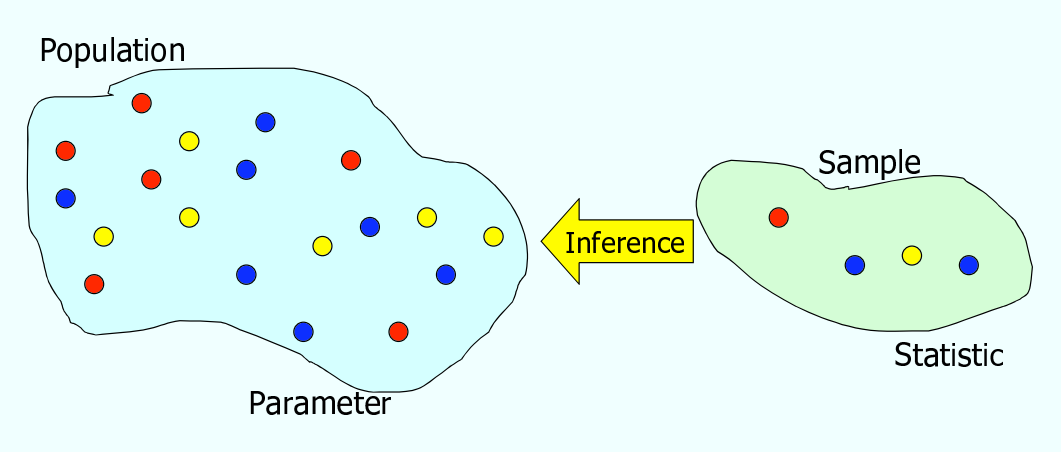
Introduction to Estimation

OPRE 6301

Statistical Inference ...

Statistical inference is the process by which we infer population properties from sample properties.



There are two types of statistical inference:

- Estimation
- Hypotheses Testing

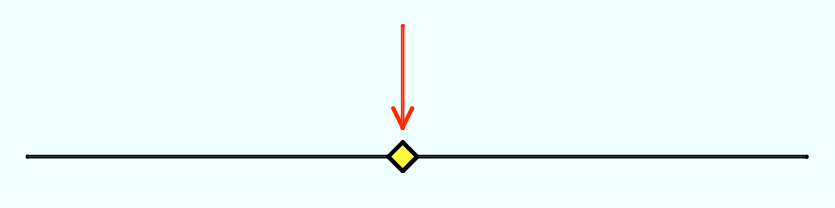
The concepts involved are actually very similar, which we will see in due course. Below, we provide a basic introduction to estimation.

Estimation . . .

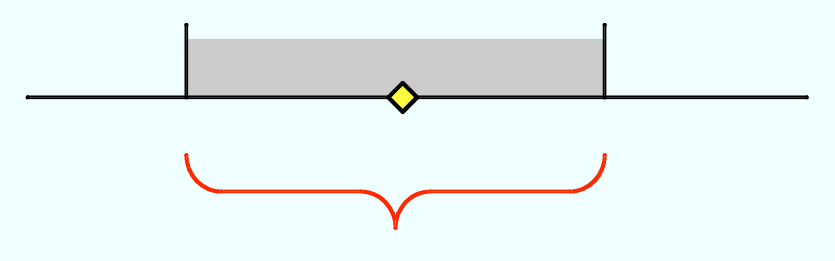
The objective of estimation is to approximate the value of a population parameter on the basis of a sample statistic. For example, the sample mean $\bar{X}$ is used to estimate the population mean μ.

There are two types of estimators:

- Point Estimator

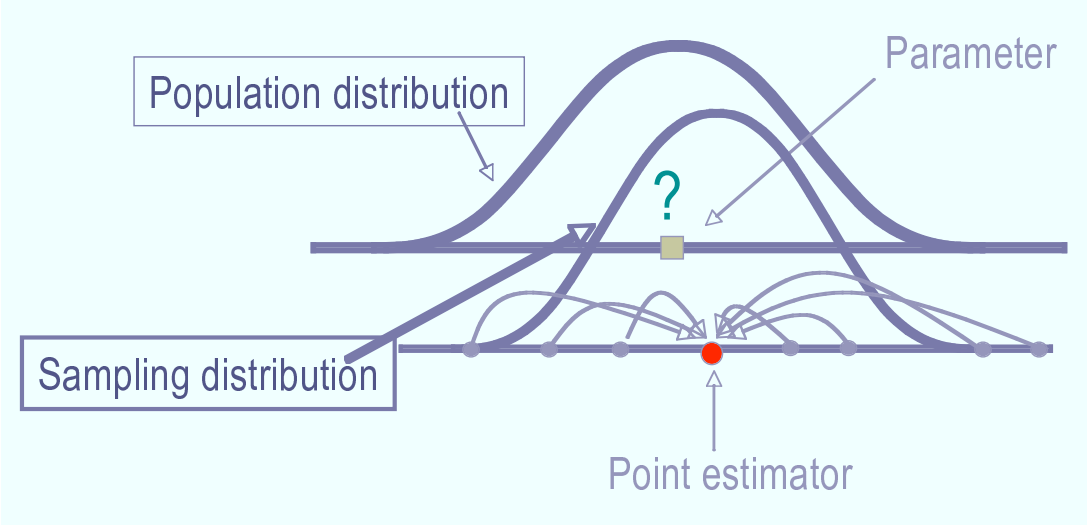


- Interval Estimator



Point Estimator . . .

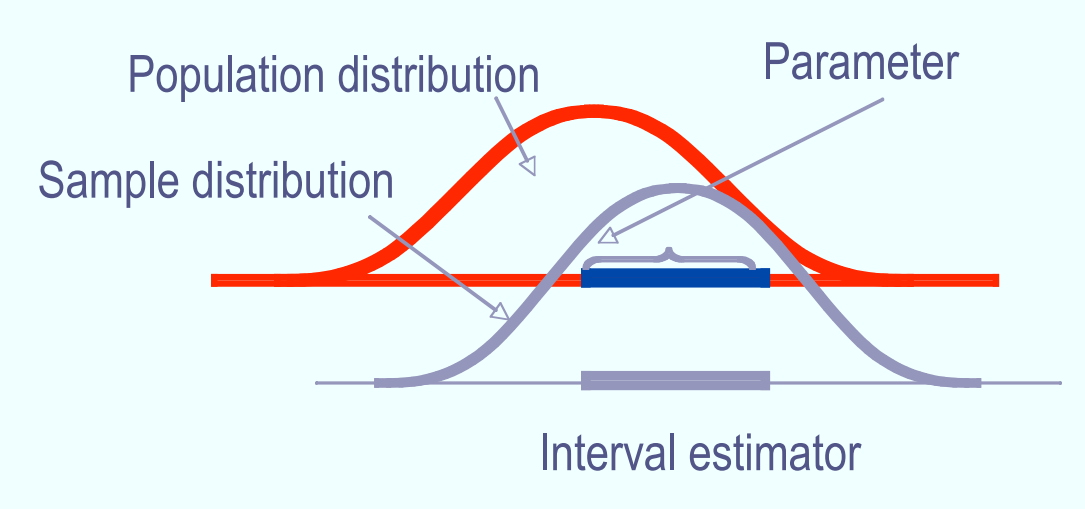
A **point estimator** draws inferences about a population by estimating the value of an unknown parameter using a single value or *point*.



Recall that for a continuous variable, the probability of assuming any particular value is zero. Hence, we are only trying to generate a value that is close to the true value. Point estimators typically do *not* reflect the effects of larger sample sizes, while **interval estimator** do . . .

Interval Estimator . . .

An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval. Here, we try to construct an *interval* that "covers" the true population parameter with a specified probability.



As an example, suppose we are trying to estimate the mean summer income of students. Then, an interval estimate might say that the (unknown) mean income is between \$380 and \$420 with probability 0.95.

Quality of Estimators ...

The desirability of an estimator is judged by its characteristics. Three important criteria are:

- Unbiasedness
- Consistency
- Efficiency

Details ...

Unbiasedness ...

An **unbiased estimator** of a population parameter is an estimator whose *expected value* is equal to that parameter. Formally, an estimator $\hat{\mu}$ for parameter μ is said to be unbiased if:

$$E(\hat{\mu}) = \mu\,. \tag{1}$$

Example: The sample mean $\bar{X}$ is an unbiased estimator for the population mean μ, since

$$E(\bar{X}) = \mu\,.$$

It is important to realize that other estimators for the population mean exist: maximum value in a sample, minimum value in a sample, average of the maximum and the minimum values in a sample ...

Being unbiased is a minimal requirement for an estimator. For example, the maximum value in a sample is *not* unbiased, and hence should not be used as an estimator for μ.

Consistency ...

An unbiased estimator is said to be **consistent** if the difference between the estimator and the target population parameter becomes smaller as we increase the sample size. Formally, an unbiased estimator $\hat{\mu}$ for parameter μ is said to be consistent if $V(\hat{\mu})$ approaches zero as $n \to \infty$.

Note that being unbiased is a precondition for an estimator to be consistent.

Example 1: The variance of the sample mean $\bar{X}$ is σ^2/n, which decreases to zero as we increase the sample size n. Hence, the sample mean is a consistent estimator for μ.

Example 2: The variance of the average of *two* randomly-selected values in a sample does *not* decrease to zero as we increase n. This variance in fact stays constant!

Efficiency ...

Suppose we are given two unbiased estimators for a parameter. Then, we say that the estimator with a smaller variance is more **efficient**.

Example 1: For a *normally* distributed population, it can be shown that the sample median is an unbiased estimator for μ. It can also be shown, however, that the sample median has a greater variance than that of the sample mean, for the same sample size. Hence, $\bar{X}$ is a more efficient estimator than sample median.

Example 2: Consider the following estimator. First, a random portion of a sample is discarded from an original sample; then, the mean of the retained values in the sample is taken as an estimate for μ. This estimator is unbiased, but is not as efficient as using the *entire* sample. The intuitive reasoning is that we are not fully utilizing available information, and hence the resulting estimator has a greater variance.

Estimating μ When σ^2 is Known ...

Constructing point estimates using the sample mean $\bar{X}$ is the "best" (according to our criteria above) estimator for the population mean μ.

Suppose the variance of a population is "known." How does one construct an *interval* estimate for μ?

The key idea is that from the central limit theorem, we know that when n is sufficiently large, the standardized variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows the standard normal distribution. It is important to realize that this is true even though we do *not* know the value of μ. The value of σ, however, is assumed to be given (this assumption, which could be unrealistic, will be relaxed later).

It follows that for a given α, we have

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,.$$

Since our "unknown" is actually μ, the above can be rearranged into:

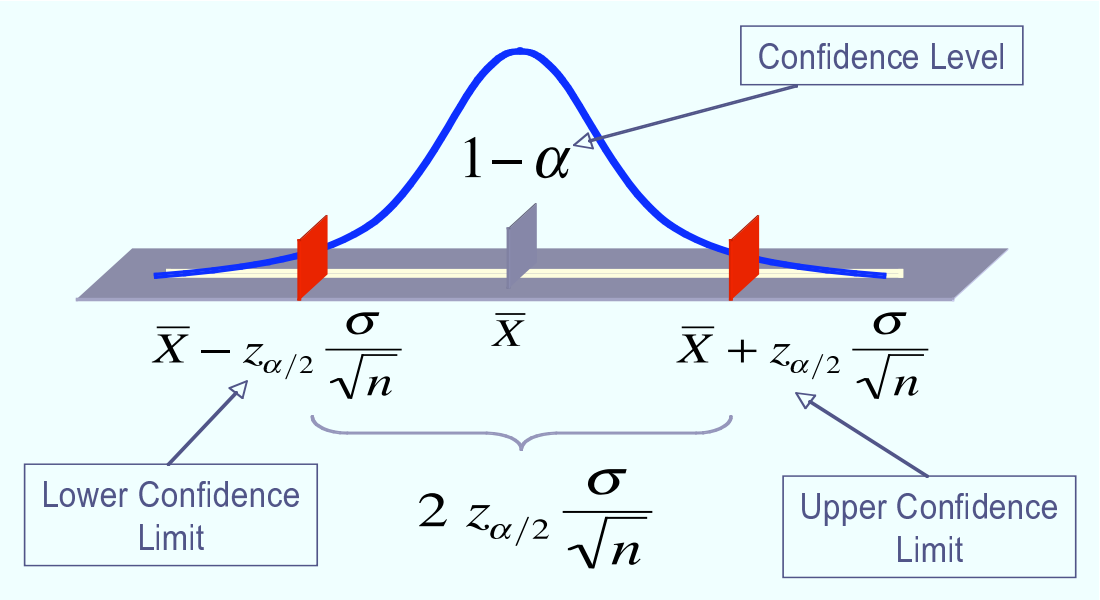
$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha\,.$$

That is, the *probability* for the interval

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{2}$$

to contain, or to cover, the unknown population mean μ is $1 - \alpha$; and we now have a so-called **confidence interval** for μ. Note that the interval estimator (2) is constructed from $\bar{X}$, $z_{\alpha/2}$, σ, and n, all of which are known. The user-specified value $1 - \alpha$ is called the **confidence level** or **coverage probability**.

Pictorially, we have



Interpretation:

If the interval estimator (2) is used *repeatedly* to estimate the mean μ of a given population, then $100(1-\alpha)\%$ of the constructed intervals will cover μ.

The often-heard media statement "19 times out of 20" refers to a confidence level of 0.95. Such a statement is good, since it emphasizes the fact that we are correct only 95% of the time.

Example: Demand during Lead Time

A computer company delivers computers directly to customers who order via the Internet. To reduce inventory cost, the company employs an inventory model. The model requires information about the mean demand during delivery lead time between a central manufacturing facility and local warehouses.

Past experience indicates that lead-time demand is normally distributed with a standard deviation of 75 computers per lead time (which is also random).

Construct the 95% confidence interval for the mean demand. Demand data for a sample of 25 lead-time periods are given in the file Xm10-01.xls.

Solution: Since $1-\alpha = 0.95$, we have $\alpha = 0.05$ and hence $\alpha/2 = 0.025$, for which $z_{0.025} = 1.96$. From the given data file, we obtain the sample mean $\bar{X} = 370.16$. The confidence interval is therefore (see (2))

$$\left(370.16 - 1.96\frac{75}{\sqrt{25}},\ 370.16 + 1.96\frac{75}{\sqrt{25}}\right)$$

or simply (340.76, 399.56).

Width of Confidence Interval . . .

Suppose we are told that with 95% confidence that the average starting salary of accountants is between \$15,000 and \$100,000. Clearly, this provides little information, despite the high "confidence" level.

Now, suppose instead: With 95% confidence that the average starting salary of accountants is between \$42,000 and \$45,000.

The second statement of course offers more precise information. Thus, for a given α, the width of a confidence interval conveys the extent of precision of the estimate. To reduce the width, or to increase precision, we can increase the sample size.

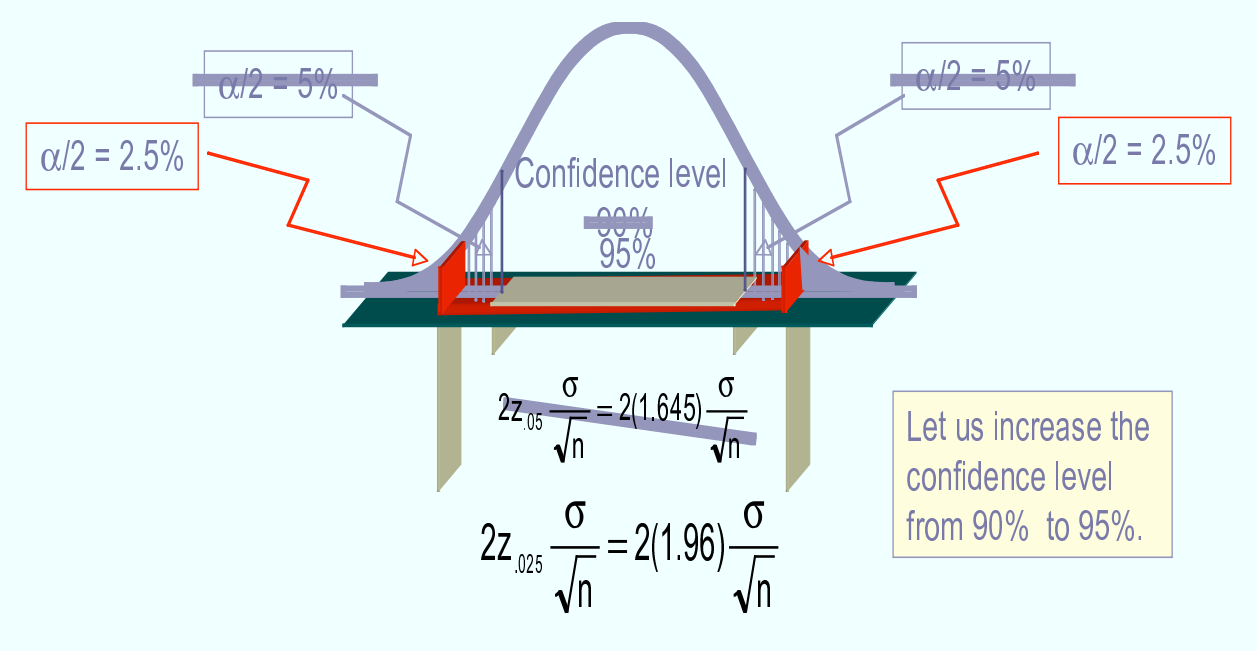
In general, recall that the upper and lower confidence limits are:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

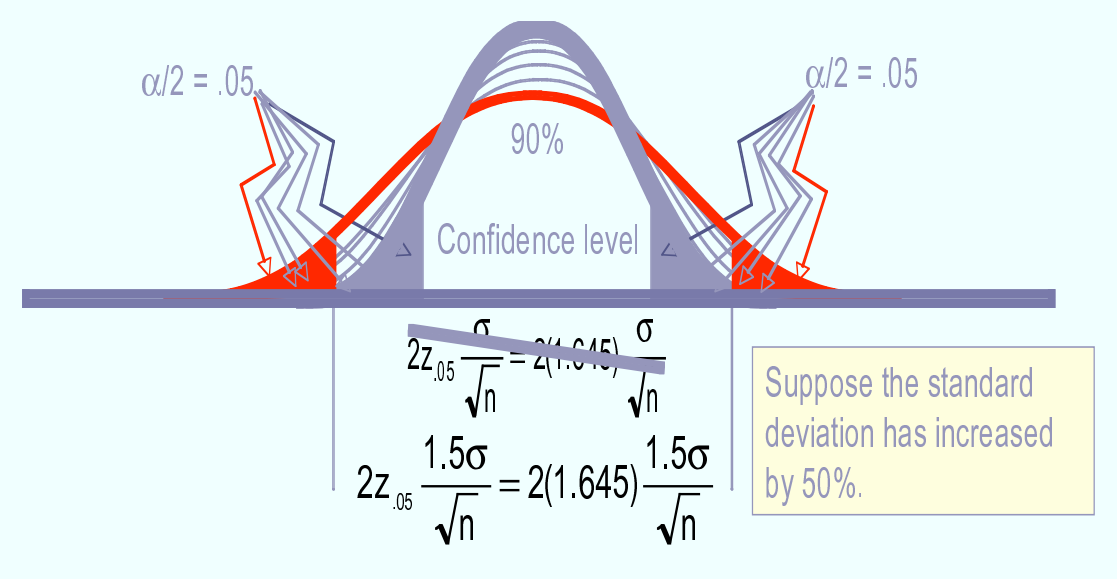
Hence, the width of the confidence interval is $2\, z_{\alpha/2}\, \sigma/\sqrt{n}$. It follows that precision depends on α, σ, and n.

Details ...

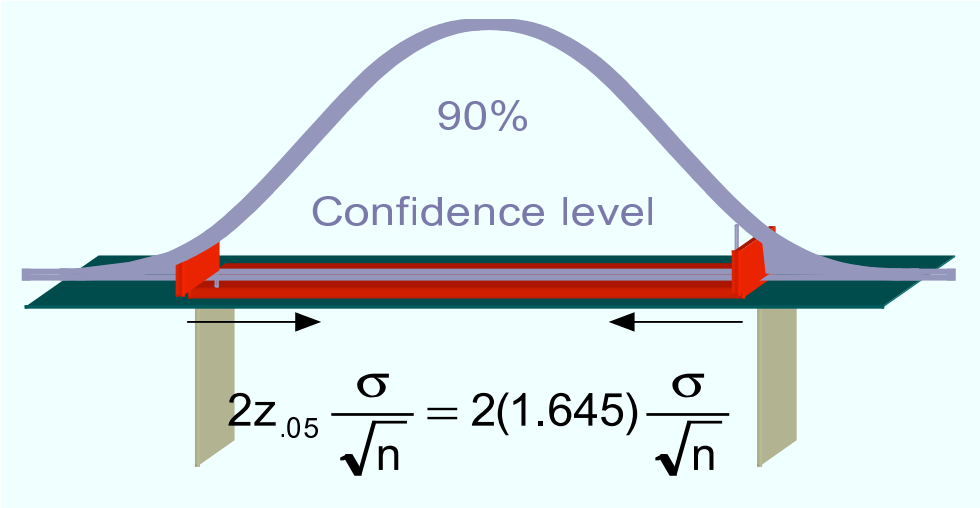
— A smaller α implies a wider interval:



— A larger σ implies a wider interval:



— A larger n implies a narrower interval:



Selecting the Sample Size . . .

To control the width of the confidence interval, we can choose a necessary sample size. Formally, suppose we wish to "estimate the mean to within w units." This means that we wish to construct an interval estimate of the form $\bar{X} \pm w$.

By solving the equation

$$w = z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

we obtain

$$n = \left(\frac{z_{\alpha/2}\, \sigma}{w} \right)^2,$$

the required sample size.

Example: Tree Diameters

A lumber company must estimate the mean diameter of trees in an area of forest to determine whether or not there is sufficient lumber to harvest. They need to estimate this to within 1 inch at a confidence level of 99%. Suppose the tree diameters are normally distributed with a standard deviation of 6 inches. What sample size is sufficient to guarantee this?

Solution: The required precision is ± 1 inch. That is, $w = 1$. For $\alpha = 0.01$, we have $z_{\alpha/2} = z_{0.005} = 2.575$. Therefore,

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{w}\right)^2 = \left(\frac{2.575 \cdot 6}{1}\right)^2 = 239\,.$$

Thus, we need to sample at least 239 trees to achieve a 99% confidence interval of $\bar{X} \pm 1$.