etc.). The parameter(s) in

which we are interested might be general population quantities, such as the population median or mean, or, in cases where we assume a specific pdf family for our variable(s) of interest, the distributional parameters of that family. Turning from the underlying population to the sample of data drawn from that population, we remind ourselves that an "estimator" is a 'guess' at the specific population attribute(s) of interest that is calculated from the data sample.

For a given population attribute of interest, many different possible 'guesses' or estimators may exist. Here, we should remind ourselves that an estimator can be constructed using either an analytical or a simulation-based approach. Each of the possible estimators for a given attribute has a number of associated properties; these properties will be described later in this lecture. Perhaps the most important property of an estimator is its "error," which is used to give an indication of the precision and reliability of the estimator; in fact, in many scientific papers, estimates of population attributes of interest must be accompanied by their errors. An estimator's error, as well as some of its other properties, provide more technical criteria for judging which of the possible estimators for an underlying attribute is the 'best' one.

Throughout this lecture, we will use an example to illustrate both the construction of estimators and the determination of their various properties. In this example, we are interested in estimating the location parameter for one underlying random variable. For now, we will assume only that the population random variable of interest has a distribution (pdf) that is symmetric in form; for this reason, the location parameter in which we are interested can be thought of as the 'centre' of underlying population distribution or, alternatively, as the variable's population mean. Also, we will assume, as is almost always done throughout this course, that the underlying population is virtually infinite. Lastly, we will assume that we have a sample consisting of n realisations of the random variable of interest. In other words, to use a term that will be defined momentarily, we have a "random sample" of size n for our variable of interest.

(II.) Sampling

Before proceeding to a discussion of estimators and their various properties, we return to the dichotomy between populations and samples, which is reflected in the link between probability and sampling. In general, we will not know the underlying population completely, and, therefore, we will have to make inferences based on a sample. In practice, we will often take a "random sample" from the underlying population of interest; in fact, many statistical inference methods require that the data be a random sample from the underlying random variable(s) of interest. A "random sample" has already been defined in Lecture 1. However, here, we offer an alternate definition: a "random sample" of size n from a variable X (with pdf f) is a set of n random variables that each have pdf f and are (statistically) independent of each other. We will denote a random sample of X as $\{X_1, X_2, ..., X_n\}$. Note that we have used capital letters (which imply population) since a random sample is defined as a set of random variables, NOT as

a set of values of those variables. The values of these n random variables that are observed in a sample are denoted by $\{x_1, x_2, ..., x_n\}$.

Returning to the link between probability and sampling, suppose that a particular random variable, X, is assumed to have a normal distribution with mean 0 and variance 4. This is a plausible distributional assumption for many variables, an example being measurement errors, which one would hope would be centred around 0. In general, making a distributional assumption means selecting a model that describes the range of possible values of X that are likely to occur. For our specific $N(0,4)$ distributional assumption, it is easy to see that most values of X will lie between –6 and 6, although the most likely values will be between –1 and 1, say. Note that although in this case we know the exact form (i.e., the pdf) of the underlying variable of interest, in the practice of data analysis we almost never know the underlying population pdf and instead have only a sample of realisations of the underlying variable of interest. However, rather than using an observed sample to deduce the properties of an unknown underlying distribution as is done in practice, in this case, we can do the reverse and generate samples from the known underlying normal distribution for X. Suppose that we extract four samples of size n=10 from this underlying normal distribution; four such samples are shown in the following figure. In this figure, the curve represents the $N(0,4)$ pdf for X, and the triangles represent the realisations of X that occur in the samples.

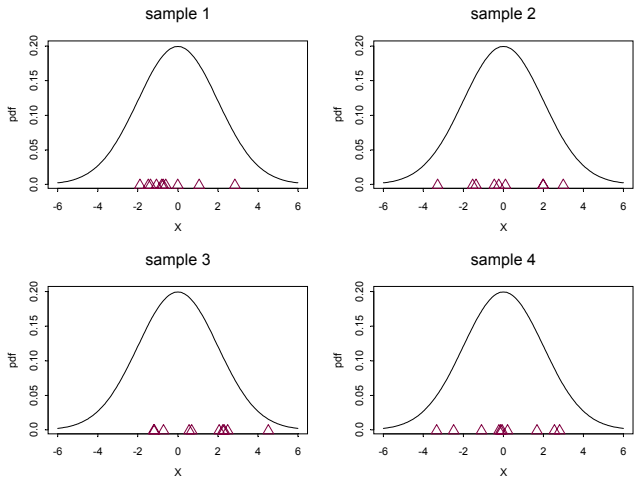

***Figure 1**: Four samples of size 10 from N(0,4)*

Alternatively, extracting four samples of size n=100 from this underlying normal distribution might result in the following four samples, which are shown in the figure below.

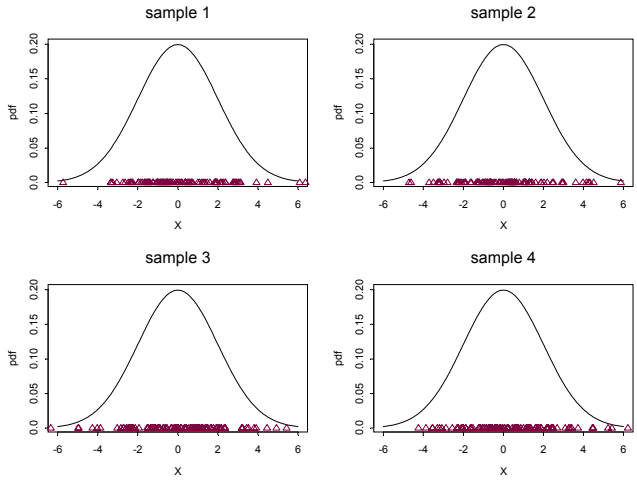

Figure 2: *Four samples of size 100 from N(0,4)*

A comparison of the two above figures shows that the bigger the sample size, the better the values in the underlying population are represented. It is also obvious that, for a fixed sample size, different sets of population values can be represented in the various samples that can possibly be drawn. It is a fact, however, that in most cases this sampling variability decreases as n increases. For instance, for a very small n, the various samples that can be extracted from a population may differ greatly from each other; this is not generally true for large n.

Leaving this example, we know that, in practice, we will not know the exact form of the underlying population pdf, and there will usually be only **one** sample of size n available for analysis. As stated many times before, this sample is usually statistically analysed in order to make inferences about the underlying population attribute(s) of interest. However, the resulting conclusions may be subject to two types of error, both of which arise through the process of sampling. The first type of error is termed "sampling bias" and refers to the quality of the observed values in a sample in terms of how adequately they represent the underlying population. Obviously, if the values in a sample are unrepresentative of the values in the population, then the conclusions drawn from the sample values using inference techniques will also be biased. Sampling bias results when the data sample is selected in an unrepresentative manner from the underlying population. A famous example of sampling bias occurred during the 1948 U.S. presidential election when a sample survey of voter opinion, conducted by telephone, projected that Dewey, the Republican candidate, would win the election. Of course, since the telephone was still somewhat of a novelty then, the survey underrepresented less wealthy voters, who tended to be Democrats, and thus incorrectly predicted the winner of the election since the majority of Americans actually favoured Truman, the Democratic candidate. The other source of potential error is termed "sampling error" and is caused by "sampling variation." Sampling variation refers to the fact that samples of the same size from the same underlying population differ in terms of the values (realisations) they contain, which therefore means that the results of using a certain inference method with the data will vary from sample to sample. [This phenomenon of sampling variation was

just illustrated by the *N(0,4)* example above.] In fact, it might be the case that the (only) sample we took was, by pure chance, a sample that did not represent the population in a fair way. As an extreme example, suppose that the variable of interest is the height of male undergraduate students at Oxford and that, by pure chance, our sample happened to contain numerous members of the University basketball team! This sample would result in misleading conclusions about the mean male undergraduate height; thankfully, however, this sample is very unlikely to occur. Note that in the case of sampling error, the values in our sample might not represent the underlying population well purely by chance, but in the case of sampling bias, the values in our sample might be unrepresentative because of the way in which the sample was collected (i.e., the "design of the experiment").

In practice, there is no way to check whether the values in our sample are representative enough because doing so would require knowing the underlying population, in which case sampling would be completely unnecessary.

(III.) Estimators

A "statistic" is simply any mathematical function of the data in a sample, or, identically, any mathematical function of the realisations of the random variables in a sample. Note that, because a statistic is based on random variables, it is itself a random variable, and, therefore, has an associated probability distribution. Examples of statistics include the sample skew, the sample median, and the sample maximum for a certain data set.

An "estimator" is a statistic that is specifically designed to measure or to be a 'guess' at a particular parameter of a population. Since estimators are a special case of statistics, they are also random variables and therefore have associated probability distributions. In general, an estimator is designated by either the Greek letter for the corresponding population parameter, topped by a hat (^) or a twiddle (~), or by a Roman letter. For instance, the estimator for the population mean, μ, is often denoted by $\hat{\mu}$; alternatively, S^2 denotes a particular estimator of the population variance. Here, we should note an important dichotomy: the distinction between an "estimator" and an "estimate." In general, the *form* of an expression for estimating an unknown parameter of interest is termed an "estimator"; however, an "estimate" is the particular realisation or value of an estimator that occurs once the sample data is plugged into the estimator expression. If an uppercase Roman letter is used to denote an estimator for an attribute of interest, then the corresponding lowercase Roman letter is used to designate a particular realisation of that estimator for a given sample of data. For instance, as stated above, S^2 denotes one possible estimator of the population variance, and s^2 denotes the value taken by this estimator for a particular sample of data from the underlying population.

In our main example, there are clearly several different estimators that could possibly be used to 'guess at' the centre of the underlying distribution; examples include the sample mean, the sample median, and the 'trimmed mean," which is calculated by taking the

average of all the observed sample values of the variable except for the very largest and very smallest values. Although use of the sample median or the trimmed mean may proffer certain advantages, such as robustness to outliers in our data sample, in this lecture, we will use the sample mean as the estimator for the population mean in our central example. The formula for the sample mean is:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

where X_i denotes one of the n realisations of X that occurs in our sample (i.e., one of the n components of our random sample). Note that $\overline{X}$ is defined in terms of the random variables $\{X_1, X_2, \cdots, X_n\}$ and **not** in terms of the sample values or realisations of those variables that occur in a sample (i.e., $\{x_1, x_2, \cdots, x_n\}$). Using the latter set of values in the formula for $\overline{X}$ would yield the sample value of $\overline{X}$, namely $\bar{x}$. Note also that the sample mean estimator was constructed analytically, as can be seen by the fact that it takes the form of an explicit mathematical formula into which the data values need merely be plugged. Although it is possible to construct an estimator using simulation techniques, no examples of doing so will be presented in this course because of the level of statistical sophistication required to understand the motivation for and the process of doing so.

(IV.) The Sampling Distribution of an Estimator

Clearly, the values of a statistic change from sample to sample since the values contained in different samples of the same size vary, especially when n is small; the same is true for the values of estimators since an estimator is a statistic. The "sampling distribution" of a statistic (estimator) is a probability distribution that describes the probabilities with which the possible values for a specific statistic (estimator) occur. The exact form of the sampling distribution for a given estimator will depend on the underlying population distribution from which the data were drawn. In general, knowing an estimator's sampling distribution is both useful and often necessary for constructing confidence intervals and hypothesis tests based on that estimator in order to draw inference about its corresponding population parameter. In constructing those entities, a statistician will be often interested in determining the probability that the distance between an estimator and the true parameter it seeks to estimate is smaller than a certain amount. Although there are mathematical results, such as Chebyshev's Inequality, that allow one to get an idea of this probability even if the estimator's exact sampling distribution isn't known, the sampling distribution for the estimator is usually necessary if one wants to determine this probability precisely. Note that this probability will depend not only on the form taken by the sampling distribution of the estimator, but also on the particular population parameter being measured by the estimator.

We will explore the concept of a sampling distribution using the following examples, both of which involve known populations of finite size ($N < \infty$). These examples are just for the purpose of illustration because we almost never know the exact form of the underlying population and because, in this course, we will almost always assume that the population

size is, for all practical purposes, infinite. This said, turning to our first example of sampling distributions, suppose that we have a population consisting of the values {2,3,5,7,9,10}. Suppose further that we want to take a sample of size n=2, "without replacement," from this population; sampling "without replacement" means that each unit in the population can appear only once in a sample. There are $\binom{6}{2} = \frac{6!}{2!\,(6-2)!} = 15$ different samples of size n = 2 that can be taken without replacement. If we calculate the sample mean, minimum, and maximum for each of the 15 possible samples, we get the following results:

sample	mean	min	max	sample	mean	min	max
(2,3)	2.5	2	3	(3,10)	6.5	3	10
(2,5)	3.5	2	5	(5,7)	6.0	5	7
(2,7)	4.5	2	7	(5,9)	7.0	5	9
(2,9)	5.5	2	9	(5,10)	7.5	5	10
(2,10)	6.0	2	10	(7,9)	8.0	7	9
(3,5)	4.0	3	5	(7,10)	8.5	7	10
(3,7)	5.0	3	7	(9,10)	9.5	9	10
(3,9)	6.0	3	9	-	-	-	-

From the above table we see that the sampling distributions of the sample mean, the sample minimum, and the sample maximum are, respectively:

Mean:

value	2.5	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.5
prob.	1/15	1/15	1/15	1/15	1/15	1/15	3/15	1/15	1/15	1/15	1/15	1.15	1/15

Minimum:

value	2	3	5	7	9
prob.	5/15	4/15	3/15	2/15	1/15

Maximum:

value	3	5	7	9	10
prob.	1/15	2/15	3/15	4/15	5/15

Another example of sampling distributions involves a larger number of population values, in this case the failures times (in hours) of 107 units of a piece of electronic equipment. Here, we will pretend that these times comprise the complete population. The following graph shows a histogram of the population values.

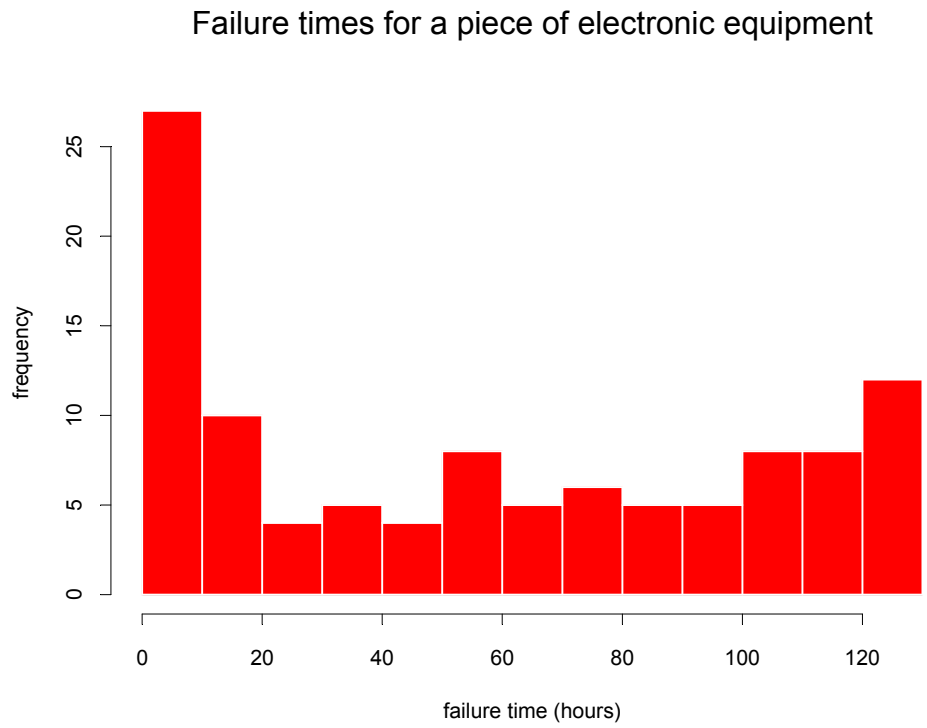

***Figure 3**: Histogram of 107 failure times*

There do not seem to be any outliers (i.e., values that are surprising relative to the majority of the population) in our population of times. In addition, the right-hand tail of the population distribution is very long. Several descriptive parameters for this population are:

n	Min	Q1	Med	Mean	Q3	Max	SD	CV
107	1	10	53.3	56.99	101.7	129.9	44.87	0.79

Note that the median and the mean are quite similar, even though the distribution of times is clearly asymmetric. Next, suppose that we extract multiple samples of several sizes (without replacement) from this population and that, for each of the samples, we calculate the sample mean and the sample maximum. Specifically, suppose that we generate 1,000 random samples of each of several sample sizes (e.g., 2, 4, 8, 16, 32, and 64) from the population of 107 failure times; note that, for almost any sample size, there are many, many different possible samples that can be generated, even though the underlying population size is reasonably small. [Of course, in practice, we would have only one sample from the underlying population.] Lastly, suppose that after generating the samples, we calculate the sample mean and sample maximum for each sample. Figure 4 uses true histograms to show the empirical sampling distributions of the sample mean for each of the various sample sizes.

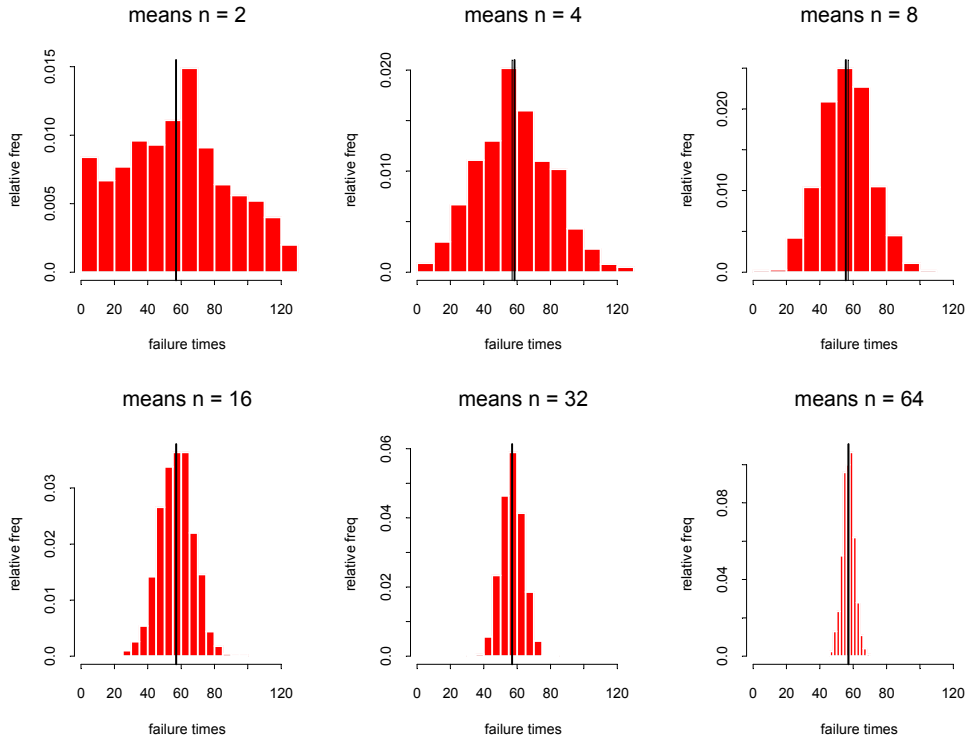

***Figure 4**: Sample means for 1000 samples from a population of 107 failure times*

In the above graph, the darker vertical lines show the population mean (56.99) and the mean of the 1,000 sample means for each sample size. Note that, even for samples of size 2, both lines are virtually indistinguishable. Further, for each of the sample sizes, the empirical sampling distribution of the sample mean is more or less symmetrically distributed around the true population mean. In addition, note that as the sample size increases, the empirical distribution of the sample mean becomes more tightly clustered around the true population mean; in other words, as the sample size increases, the spread of the (empirical) sampling distribution decreases.

Next, we turn to the empirical sampling distributions of the sample maximum. Figure 5 presents, for each of the six different samples sizes, a true histogram of the 1000 sample maxima. For each of these histograms, the population maximum (129.9) is marked as a dark vertical line at the right hand side of the histograms, and the other vertical line corresponds to the mean of the 1000 sample maxima for that sample size.

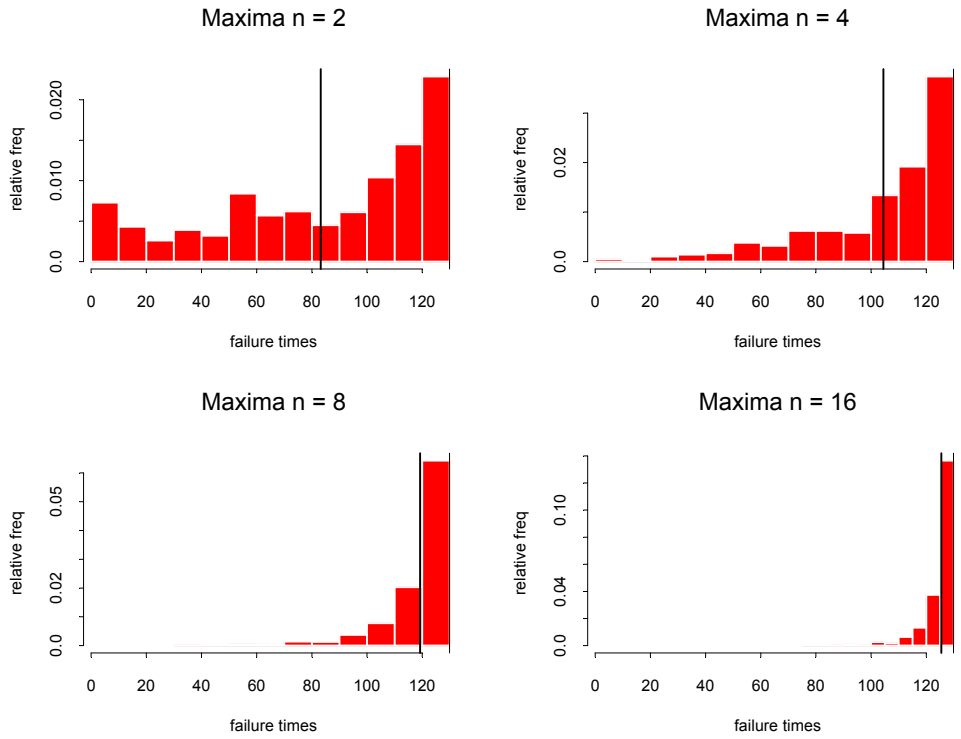

Figure 5: *Sample maxima for 1000 samples from a population of 107 failure times*

Note that, in the above figure, the two lines do not coincide as closely as they did for the sample mean histograms, especially for small sample sizes. Further, note that, for any sample size, the empirical sampling distribution of the sample maximum is not symmetrical around the true population maximum (or around any value, for that matter); in fact, these empirical sampling distributions are decidedly asymmetrical.

Now, let us return to the central example in this lecture, that of estimating the population mean of a random variable. Previously, we had decided to use the sample mean as an estimator of the variable's population mean. Thus, we may want to know the sampling distribution of the sample mean. However, in our example, as is generally true in practice, we only have a sample from the underlying population and do not know all the values in the underlying population; thus, we cannot generate samples from the known population in order to get an idea of the sampling distribution of a certain statistic, as was done in the previous two examples. Therefore, we will have to use the statistical theory of the sample mean in order to describe its sampling distribution. This theory is outlined in the following section, which represents a slight digression from our central topic of estimation. Note that, for the sample mean, we can determine its sampling distribution analytically using mathematics and statistical theory in certain special cases; in other words, in these certain cases, we will be able to write down an explicit mathematical formula for the pdf of the sampling distribution of the sample mean. However, for certain other statistics that can serve as estimators of underlying population parameters, such as the sample median and sample maximum, it will often be much easier to use simulation techniques, like the ones employed in the above two examples, to find their sampling distribution. These techniques will be discussed in Lecture 6.

(V.) The Sampling Distribution of the Sample Mean

Special Case

As for the sampling distribution of the sample mean of the variable X, we first examine the specific case where X is known to have a $N(\mu,\sigma^2)$ distribution. Here, since the sample mean can be thought of as a sum of n random variables that is then scaled by a factor of $1/n$, we can employ various properties of population expected values and variances that address summing and scaling random variables as well as the fact that the sum of normal random variables is also a normal random variable. Using these facts and properties, we can easily show that the sampling distribution of $\bar{X}$ takes the form $N(\mu,\sigma^2/n)$ in the specific case where X is known to have a $N(\mu,\sigma^2)$ distribution. The above pdf for $\bar{X}$ tells us that its sampling distribution is distributed symmetrically around the true population mean, as we might hope, and that, since σ^2/n is generally much smaller than σ^2, the values of $\bar{X}$ that can occur for a sample of size n are much more tightly clustered around the true mean value than are the values of X that occur.

However, since this specific case only applies when X is known to have a $N(\mu,\sigma^2)$ distribution, what if, as is likely to be true in practice, we either don't believe that X's underlying distribution is normal or don't know anything about X's underlying distribution at all? For instance, in our central example, in which we are interested in using the sample mean to estimate the underlying population mean, we only know that the underlying population distribution is symmetric. For cases even more general than ours (i.e., for situations in which X's distribution is unknown and possibly highly asymmetric), the following important and oft-cited theorem states that the sampling distribution of $\bar{X}$ is approximately the same as in the $X\sim N(\mu,\sigma^2)$ case, provided that our sample of X values is sufficiently large in size. This result is true no matter what distribution X has, as long as X's population mean and variance fulfil certain conditions, which are stated below.

The Central Limit Theorem

As just mentioned above, the approximate sampling distribution of the r.v. X can be stated in an exact mathematical form, provided the sample size is large, by invoking one of the most remarkable results of mathematical statistics: the "Central Limit Theorem" ("CLT").

The CLT establishes that, whatever the shape of the pdf of X, as long as its mean (μ) and its variance (σ^2) are finite, the sampling distribution of the sample mean $\bar{X}$ for sample size n tends to become $N(\mu,\sigma^2/n)$ as n increases. In other words, for large values of n, the probability that $\bar{X}$ lies between any two values, say a and b, approaches the probability

that a random variable $Y \sim N(\mu, \sigma^2/n)$ lies in [*a*,*b*]. This probability can be easily calculated using a computer or, after transforming Y to a standard normal variable, standard normal tables. Being able to determine the probabilities of $\overline{X}$ in this way is very useful if one wants to create a confidence interval for the underlying population mean, as will be demonstrated in Lecture 5.

Here, we should note that it is unnecessary to use the CLT to find the sampling distribution for the sample mean in cases where X is known to be normally distributed. In these cases, the sample mean's sampling distribution is exactly, not approximately, $N\left(\mu, \sigma^2/n\right)$, no matter how small the sample size is.

For cases in which X is not known to have a normal distribution, one question that has no general answer is *how large must n be for the normal approximation to the sampling distribution of* $\overline{X}$ *to be a good one?* A general guideline is that the more asymmetric the original distribution of X is, the larger n has to be. However, for most practical purposes, moderately large sample sizes are sufficient to make the use of the normal approximation appropriate.

We now give three illustrations of the CLT, all of which assume that we know the values of the underlying population, which is finite in size. Again, these examples are just for the purpose of illustration because, in practice, we almost never know the exact form of the underlying distribution and because, in this course, we usually assume that $N=\infty$. In the first example, the underlying distribution of X is apparently symmetric, without being normal. The population in this example is the heights (in mm) of a sample of 199 British married couples who were measured in 1980; although these heights represent a sample from the overall British population, in our example, we will pretend that they comprise the entire population. In our example, these heights are pooled, so that there are 398 units in the population. Incidentally, if the heights are treated separately for each sex, then they can be regarded as normally distributed; however, pooling the heights for the two genders results in a "bimodal" distribution, which obviously cannot be normal since the normal distribution is characterised by having only one peak. A true histogram of the pooled heights appears below in Figure 6.

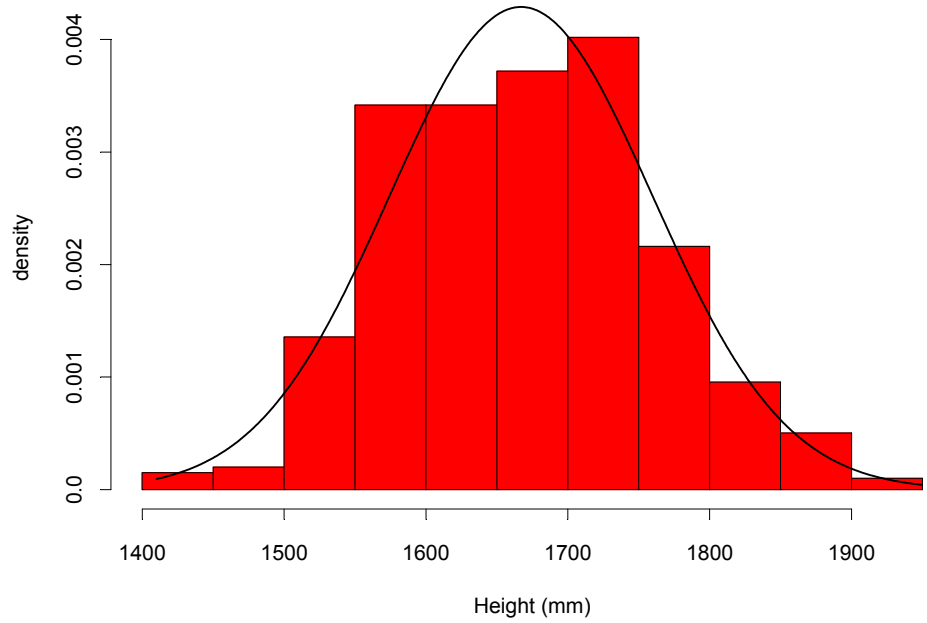

***Figure 6**: True histogram for heights of married couples*

Judging from the above figure, it appears that the distribution of heights may be flatter than the normal pdf, as is also evidenced by the following descriptive parameters for the population of heights:

Min	Q1	Med	Mean	Q3	Max	SD	CV	b1	b2
1410	1590	1670	1667	1730	1949	92.59	0.05	0.11	2.66

Now, suppose that we take 1,000 samples of size 3 and 1,000 samples of size 10, without replacement, from the underlying population of 398 heights. For each of the two sample sizes, Figure 7 shows a true histogram of the sample means for the 1,000 samples; these histograms provide a pictorial description of the empirical sampling distribution of the sample mean for the two sample sizes. For each sample size, Figure 7 also shows the appropriate normal approximation curve, as derived from the CLT and the known population distribution; note that in this example, the appropriate normal approximation curve is $N(1{,}667,\ 92.59^2/n)$, where n is the relevant sample size (3 or 10).

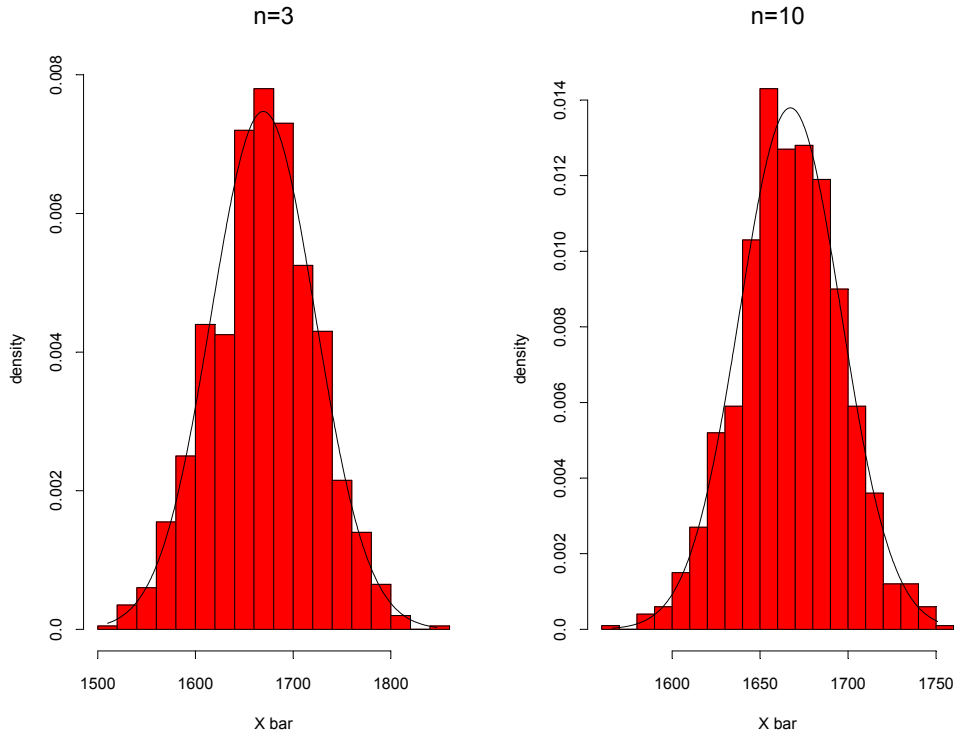

***Figure 7**: Normal approximations to the mean of husbands & wives data*

As we can see, the normal approximations to the sampling distributions are very good, even for these very small sample sizes, despite the fact that the original distribution was not normal.

In our next example, we show how a simple transformation can improve the normal approximation to the sampling distribution of the sample mean. In this example, the population consists of 94 dorsal lengths (in mm) of taxonomically distinct octopods measured in 1929. The left hand side graph in Figure 8 below shows a true histogram of the values in the original population. Immediately, the skewness of the underlying population distribution is evident. The other graph in Figure 8 shows a true histogram for the logarithms of all 94 values in the original population. Once the population values have been transformed, the normal distribution (or at least a symmetric distribution) appears to be a reasonably good population model.

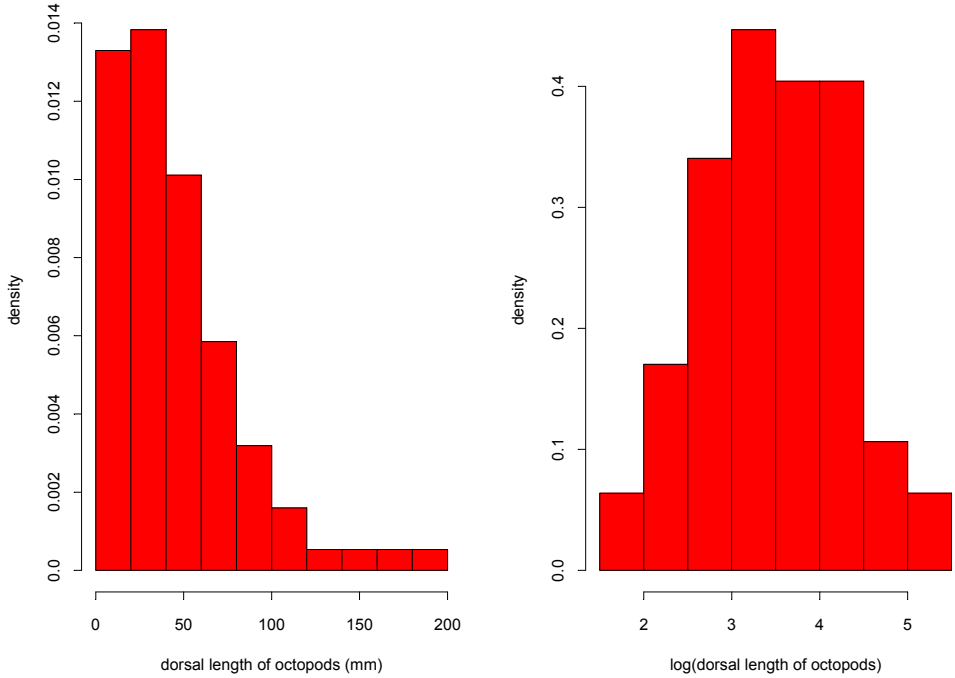

***Figure 8**: True histograms for dorsal length of octopods and its logarithm*

Next, suppose that we take multiple samples of size 5 and multiple samples of size 10 from the 94 population values on their original scale and then calculate the sample means of the resulting samples. The results of this process, in addition to the corresponding normal approximations, are shown in the first row of Figure 9. Additionally, suppose that we transform the observed values in each of the aforementioned samples to the logarithmic scale before calculating the sample mean. The bottom row in Figure 9 shows the results of doing this, as well as the appropriate normal approximations derived from the underlying population of logarithmic values shown on the right side of Figure 8.

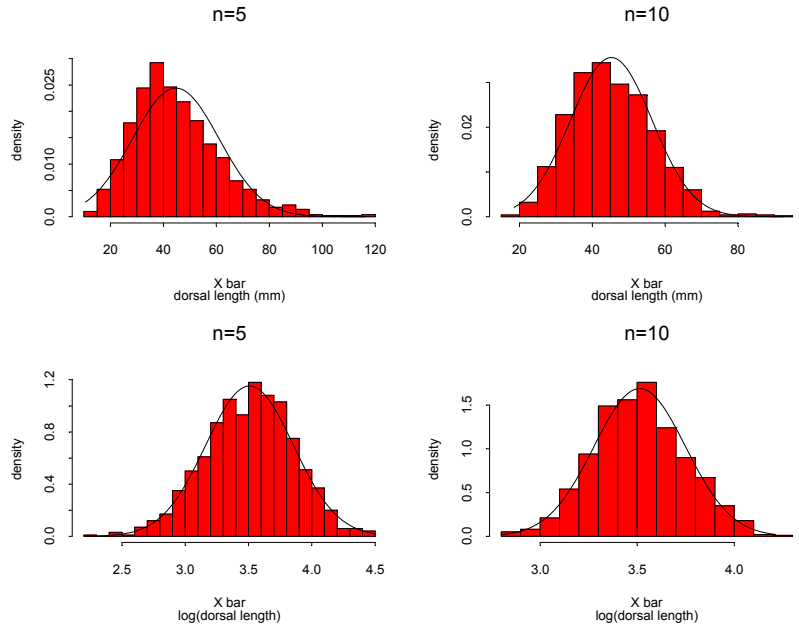

***Figure 9**: Normal approximations to the means of dorsal length of octopods data*

Figure 9 shows the difference that the logarithmic transformation can make for very small sample sizes in terms of the quality of the normal approximation. The reason for this is that the underlying population distribution for the logarithmically transformed values is much closer to symmetric than the distribution of the untransformed values, which means that, for the former values, the normal approximation to the sampling distribution of the sample mean becomes a good one at a smaller n than for the latter values. The only snag in transforming the dorsal length observations is that the unit of measurement would then be log (millimetres), whatever that is! However, this should not deter us from using a transformation for the purpose of improving the CLT normal approximation to the sampling distribution of the sample mean. This is the case because we can simply apply the inverse transformation (in this case, the exponential transformation) in order to return to the original scale of measurement after inferences about the population mean have been made using the CLT for the variable on the logarithmic scale.

Our final example of the CLT involves a very skewed distribution. The population consists of the number of lice found on 1,070 male prisoners in Cannamore, South India, from 1937-1939. Some descriptive parameters for the population are:

Min	Q1	Med	Mean	Q3	Max	SD	CV	b1	b2
0	0	0	6.389	4	187	18.02	2.82	5.18	36.43

Note the large values of the population skewness and kurtosis coefficients, as well as the fact that the standard deviation is almost three times larger than the mean, as seen by the population CV. Clearly, we have a population whose distribution cannot be described by only its mean and variance (which is, in effect, what we do when we describe a distribution as normal with mean μ and variance σ^2). A true histogram for the 1,070 population values appears in Figure 10. Note that this histogram uses bin intervals of different lengths, as was discussed in Lecture 1.

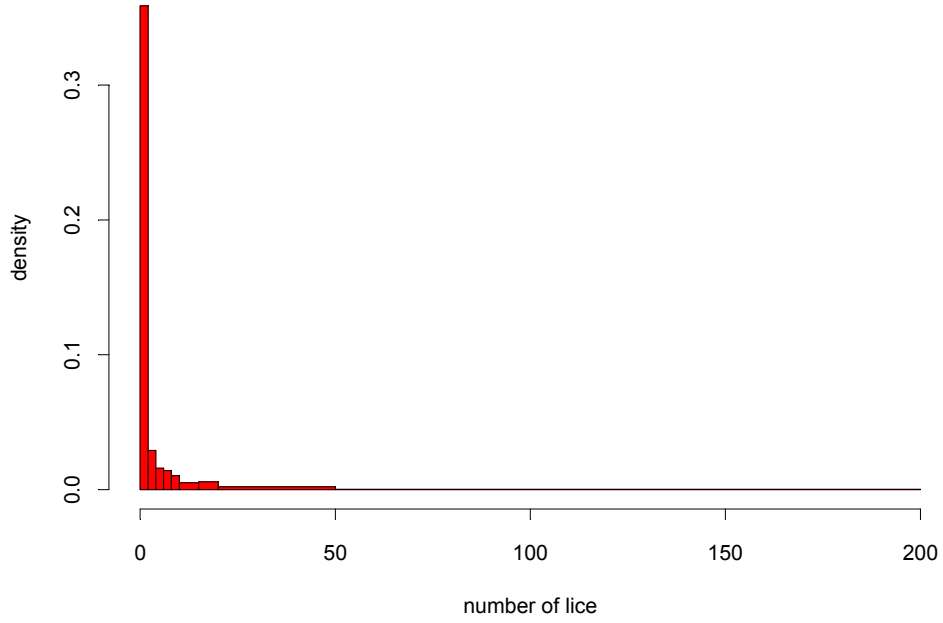

Figure 10: *True histogram for lice data*

Next, suppose that, as usual, we simulate 1,000 samples of each of various sample sizes and then calculate the sample means for the resulting samples. Figure 11 shows the empirical sampling distribution of $\overline{X}$ for each of the various sample sizes, over which is superimposed the appropriate normal approximation curve. It is clear that, in this example, the normal approximation is not a very good one for the sampling distributions, even for sample sizes as large as 500. In this case, using a logarithmic transformation does not help; thus, we would have to employ more complicated methods in order to use the CLT to make inferences about the mean of the underlying population.

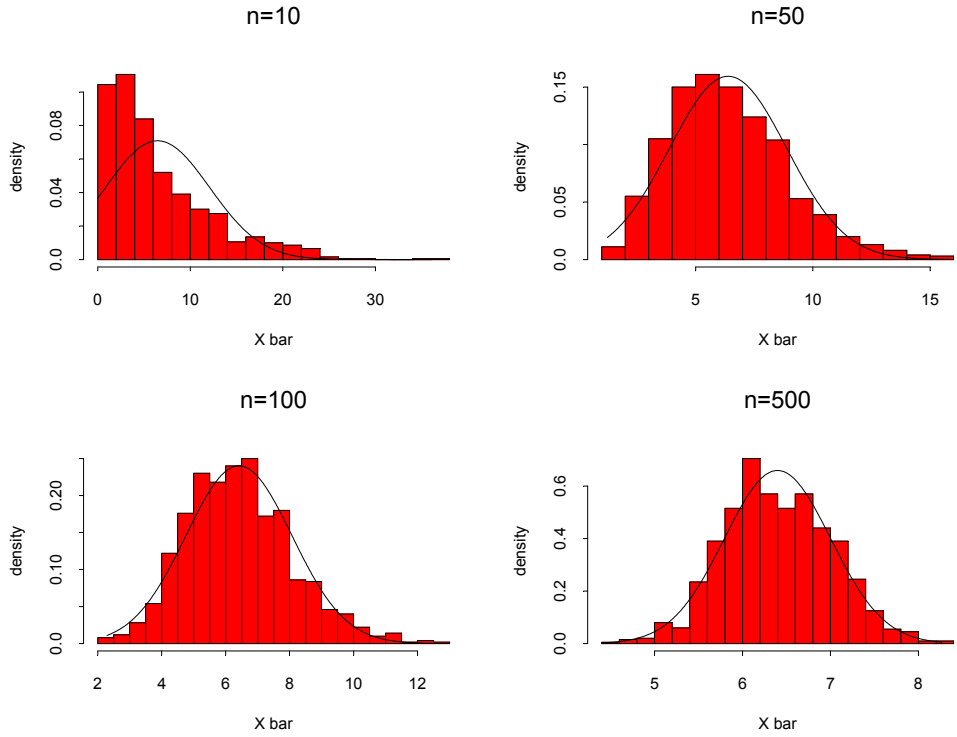

Figure 11: *Normal approximations for the means of lice data*

An even more extreme example of a distribution that needs very large sample sizes for the CLT to be appropriate is that of the gains won from National Lottery tickets. In this case, the population consists of literally millions of tickets whose gain is £0, and very few tickets with gains of millions of pounds. The CLT (which is, of course, true in any case as long as the mean and the variance are finite) will apply to samples from this population only if they contain millions of observations of ticket gains.

These examples prove that there are cases, albeit fairly extreme ones, for which the CLT is not appropriate if the sample size is not enormous; in these cases, it will be more difficult to make inferences about the population mean. However, for most situations, the CLT provides us with an explicit pdf for the approximate sampling distribution of the sample mean for reasonably large sample sizes, thus making inference about the mean (via confidence intervals and hypothesis tests) quite straightforward. In our central example, for instance, we can assume that the CLT would most likely provide a very good approximation to the sampling distribution of the sample mean because we know that the underlying population distribution of X is symmetric. As a result, we would not have to be sceptical of any conclusions drawn by using the CLT to make inference about the population mean.

(VI.) Other Properties of Estimators

Having just finished a lengthy discussion of the sampling distribution for the estimator in our central example (i.e., the sample mean), we now return to the more general topic of

the various properties of estimators. In order to illustrate these properties, let us return to the simulation example, introduced in Section IV, where 1,000 samples of each of various sample sizes were generated and the sample mean and maximum for each sample was then calculated. Consider, once again, the empirical sampling distributions for the sample mean shown in Figure 4. The graphs in Figure 4 suggest that when using the sample mean as an estimator for the underlying population mean, in general, we would expect the value of our estimate to be very near the true value of the parameter, especially for large sample sizes. This fact derives from two properties of the sample mean:

(1.) The expected value of the sample mean is, in fact, the true population mean. Before proceeding, we should note that the "expected value of an estimator" is the mean (in the population sense) of the estimator's sampling distribution. This property can be illustrated by examining, for a given graph, the vertical lines indicating the population mean and the mean of the 1,000 sample means, the latter of which can itself be viewed as an estimate of the expected value of the sample mean in our example. The fact that these two lines are almost always coincident suggests the aforementioned property. For the sake of comparison, consider the sampling distributions for the sample maximum shown in Figure 5, in which the lines indicating the population maximum and the mean of the 1,000 sample maxima are not particularly near to each other, even for large sample sizes. This phenomenon suggests that the expected value of the sample maximum is probably not the true population maximum.

(2.) The spread of the sample mean's distribution around its true value (i.e., around the population mean) decreases as the sample size decreases. This property is illustrated by the fact that the sample mean values are much more tightly clustered around the population mean (i.e., the true histograms are narrower) for larger sample sizes. In fact, if we had observed a particularly large sample of, say, size 64, it is very unlikely that the distance between the value of the sample mean for that sample and the true value of the parameter (56.99) would be larger than 10. Indeed, we would be very unlucky if that distance were larger than 5. However, if our resources did not allow us to sample as many as 64 individuals, then we would have to settle with more uncertainty, as is witnessed by the fact that the spread of the sampling distributions for the sample mean is larger for smaller sample sizes. For instance, if our sample size were only 4, then we could possibly do as badly as getting values of either 30 or 100 as estimates of the true mean: patently wrong values such as these would almost never occur in a sample of size 16 or more. This property, whereby a bigger sample size increases the probability that an estimate from that sample will be closer to its true population counterpart, also appears to hold when the sample maximum is used as an estimate for the population maximum. This can be seen by examining Figure 5, in which the values taken by the sample maximum stray less and less far from the true population maximum as n increases.

The Bias of an Estimator

Property 1 above addresses the concept of the "bias" of an estimator, which is defined as the difference between the expected value of an estimator and the corresponding population parameter it is designed to estimate. Further, an estimator is said to be "unbiased" for a parameter, θ, if its expected value is precisely θ. As suggested by the aforementioned proximity of the two vertical lines in the sample mean distributions in Figure 4, the sample mean is an unbiased estimator for the population mean. On the other hand, the sample maximum is not an unbiased estimator for the sample maximum, as can be seen by the distance between the two vertical lines in the distributions in Figure 5. Further, the fact that these two lines in Figure 5 are closer for larger sample sizes demonstrates that, when using the sample maximum to estimate the population maximum, its bias decreases as the sample size grows.

Unbiasedness is generally a desirable property for an estimator to have. Thus, the unbiasedness of the sample mean as an estimator for the population mean may give us some incentive to use it as the estimator in our central example. Often, an estimator for a given population parameter is constructed so that it will be unbiased. As an illustration of this, we return to the sample variance, for which a formula was given in Lecture 1. When this formula was given, we noted that the sum of the squared differences between the observations and the sample mean is usually divided by $n-1$ rather than by n, as would be done for a true average. The reason for doing this is that dividing by the former quantity makes the sample variance an unbiased estimator for the true population variance. If, instead, the sum of squared deviations were divided by n, the expected value of the resulting sample variance would differ from the true population variance by a multiplicative factor of $n/(n-1)$.

Note that the bias of an estimator can be calculated analytically or using simulation techniques. For instance, the bias of the sample mean as an estimator for the population mean (i.e., 0) was calculated analytically since mathematical calculation and statistical theory can be used to derive an explicit expression for the expected value of the sample mean (i.e., μ). However, it is not as easy to use analytical techniques to find the bias of the sample median as an estimator for the true population mean; instead, the bias of the sample median can be estimated using simulation techniques, as will be demonstrated in Lecture 6.

The Error of an Estimator

Property 2, above, mentions the spread of an estimator's sampling distribution. In general, the most common measure of the spread or dispersion of a distribution, relative to its mean, is the standard deviation. When we are talking specifically about the sampling distribution of an estimator, its standard deviation is referred to as the "standard error of the estimator." The standard error is used as a measure of the "precision" of the estimator, where saying that the estimator is "precise" merely means that the values of its sampling distribution are tightly clustered around the true value of the parameter it seeks to estimate. However, the standard error will, of course, depend on the scale or units of measurement of the variable of interest. For this reason, the standard

error is usually used to assess the precision of its corresponding estimate in a way that makes the units of measurement irrelevant. Specifically, it is common to eliminate the units of measurement by comparing the value of an estimator with that of its standard error via the ratio $\hat{\theta}/s.e.(\hat{\theta})$, which is unit-free; in general, the larger this ratio is, the more we would trust our estimate. For instance, finding that an estimate is bigger than its standard error (usually, 'bigger' means at least twice as big) is an indication that the estimate is reliable (i.e., precise) enough. If this is not the case, one might conclude that there is a large amount of random noise in the data that is resulting in an imprecise estimate of the parameter of interest. In addition to being used to judge the precision of an estimate, standard errors are also employed in the construction of confidence intervals, as will be demonstrated in Lecture 5.

The standard error of an estimator can be determined using either simulation techniques or analytical techniques. In some cases, it is hard to find an explicit formula for the standard error of an estimator using mathematical calculations and statistical theory. In these cases, it is possible to get an estimate of the standard error of an estimator by employing simulation techniques, which will be more fully detailed in Lecture 6. For now, let us say that most of these simulation approaches involve using a computer to simulate many different samples (from an appropriate distribution) in order to get an idea of an estimator's sampling distribution and then calculating an estimate of the estimator's standard error from its simulated sampling distribution. However, in some instances, we will not need to use such simulation techniques because we can calculate an explicit formula for the standard error of an estimator using mathematical calculation and statistical results. In cases where an estimator's standard error can be calculated analytically, the sample data need merely be plugged into the resulting explicit formula for the standard error in order to get an estimate of the estimate's standard error.

As an illustration of an estimator for which a standard error formula can be derived analytically, let us return once again to our central example. Recall that we had decided to use the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, as an estimator for μ, the underlying population mean. Using mathematics and statistical theory, it can be shown that the standard error of $\overline{X}$ is

$$s.e.\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}},$$

where n is the sample size and σ is the (finite) population standard deviation for X (i.e., the square root of the variance of the underlying population variable X). Note that the derivation of the above standard error of the mean did <u>not</u> require a knowledge of the distribution from which the data were drawn (i.e., we did not have to know the pdf for X). However, the above formula <u>does</u> require that we know σ. Unfortunately, if we do not know the true underlying mean for a certain random variable, it seems unlikely that we will know its true population variance. Thus, we often replace σ with its sample counterpart, the sample standard deviation. More specifically, we would calculate the sample variance,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

(which is an unbiased estimator for the population variance, as stated above) for our sample, take the square root of the calculated sample variance in order to obtain the sample standard deviation, and lastly, plug the sample standard deviation into the above formula for the standard error of the sample mean. The resulting estimate of the standard error of the sample mean would be:

$$s.e.\left(\overline{X}\right) = \frac{S}{\sqrt{n}}.$$

Note once again the use of capital letters in the definition of this estimate, which indicates that we are discussing the form of the estimators $\overline{X}$ and S^2, rather than realisations of them, as no data have been plugged in yet.

As an example of the use of this formula, consider the data in the table below, which come from two similar experiments designed to measure acceleration due to gravity. In the following table, the unit of measurement is $10^3 \times (\text{cm/sec}^2 - 980)$.

Experiment 1: 95 90 76 87 79 77 71
Experiment 2: 82 79 81 79 79 78 79 82 76 73 64

Some descriptive sample statistics for these samples are:

	mean	**std. dev.**	**std. err.**	**CV**	**n**
experiment 1	82.14	8.649	3.269	10.52%	7
experiment 2	77.45	5.165	1.557	6.67%	11

It is clear that the standard errors (of the two sample means) are small relative to the magnitude of the respective means. Note, however, that the standard error of the mean in the second experiment is almost half the size of the standard error in the first, which implies that the values in the first experiment have greater dispersion.

(VII.) The Trade-off Between Standard Error and Bias for Estimators

Suppose that we are considering two different estimators for the same population parameter. For instance, we might compare the sample median and the sample mean as estimators for the underlying population mean. In general, one of the two possible estimators might have a smaller standard error than the other. If this were true, then we would say that the estimator with smaller standard error was "more efficient" than the other estimator. In addition, we would call an estimator "efficient" if it achieved the smallest standard error possible for the estimation of a given parameter; the smallest possible standard error for an estimator of a certain parameter can be found using

mathematical results such as the Cramer-Rao lower bound. Greater efficiency is a generally attractive property for an estimator to have since it means that the estimator has a small standard error and, thus, is a more precise estimator for the underlying parameter. For this reason, we might be tempted to say that, of two estimators, the one with the smaller standard error is 'better.'

However, we stated above that unbiasedness is also a generally desirable property for an estimator to have, which might lead us to say that, for one unbiased estimator and one biased estimator of the same population parameter, the former estimator is 'better.' Ideally, then, we would prefer an unbiased estimator with the smallest standard error possible. However, quite frequently, no such estimator exists, and we are faced with a dilemma in which our two previous definitions of 'best' conflict because we must choose between two estimators (for the same parameter) where only the former estimator is unbiased but the latter has a smaller standard error. We frequently face dilemmas of this nature because, for estimators, there is a trade-off between bias and variance (i.e., the square of an estimator's standard error), as is often the case more generally in statistics. As an illustration of the bias-variance trade-off for three different estimators, consider the oft-cited example of three cannons, each of which is aimed at a fixed target and repeatedly fired. The resulting shots will have two sources of variation. First, for each cannon, its shots will vary randomly due to many factors (e.g., windspeed, temperature, number of previous shots, etc.). Secondly, a cannon's shots might not hit the centre of the target because of some systematic problem, whether caused by the inferior quality of the cannon's parts or by the quality of the operator's eyesight; for instance, one cannon's shots might consistently end up on the right side of the target. The first of these sources of error, the precision of the cannon, can be likened to the variance or standard error of an estimate, and the second, the accuracy of the cannon, to the bias of an estimator. The trade-off between precision and accuracy for three possible sets of cannon shots is illustrated below in Figure 12; in these graphs, the intended target is the point (x=0, y=0).

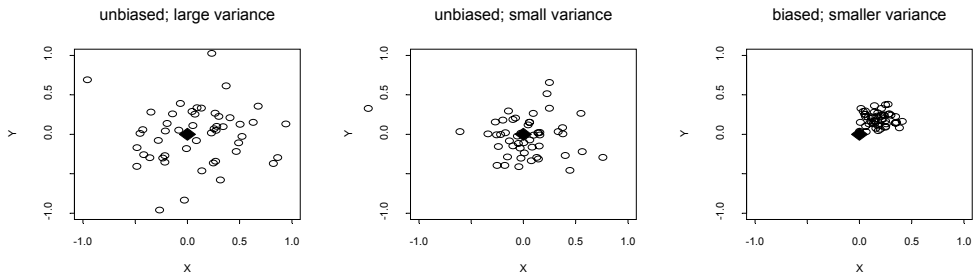

***Figure 12**: Three estimators with different properties.*

Ideally, we would like to have a cannon that is both accurate and precise. As is often true for estimators, this is not possible for the three cannons, and we are forced to choose between cannons with greater accuracy (e.g., Cannon 1) and cannons with greater

precision (e.g., Cannon 3). The choice of cannon will probably depend on the situation in which it will be used.

Similarly, there is no rule that allows us to choose between several estimators for the same population parameter since, in different situations, a different choice of estimator might be preferable. However, unlike in the cannon example, with statistical estimators, we will not usually know the exact bias and variance of each possible estimator. Thus, we will have to rely on results from mathematical statistics to get an idea of these quantities and, therefore, to ensure that we are using an estimator that is optimal in terms of these two quantities.

As for a rule for deciding between bias and variance in estimators, in general, we can only say that it is a good idea to select an estimator that achieves a balance between bias and standard error (i.e., Cannon 2). Obviously, we can find an estimator of a certain parameter that has extremely small (i.e., 0) variance, such as the estimator described by the rule: 'give the value of 5 as an estimate of the population parameter of interest regardless of the sample values observed.' Clearly, the sampling distribution of this estimator has no variation. However, this estimator is obviously very biased, and everybody would agree that it is a useless estimator. This example demonstrates that, in general, it makes sense to choose estimates with the smallest possible variation, given that they are unbiased or nearly unbiased. In fact, this practice is often followed in statistics, leading to a near obsession with estimators that are "B.L.U.E"; here, B.L.U.E. refers to the Best (in terms of efficiency) estimator in the class of Linear (i.e., estimators that are linear functions of the data) Unbiased Estimators for a parameter of interest.

Lastly, in addition to desiring an estimator for a parameter that is both unbiased and efficient, we may also want our estimator to be robust to outliers in our data sample. For instance, we might choose the sample median rather than the sample mean as an estimator for the population mean in our central example because the former estimator is robust to outliers, whereas the latter is not. Note, however, that robustness will often (but not always) come at the expense of higher error.

(VIII.) Advantages and Disadvantages of Increased Sample Size

Recall that in Section IV, we noted that the spread of the sample mean's distribution around its true value (i.e., the population mean) decreases as the sample size increases. This fact is illustrated pictorially by Figure 4. This fact can also be seen by examining the expression for the CLT normal approximation to the sample mean's distribution since the approximate sampling distribution has a variance of σ^2/n, which obviously decreases as n increases. Recall that this phenomenon, whereby the spread of an estimator's sampling distribution decreases as sample size increases, was also observed to occur for the sample maximum in Figure 5.

In general, it is true that the precision of an estimator increases (i.e., its standard error decreases) as the sample size increases; in other words, estimators will be more precise when the sample size is large. In fact, having a larger sample is generally advantageous; for instance, in the case where we are trying to detect whether the true means of two underlying populations are the same, having large samples from both populations will allow us to detect even very small differences in the true means. However, the advantages associated with large sample size must be balanced against the cost and time required to collect a larger sample. In addition, if we decide to collect a larger sample, machine or personnel constraints may mean that our measurements of the variable(s) of interest will be of lower quality; in this case, a smaller sample with higher quality measurements might be preferable.

MCB (I-2000), *KNJ* (III-2001), *JMcB* (III-2001)