Chapter 4

Part I. Sampling Distributions and Confidence Intervals

Section 1.
Sampling Distribution

Using Statistics

- Statistical Inference:
  - ✓ Predict and forecast values of ***population parameters...***
  - ✓ Test hypotheses about values of population parameters...
  - ✓ Make decisions...

On basis of ***sample statistics*** derived from limited and incomplete sample information

Make generalizations about the characteristics of a ***population...***

On the basis of observations of a ***sample***, a part of a population

Sample Statistics as Estimators of Population Parameters

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.

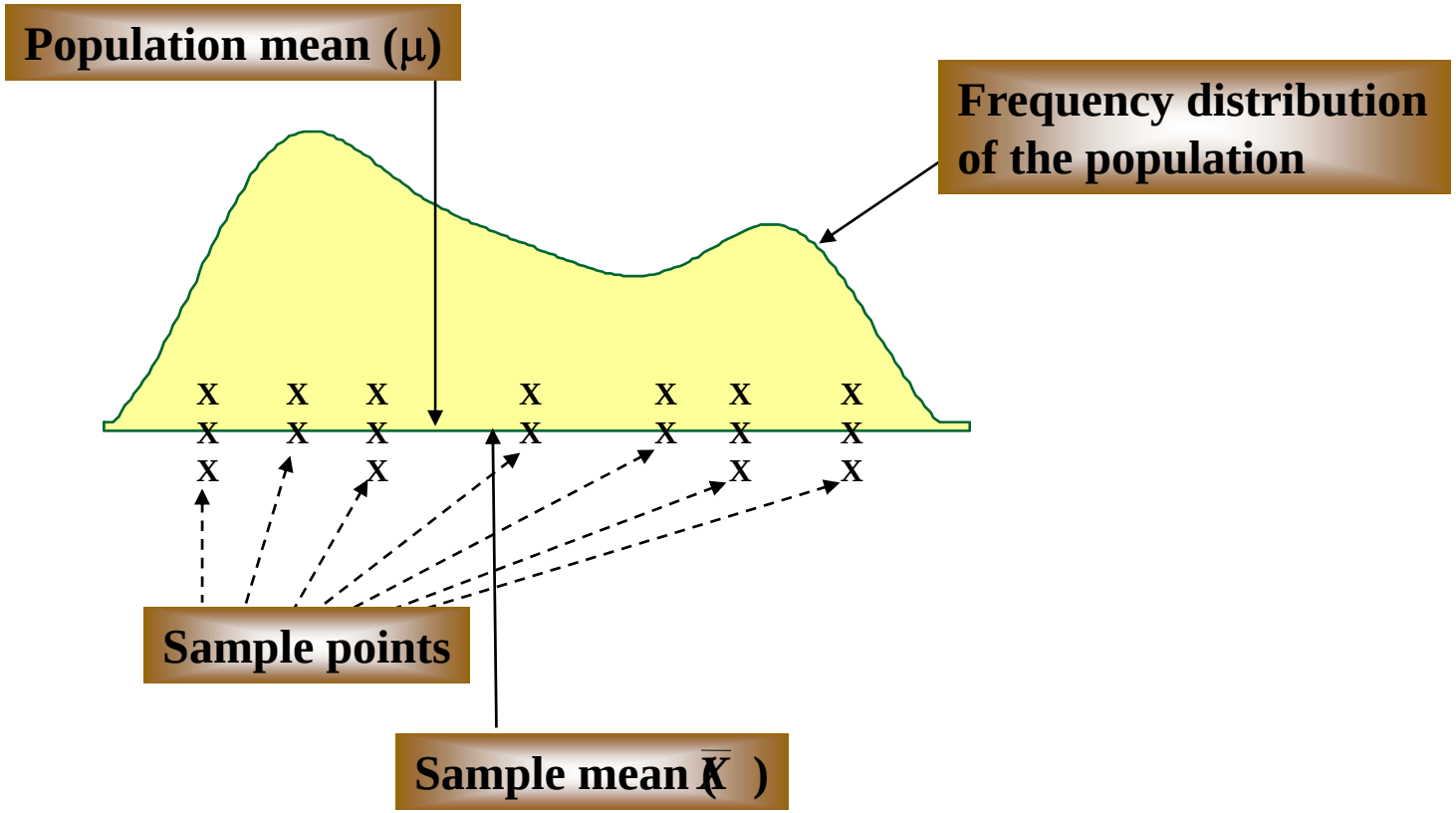
A **population parameter** is a numerical measure of a summary characteristic of a population.

- An **estimator** of a population parameter is a sample statistic used to estimate or predict the population parameter.
- An **estimate** of a parameter is a *particular* numerical value of a sample statistic obtained through sampling.
- A **point estimate** is a single value used as an estimate of a population parameter.

Estimators

- The sample mean, $\bar{X}$, is the most common estimator of the population mean, μ.
- The sample variance, s^2, is the most common estimator of the population variance, σ^2.
- The sample standard deviation, s, is the most common estimator of the population standard deviation, σ.
- The sample proportion, $\hat{p}$, is the most common estimator of the population proportion, p.

A Population Distribution, a Sample from a Population, and the Population and Sample Means



Other Sampling Methods

- **Stratified sampling:** in stratified sampling, the population is partitioned into two or more subpopulation called strata, and from each stratum a desired sample size is selected at random.

- **Cluster sampling:** in cluster sampling, a random sample of the strata is selected and then samples from these selected strata are obtained.

- **Systemic sampling:** in systemic sampling, we start at a random point in the sampling frame, and from this point selected every k^{th}, say, value in the frame to formulate the sample.

Sampling Distributions

- The **sampling distribution** of a statistic is the probability distribution of all possible values the statistic may assume, when computed from random samples of the same size, drawn from a specified population.

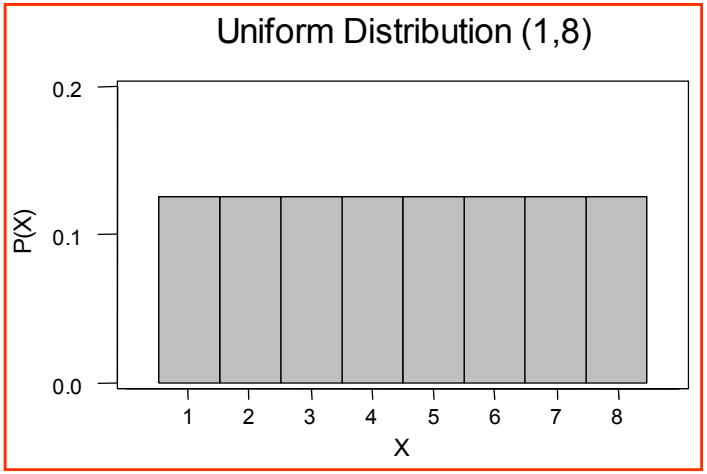
- The **sampling distribution of** $\overline{X}$ is the probability distribution of all possible values the random variable $\overline{X}$ may assume when a sample of size *n* is taken from a specified population.

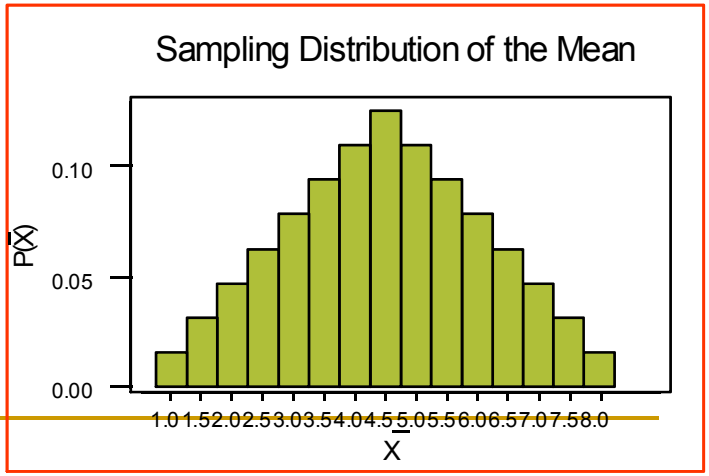
Properties of the Sampling Distribution of the Sample Mean

- Comparing the population distribution and the sampling distribution of the mean:
  - ✓ **The sampling distribution is more bell-shaped and symmetric.**
  - ✓ **Both have the same center.**
  - ✓ **The sampling distribution of the mean is more compact, with a smaller variance.**

$$E(\bar{X}) = \mu_{\bar{X}} = \mu_X$$

$$Var(\bar{X}) = \frac{\sigma_X^2}{n}$$



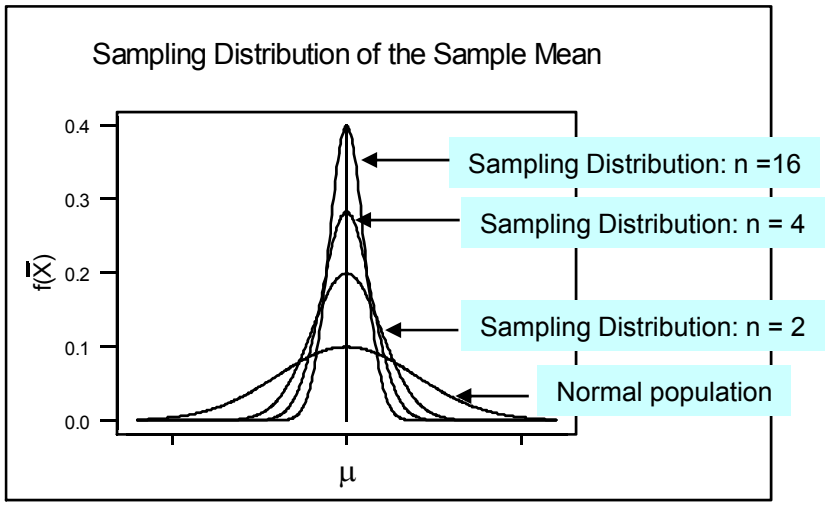


Sampling from a Normal Population

When sampling from a **normal population** with mean μ and standard deviation σ, the sample mean, X, has a **normal sampling distribution:**

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

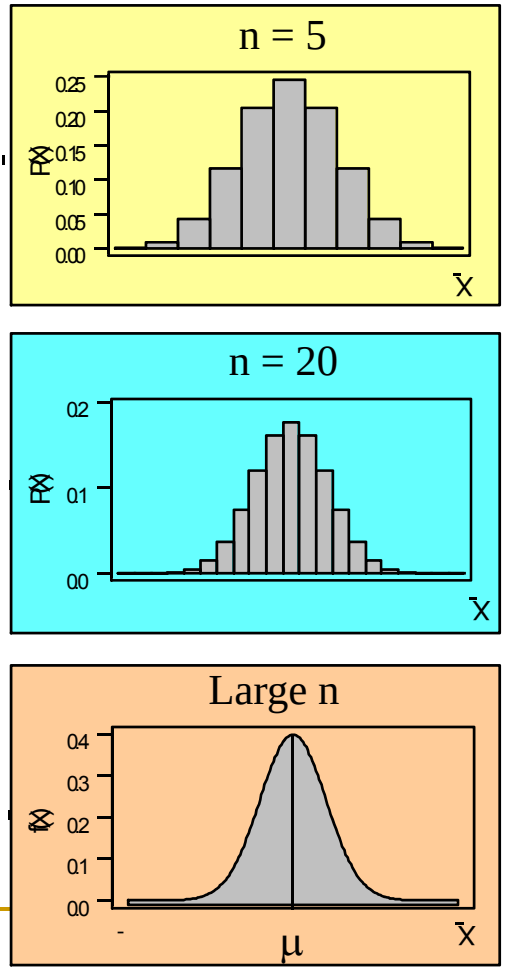
This means that, as the sample size increases, the sampling distribution of the sample mean remains centered on the population mean, but becomes more compactly distributed around that population mean



The Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean will tend to a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large (n >30).

For "large enough" n: $\bar{X} \sim N(\mu, \sigma^2 / n)$



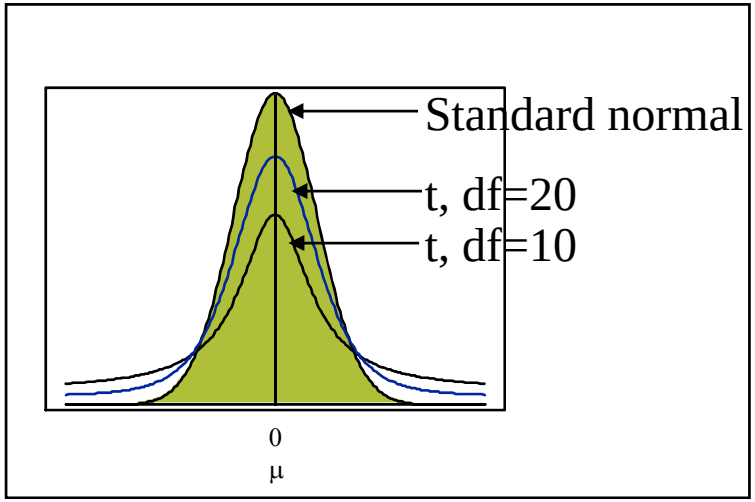
Student's *t* Distribution

If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, *s*. If the population is normal, the resulting statistic:

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

has a ***t distribution* with *(n - 1) degrees of freedom***.

- The *t* is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of *t* is 0.
- The *t* distribution approaches a standard normal as the number of degrees of freedom increases.

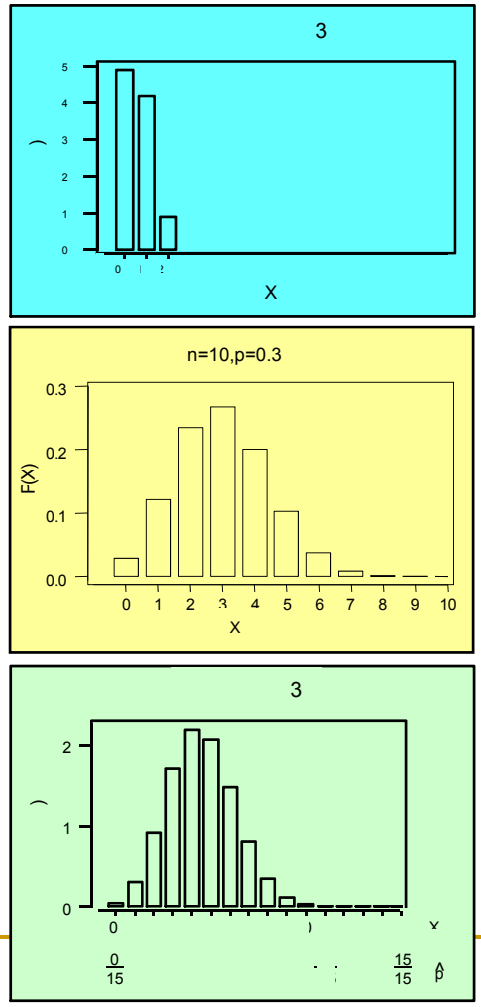


The Sampling Distribution of the Sample Proportion, $\hat{p}$

The **sample proportion** is the percentage of successes in n binomial trials. It is the number of successes, X, divided by the number of trials, n.

Sample proportion: $\hat{p} = \frac{X}{n}$

As the sample size, n, increases, the sampling distribution of $\hat{p}$ approaches a **normal distribution** with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$



Estimators and Their Properties

An **estimator** of a population parameter is a sample statistic used to estimate the parameter. The most commonly-used estimator of the:

Population Parameter		Sample Statistic
Mean (μ)	is the	Mean ($\overline{X}$)
Variance (σ^2)	is the	Variance (s^2)
Standard Deviation (σ)	is the	Standard Deviation (s)
Proportion (p)	is the	Proportion ($\hat{p}$)

- Desirable properties of estimators include:

  Unbiasedness **Efficiency**
  Consistency **Sufficiency**

Types of Estimators

- Point Estimate
  - ✓ A single-valued estimate.
  - ✓ A single element chosen from a sampling distribution.
  - ✓ Conveys little information about the actual value of the population parameter, about the accuracy of the estimate.

- Confidence Interval or Interval Estimate
  - ✓ An interval or range of values believed to include the unknown population parameter.
  - ✓ Associated with the interval is a measure of the ***confidence*** we have that the interval does indeed contain the parameter of interest.

Section 2. Confidence Intervals for Population Means (Z-CI and t-CI)

Confidence Interval for μ when σ is known

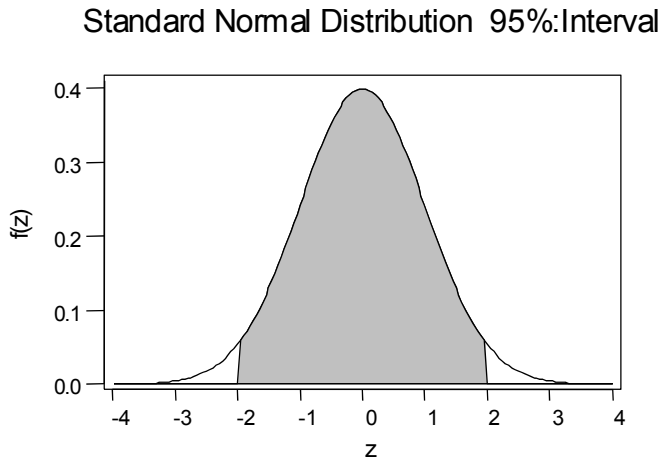
- If the population distribution is normal, *the sampling distribution of the mean is normal*.
- If the sample is sufficiently large (≥30), regardless of the shape of the population distribution, *the sampling distribution is normal* (Central Limit Theorem).

In either case :

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

or

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$



Confidence Interval for μ when σ is known

Before sampling, there is a 0.95probability that the interval

$$\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$$

will include the sample mean (and 5% that it will not).

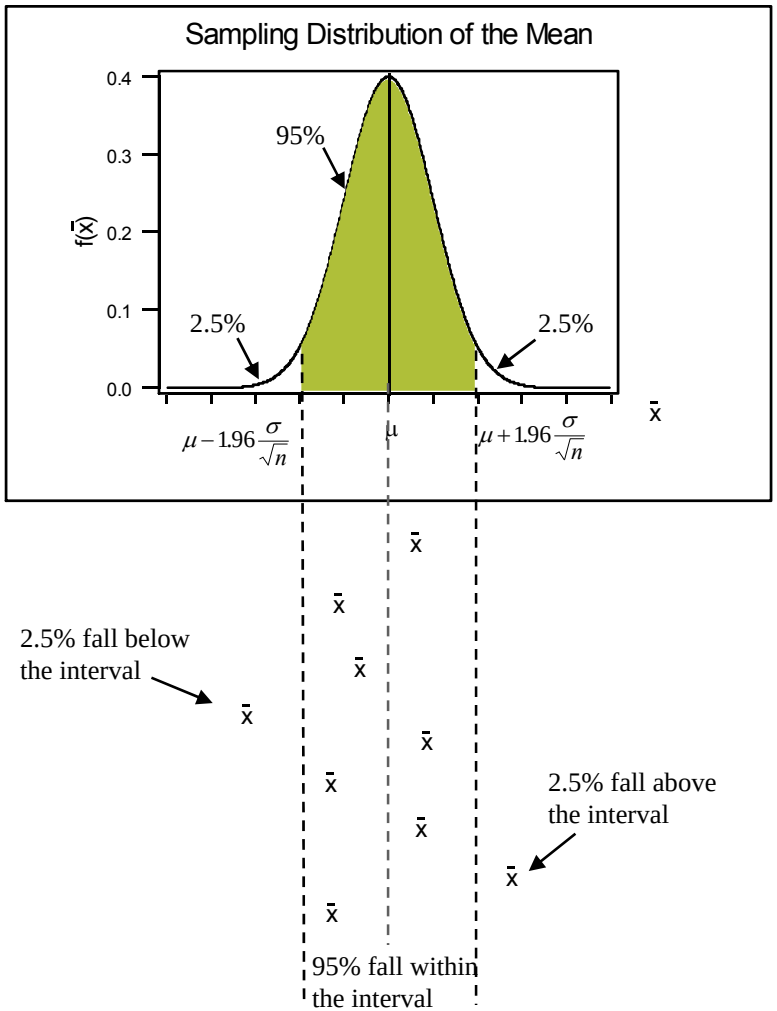
Conversely, after sampling, approximately 95% of such intervals

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

will include the population mean (and 5% of them will not).

That is, $\mathbf{\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}}$ **is a 95% confidence interval for** μ.

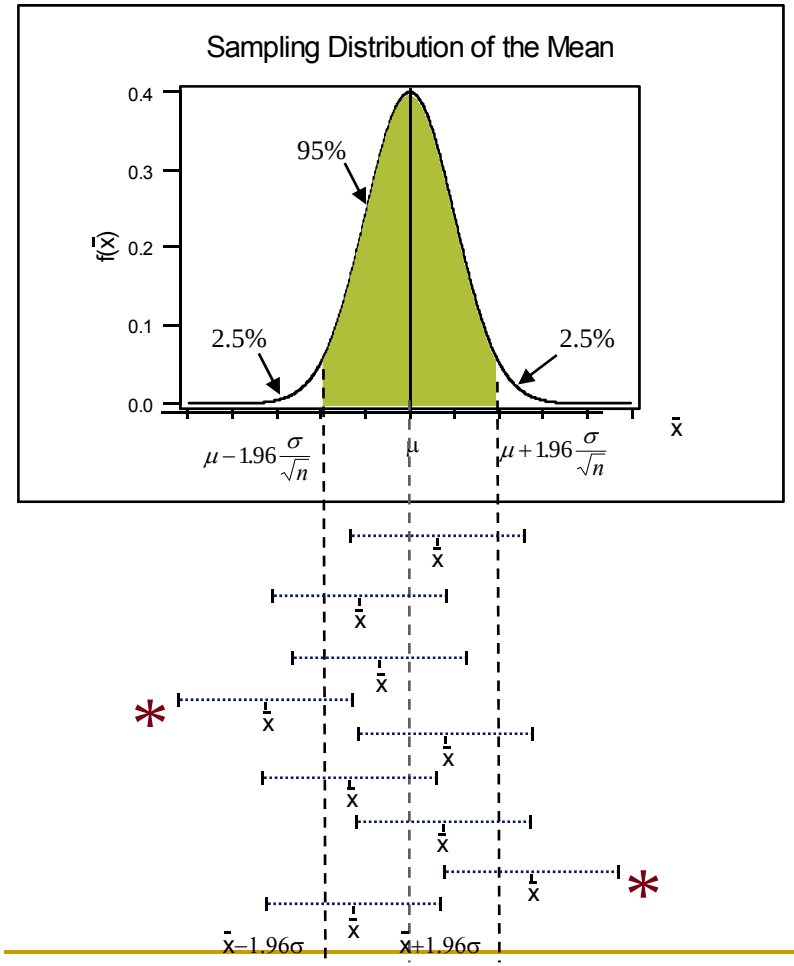
A 95% Interval around the Population Mean



Approximately 95% of sample means can be expected to fall within the interval $\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$.

So 5% can be expected to fall outside the interval $\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$.

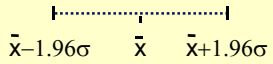
95% Intervals around the Sample Mean



Approximately 95% of the intervals

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

around the sample mean can be expected to include the actual value of the population mean, μ. (When the sample mean falls within the 95% interval around the population mean.)



The 95% Confidence Interval for μ

A 95% confidence interval for μ when σ is known and sampling is done from a normal population, or a large sample is used:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

The quantity $1.96 \frac{\sigma}{\sqrt{n}}$ is often called the ***margin of error*** or the ***sampling error***.

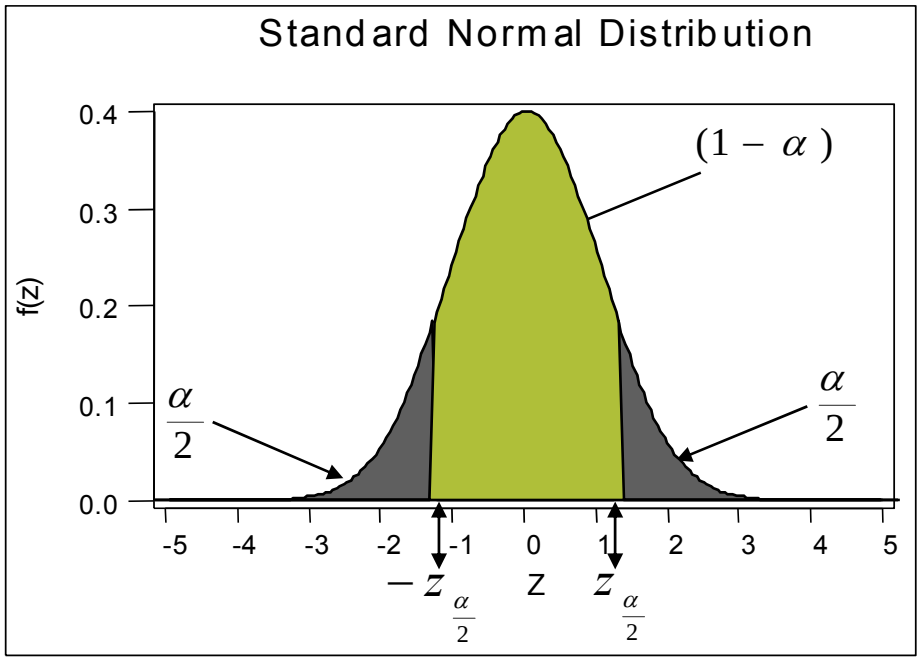
For example, if: n = 25
σ = 20
$\bar{x} = 122$

A 95% confidence interval:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 122 \pm 1.96 \frac{20}{\sqrt{25}}$$
$$= 122 \pm (1.96)(4)$$
$$= 122 \pm 7.84$$
$$= [114.16, 129.84]$$

A (1-α)100% Confidence Interval for μ

We define $z_{\frac{\alpha}{2}}$ as the z value that cuts off a right-tail area of $\frac{\alpha}{2}$ under the standard normal curve. (1-α) is called the ***confidence coefficient***. α is called the ***error probability***, and (1-α)100% is called the ***confidence level***.



$$P\left(z > z_{\frac{\alpha}{2}}\right) = \alpha/2$$

$$P\left(z < -z_{\frac{\alpha}{2}}\right) = \alpha/2$$

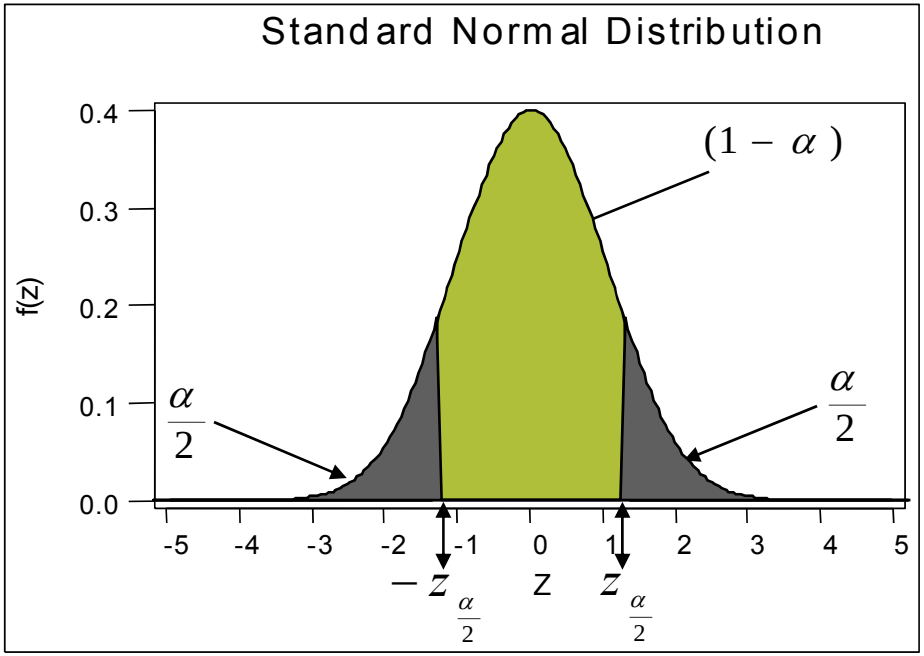
$$P\left(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}\right) = (1-\alpha)$$

(1- α)100% Confidence Interval:

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

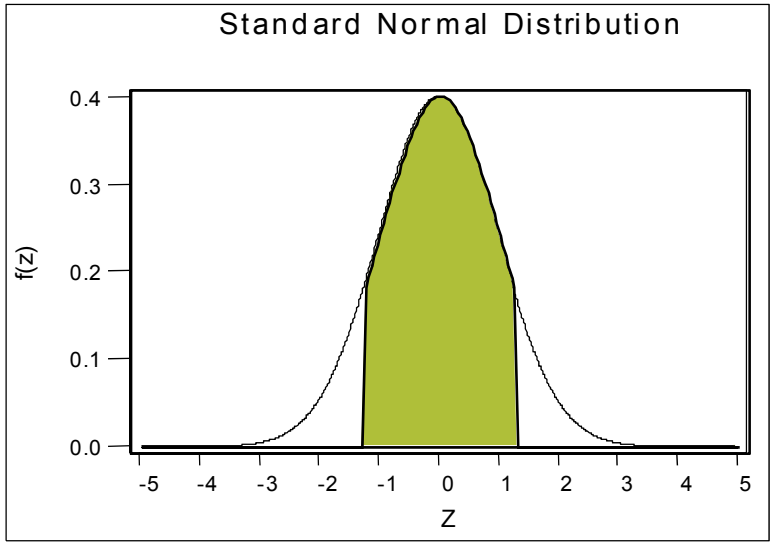
Critical Values of z and Levels of Confidence

$(1-\alpha)$	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
0.99	0.005	2.576
0.98	0.010	2.326
0.95	0.025	1.960
0.90	0.050	1.645
0.80	0.100	1.282



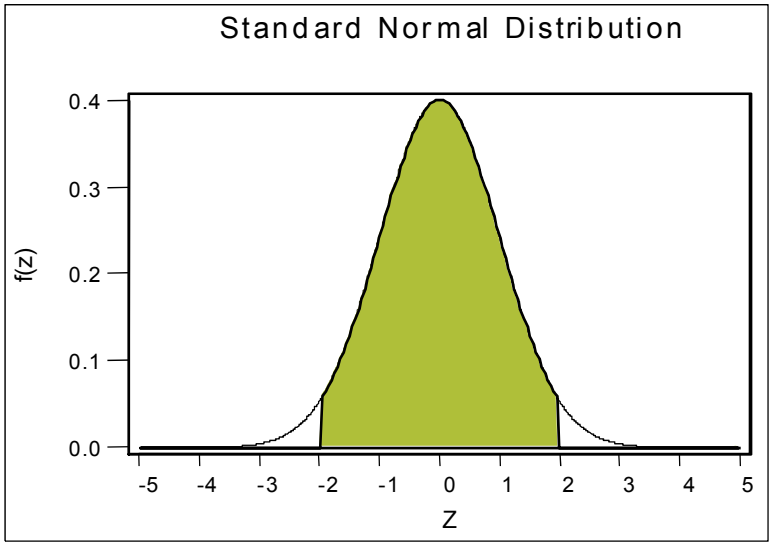
Level of confidence and width of the confidence interval

When sampling from the same population, using a fixed sample size, the ***higher the confidence level, the wider the confidence interval.***



80% Confidence Interval:

$$\bar{x} \pm 1.28 \frac{\sigma}{\sqrt{n}}$$

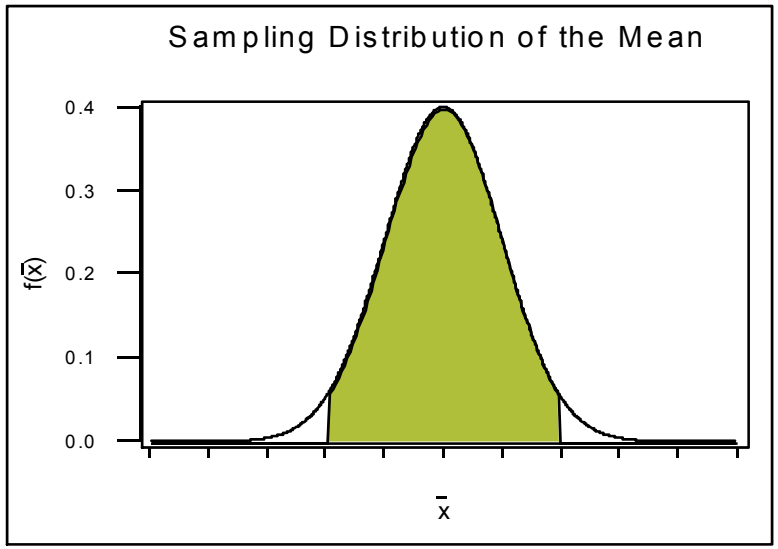


95% Confidence Interval:

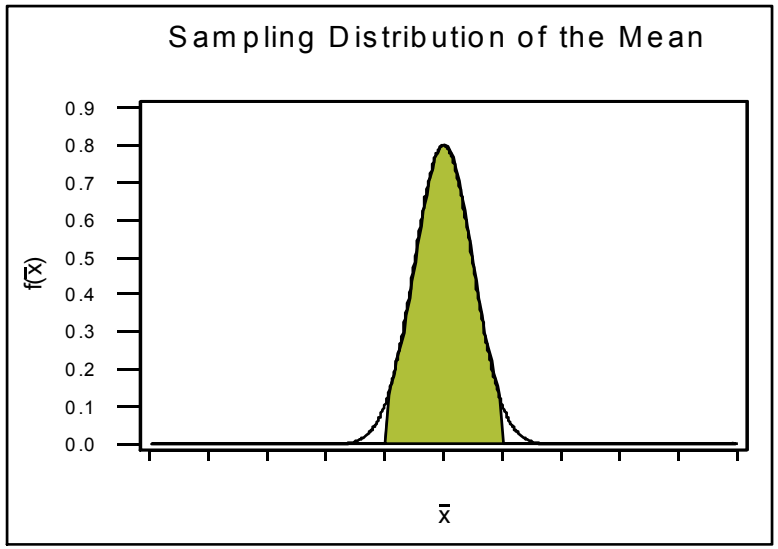
$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

The Sample Size and the Width of the Confidence Interval

When sampling from the same population, using a fixed confidence level, the ***larger the sample size, n, the narrower the confidence interval***.



95% Confidence Interval: **n = 20**



95% Confidence Interval: **n = 40**

Note: The width of a confidence interval can be reduced only at the price of: a **lower level of confidence**, or a **larger sample**.

Example 1

Population consists of the Fortune 500 Companies (Fortune Web Site), as ranked by Revenues. You are trying to to find out the average Revenues for the companies on the list.

The population standard deviation is $15,056.37. A random sample of 30 companies obtains a sample mean of $10,672.87. Give a 95% and 90% confidence interval for the average Revenues

Chi-square Distribution

The random sample $X_1, X_2, \ldots, X_n$, is from a normal distribution N(μ, σ^2), $\overline{X}$ is the sample mean and S^2 is the sample variance, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

where $\chi^2(n-1)$ is the chi - square distribution with degrees of freedom (n - 1).

Property of Chi - square distribution $W \sim \chi^2(r) = \text{Gamma}(r/2,\ 1/2)$:

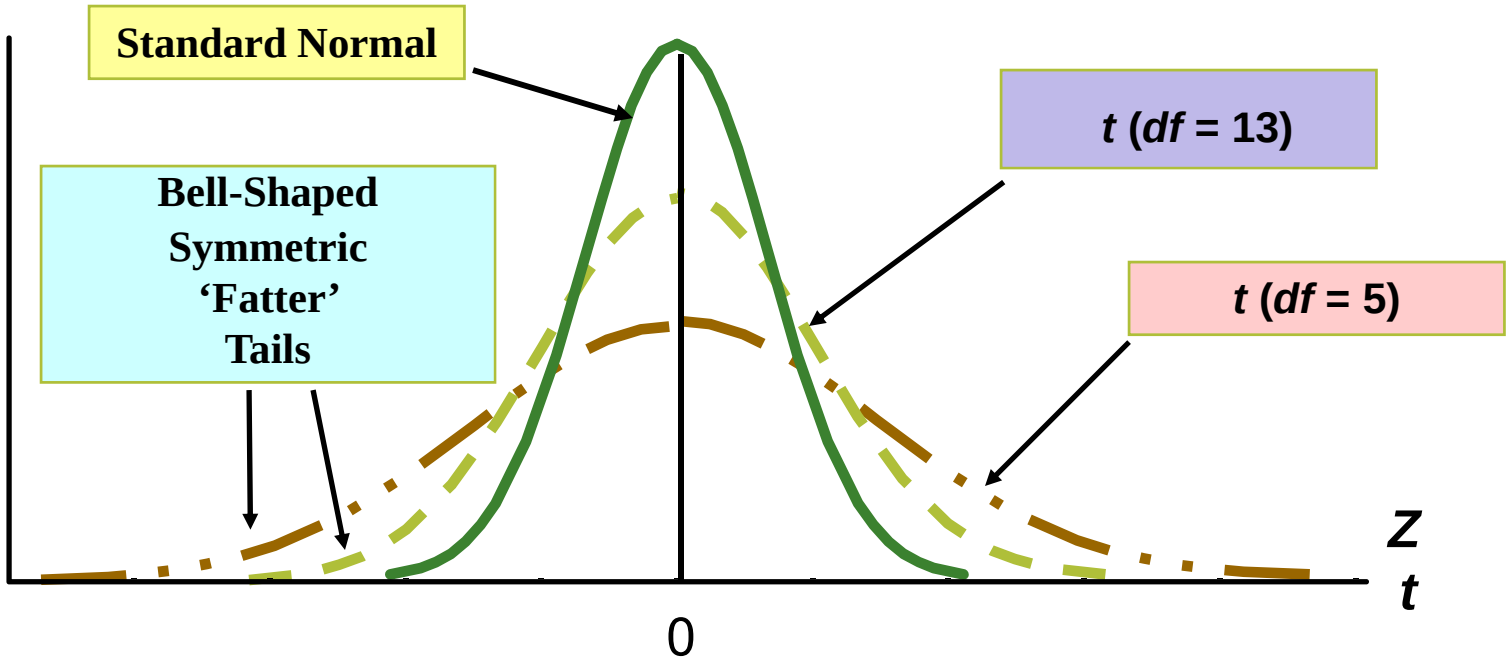
1. $E(W) = r,\ \ Var(W) = 2r$
2. If $Z \sim \text{N}(0,1)$, then $Z^2 \sim \chi^2(1)$.
3. Additive Property : Independent Chi - square r.v. $W_i \sim \chi^2(r_i)$

   then $W_1 + W_2 + \ldots + Wm \sim \chi^2(r_1 + \cdots + r_m)$

t distribution

- Assume Z~N(0, 1), W ~ $\chi^2(r)$, Z and W are independent, then $\frac{Z}{\sqrt{W/r}} \sim t(r)$

- The statistic $T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$ degrees of freedom=(n-1)

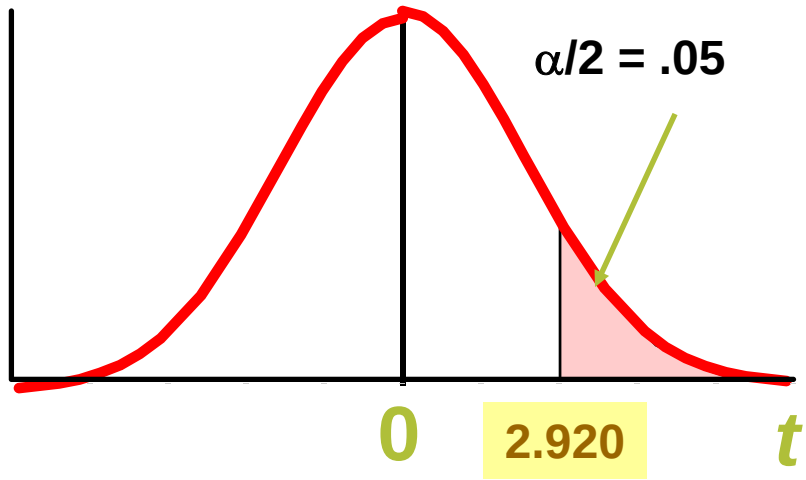


Student's *t* Table

	Upper Tail Area		
r	.25	.10	**.05**
1	1.000	3.078	6.314
2	0.817	1.886	**2.920**
3	0.765	1.638	2.353

t Values

Let: n = 3
df = *n* - 1 = 2
α = .10
α/2 =.05



Find t values:

1. α=0.10, n=20
2. α=0.01, n=8
3. α=0.025, n=10

Confidence intervals for μ when σ is unknown (t distribution)

A *(1-α)*100% confidence interval for μ when σ is not known (assuming a normally distributed population):

$$\bar{x} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2}}(n-1)$ is the value of the *t* distribution with *n-1* degrees of freedom that cuts off a tail area of $\frac{\alpha}{2}$ to its right.

Example 2:

A stock market analyst wants to estimate the average return on a certain stock. A random sample of 15 days yields an average (annualized) return of $\bar{x} = 10.37\%$ and a standard deviation of *s* = 3.5%. Assuming a normal population of returns, give a 95% confidence interval for the average return on this stock.

Section 3.
Confidence Interval for Proportions

Large-Sample Confidence Intervals for the Population Proportion, p

A large-sample $(1-\alpha)100\%$ confidence interval for the population proportion, p:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where the sample proportion, $\hat{p}$, is equal to the number of successes in the sample, x, divided by the number of trials (the sample size), n, and $\hat{q} = 1 - \hat{p}$.

For estimating p, a sample is considered large enough when $np > 5$ and $n(1-p) > 5$

Example 3

A marketing research firm wants to estimate the share that foreign companies have in the American market for certain products. A random sample of 100 consumers is obtained, and it is found that 34 people in the sample are users of foreign-made products; the rest are users of domestic products. Give a 95% confidence interval for the share of foreign products in this market.

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.34 \pm 1.96 \sqrt{\frac{(0.34)(0.66)}{100}}$$
$$= 0.34 \pm (1.96)(0.04737)$$
$$= 0.34 \pm 0.0928$$
$$= [0.2472, 0.4328]$$

Thus, the firm may be 95% confident that foreign manufacturers control anywhere from 24.72% to 43.28% of the market.

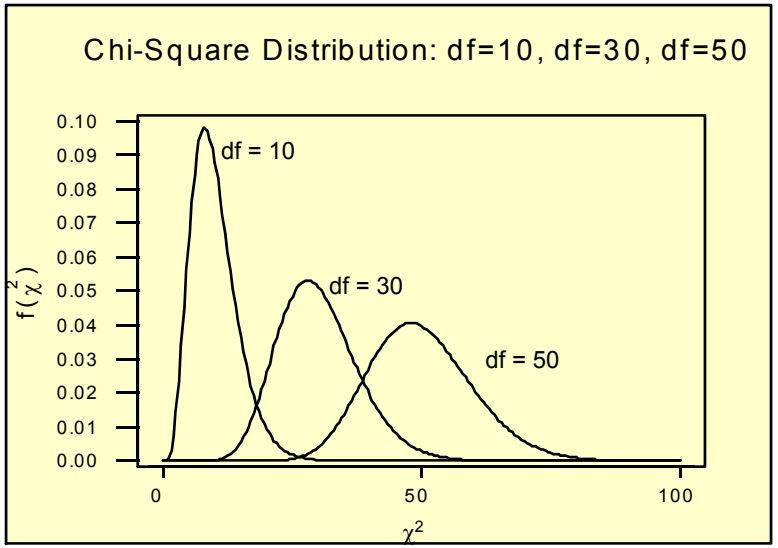
Confidence Intervals for the Population Variance: The Chi-Square (χ^2) Distribution

- The sample variance, s^2, is an unbiased estimator of the population variance, σ^2.

- Confidence intervals for the population variance are based on the chi-square ($\chi^2(r)$) distribution.
  - ✓ The **chi-square distribution** is the probability distribution of the sum of several independent, squared standard normal random variables.
  - ✓ The mean of the chi-square distribution is equal to the degrees of freedom parameter, **($E[\chi^2] = r$).** The variance of a chi-square is equal to twice the number of degrees of freedom, **($Var[\chi^2] = 2r$)**.

The Chi-Square (χ^2) Distribution

- The chi-square random variable cannot be negative.
- The chi-square distribution is skewed to the right.
- The chi-square distribution approaches a normal as the degrees of freedom increase.



In sampling from a normal population, the random variable:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with (n-1) degrees of freedom.

Confidence Interval for the Population Variance

A (1-α)100% confidence interval for the population variance * (where the population is assumed normal) is:

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]$$

where $\chi^2_{\frac{\alpha}{2}}$ is the value of the chi-square distribution with *n* - 1 degrees of freedom that cuts off an area $\frac{\alpha}{2}$ to its right and $\chi^2_{1-\frac{\alpha}{2}}$ is the value of the distribution that cuts off an area of $\frac{\alpha}{2}$ to its left (equivalently, an area of $1-\frac{\alpha}{2}$ to its right).

* Note: Because the chi-square distribution is skewed, the confidence interval for the population variance is not symmetric

Confidence Interval for the Population Variance – Example 4

In an automated process, a machine fills cans of coffee. If the average amount filled is different from what it should be, the machine may be adjusted to correct the mean. If the *variance* of the filling process is too high, however, the machine is out of control and needs to be repaired. Therefore, from time to time regular checks of the variance of the filling process are made. This is done by randomly sampling filled cans, measuring their amounts, and computing the sample variance. A random sample of 30 cans gives an estimate $s^2 = 18{,}540$. Give a 95% confidence interval for the population variance, σ^2.

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right] = \left[\frac{(30-1)18540}{45.7}, \frac{(30-1)18540}{16.0}\right] = [11765, 33604]$$

Sample-Size Determination

Before determining the necessary sample size, three questions must be answered:

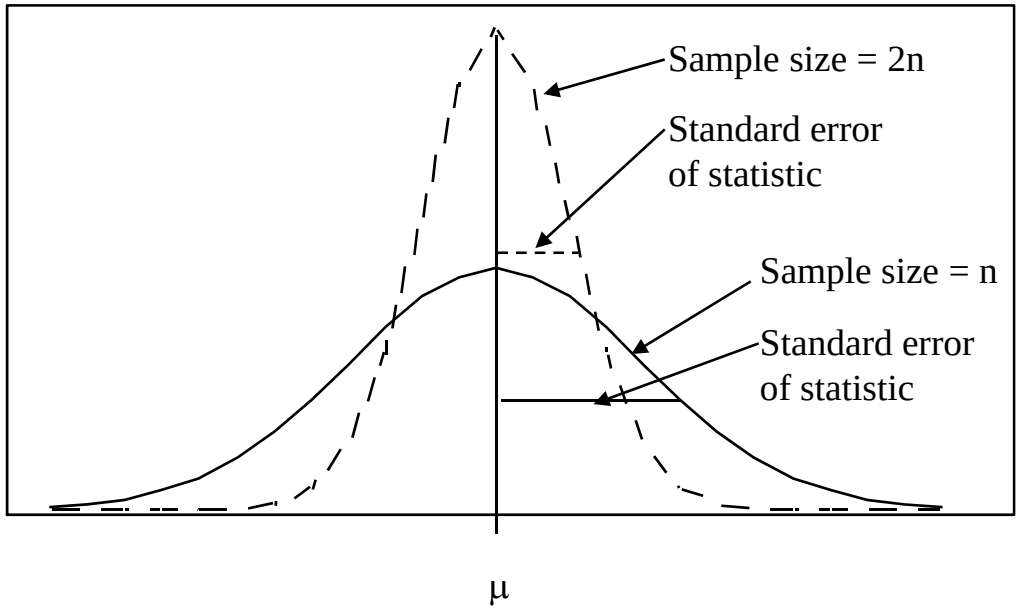
- How close do you want your sample estimate to be to the unknown parameter? (What is the desired ***bound*, B**?)
- What do you want the desired confidence level **(1-α)** to be so that the distance between your estimate and the parameter is less than or equal to B?
- What is your estimate of the variance (or standard deviation) of the population in question?

For example: A $(1-\alpha)$ Confidence Interval for μ: $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Bound, B

Sample Size and Standard Error

The sample size determines the bound of a statistic, since the standard error of a statistic shrinks as the sample size increases:



Minimum Sample Size: Mean and Proportion

Minimum required sample size in estimating the population mean, μ:

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{B^2}$$

Bound of estimate:

$$B = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Minimum required sample size in estimating the population proportion, $\hat{p}$

$$n = \frac{z_{\frac{\alpha}{2}}^2 pq}{B^2}$$

Sample-Size Determination: Example 5

A marketing research firm wants to conduct a survey to estimate the average amount spent on entertainment by each person visiting a popular resort. The people who plan the survey would like to determine the average amount spent by all people visiting the resort to within $120, with 95% confidence. From past operation of the resort, an estimate of the population standard deviation is s = $400. What is the minimum required sample size?

$$n = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{B^2}$$

$$= \frac{(1.96)^2 (400)^2}{120^2}$$

$$= 42.684 \approx 43$$

Sample-Size for Proportion: Example 6

The manufacturers of a sports car want to estimate the proportion of people in a given income bracket who are interested in the model. The company wants to know the population proportion, p, to within 0.01 with 99% confidence. Current company records indicate that the proportion p may be around 0.25. What is the minimum required sample size for this survey?

$$n = \frac{z_{\frac{\alpha}{2}}^2 pq}{B^2}$$

$$= \frac{2.576^2 (0.25)(0.75)}{0.10^2}$$

$$= 124.42 \approx 125$$