A random sample of size \( n \) from an infinite population or a finite population where we sample with replacement means a collection of \( n \) observations \( X_1, X_2, \ldots, X_n \) where the random variables \( X_i \) are I.I.D. (independent and identically distributed meaning each has the same distribution \( f(x) \) = density in the continuous case, probability mass function in the discrete case)

A random sample of size \( n \) from a finite population of size \( N \) (sampling without replacement is understood) means each subset of size \( n \) has the same probability of being selected. For the case of unordered selections this means each of the \( \binom{N}{n} \) possible selections has equal probability \( 1/\binom{N}{n} \) of being chosen. Depending on the problem there may be times when ordered selections are desired in which case there are \( _N P_n = N (N-1) \cdots (N-n+1) \) such permutations each with equal probability 1 over that number.

A statistic is any quantity which can be calculated from the observed random sample, that is to say a function of the observations \( X_1, X_2, \ldots, X_n \) such as the sample mean and variance \( \bar{X} \) and \( S^2 \). The later are often used to estimate the actual constant population parameters mean \( \mu \) and variance \( \sigma^2 \).

Some of the pitfalls that can occur when trying to construct a random sample are well illustrated by Johnson's interesting example of trying to determine a random sample of lengths of logs being fed into a sawmill having a constant speed conveyor belt. The seemingly obvious approach of stopping the belt at a certain location every ten minutes to measure a log is flawed since longer logs spend more time traveling from one end of the log to the other down the conveyor belt hence are more likely to be selected than shorter logs are. For a related sort of problem consider a study of a large sample of men and women designed to examine the relationship between the average weight of a person according to age group and height of the person. The headline of a Midwest newspaper read “Men loose weight earlier’” In fact what was likely happening was that obese individuals were dying off giving the impression that the average weight of males is decreasing with age since the sample only included weights of individuals who were still alive, and not the weights of those who died.

EXAMPLE 1 : Problem 6.3 of text (similar to HW problem 6.2) Explain why the following will not lead to random samples from the data.

a) To determine what the average person spends on a vacation, a market researcher interviews passengers aboard a luxury cruise.

It is not clear whether average person includes those who can afford a vacation on a luxury cruise. Nor should we count on people on the cruise being willing to answer the question honestly. Perhaps those who spend more don't want to admit it while the same could be true for those who are ashamed of being poor.

b) To determine the average income of its graduates 10 years after graduation the alumni office of a university sends questionnaires in 2005 to all the members of the class of
Again the rich may not want to answer due to privacy or security reasons while the poor may feel ashamed to answer. So unless we can guarantee all the questionnaires are returned and are answered honestly we don't know what kind of a sample we are getting. Perhaps it was harder to locate wealthy individuals or poor individuals not currently employed so that the questionnaires were less likely to reach those people.

c) To determine public sentiment about certain import restrictions, an interviewer asks voters “Do you feel that this unfair practice should be stopped?”

Here the strong wording of the word “unfair” could bias the answers in one direction or another in a manner so that the answers might be different if the word were left out of the question.

**Sampling distribution of the (sample) mean** \( \bar{X} : \sigma \text{ known} \) Theorem 6.1 tells us the mean and variance of the sample mean but provides no further detail into the nature of its probability distribution. The central limit theorem, Theorem 6.2 tells us the distribution of the sample mean is approximately normal when \( n \) is large. To apply this result requires knowledge of the population standard deviation \( \sigma \). The central limit theorem result extends to small \( n \) provided we are sampling from an (approximately) normal population.

**Theorem 6.1** of the text: If a random sample of size \( n \) is taken from a population having mean \( \mu \) and variance \( \sigma^2 \) then the sample mean \( \bar{X} = \frac{1}{n} \sum X_i \) is a random variable whose distribution also has mean

\[ E[\bar{X}] = \mu \]

and for samples from an infinite populations its distribution has variance

\[ V[\bar{X}] = \frac{\sigma^2}{n} \]

while for samples from a finite population of size \( N \) the variance is

\[ V[\bar{X}] = \frac{\sigma^2}{n} \frac{N-n}{N-1} \]

The factor of

\[ \frac{N-n}{N-1} \]

is called the **finite population correction factor**. It will be close to 1 and so may be safely ignored if \( n \) is small compared to \( N \).

Proof of Theorem 6.1: We have earlier calculated using the laws of expectations the mean of the sample mean \( \bar{X} \) (it is just a linear combination of the \( X_i \)'s where the constants are all equal to \( 1/n \) ) and computed its variance in the case of sampling from an infinite population (i.e. using the independent, identically distributed assumption).

The above general result for sampling from a finite population is stated without proof but we note that we have earlier derived this finite population correction factor in the special case of the hypergeometric random variables which we viewed as a sum of dependent Bernoulli 0 or 1 valued random variables (taking value 1 if a defective item is selected on that particular trial and 0 else).
EXAMPLE 2 (like problem 6.5) How many samples of size \( n=3 \) can be chosen from a finite population of size \( N=12 \)? If the samples are unordered we know there are \( \binom{N}{n} = \binom{12}{3} = 220 \) ways to choose 3 items from 12.

EXAMPLE 3 (like problem 6.9) Given the infinite population whose probability distribution is given by

<table>
<thead>
<tr>
<th>( x )</th>
<th>( f(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/3</td>
</tr>
<tr>
<td>2</td>
<td>1/3</td>
</tr>
<tr>
<td>3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

the table, list the 9 possible samples of size 2 and determine the mean and variance of the sample mean. Show that it agrees with Theorem 6.1

The 9 possible samples \( (X_1, X_2) \) of size 2 are

\[(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)\]

having

\[ \bar{x} = (x_1 + x_2) / 2 = 1, 3/2, 2, 3/2, 2, 5/2, 2, 5/2, 3 \]

If samples are random this means that \( X_1, X_2 \) are independent random variables so that the joint distribution factors as the product of the marginal probabilities each of which are 1/3. Thus each of the above 9 pairs has equal joint probability \( (1/3)(1/3) = 1/9 \). For example \( p(2,3) = 1/9 \) etc

The event that \( \{ \bar{x} = 2 \} = \{ (X_1, X_2) \in (1,3), (2,2), (3,1) \} \) has probability 3/9 = 1/3 since it is composed of 3 disjoint events each having the joint probability 1/9. In this way we find that \( \bar{x} \) has distribution given by the table below from which we can directly compute the mean and variance of the sample mean \( \bar{x} \) and show we get \( E[\bar{X}] = 2 = \mu = E[X_1] = E[X_2] \) since

\[ 1 \cdot (1/9) + (3/2) \cdot (2/9) + 2 \cdot (3/9) + (5/2) \cdot (2/9) + 3 \cdot (1/9) = 2 \]

and

\[ V[\bar{X}] = (1-2)^2 \cdot \frac{1}{9} + (\frac{3}{2} - 2)^2 \cdot \frac{2}{9} + (0^2 \cdot (3/9)) + (\frac{5}{2} - 2)^2 \cdot \frac{2}{9} + (3-2)^2 \cdot \frac{1}{9} = 1 = \frac{\sigma^2}{2} \]

where \( \sigma^2 = E[(X-2)^2] = \sum_x (x-2)^2 f(x) = (1-2)^2 \cdot \frac{1}{9} + 0^2 \cdot \frac{2}{9} + (3-2)^2 \cdot \frac{1}{9} = \frac{2}{3} \)

We would arrive at the same answers using the formula for expectation of a function

\[ h(X_1, X_2) = (X_1 + X_2) / 2 = \bar{x} \] in terms of the joint probability so that

\[ E[X] = \sum_{x_1, x_2} (x_1 + x_2) / 2 \cdot p(x_1, x_2) = (1+3/2+2+3/2+2+5/2+2+5/2+3) \cdot (1/9) = 2 \]

and similarly using \( p(x_1, x_2) = 1/9 \), \( V[\bar{X}] = E[(\bar{X} - 2)^2] = \sum_{x_1, x_2} ((x_1 + x_2) / 2 - 2)^2 p(x_1, x_2) \)

\[ = ((1-2)^2 + (3/2 - 2)^2 + 0^2 + (3/2 - 2)^2 + 0^2 + (5/2 - 2)^2 + 0^2 + (5/2 - 2)^2 + (3-2)^2) \cdot (1/9) = 1/3 \]

EXAMPLE 4 (like problem 6.11) When we sample from an infinite population, what happens to the standard error of the mean \( \sigma / \sqrt{n} \)

a) if the sample size is increased from 90 to 360:

\[ \frac{\sigma}{\sqrt{360}} = \frac{\sigma}{\sqrt{4 \cdot 90}} = \frac{1}{2} \cdot \frac{\sigma}{\sqrt{90}} \]

so decreases by factor of 2

b) if the sample size is decreased from 90 to 10:

\[ \frac{\sigma}{\sqrt{90}} = \frac{\sigma}{\sqrt{9 \cdot 10}} = \frac{1}{3} \cdot \frac{\sigma}{\sqrt{10}} \]

so increases by factor of 3
EXAMPLE 5 (like problem 6.12) What is the value of the finite population correction factor \( \frac{N-n}{N-1} \)

a) when the sample size is \( n = 20 \), the population size is \( N = 800 \):

\[
\frac{800 - 20}{800 - 1} = 0.976
\]

b) \( n = 200, \ N = 10,000 \):

\[
\frac{10000 - 200}{10000 - 1} = 0.98
\]

The Central Limit Theorem (Theorem 6.2 of text) says that for large \( n \) (in practice for \( n \geq 25 \) or 30) the sample mean \( \bar{X} \) will be approximately normally distributed (exactly so in the limit as \( n \to \infty \)) provided the population variance \( \sigma^2 < \infty \) is finite. Thus the standardized variable

\[
Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (= \frac{S_n - n \mu}{\sqrt{n} \sigma})
\]

where \( S_n = X_1 + X_2 + \ldots + X_n \) is approximately standard normal.

The Central limit theorem applies for large \( n \). To say more about the distribution of \( \bar{X} \) for small \( n \) beyond just giving its mean and variance as in Theorem 6.1 we have to make strong assumptions about the distribution of the population from which we are getting the sample of individual observations.

In particular if the population is normally distributed, meaning each observation \( X_i \) is normal then the sample mean \( \bar{X} \) is normal. In fact any linear combination of independent normal random variables is normal no matter how small the sample size \( n \), implying that the standardized variable will be standard normal. This is not hard to show but we will not do so here. Another interesting property is:

For a random sample from a normal population by examining the joint distribution one can show that the sample mean and sample variance \( \bar{X} \) and \( S^2 \) are independent random variables and only for a normal population is this true!

EXAMPLE 6 (like problems 6.15 + 6.17) The weight of passengers on an airline flight can be modeled as a random variable with mean 150 lbs and variance 225 lbs².

a) On a flight with \( n = 144 \) passengers, what is the probability that the sample mean lies between 148 lbs and 151 lbs?

Note with \( \sigma = \sqrt{225} = 15 \text{ lbs} \), \( \mu = 150 \),

\[
P(148 \leq \bar{X} \leq 151) = P\left(\frac{148 - 150}{15/\sqrt{144}} \leq Z = \frac{\bar{X} - 150}{\sigma / \sqrt{n}} \leq \frac{151 - 150}{15/\sqrt{144}}\right)
\]

\[
= P(-1.6 \leq Z \leq .8) = F(.8) - F(-1.6) = .7881 - .0548 = .7333
\]

b) (like problem 6.17) What is the probability that the total weight of all 144 passengers lies between 21312 lbs and 21744 lbs?

Writing \( S_{144} = \sum_{i=1}^{144} X_i = 144 \cdot \bar{X} \) one has for \( Z = \frac{S_n - n \mu}{\sqrt{n} \sigma} \):
\[ P(21312 \leq S_{144} \leq 21744) = P\left(\frac{21312 - 144 \cdot 150}{\sqrt{144 \cdot 15}} \leq Z = \frac{S_{144} - 144 \cdot 150}{\sqrt{144 \cdot 15}} \leq \frac{21744 - 144 \cdot 150}{\sqrt{144 \cdot 15}}\right) \]

\[ = P(-1.6 \leq Z \leq 8) = F(8) - F(-1.6) = .7881 - .0548 = .7333 \]

Since the total weight of all 144 passengers divided by 144 gives the sample mean we could have phrased the problem initially as the same as the probability that the sample mean lies between 
21312/144 = 148 and 21744/144 = 151 which would have made it obvious this is the same question as part a) in disguise.

However the central limit theorem is often stated in terms of the sum
\[ S_n = n \bar{X} \]

Sampling distribution of the (sample) mean \( \bar{X} \) for \( \sigma \) unknown: For large sample size \( n \) say \( n > 40 \) then in practice since the sample standard deviation \( s \) approaches the population standard deviation \( \sigma \) as \( n \to \infty \), replacing the unknown \( \sigma \) by the sample standard deviation \( s \) in the statement of the central limit theorem still yields an approximately standard normal distribution.

For \( n \) small however we must assume that our original population is (approximately) normal with mean \( \mu \) and variance \( \sigma^2 \) in which case for \( \bar{X} \) the sample mean of a sample of size \( n \) one has

Theorem 6.3 under the above assumptions the statistic
\[ t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \]

where \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \)

is a random variable having a t-distribution with parameter
\[ \nu = n-1 \]
the number of “degrees of freedom”.

The t-distribution is symmetric about the origin just like the standard normal distribution but is a bit wider than the normal distribution. In practice the t-distribution will be approximately standard normal if \( n \geq 30 \) Note that this normality holds with \( n \geq 30 \) when sampling from an approximately normal population whereas for \( n > 40 \) \( Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \) will in practice be approximately standard normal no matter what the original distribution of the population we are sampling from.

EXAMPLE 7 A certain process for making washers is under control if the (population) mean diameter of the washers is \( \mu = 2 \) cm. What can we say about this process if a sample of 12 of these washers has a (sample) mean diameter of 2.004 cm and a (sample) standard deviation of .007 cm?

We can't say much unless we assume that the population we are sampling from is approximately normal. Let us assume this. The standard deviation (standard error) of the mean will be
\[ s/\sqrt{12} = .007/\sqrt{12} = .002 \]

so that for this sample, the random variable with a t-distribution with \( \nu = n-1 = 11 \) degrees of freedom takes the observed value :
\[ t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{2.004 - 2}{.002} = 2 \]

Assuming the null hypothesis \( H_0: \mu = 2 \) that the process is under control, the P-value which is the probability of observing evidence as strong or stronger than the observed data in favor of rejecting the null hypothesis, is by the symmetry of the t-distribution equal to the probability
\[ P(t \geq 2) \]

This assumes a two-sided hypothesis test in which we reject the null hypothesis \( H_0: \mu = 2 \) in favor of the alternative hypothesis \( H_a: \mu \neq 2 \) if the observed t-value is sufficiently larger or sufficiently
smaller than 2. (For a **one sided test** the alternative hypothesis would be of the form \( H_a: \mu > 2 \) in which case we'd reject the null hypothesis if the observed t-value is sufficiently larger than 2. Or it could be of the form \( H_a: \mu < 2 \) which would be supported if the observed t-value is smaller than 2.)

Now from the t-table we have the **t-critical values** for 11 degrees of freedom are \( t_{.05} = 1.796 \) and \( t_{.025} = 2.201 \)

so since 2 is roughly halfway in between these we would guess that the probability \( P(t \geq 2) \) of seeing a t value greater than or equal to 2 which is the area under the t-curve to the right of 2 is approximately halfway between 5% and 2.5% or around 3.7% = .037 and thus the P-value which is the **statistical significance level** of the data or the strength of the evidence for rejecting the null hypothesis is about 7.4% for a two sided test (or at least less than 10% and greater than 5%).

**The smaller this probability, the stronger the evidence in favor of the alternative hypothesis.** If we were to reject the null hypothesis at this level of significance we would have around a 7.4% chance of making a **type I error**, that is of **concluding the null hypothesis is false when it is in fact true** or in other words concluding that the process is out of control when really it isn’t.

While the observed data is not that unlikely, still roughly 92% of the time that the process is under control we would not observe t values this far from 0, the mean of the t distribution. There is at least fairly good reason to suspect that the process might be out of control.

**The sampling distribution of the (sample) variance** \( S^2 \). As with the t-distribution to be able to say something about the distribution of the sample variance we assume the strong assumption that we are sampling from a normal population having population variance \( \sigma^2 \). In this case one can show

**Theorem 6.4** If \( S^2 \) is the (sample) variance of a sample of size \( n \) taken from a normal population with (population) variance \( \sigma^2 \) then

\[
X^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

is a random variable having the **chi-squared** distribution with the parameter \( \nu = n-1 \) = the number of “degrees of freedom”.

**The chi-squared distribution is a special case of the gamma distribution** with \( \alpha = \nu/2 \) and \( \beta = 2 \).

Unlike the normal or the t-distribution, the chi-squared distribution is **not symmetric** so for small \( \alpha \) one often needs to know both \( X^2 \) critical values : \( X^2_{\alpha} \) and \( X^2_{1-\alpha} \). Here the probability (area) to the right of the value \( X^2_{\alpha} \) is \( \alpha \).

**EXAMPLE 8** (like problem 6.23) The claim (null hypothesis ) \( H_0: \sigma^2 = 12.4 \)

that for a sample of size \( n=17 \) from a normal population the population variance equals 12.4 is rejected in favor of the (one-sided) alternative hypothesis that

\( H_a: \sigma^2 > 12.4 \)

if the (sample) variance \( S^2 \) lands in the **rejection region** meaning it satisfies

\( S^2 \geq 24.8 \).

(In the one-sided test case in this example we interpret the null hypothesis \( H_0: \sigma^2 = 12.4 \) as really meaning \( H_0: \sigma^2 \leq 12.4 \) )
What is the probability of a type I error meaning that the null hypothesis claim $H_0: \sigma^2 = 12.4$ is rejected when in fact it is true?

Here $\nu = n - 1 = 17 - 1 = 16$ is the number of degrees of freedom and we want to know the probability that $H_0: \sigma^2 = 12.4$ is true but we observe $S^2 \geq 24.8$. This means we want to know

$$P(\chi^2(\nu = 16) = \frac{16S^2}{12.4} \geq \frac{16 \cdot 24.8}{12.4} = 32) = P(\chi^2 \geq 32).$$

From the chi-squared table 5 of appendix B, we have the chi-squared critical value for 16 degrees of freedom $\chi^2_{0.01} = 32$ meaning that our type I error probability for the one-sided test is

$$P(\chi^2 \geq 32) = .01$$

which is the probability (area) to the right of 32 under the chi-squared curve. Note that if the test were two-sided we would also reject the null hypothesis for sufficiently small values of $s^2$ and hence for small values of $\chi^2 = 16s^2/\sigma^2$. Thus since the chi-squared distribution is not symmetric we would then want to also look up $\chi^2_{0.99} = 5.812$

which means the rejection region would also include the region where

$$S^2 = 12.4 \frac{\chi^2}{16} \leq 12.4 \cdot \frac{5.812}{16} = 4.5043 \quad \text{or simplifying, where}$$

$$S^2 \leq 4.5043.$$  

For the two sided test the type one error probability for the two sided rejection region

$$S^2 \geq 24.8 \quad \text{or} \quad S^2 \leq 4.5043$$

or equivalently if

$$\chi^2 \geq 32 \quad \text{or} \quad \chi^2 \leq 5.812$$

is twice as much as for the one sided test or a type I error probability of .02 instead of .01.

**Distribution of the ratio of sample variances:**

To determine if two random samples come from normal populations with the same or nearly the same variance one wants to look at the random variable with **F-distribution**;

**Theorem 6.5** If $S_1^2$ and $S_2^2$ are the variances of independent random samples of size $n_1$ and $n_2$ respectively, taken from two normal populations with the same variance then

$$F = \frac{S_1^2}{S_2^2}$$

is a random variable having the F-distribution with parameters $\nu_1 = n_1 - 1$ = denominator degrees of freedom and $\nu_2 = n_2 - 1$ = denominator degrees of freedom.

**EXAMPLE 9** (like problem 6.25) If independent random samples of size $n_1 = 26$ and $n_2 = 12$ come from two normal populations having the same variance, what is the probability that either sample variance will be at least 4 times as large as the other?

This is a two sided test of the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$. We use the property of the F distribution that its F-critical values for the left and right tails of the distribution are related by
where $F_{1-\alpha}$ is the F-value which has area (probability) $\alpha$ to the left of it under the F curve (for the given two parameters). This is just a consequence of interchanging the roles of the sample variances. Note that the number of degrees of freedom has also been swapped. Thus we want to find $P\left(F = \frac{S_1^2}{S_2^2} \geq 4 \text{ or } F \leq \frac{1}{4}\right)$.

From the F-table with parameters 25 and 11 we find the approximate F-critical values $F_{.01}(25, 11) = 4.01 \approx 4$ and so $F_{.99}(11, 25) = \frac{1}{4}$.

However it is $F_{1-\alpha}^{25, 11} = \frac{1}{F_{\alpha}(11, 25)}$ that we are interested in. This is not an issue in the corresponding homework problem since there the two sample sizes are equal so one can flip things easily.

Interpolating we find that $F_{.99}(25, 11) = \frac{1}{F_{.01}(11, 25)} = \frac{1}{3.06}$. Since $\frac{1}{4}$ is less than this value the probability is less than .01 that $F < \frac{1}{4}$. Thus the probability is roughly .01 that the F value falls to the right of 4 and less than .01 that it falls to the left of $\frac{1}{4}$ so without access to a computer program to calculate arbitrary F critical values, probability $\alpha < .02$ is all we know about the type I error probability that the null hypothesis that the variances are the same $H_0: \sigma_1^2 = \sigma_2^2$ is rejected based on the rejection region $F \geq 4$ or $F \leq 1/4$.

\section*{Chapter 7: Inferences concerning means.}

\textbf{Point estimation}: The word point here refers to a single number estimate $\hat{\theta}$ of a population parameter $\theta$. Such a statistic (meaning a function of the sample data) is said to be an \textbf{unbiased estimator} of $\theta$ if the average value of the estimator is the actual parameter so that $E[\hat{\theta}] = \theta$, otherwise the estimator is said to be \textbf{biased} (with bias $E[\hat{\theta}] - \theta$).

We have seen that the sample mean $\bar{x}$ of a random sample is an unbiased estimator for the actual population mean $\mu$ where for infinite populations random means independent identically distributed and for finite it means with all (same size) samples equally likely.

For normal populations the sample mean and the sample median are both unbiased estimators of the population mean but it can be shown that the variance of the sample median is more than 1.5 times as large as the variance of the sample mean:

$$V(\text{median}) > 1.5V(\bar{x}) \text{ where } V(\bar{x}) = \sigma^2/n$$

We say that in this case that the sample mean is a \textbf{more efficient unbiased estimator} of $\mu$ than the sample median, meaning that (at least for normal populations) both estimators are unbiased but the variance of the sample mean is no larger than the variance of the sample median and is less than it for at least one value of the parameter being estimated (in this case the variance is less for all values of the parameter $\mu$).

\textbf{Maximum error of estimate}: When using $\bar{x}$ to estimate $\mu$ we would like to know how to bound the maximum error of our estimate with high probability. From the central limit theorem (Theorem 6.2) we know that for large samples ($n \geq 25$ or 30)

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad (n \geq 25 \text{ or 30 or for small } n \text{ when population is normal})$$

is approximately normal (exactly normal even for small $n$ when population is normal) so that with
probability \( 1 - \alpha \)

\[-z_{\alpha/2} \leq \bar{X} - \mu \leq z_{\alpha/2} \leq \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \text{ or } \text{Error} = |\bar{X} - \mu| \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = E = \text{maximum error}\]

gives the maximum error of our estimate with probability \( 1 - \alpha \). Note both of the above statements also hold for small samples when sampling from a normal population since then \( \bar{X} \) is exactly normal.

**Determining sample size:** Inverting, if we want to know how large a sample size we need to guarantee a bound on the maximum error by amount \( E \) we find that we must have

\[n \geq \left[ \frac{z_{\alpha/2} \sigma}{E} \right]^2\]

For somewhat larger samples say \( n > 40 \) we are fairly safe in replacing the population standard deviation \( \sigma \) by the sample standard deviation \( s \) so that

\[\text{Error} = |\bar{X} - \mu| \leq z_{\alpha/2} s \frac{1}{\sqrt{n}} = E = \text{maximum error} \quad (n > 40)\]

For small samples from a normal or approximately normal population the above becomes

\[t = \frac{\bar{X} - \mu}{S/\sqrt{n}}\]

has a t-distribution with \( n-1 \) degrees of freedom so that (for small \( n \) samples from normal population)

\[\text{Error} = |\bar{X} - \mu| \leq t_{\alpha/2} S \frac{1}{\sqrt{n}} = E = \text{maximum error}\]

**Confidence intervals:** The above considerations can be recast in the language of confidence intervals. A \( 100 (1 - \alpha)\% \) **large sample confidence interval (CI)** for the population mean \( \mu \) is

\[\left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] 100(1 - \alpha)\% \quad \text{CI for } \mu \quad (n \geq 25 \text{ or } 30)\]

The above also holds for **small samples taken from a normal population** since then \( \bar{X} \) is exactly normal.

**Interpretation of CI:** The above means that the probability is \( 100(1 - \alpha)\% \) (our confidence level) that \( \mu \) will belong to this interval prior to collecting the sample. Once the sample mean has been determined for a particular sample, there is no probability involved. Either \( \mu \) belongs to the interval or it doesn't. Thus the interpretation is that if we were to do this determination for many randomly collected samples, then for on average \( 100(1 - \alpha)\% \) of them \( \mu \) would belong to the interval.

**Large sample confidence interval when \( \sigma \) is unknown:**

For sufficiently large samples \( n > 40 \) we can extend the interval to the case where \( \sigma \) is unknown

\[\left[ \bar{X} - z_{\alpha/2} S \frac{1}{\sqrt{n}}, \bar{X} + z_{\alpha/2} S \frac{1}{\sqrt{n}} \right] 100(1 - \alpha)\% \quad \text{CI for } \mu \quad (n > 40)\]

**Small sample CI** when sampling from an (approximately) normal population uses the t-distribution

\[\left[ \bar{X} - t_{\alpha/2} S \frac{1}{\sqrt{n}}, \bar{X} + t_{\alpha/2} S \frac{1}{\sqrt{n}} \right] 100(1 - \alpha)\% \quad \text{CI for } \mu \quad \text{for small } n\]

with \( n-1 \) degrees of freedom.

**EXAMPLE 1** (for chapter 7, like problem 7.1) Consider the finite population consisting of the values 3, 9, 15, and 21 (which might be written on 4 slips of paper).

a) List all possible samples of size 2 that can be taken without replacement from this finite population:

There are 4 choose 2 which equals 6 ways to select unordered samples from these 4 elements. They are the subsets \( \{3,9\}, \{3,15\}, \{3,21\}, \{9,15\}, \{9,21\}, \{15,21\} \)

b) Calculate the mean of each sample listed in part a) and assigning each sample an equal probability of 1/6 verify the mean of the sample mean agrees with the population mean. (Recall a random sample for
a finite population means each same size sample is equally likely). WEEK 8 page 10

The mean of each sample in the same order as our list is 6, 9, 12, 12, 15, 18. These are the 6 possible values of the random variable \( \bar{X} \) each occurring with equal probability 1/6 so by definition, the expected value of \( \bar{X} \) is the sum of the values times their probabilities

\[
\frac{(6+9+12+12+15+18)}{6} = \frac{72}{6} = 12
\]

But one easily checks the mean of our population of size 4 is the same value:

\[
\frac{(3+9+15+21)}{4} = \frac{48}{4} = 12
\]

In the general case one has for the expected value of \( \bar{X} \) with \( j \) ranging over all \( N \choose k \) subsets of size \( k \):

\[
\left( \frac{1}{N} \right) \sum_{i=1}^{N} \sum_{j=1}^{\left| k \right|} k_{ij} X_i/k \text{ where } k_{ij} = 1 \text{ if } i \in j^{th} \text{ subset of size } k \text{ } ( k_{ij} = 0 \text{ else })
\]

But the number of subsets of size \( k \) that a given element \( X_i \) belongs to is the same as all sets of the remaining \( k-1 \) elements in the subset of size \( k \) chosen from the remaining \( N-1 \) elements of the population or in other words using the properties of binomial coefficients, the above sum is just the population mean

\[
\left( \frac{1}{N} \right) \sum_{i=1}^{N} \left( \frac{N-1}{k-1} \right) X_i = (1/N) \sum_{i=1}^{N} X_i = \mu
\]

**EXAMPLE 2** (like 7.2) Show that for a sample of size \( n \) of Bernoulli 0-1 valued random variables with parameter \( p = \) success probability = \( E[X_i] \) and constants \( k_i \) with \( \sum_{i=1}^{n} k_i = 1 \) the random variable which is the linear combination \( \sum_{i=1}^{n} k_i X_i \) of the Bernoulli random variables is an unbiased estimator for the parameter \( p \):

We wish to show

\[
E\left[ \sum_{i=1}^{n} k_i X_i \right] = E[k_1 X_1 + k_2 X_2 + ... + k_n X_n] = p
\]

but by the properties of expectations this is just

\[
\sum_{i=1}^{n} k_i E[X_i] = \sum_{i=1}^{n} k_i p = p \sum_{i=1}^{n} k_i = p
\]

as claimed since the constants \( k_i \) sum to 1. Note in particular when we choose \( k_i = 1/n \) the fixed factor 1/n then comes outside the sum leaving the sum of the Bernoulli random variables which gives a binomial random variable

\[X = \text{number of successes in } n \text{ Bernoulli trials} = \sum_{i=1}^{n} X_i \]

Then \( \hat{p} = X/n \) just gives the sample mean of the Bernoulli random variables which by the above is an unbiased estimator of the parameter \( p \).

**EXAMPLE 3** (like 7.4, 7.5, 7.13) For a sample of size \( n = 64 \) with sample mean and standard deviation

\[ \bar{x} = 20,432 \text{ and } s = 18,064 \]

a) what can one assert with 98% confidence about the maximum error if the sample mean is used as an
This is a large sample so we can safely replace the population standard deviation by the sample standard deviation and then the central limit theorem gives us the approximation (without knowing the distribution of the population we are sampling from):

\[
Error = |\bar{X} - \mu| \leq z_{\alpha/2} s / \sqrt{n} = E = z_{0.01} \sqrt{(18,064) / 64} = 2.327 \cdot (18064) / 8 = 5254.36
\]

b) The corresponding 98% confidence interval for the population mean \(\mu\) is

\[
[\bar{X} - z_{\alpha/2} s / \sqrt{n}, \bar{X} + z_{\alpha/2} s / \sqrt{n}] = [20,432 - 5254.36, 20,432 + 5254.36] = [15177.6, 25686.36].
\]

c) (like 7.13) What could we say about the error with 98% confidence if we estimate the population standard deviation with the sample standard deviation \(s\) and we know that the population is of finite size \(N = 160\). In this case our estimate for the standard error of the mean is no longer \(s / \sqrt{n}\) but rather the variance gets multiplied by the finite population correction factor so the standard deviation uses the square root of this correction factor hence now we have

\[
E = z_{\alpha/2} \sqrt{n} \sqrt{\frac{N - n}{N - 1}} = 2.327 \cdot (18064 / 8) \sqrt{\frac{160 - 64}{159}} = 5254.36 \sqrt{\frac{96}{159}} = 5254.36 \cdot 0.777029 = 4082.79
\]

for our error bound.

**EXAMPLE 4** determine sample size (problem 7.11, like 7.12) The dean of a college wants to use the mean of a random sample to estimate the average amount of time it takes students to get from one class to the next, and she wants to be able to assert with 99% confidence that the error is at most .25 minute. If it is known from experience that \(\sigma = 1.40\) minutes, how large a sample will she have to take?

We want the maximum error \(E = z_{\alpha/2} \sigma / \sqrt{n} = 2.575 (1.4) / \sqrt{n} \leq 0.25\) with \(\alpha = .01\)

or \(n \geq (z_{.005} \sigma / E)^2 = (2.575 (1.4) / .25)^2 = 207.9\) so \(n \geq 208\)

**EXAMPLE 5** (like 7.22 and 7.23 ) Inspecting computer central processing units (cpu's), a quality control engineer detects 32, 35, 36, 38 and 39 defective cpu's in 5 production runs of size 300 cpu's each.

a) What can he assert with 95% confidence about the maximum error if he uses the mean of the sample of size 5 as an estimate of the true average number of defectives in a production run of size 300?

The number of defectives being binomial random variable are actually sums of 300 Bernoulli 0-1 valued random variables and as this is a large number are well approximated by the normal approximation to the binomial. Thus we can regard our 5 observations as coming from an approximately normal population and calculating the sample mean and sample variance we have

\[
\bar{x} = \frac{32 + 35 + 36 + 38 + 39}{5} = 30 + \frac{2 + 5 + 6 + 8 + 9}{5} = 36
\]

and

\[
s^2 = \frac{1}{4}(( -4)^2 + ( -1)^2 + 0^2 + 2^2 + 3^2) = \frac{30}{4} = \frac{15}{2}
\]

so \(s = \sqrt{\frac{15}{2}}\).

Then for a t random variable with parameter 4 degrees of freedom we have
Error = \mid \bar{X} - \mu \mid \leq t_{\alpha/2} s / \sqrt{n} = E = t_{0.025} \sqrt{ \frac{15}{2} / \sqrt{5} } = 2.776 \sqrt{ \frac{3}{2} } = 3.3999 \quad \text{WEEK 8 page 12}

is the maximum error.

b) What is the corresponding 95% confidence interval for the population mean?
This is just the sample mean plus or minus the above maximum error so
\[ [ 36 - 3.3999, 36 + 3.3999 ] = [ 32.6001, 39.3999 ] \]

Instead of viewing this as a t-distribution with sample size 5 so 4 degrees of freedom, notice that the total number of defectives is 180 in 1500 Bernoulli trials so our estimate for \( p \) is 180/1500 = 3/25 (which is just the sample mean of the 1500 Bernoulli 0-1 random variables) as the probability that any one cpu is defective with variance for a single cpu being \( p(1-p) \). The expected number of defectives in a sample of size 300 is \( np = 300p \) for a binomial random variable with \( n=300 \). Then 300(3/25) = 36 is our estimate of this. We are 95% confident that
\[-z_{\alpha/2} = -z_{0.025} = -1.96 \leq Z = \frac{3/25 - p}{\sqrt{p(1-p)/\sqrt{1500}}} \leq 1.96 .\]

The simplest approximation here is to estimate the standard deviation in the denominator by using our estimate 3/25 for \( p \). This is not necessarily as accurate as exactly solving the above quadratic inequality for \( p \) which we convert into two quadratic equations for (25/3) \( p \):
\[ ((25/3)p)^2(1+(1.96)^2/1500) - 2(1+(25/3)(1.96)^2/1500)((25/3)p)+1 = 0 \]
which could then be solved by the quadratic formula. The upper bound for \( p \) so obtained would give an upper bound when plugged into the variance \( p(1-p) \). Then our max error would take the form
\[ 300/3/25 - p \mid < 300 \times 1.96 \times \text{square root ( upper bound for } p(1-p) \text{) / square root of 1500} .\]
I am not sure how this would compare to the above t-distribution calculation.
It is also interesting to note that the sample variance of a single Bernoulli obtained from the sample of 1500 Bernoulli r.v.’s which should be a fair estimate of the actual variance \( p(1-p) \) is a factor of (1500/1499) (nearly identical to 1) times the approximation obtained by replacing \( p \) by its estimator 3/25 or \( (3/25)(1-(3/25)) \) pooling the 1500 defectives together may overlook the various 5 constraints specified by giving the number of defectives in each of the 5 production runs.

EXAMPLE 6 (problem 7.26 similar to 7.25) Find the maximum likelihood estimator

a) for \( \beta = 1/\lambda \) when \( f(x;\beta) \) is the exponential distribution:
With \( f(x;\beta) = (1/\beta)e^{-x/\beta} \) we want to find the parameter which maximizes the joint density
\[ L(\beta) = f(x_1;\beta)f(x_2;\beta) \cdots f(x_n;\beta) = \beta^{-n}e^{-(x_1+x_2+\ldots+x_n)/\beta} \]
for a fixed sample of observations (i.e. which maximizes the so called likelihood function).
Since the log function is increasing everywhere, this is the same as maximizing the log likelihood (logarithm of the likelihood)
\[ \log L(\beta) = -n\log \beta - (x_1+x_2+\ldots+x_n)/\beta .\]
Setting the derivative with respect to the parameter equal to zero gives
\[ -n/\beta + (x_1+x_2+\ldots+x_n)/\beta^2 = 0 \]
or \( \beta = (x_1+x_2+\ldots+x_n)/n = \bar{x} \).
Thus the sample mean is the maximum likelihood estimator for the parameter \( \beta = \mu \) which for exponential random variables is the population mean.

b) for \( \mu \) when \( f(x;\mu) \) is a normal distribution with variance 1. Maximizing the likelihood
is the same as maximizing the log likelihood

\[ \log L(\mu) = \log (\sqrt{2\pi})^{-n} e^{-\frac{1}{2}[(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2]/2} \]

Setting the derivative with respect to \( \mu \) equal to zero gives

\[ (x_1-\mu)+(x_2-\mu)+\ldots+(x_n-\mu) = 0 \quad \text{or} \quad \hat{\mu} = \bar{x} \]

so that the maximum likelihood estimator for the population mean \( \mu \) is again the sample mean.