

## Notes for MSIT 431 for Fall 2011

### Statistical Methods<sup>1</sup> by A. H. Haddad

#### 6. Inference Basics

##### 6.1 Introduction

So far we considered a variety of probabilistic models to approximate various applications in a variety of communication systems or networks. However, these models require knowledge of the values of some parameters such as the mean, variance, correlation coefficient, and sometimes even the shape of the distribution if it is not assumed in advance. Inference deals with the subject of confirming the validity of a probabilistic model by using empirical data, and that of obtaining a reasonably accurate estimate of the parameters of probabilistic models from experimental data. This section deals with the basic inference problem and it is not intended as a comprehensive coverage of statistical inference. We shall first consider the problem of estimating the mean of a distribution from data samples, when the variance is known. Then we shall address in principle the issue of how to handle more complex cases or models.

##### 6.2 Sampling Basics

We discussed in Section 5.7 how the law of large number helps us get more accurate information from data samples as the number of samples increases. However, the key to the successful utilization of data is that the data samples are independent else we shall not have an accurate estimate of any parameter of interest. In addition, we must insure that all the samples follow the same model, so that we cannot mix different data models, since then the results may not be useful. An example of such a case is if we wish to model the lengths of messages, we may have to sample separately voice messages, data messages, and internet-browsing messages. If we mix all our samples together we shall obtain incorrect model parameters.

We can summarize the basic principles of obtaining data samples as follows:

- a. **Random Sample.** The sample should be random with each data point independently selected and each member of the group having equal probability of being selected. In such a case a number can be assigned to each member of the group to be sampled, and then use a random number generator to pick the sample points from among the group.
- b. **Replication.** It should be possible to replicate the experiment, with a different random sample and the result will be within the same tolerance parameters. If replication yields answers outside the expected tolerances, then our sample and the results are not correct.

---

<sup>1</sup>Copyright 2011 A. H. Haddad

- c. **Stratified Sampling.** If we have different groups or different types, the fraction of the number of samples among each group must be equal to the number among the population. As an example consider the different types of messages. If the system has certain fraction of messages of each type, then the same fraction needs to be included in each sample.

We turn now to the actual problem of inference, and we start by the estimate of the mean, but before we do so, we address the general problem of inference.

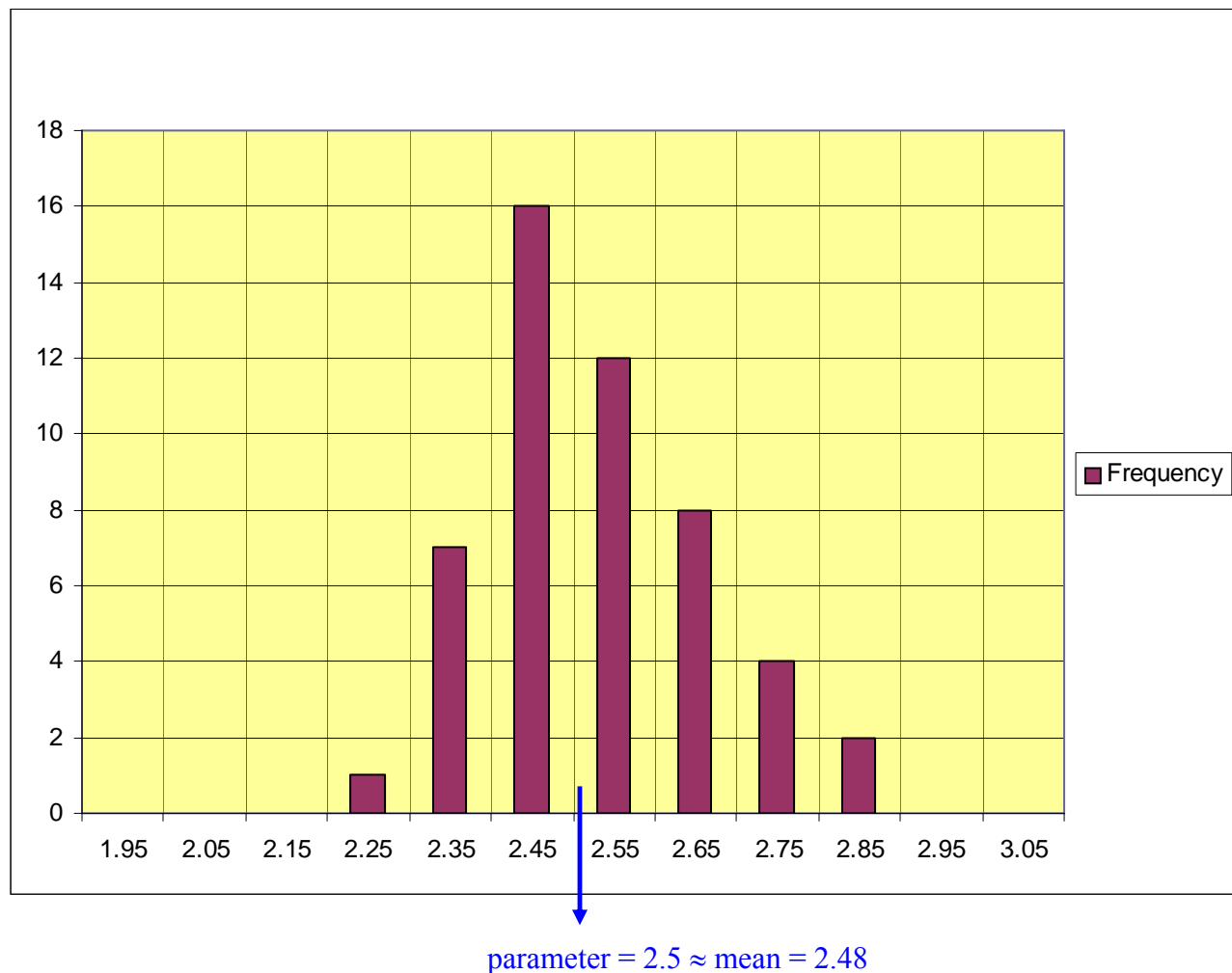
### 6.3 Properties of Estimators

When we estimate a parameter from data samples, we obviously may obtain a different answer if we were to repeat the process at another time or with a different data sample from the same population. This means that for us to trust a parameter obtained from a sample, we need to find the properties of the resulting estimate if the sample is changed. Since we are dealing with random variables, we can in theory repeat our process and check what happens as we do so. Suppose we are estimating an unknown parameter,  $\lambda$ , of some distribution, and we collected  $n$  samples and obtained one value for the estimate of the parameter,  $\hat{\lambda}$ . If we now repeat the same process with a variety of data samples we should obtain several such estimates. We can then plot the histogram of these estimates, and we call the resulting histogram, the *sampling distribution* of the estimate. The key point here is that every time we repeat the experiment we obtain a different estimate and our objective is to find the property of all such estimates.

Figure 6.1 shows the sampling distribution of estimating such a parameter using  $n = 100$  samples. The sampling distribution was obtained by using 50 replication of the experiment with each having different 100 samples. When we say different, we mean that the sample was selected independently of the others, but that does not mean it did not include some of the same subjects or data points. It means that if we had a population of 10,000 and we picked a random sample of size 100, we had to repeat the experiment with 50 such different samples. In this case the replications may contain some of the same individuals, but were picked entirely independently from each other.

The true value of the parameter we were estimating was 2.5. We see that since we used only 100 samples to estimate its value, we obtained answers as low as 2.2 and as high as 2.9. The sampling distribution shows how the estimate is distributed as the data samples change. Since the sampling distribution is a distribution, we can speak in such cases about the mean of the estimate and the variance (or the spread) of the estimate. These will be the *mean* and the *variance* of the *sampling distribution*, and they will characterize how good our estimate is. For the example shown above, we computed the sample mean and the sample standard deviation for the 50 replicated estimates of the parameter. We obtained a sample mean of 2.48 (not too far from our true value) and a sample standard deviation of 0.136. Since in practice we cannot replicate the experiment as it defeats the purpose of sampling in the first place, the question is how to characterize the sampling distribution without actually repeating the experiment. In other words, how do we obtain the variance of the sampling distribution without doing any replications?

We can try to approximate the sampling distribution. For example, in this case we tried to approximate it with a normal distribution and we obtained for this case Figure 6.2.



**Figure 6.1** A sampling distribution for the case of  $n = 100$  using 50 replications

However, in order to use such an approximation, we need to obtain the relevant parameters for such an approximation, which are the mean and variance.

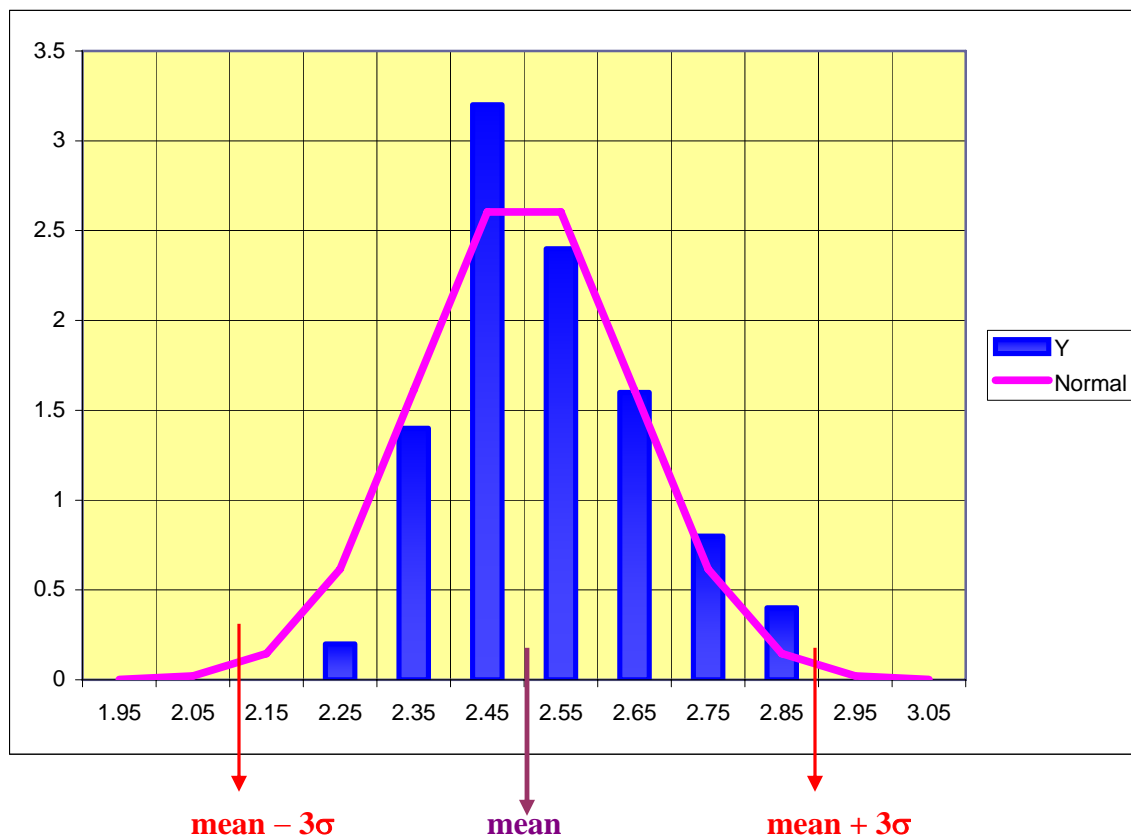
The mean and variance also identify two basic characteristics for the estimate:

### **Bias:**

The first characteristic is the difference between the true value of the parameter we are estimating and the mean of the estimate we derived from the sampling distribution. This difference is called the *bias* of the estimator. In the example above the bias is

$$2.48 - 2.5 = -0.02$$

If the bias is zero we say the estimator is *unbiased*. In our example the estimator is biased but by a very small bias.



**Figure 6.2** Normal approximation of the sampling distribution of Figure 6.1

### Variance:

The second characteristic of the estimator is its spread about the mean; in this case we use as a measure of the spread the standard deviation of the estimate, as obtained from the sampling distribution, which for our example was 0.136. We prefer to have as small a spread as possible, since the spread determines the margin of error in our estimator relative to the true value we are estimating. The square of the standard deviation is called (obviously) the sample *variance* of the estimator. A good estimator is one with zero bias and the smallest variance of all such estimators based on the data available. We call such an estimator *unbiased minimum variance* estimator.

Desired characteristics of an estimator are small bias and small variance. In most cases we prefer to have an unbiased estimator, but if the unbiased estimator will have too large a variance, we may settle for an estimator with small bias (even if non-zero) but whose variance is also small. We illustrate these with pictorial examples.

The three graphs given in Figure 6.3 show three sampling distributions for the same parameter we discussed earlier (whose true value is 2.5) and we see that the first in Figure 6.3(a) has a small variance but a large bias, the second in Figure 6.3(b) has a small bias but a large variance, and the third in Figure 6.3(c) has both a large variance and a large bias.

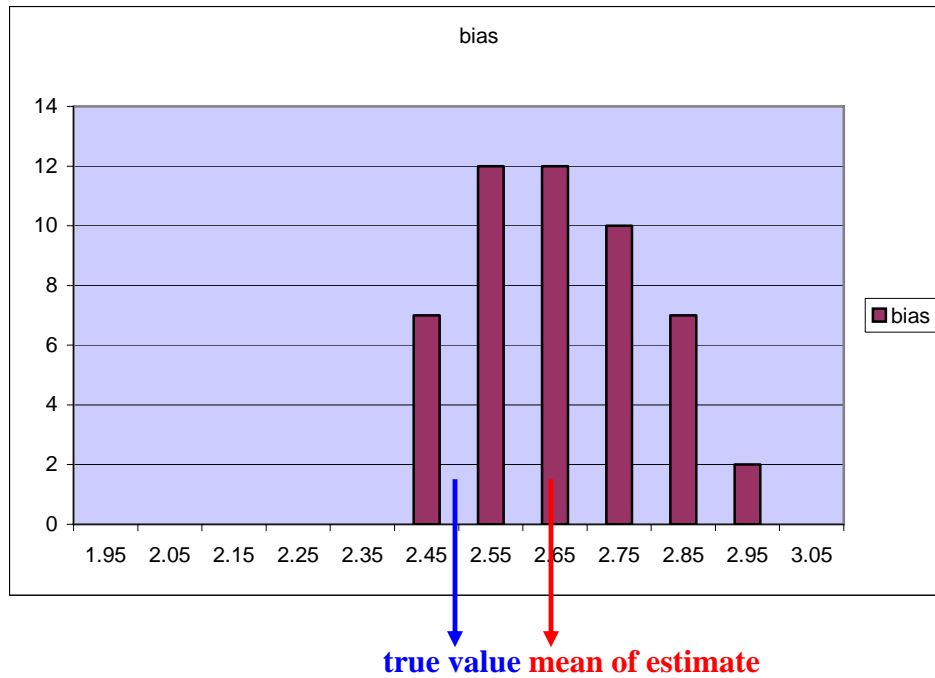


Figure 6.3(a) A biased estimator with small variance for the case of Figure 6.1

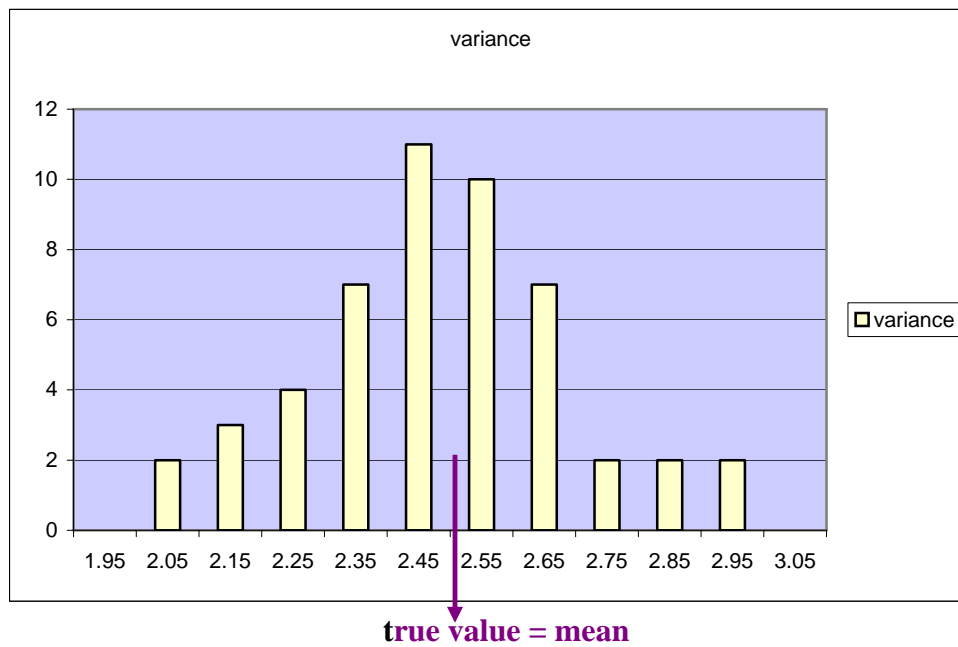
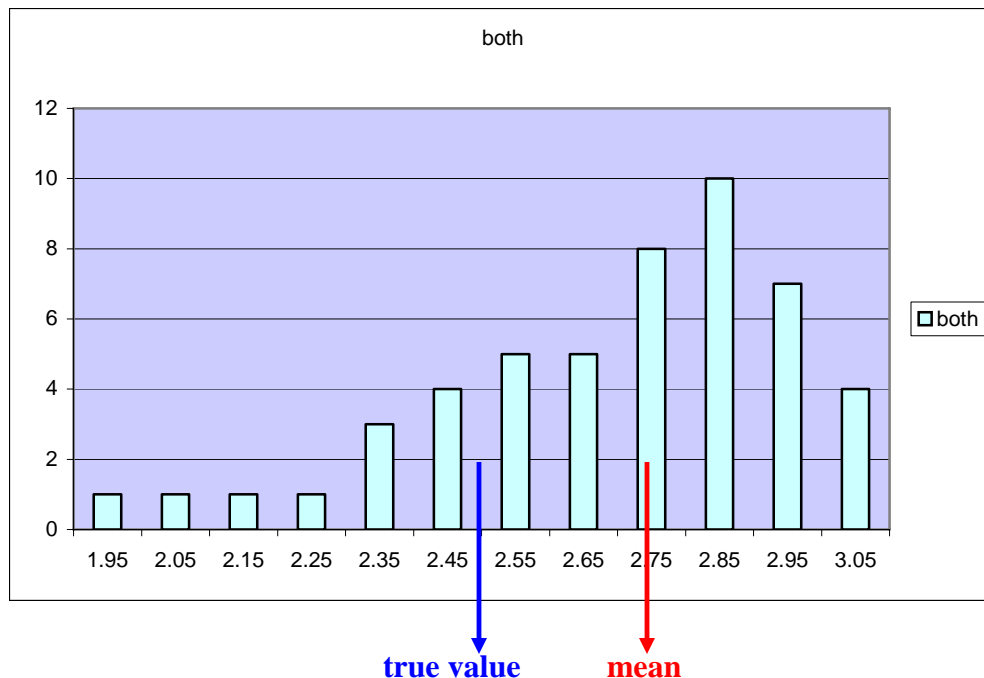


Figure 6.3(b) An unbiased estimator with large variance for the case of Figure 6.1



**Figure 6.3(c) An estimator with large bias and variance for the case of Figure 6.1**

We now consider how we can provide a quantitative measure for the estimate of the mean of a distribution, which provides information as to the accuracy of the estimator.

## 6.4 Inference for the Mean

In discussing inference for the mean of a distribution, we shall assume that we have an independent random sample  $\{x_i\}$  of size  $n$  from a population, which is assumed to have the same distribution. We also assume that we know the variance of the true distribution,  $\sigma^2$ , but we do not know the mean  $\mu$  of the distribution. If we take the sample average as the estimate of the mean, we know by the law of large numbers that the estimate approaches the mean as the number of samples approaches infinity. The question is how to derive a measure of the accuracy if  $n$  is finite, which it is in practice? In order to answer this question, we must ensure that the samples obtained are independent.

An example of where we know the variance (or know some bound on the variance) is the case where we are measuring a variable and our measurement device is calibrated so that we know its error variance. In the binomial case where we do not know the probability of success  $p$ , we know that the variance is  $p(1-p)$  for each trial, but since  $0 < p < 1$ , we can show that the variance is less than 0.25, so that the standard deviation is less than 0.5. It should be noted that the variance,  $\sigma^2$ , we speak of is the variance of a **single** measurement or a **single** observation. This is the variance of the **original** distribution whose mean we wish to estimate. This is **not** to be confused with the variance of the **estimate**, representing the accuracy of the estimation process.

As mentioned above, we are using the sample average as our estimate of the mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i \quad (6.1)$$

First note that the data points in the sample average (6.1) are assumed to be the outcomes of random variables  $\{X_i\}$ , whose mean we are trying to estimate. In this case we also know that the sample average has mean and variance obtained by the rules of mean of sums of random variables as we derived earlier:

$$E\{\bar{x}\} = \frac{1}{n} \sum_{i=1}^n E\{X_i\} = \mu \quad (6.2)$$

The reason we are using the random variables instead of their outcomes to analyze the result of the sample mean, is that we wish to find out what happens if we were to repeat the experiment with several different data sets. We obviously may get a different estimate every time we repeat the process. From Equation (6.2) we know that the average of these estimates will indeed be equal to the mean we are trying to estimate. However, we need to find the size of the spread of the sample average from the true mean  $\mu$  we are trying to estimate. Since the mean of the sample average is indeed equal to the parameter,  $\mu$ , we next need to find its variance, so as to know the spread of our estimate about the true mean as we discussed in the previous section. The variance of the sample average is found by using the fact that the variance of a sum of independent random variables is equal to the sum of the variances and we already derived that in Section 5.7 where we denoted the variable as  $Y$ . It is apparent now why we insisted that the samples we use are independent, else the result for the variance of the estimate will not hold.

$$\text{Var}\{\bar{x}\} = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n} \quad (6.3)$$

We now apply the central limit theorem to approximate the probability density of the sample average (which is the sampling distribution we mentioned earlier). In such a case we know for example, that the estimate will be within 3 standard deviations of the true mean with probability 99.7%, and within 2 standard deviations with probability 95.45%.

In general if we are to pick a confidence level  $C$  (say 98%), what is the margin of error of our estimate about the true mean  $\mu$  with that confidence level? We call this margin of error “*confidence interval*” and we denote it by  $m$ , then we wish to have a measure of how confident we are in the answer of finding the true mean within a margin of error of the estimate obtained:

$$P\{|\bar{x} - \mu| \leq m\} = C \quad (6.4)$$

By using the central limit theorem we find that this probability is equal to:

$$C = P\{|\bar{x} - \mu| \leq m\} = P\left\{-\frac{m\sqrt{n}}{\sigma} \leq Z \leq \frac{m\sqrt{n}}{\sigma}\right\} \quad (6.5)$$

Here  $Z$  represents the unit normal random variable with zero mean and unit variance. We therefore can evaluate the expression for the confidence level by using the normal cumulative distribution table as:

$$C = P\{|\bar{x} - \mu| \leq m\} = \Phi\left(\frac{m\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{m\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{m\sqrt{n}}{\sigma}\right) - 1 \quad (6.6)$$

Figure 6.4 shows the sample distribution of the estimate of the mean, its mean and standard deviation, and the margin of error  $m$ , and indicates the region, which should define the value of  $C$  = the area between the two vertical lines in the graph.

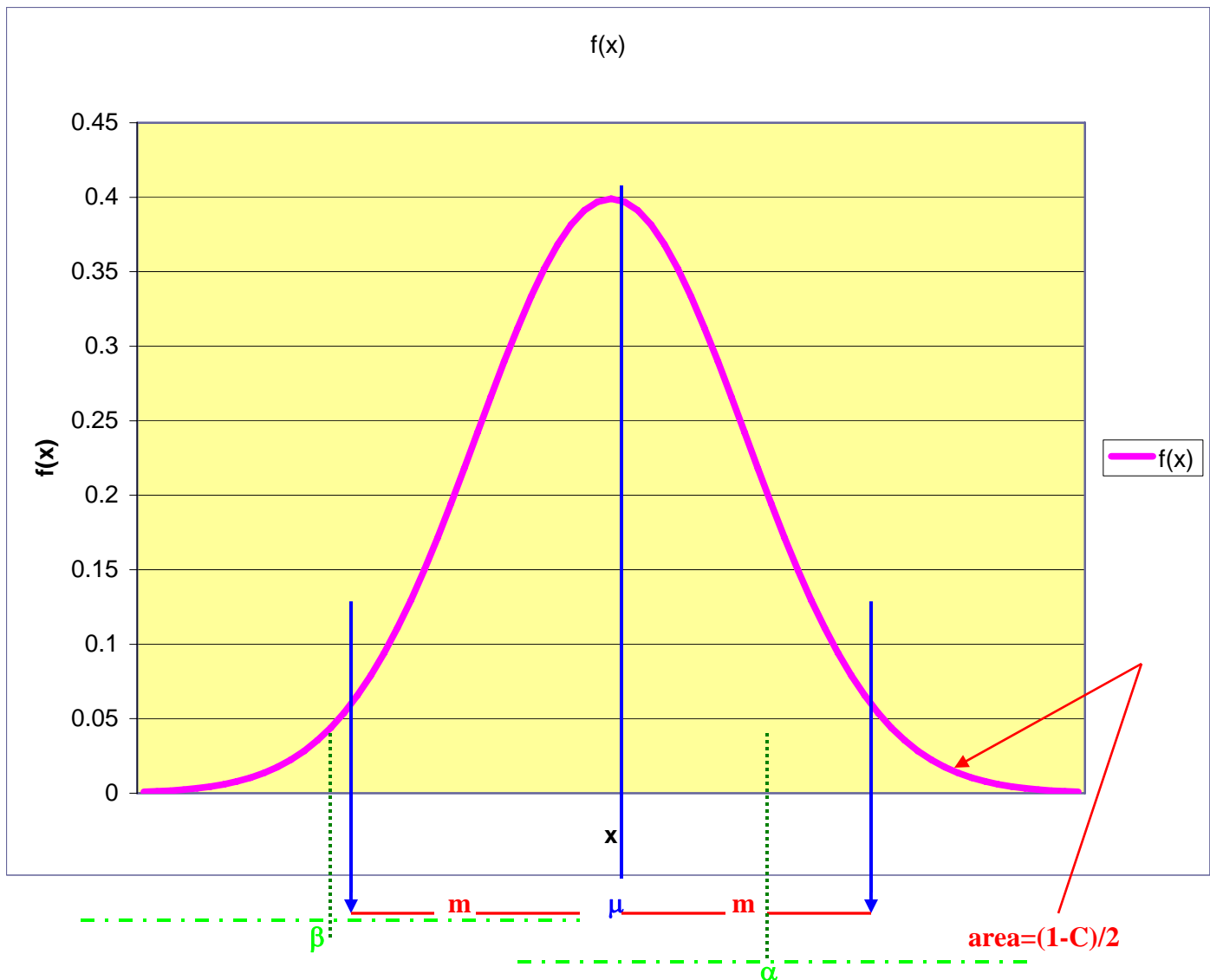


Figure 6.4 Illustration of the confidence interval for inference of the mean

We now have three choices:

1. If we fix  $m$  and  $n$ , then using the expression for  $C$ , we can find the confidence level  $C$  from (6.6).
2. If we fix  $C$  and  $n$ , we can then find the **confidence interval  $m$**  as follows: From  $C$  we obtain the value  $Z_c$  that satisfies the expression:

$$2 \Phi(Z_c) - 1 = C, \text{ or } 1 - \Phi(Z_c) = \frac{1-C}{2}, \text{ or } \Phi(Z_c) = \frac{1+C}{2} \quad (6.7)$$

The value of  $Z_c$  can be found directly from the unit normal table, or from special tables that provide the value of  $Z_c$  for every value of  $(1-C)/2$ . The expression for  $m$  becomes:

$$m = \frac{Z_c \sigma}{\sqrt{n}} \quad (6.8)$$

3. Finally, if we desire both a specific  $C$  and a specific confidence interval  $m$ , then the only way we can satisfy both is to change the number of samples  $n$ . First we have to find the value of  $Z_c$  from  $C$  as we have done above and then evaluate  $n$  as in the following expression:

$$n = \left( \frac{Z_c \sigma}{m} \right)^2 \quad (6.9)$$

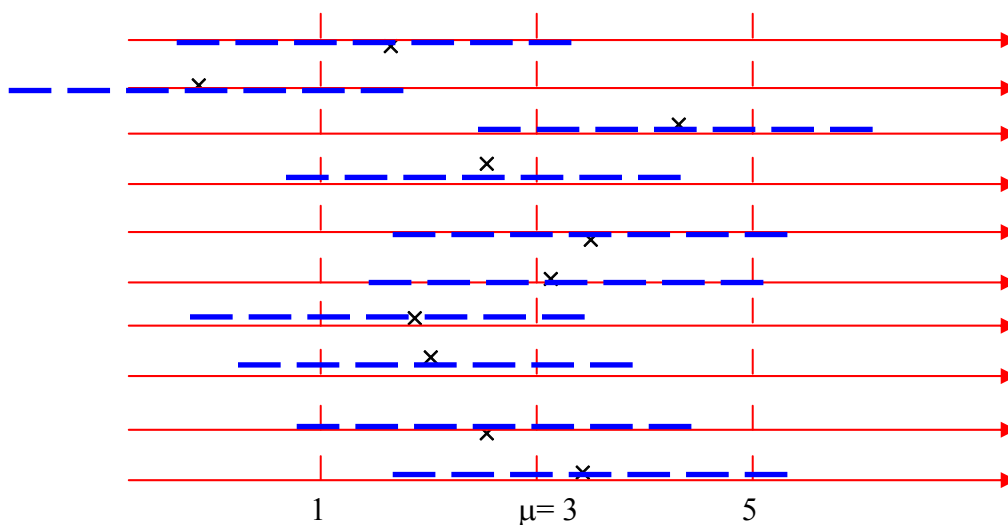
What does the margin of error and confidence level tell us? Suppose we have a confidence level of  $C=98\%$  and margin of error of  $m=5$  in the measurement of some unknown quantity,  $\mu$ . Then we carry out the experiment one time to obtain the value  $\bar{x}$  as an estimate of  $\mu$ . If we repeat this experiment 100 times we may obtain different values of  $\bar{x}$ , but 98 of these values (**on average**) will be within the interval:

$$(\mu-m, \mu+m) = (\mu-5, \mu+5)$$

Only 2 (**on average**) of these 100 will be outside the range described above. That relation of the value of  $\bar{x}$  to the interval  $(\mu-5, \mu+5)$  implies that the unknown constant  $\mu$ , which we are trying to estimate, will fall in the interval  $(\bar{x}-5, \bar{x}+5)$ , 98 times and only 2 times the true value will be outside this interval, which we called the confidence interval.

A pictorial illustration of this observation for the case of  $C=90\%$  is shown in Figure 6.5. In this Figure we assumed that the unknown value was indeed  $\mu=3$ , and we obtained  $m=2$  for the assumed measurement standard deviation and number of measurements. Of 10 times that the experiment was repeated only once the true value was outside the confidence interval as shown in the figure.

In the figure we repeated the estimation 10 times and we obtained the following quantities for the value of  $\bar{x}$  : [1.56, 0.32, 4.20, 2.37, 3.40, 3.03, 1.78, 1.85, 2.55, 3.28]. We marked the values with an  $\times$  in 10 separate lines that show the true value as well as the confidence interval limits. We see that only for the value of  $\bar{x} = 0.32$  does the true value of the mean (in this case 3) fall outside the range  $0.32 \pm 2$  obtained by the estimate. In the other 9 experiments the true value falls within  $\pm 2$  of the value provided by the estimate,  $\bar{x}$ . A dashed line of length  $2m$  is shown centered around each result for  $\bar{x}$ , which illustrates the observation mentioned in the last sentence.



**Figure 6.5 Graphical illustration of the confidence interval**

Another way to illustrate this fact is to show the probability density function of the random variable  $\bar{X}$ , and draw an interval of  $\pm m$  around the true mean,  $\mu$ , of the resulting density. The probability of being outside this range is shown in Figure 6.4 and it is equal to  $[1-C]$ , so that if a sample  $\bar{x}$  of  $\bar{X}$  is obtained inside the marked range of values it will occur with probability  $C$ . In this case the true mean will be within the same margin of error  $\pm m$  around the sample value so obtained. This observation is also illustrated in Figure 6.4.

In the figure two values  $\bar{x} = \alpha$ , that is larger than the true mean, and  $\bar{x} = \beta$ , that is smaller than the true mean, are also shown to illustrate that if these values fall within the margin of  $\pm m$  from the true mean, then the true mean falls with the same margin from the resulting sample mean value. This last fact is shown by the margin of error displayed around these two values by dashed-dotted green lines of length  $2m$  (the  $2m$  value of the length is not noted in the figure in order not to clutter the picture).

We shall consider an example that illustrates all three possible choices. The first example is a simple one of measuring an unknown quantity when the instrument has measurement errors with zero mean and known standard deviation. The second example deals with a

binomial random variable: we need to find the probability  $p$  of the preference in a poll of  $n$  persons. The latter case can also apply to finding the error rate in a binary channel, or the probability of Heads in the toss of a coin, or in the probability of failures of components, or the probability of blocked calls, or the probability of dropped packets, among others.

### Example 6.1:

Consider the case of taking  $n$  measurements of the mean of a message length in a communication system, which is given as  $\mu = (1/\lambda)$ . Since the model is assumed to be exponential with parameter,  $\lambda$ , then the standard deviation is also equal to  $\sigma = (1/\lambda)$ . Suppose we take  $n$  independent measurements of the length of messages and we estimate the mean by the sample average. Suppose the sample average provides us with a numerical result of

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i = 5 \text{ seconds} \quad (6.10)$$

Since we are primarily interested in estimating the mean length, we shall assume as a first-order approximation that the standard deviation in each sample is known and is also equal to:

$$\sigma = (1/\lambda) = 5 \text{ seconds} \quad (6.11)$$

In order to see how the three choices discussed above are used, we shall consider each case.

1. We fix  $n$  at 100, and we would like a margin of error of  $\pm 1.0$  second. What should be our confidence level in this estimate? In order to find the confidence level in our choice of error margin  $m = 1.0$  we have to find  $m$  as a fraction of the standard deviation of the sampling distribution:

$$Z_c = \frac{m}{(\sigma/\sqrt{n})} = \frac{m\sqrt{n}}{\sigma} = \frac{1.0\sqrt{100}}{5} = 2.0 \quad (6.12)$$

Hence by using the normal table we find

$$C = 2 \Phi(Z_c) - 1 = 2 \Phi(2) - 1 = 2(0.97725) - 1 = 0.9545 = 95.45\% \quad (6.13)$$

In this case we can state that the value of the mean message length is equal to  $5 \pm 1.0$  second, with  $C = 95.45\%$  confidence.

2. We fix  $n$  at 100 again, but we would like a confidence level of  $C = 98\%$ . What is the margin of error in the estimate of the value of  $\mu$ ?

From the table of the normal distribution, for  $C = 0.98$  we find the value of  $Z_c = 2.326$ .

We therefore have the expression for the margin of error as:

$$m = \frac{Z_c \sigma}{\sqrt{n}} = \frac{2.326 \times 5}{\sqrt{100}} = 1.163 \quad (6.14)$$

We therefore can state that the value of the unknown mean message length is  $5 \pm 1.163$  seconds with 98% confidence.

3. Finally, we would like a margin of error of  $\pm 1.0$  second and confidence level of 98%. How many samples should we use? Here since we require confidence level  $C = 98\%$ , the inverse of the normal distribution provides us with the same value of  $Z_c = 2.326$  that we obtained in case 2. We now apply the expression relating  $n$  to the margin of error  $m$  and the value of  $Z_c$  to obtain the number of samples  $n$  that we require.

$$n = \left( \frac{Z_c \sigma}{m} \right)^2 = \left( \frac{2.326 \times 5}{1.0} \right)^2 = 135.3 \quad (6.15)$$

This means that if we wish to have the value of the measurement be  $5 \pm 1.0$  with confidence level  $C = 98\%$ , we need 136 samples.

The result may still be inaccurate as the error margin is 20% of the value we are measuring. What if we wanted an error margin of  $\pm 10\%$  of the measured value, namely an error margin of  $m = 0.5$  second? In this case if we still insist on confidence level of 98% then the number of sample increases by a factor of 4 as we are reducing the error margin by a factor of 2. Hence we then would require  $n = 4 \times 135.3 = 541$  samples.

### Example 6.2:

In this example, we wish to evaluate the probability  $p$  of failures of components. We test  $n$  components and we count the number of failures. We use the ratio as the estimate for  $p$ . We know that in this case the sample average (which is the number of failures divided by  $n$ ) would have mean  $p$ , and variance  $p(1-p)/n$  that is bounded by  $1/(4n)$ . (See Figure 6.6 showing the variance as a function of  $p$  for a single sample. The figure implies that we could use  $\sigma = 0.5$  as the maximum value of the standard deviation of a single observation.)

Hence the standard deviation is bounded by  $1/(2\sqrt{n})$ , which means that if we use this value our results will be on the conservative side. Again we have three possible choices in formulating the problem depending on what we wish our estimate properties to be. Suppose we sampled  $n$  components and we obtained as an answer for the estimate of  $p$  to be 0.32, so we can say that failure probability  $p = 0.32$ . However, what is the margin of error and confidence level for this result?

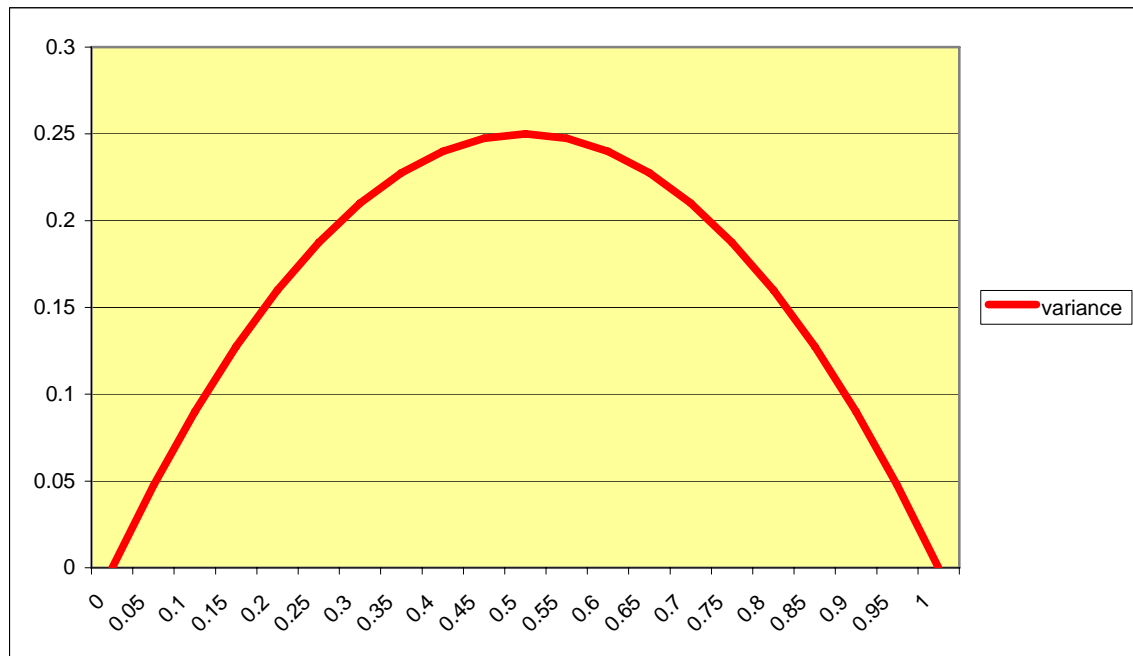
1. If we set  $n=100$  as in the previous example, and we require a margin of error of  $m = 0.08$  in the estimate of  $p$ , then what should be our confidence in this estimate? Again we first find the value of  $Z_c$  so that we can use the normal table for  $C$ :

$$Z_c = \frac{m\sqrt{n}}{\sigma} = \frac{0.08\sqrt{100}}{0.5} = 1.6 \quad (6.16)$$

That results in a value of C as follows:

$$C = 2\Phi(1.6) - 1 = 2(0.9452) - 1 = 0.8904 \quad (6.17)$$

We can now state that the failure probability is equal to  $p = 0.32 \pm 0.08$  with 89% confidence.



**Figure 6.6 Variance of the binomial random variable as a function of p**

2. If we keep  $n$  unchanged at 100 but we would like a higher confidence level of 98%, what should be the confidence interval  $m$  for our estimate? Here we obtain as we did in the previous example  $Z_c = 2.326$  which is derived from  $C = 0.98$  and the normal tables. As a result  $m$  will be given by:

$$m = \frac{Z_c \sigma}{\sqrt{n}} = \frac{2.326 \times 0.5}{\sqrt{100}} = 0.1163 \quad (6.18)$$

It means that we can state our answer as  $p = 0.32 \pm 0.1163$  with 98% confidence. We see that we have a very large margin of error, and that is due to the small number of samples and high confidence we required. Since we did not increase  $n$ , we only traded confidence level with confidence interval.

3. If we now would like to keep the same confidence interval  $m = \pm 0.08$  we wanted in part 1, but we wish  $C$  to be 98% again as in part 2, how many samples should we use? In this case we have the same  $C$  as in case 2, and hence we have the

same value of  $Z_c = 2.326$ . We now use this value together with the margin of error  $m = 0.08$ , to find  $n$  as we obtained in the previous example:

$$n = \left( \frac{Z_c \sigma}{m} \right)^2 = \left( \frac{2.326 * 0.5}{0.08} \right)^2 = 211.4 \quad (6.19)$$

We therefore can state that the value of  $p = 0.32 \pm 0.08$  with 98% confidence, provided we obtained the 0.32 value using  $n = 212$  samples.

Again we see that the error margin is about 25% of the estimated value of  $p$ . What if we dropped the error margin to  $\pm 10\%$  of  $p$ ? In this case  $m = 0.032$ , which is 0.4 (40%) of the  $m=0.08$  we had above. This means that the number of samples required will increase by a factor of  $(1/0.4)^2$  over the one obtained in this case earlier. The resulting number of sample is 6.25 times higher than the 211.4 we found above. The result is that we need  $n = 6.25 \times 211.4 = 1321$  samples!

What happens if we wish to use this approach when  $p$  is rather small? In such a case we may want to use instead of 0.5 for the standard deviation of a single sample (the parameter  $\sigma$  we use in all of our formulas), the actual estimate of the standard deviation for the binomial distribution (i.e.,  $\sqrt{p(1-p)}$ ) as we have done for the exponential case in Example 6.1. We should then obtain less conservative results and we may not need as many samples. We shall consider an example later in this section using this approach.

### Example 6.3:

Finally, let us consider polls. Usually polls try to determine the probability that the population is for one issue or against it (or for one candidate or another), so we are again trying to estimate  $p =$  the probability that a person in the population is for the issue, and then  $(1-p)$  is the probability that the person is against it. The confidence level is fixed in this case to 99.74% (assumed by pollsters as almost a certainty, being equal to three standard deviations, as they do not want to confuse the people by also stating their confidence level). For this confidence level we have the 3 standard deviation rule for the margin of error, so that  $Z_c = 3$ . The pollsters set the margin of error at  $\pm 0.04$  (4-percentage points in the resulting  $p$ ). We can therefore apply the results of Example 6.2 above using case 3 to obtain:

$$n = \left( \frac{Z_c \sigma}{m} \right)^2 = \left( \frac{3 \times 0.5}{0.04} \right)^2 = 1,406 \quad (6.20)$$

You may notice that when poll results are shown in the news they do mention the  $\pm 4$ -percentage-points margin of error and they also mention that about 1,400 or 1,500 persons were polled. They assume that the 99.7% confidence is a certainty. If we only needed a confidence level of 99% then we would have to poll only:

$$n = \left( \frac{Z_c \sigma}{m} \right)^2 = \left( \frac{2.58 \times 0.5}{0.04} \right)^2 = 1,040 \quad (6.21)$$

The answer would have to be slightly different when we have more than two possible choices in the polls. That is always the case when you include undecided or third or fourth candidates.

Three other issues that need to be considered:

- (1) The first deals with the use of the estimated probability in the binomial case;
- (2) The second deals with a single sided margin of error (in failure probability, for example, we do not care if our estimate is higher than the true value, so we do not need a margin of error on the lower side of  $p$ );
- (3) The third deals with the unknown variance case, where we use the sample to also estimate the variance or the standard deviation.

These three topics will be considered next.

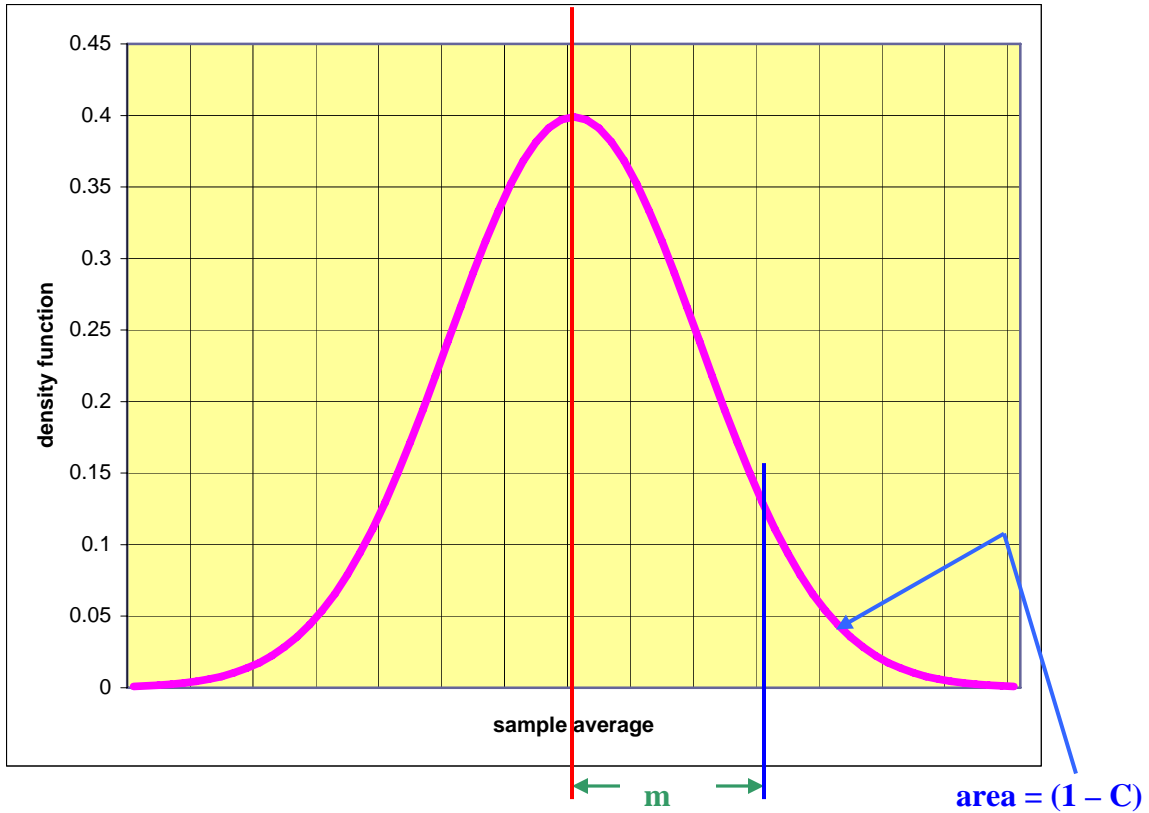
## 6.4 Inference of the Mean – additional cases

### 6.4.1 One-Sided Margin of Error

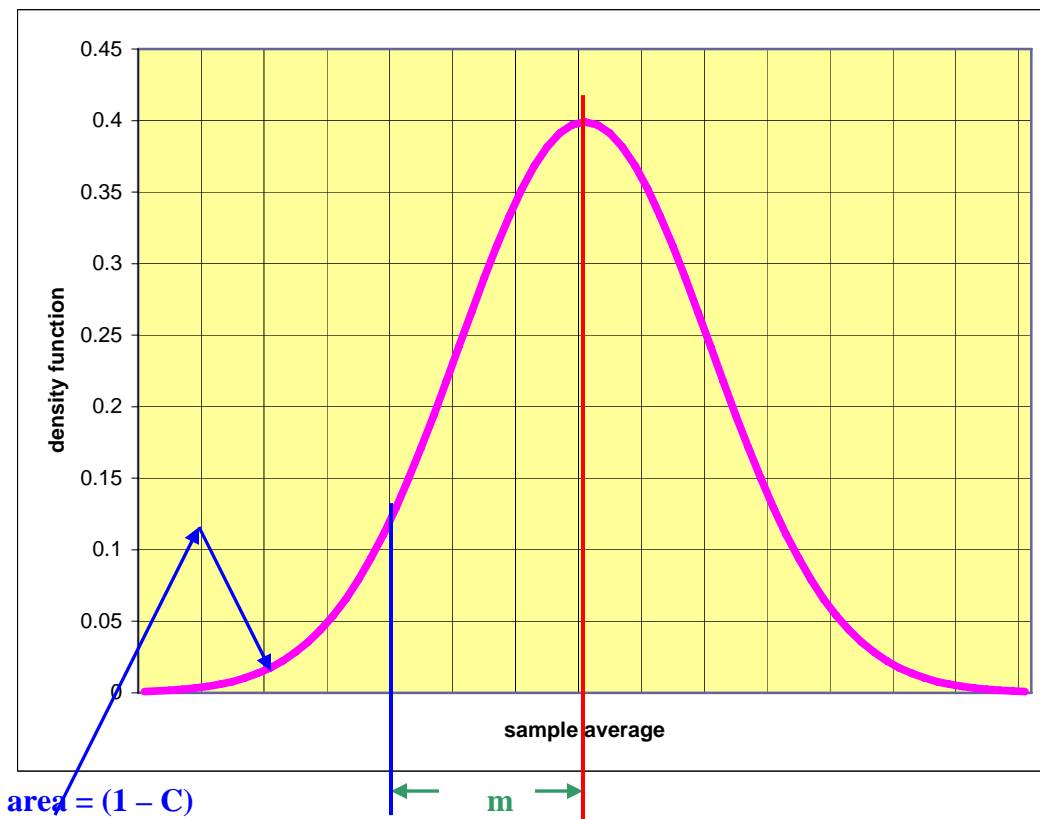
In many cases we may only be interested in a one-sided confidence interval. For such a case  $(1-C)$  need not be divided by 2 to obtain the probability in the tail of the distribution (that of being outside the interval). When do we use such a one-sided interval? Suppose we are measuring probability of error, then we do not care if the error in our estimate is on the negative side (the true error may be even smaller than our estimate), so instead of saying that the error probability is  $0.01 \pm 0.002$  with confidence  $C = 98\%$ , the user may well be happier to have us state the resulting estimate that  $p$  is less than 0.012 with confidence  $C = 99\%$ . Similarly if we measure the life of a component, rather than stating the answer that the average lifetime is  $1000 \pm 100$  with confidence  $C = 98\%$ , we can state the result as the lifetime is greater than 900 hours with confidence  $C = 99\%$ .

The process of determining the relations among  $C$ ,  $n$ , and  $m$  is the same as before except that in looking up in the tables for  $Z_c$  we look at the tail using  $\alpha = (1-C)$  rather than  $\alpha=0.5(1-C)$  which we used in the two-sided confidence interval. If we wish to keep the same confidence level when using one-sided confidence interval, then we could either obtain a smaller margin of error or use a smaller number of samples if we are satisfied with the margin of error and the confidence level.

The results are illustrated graphically for an upper or a lower confidence interval in Figures 6.7(a-b). The red line at the center of the distribution in both figures shows the estimate of the mean (i.e., the sample average). The blue line in both figures indicates the distance of the margin of error from the sample average. The area under the tail of the distribution in each figure should be equal to  $(1-C)$ . For two-sided confidence intervals the area under each tail is equal to  $0.5(1 - C)$ .



**Figure 6.7(a) Upper one-sided confidence interval**



**Figure 6.7(b) Lower one-sided confidence interval**

## 6.4.2. Unknown Variance Case

### What do we do when the variance is unknown?

It should be noted that we use the normal tables when the variances are known. We also use these tables when the problem involves a binomial random variable (or as some books call it estimating a proportion). We shall address the binomial case first.

The variance of the random variable in one trial in a binomial random variable case has a known variance if  $p$  (probability of success in one trial) is known. Therefore, once we estimated  $p$  by taking the relative frequency of the successes, we do not need to estimate the variance by using the sample variance. Consequently we have two ways of handling the variance of a binomial random variable. When  $p$  is expected to be not too far off from 0.5, we use 0.5 for the value of  $\sigma$  as the standard deviation of a single experiment, since that is the largest value of the standard deviation. This may be conservative but it provides us with more accurate answers than the t-distribution, which we shall consider later on. When  $p$  is small we can use the estimate of  $p$  in the expression for the variance of a single experiment, i.e.  $\sigma^2 = p(1-p)$ . Even if the confidence interval yields a value of  $p$  that is slightly different than the original estimate the difference in the standard deviation will not cause much change in  $C$  or  $m$ .

Consider as an example the case of estimating the error probability in a binary communication channel. Suppose we use 10,000 test bits and obtain 25 errors then we can state that our estimate of  $p$  is 0.0025. We now wish to find the confidence interval for a given  $m$  or vice versa. To that end we need to find the standard deviation  $\sigma$  of a single variable, which is equal to

$$\sigma = \sqrt{p(1-p)} = \sqrt{0.0025(1-0.0025)} \approx 0.05 \quad (6.22)$$

If we wish a two-sided confidence level of 95% then we may find  $Z_c = 1.96$  from the normal table, and then we use the approximate standard deviation to obtain the confidence interval  $m$ :

$$m = \frac{Z_c \sigma}{\sqrt{n}} = \frac{1.96 \times 0.05}{\sqrt{10000}} = 0.00098 \quad (6.23)$$

Note that if we wished a one-sided estimate in this case we would keep the same  $m$ , but the confidence level becomes 97.5%. Let us consider a one-sided margin of error for this case with the same confidence level of 95%, then the critical value  $Z_c = 1.645$  and our one-sided margin of error becomes:

$$m = \frac{1.645 \times 0.05}{\sqrt{10000}} = 0.0008225 \quad (6.24)$$

We can therefore state in the one-sided case that the true  $p$  is less than  $(0.0025+0.0008225) = 0.003325$  with confidence 95%.

If we were to correct the value of  $\sigma$  by using the upper bound on the estimate of  $p$  (0.003325) obtained from our inference process, we would see very little difference from the results obtained above.

### What do we do for non-binomial distributions?

In cases where the distribution is not binomial, then we have to estimate the variance by using the sample variance  $s_x^2$  from the same data. Since the estimate of the variance involves  $(n-1)$  degrees of freedom, we can compute the distribution of the sample mean to be what is called the t-distribution with  $(n-1)$  degrees of freedom. Tables are available for finding the value of the cut-off  $t_c$  so that  $P\{|T| < t_c\}$  is equal to  $C$ . The table shows the cut-off as a function of the confidence level (for two sided confidence interval) as well as a function of the area in the tail. Again, we can use one or two-sided tails by using  $(1-C)$  or  $0.5(1-C)$  as the area of the tail for which we wish to find the value of  $t_c$ . The variable  $T$  is again defined as

$$T = \frac{(\bar{x} - \mu)\sqrt{n}}{s_x} \quad (6.25)$$

So that it has mean zero and variance 1. Note that here we see that even though we obtain the estimate of the variance by dividing by  $(n-1)$ , the estimate of the variance of the sample mean is:

$$\text{Variance of sample mean} = \frac{s_x^2}{n} \quad (6.26)$$

Hence the standard deviation we use in the computation of the margin of error and confidence level is:

$$\frac{s_x}{\sqrt{n}} \quad (6.27)$$

Hence the same arguments in finding the estimate, its margin of error and confidence interval apply, except that we use different tables for the critical values as a function of  $C$ . As  $n$  approaches  $\infty$  the t-distribution approaches the normal and these are usually provided in the same table for various values of the number of degrees of freedom.

The expression relating the margin of error and the critical value of the t-distribution, which is obtained from the confidence level  $C$  is therefore:

$$m = \frac{t_c s_x}{\sqrt{n}} \quad (6.28)$$

While  $n$  appears in this expression, the table we use to obtain the relation between  $C$  and  $t_c$  uses  $(n-1)$  degrees of freedom!

The same tables are also applicable to the estimates of the regression line when we have two correlated data samples. The confidence level and confidence interval are applied here to the estimate of the regression line, namely both its intercept and its slope.

**What happens if we wish to estimate the variance with some confidence?**

Again we use the standard way of obtaining the estimate (the sample variance). Then we have to obtain the model for the sampling distribution of the sample variance. In such a case the distribution is a variation of the Chi-squared (it is the Chi-squared if the mean is known) with  $(n-1)$  degrees of freedom. The principle of deriving  $m$  and  $C$  is exactly the same just the tables used are different.

Other variations are also used for other models such as time to failure models where the normal approximation is not as valid as other models. We shall not address these any further.