Chapter 9:

One- and Two-Sample Estimation Problems

Introduction

• Statistical inference may be divided into two major areas: **estimation** and **tests of hypotheses**.

• Suppose we have a population with some unknown parameter(s).

Example: Normal(μ,σ)

μ and σ are parameters.

• We need to draw conclusions (make inferences) about the unknown parameters.

• We select samples, compute some statistics, and make inferences about the unknown parameters based on the sampling distributions of the statistics.

Statistical Inference

(1) Estimation of the parameters (Chapter 9)

→ Point Estimation

→ Interval Estimation (Confidence Interval)

(2) Tests of hypotheses about the parameters (Chapter 10)

Classical Methods of Estimation

Point Estimation:

A point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\bar{x}$ of the statistic $\bar{X}$ computed from a sample of size n is a point estimate of the population mean μ.

Note:

An estimator is not expected to estimate the population parameter without error. We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off.

Interval Estimation (Confidence Interval = C.I.)

An interval estimate of some population parameter θ is an interval of the form $(\hat{\theta}_L, \hat{\theta}_U)$, i.e, $\hat{\theta}_L < \theta < \hat{\theta}_U$. This interval contains the true value of θ "with probability $1-\alpha$", that is $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha$.

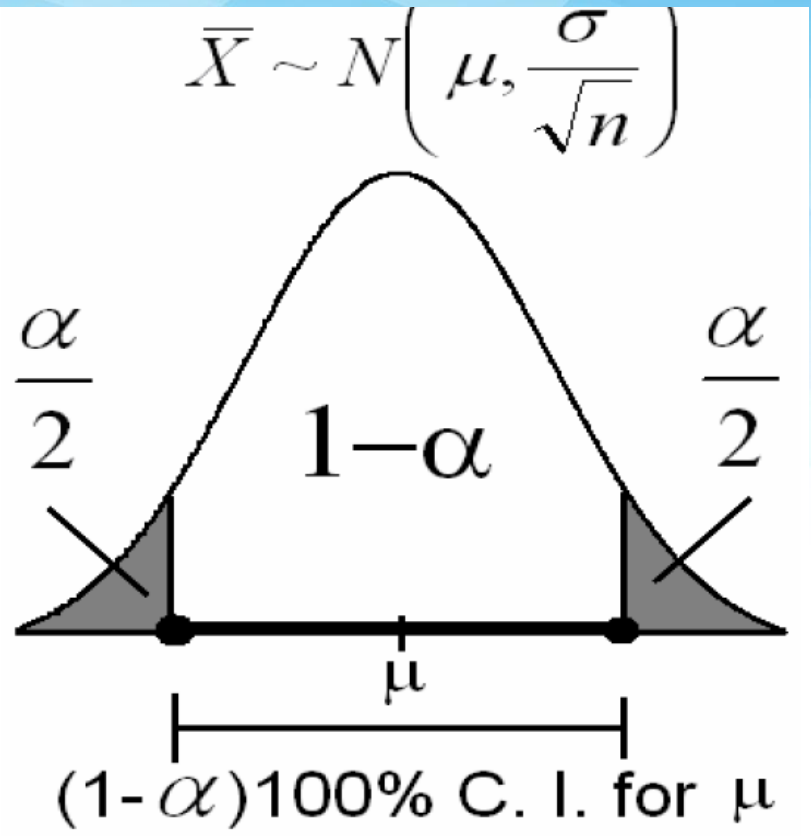
- $(\hat{\theta}_L, \hat{\theta}_U) = \hat{\theta}_L < \theta < \hat{\theta}_U$ is called a $(1-\alpha)100\%$ confidence interval (C.I.) for θ.
- $1-\alpha$ is called the confidence coefficient
- $\hat{\theta}_L$ = lower confidence limit
- $\hat{\theta}_U$ = upper confidence limit
- $\alpha = 0.1, 0.05, 0.025, 0.01$ $(0 < \alpha < 1)$

Single Sample: Estimation of the Mean (μ):

- $E(\bar{X}) = \mu_{\bar{X}} = \mu$
- $Var(\bar{X}) = \sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n}$
- $\bar{X} \sim N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$
- $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

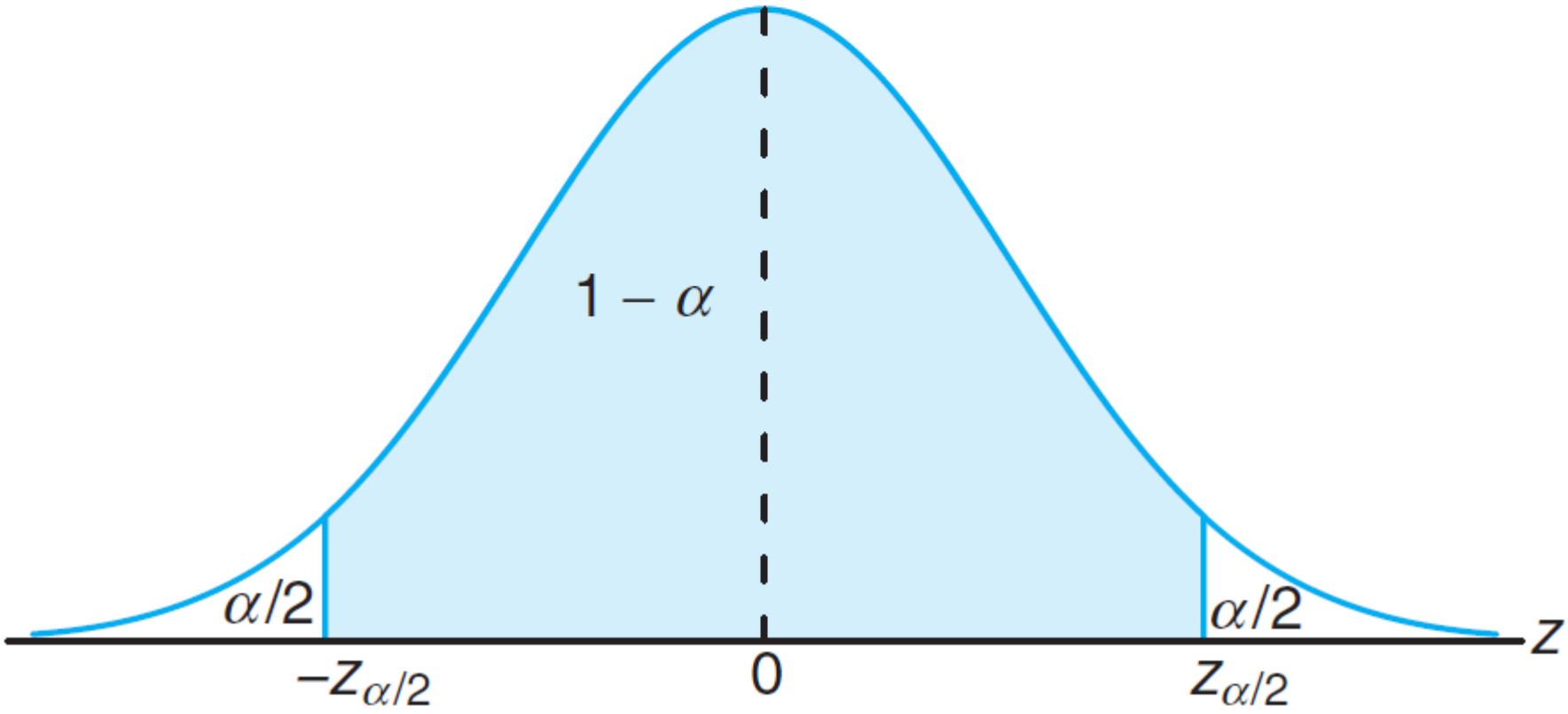
  (σ^2 is known)
- $T = \dfrac{\bar{X} - \mu}{S / \sqrt{n}} \sim t(n-1)$

  (σ^2 is unknown)



- We use the sampling distribution of $\bar{X}$ to make inferences about μ.

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



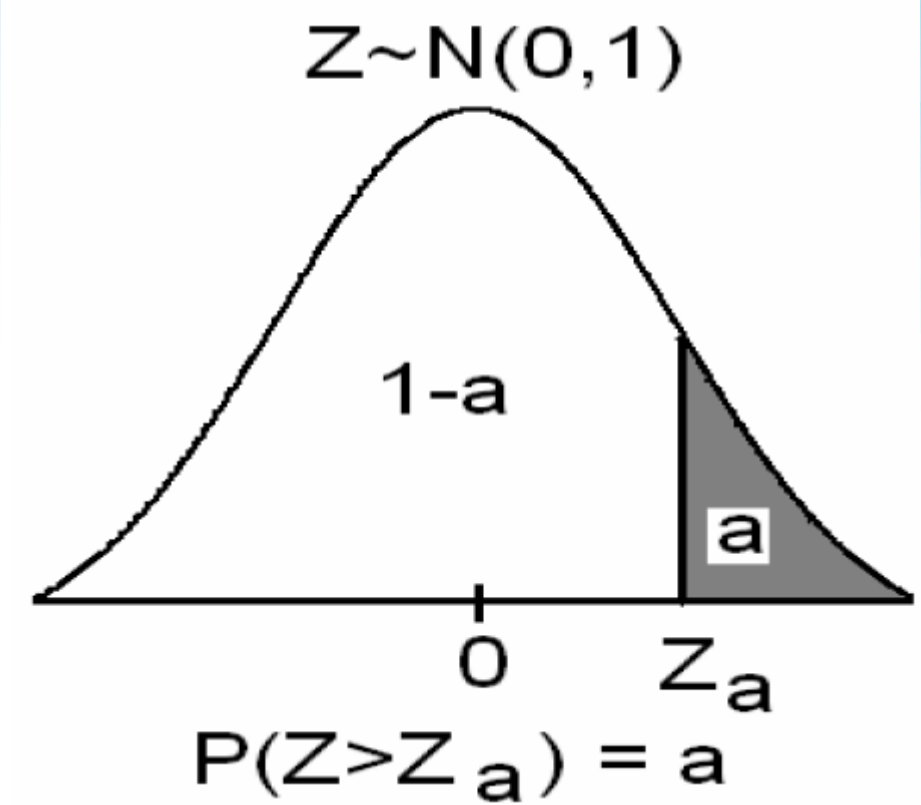
Notation:

Z_a is the Z-value leaving an area of a to the right;

i.e., $P(Z > Z_a) = a$

or equivalently,

$P(Z < Z_a) = 1 - a$



Point Estimation of the Mean (μ)

The sample mean $\bar{X} = \frac{\sum X_i}{n}$ is a "good" point estimate for μ.

Interval Estimation (Confidence Interval) of the Mean (μ)

(i) First Case: σ^2 is known:

Result:

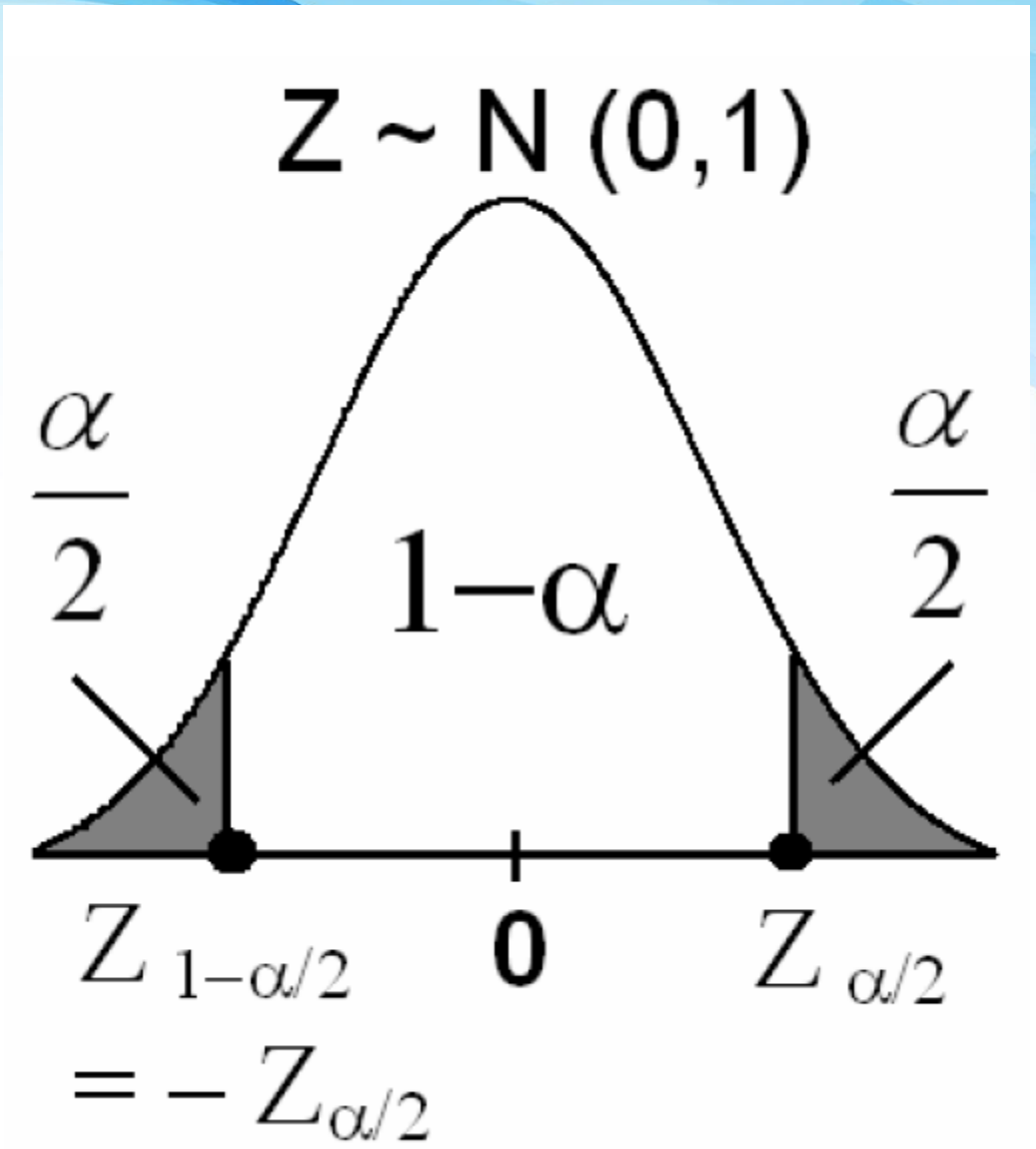
If $\overline{X} = \sum_{i=1}^{n} X_i / n$ is the sample mean of a random sample of size n from a population (distribution) with mean μ and known variance σ^2, then a $(1-\alpha)100\%$ confidence interval for μ is :

$$(\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

$$\Leftrightarrow \overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where $Z_{\frac{\alpha}{2}}$ is the Z-value leaving an area of $\alpha/2$ to the right; i.e., $P(Z > Z_{\frac{\alpha}{2}}) = \alpha/2$, or equivalently, $P(Z < Z_{\frac{\alpha}{2}}) = 1 - \alpha/2$.


Z ~ N (0,1)
α/2
1−α
α/2
Z 1−α/2
0
Z α/2
= − Zα/2

Note:

We are $(1-\alpha)100\%$ confident that

$$\mu \in (\overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} , \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

Example

The average zinc concentration recorded from a sample of zinc measurements in 36 different locations is found to be 2.6 gram/milliliter. Find a 95% and 99% confidence interval (C.I.) for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3.

Solution:

μ=the mean zinc concentration in the river. (unknown parameter)

Population	Sample
μ=??	n=36
σ=0.3	$\bar{X}$ =2.6

First, a point estimate for μ is $\bar{X} = 2.6$

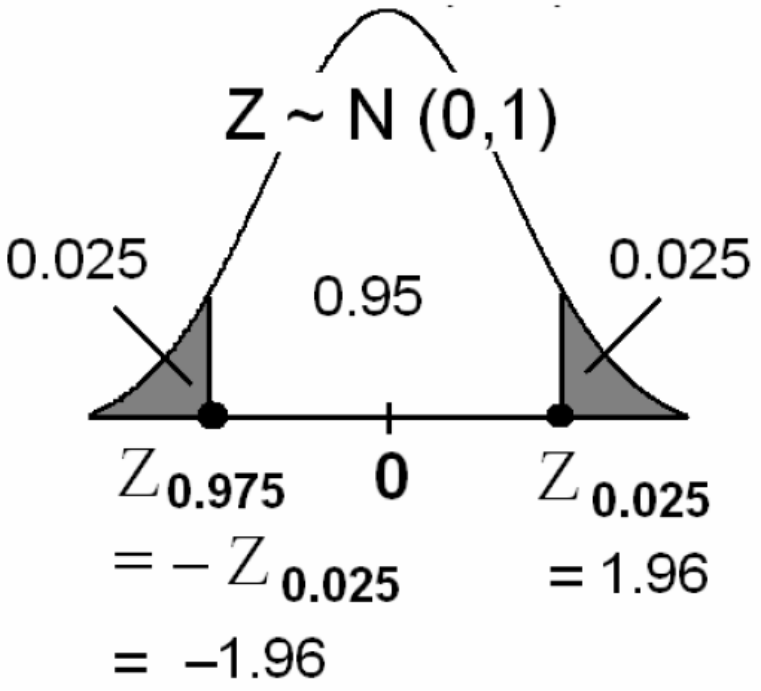
(a) We want to find 95% C.I. for μ.

$\alpha = ??$

$95\% = (1-\alpha)100\%$

$\Leftrightarrow 0.95 = (1-\alpha)$

$\Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025$



$$Z_{\frac{\alpha}{2}} = Z_{0.025}$$

$$= 1.96$$

A 95% C.I. for μ is

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow 2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$\Leftrightarrow 2.6 - 0.098 < \mu < 2.6 + 0.098$$

$$\Leftrightarrow 2.502 < \mu < 2.698$$

$$\Leftrightarrow \mu \in (2.502, 2.698)$$

We are 95% confident that $\mu \in (2.502, 2.698)$.

(b) Similarly, we can find that A 99% C.I. for μ is

$2.471 < \mu < 2.729$

$\Leftrightarrow \mu \in (2.471, 2.729)$

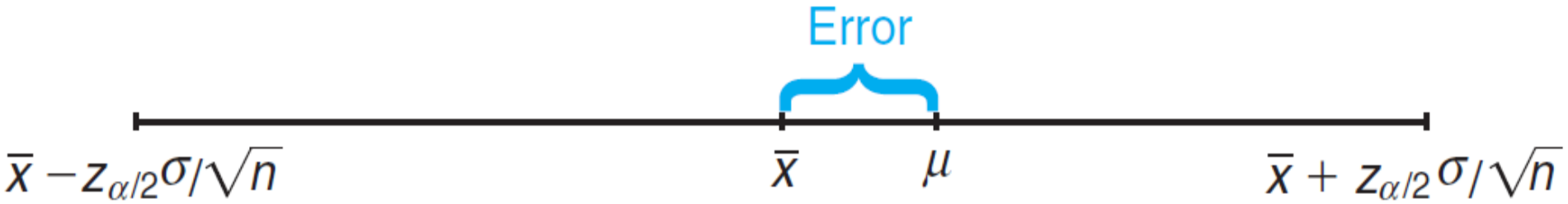
We are 99% confident that $\mu \in (2.471, 2.729)$

Notice that a 99% C.I. is wider that a 95% C.I.

See Ex 9.2 on page 271

Theorem

If $\bar{X}$ is used as an estimate of μ, we can then be $(1-\alpha)100\%$ confident that the error (in estimation) will not exceed $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.



Note:

max error of estimation = $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ with

$(1-\alpha)100\%$ confidence.

Example:

In the previous example, we are 95% confident that the sample mean $\bar{X} = 2.6$ differs from the true mean μ by an amount less than

$$Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = (1.96)\left(\frac{0.3}{\sqrt{36}}\right) = 0.098$$

Note:

Let *e* be the maximum amount of the error, that is

$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

then:

$$e = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sqrt{n} = Z_{\frac{\alpha}{2}} \frac{\sigma}{e} \Leftrightarrow n = \left(Z_{\frac{\alpha}{2}} \frac{\sigma}{e} \right)^2$$

Theorem

If $\bar{X}$ is used as an estimate of μ, we can then be $(1-\alpha)100\%$ confident that the error (in estimation) will not exceed a specified amount e when the sample size is

$$n = \left(Z_{\frac{\alpha}{2}} \frac{\sigma}{e} \right)^2$$

Note:

When solving for the sample size, n, **we round all fractional values up to the next whole number.** By adhering to this principle, we can be sure that our degree of confidence never falls below $100(1-\alpha)\%$.

Example

How large a sample is required in the previous example if we want to be 95% confident that our estimate of μ is off by less than 0.05?

Solution:

We have $\sigma = 0.3$, $Z_{\frac{\alpha}{2}} = 1.96$, $e = 0.05$. Then by Theorem 9.2,

$$n = \left(Z_{\frac{\alpha}{2}} \frac{\sigma}{e} \right)^2 = \left(1.96 \times \frac{0.3}{0.05} \right)^2 = 138.3 \approx 139$$

Therefore, we can be 95% confident that a random sample of size $n=139$ will provide an estimate $\overline{X}$ differing from μ by an amount less than $e=0.05$.

Interval Estimation (Confidence Interval) of the Mean (μ)

(ii) Second Case: σ^2 is unknown:

Recall:

- $T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

Result:

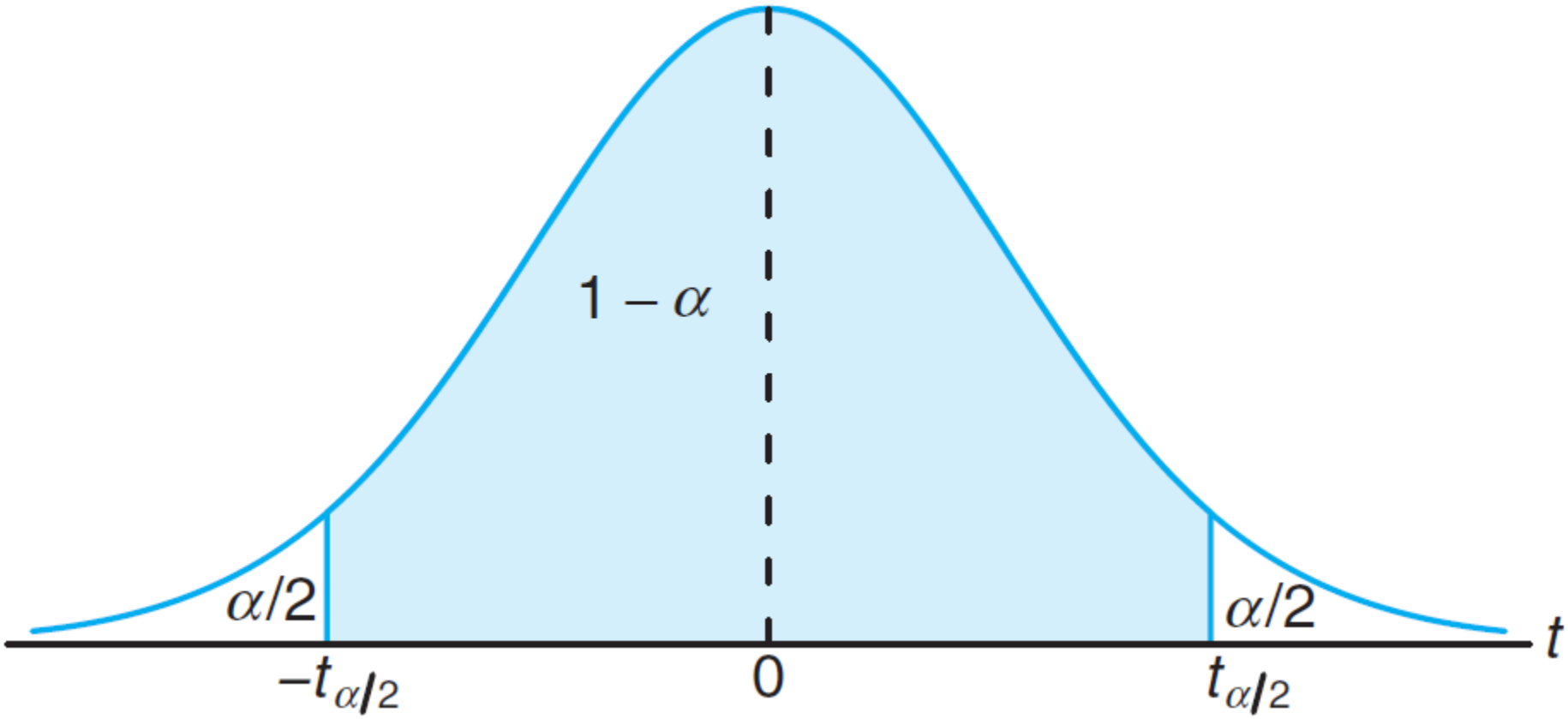
If $\overline{X} = \sum_{i=1}^{n} X_i / n$ and $S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)}$ are the sample mean and the sample standard deviation of a random sample of size n from a normal population (distribution) with unknown variance σ^2, then a $(1-\alpha)100\%$ confidence interval for μ is :

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

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



$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha.$$

Confidence Interval on μ, σ^2 Unknown

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ is the t-value with $v = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

Example

The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% C.I. for the mean of all such containers, assuming an approximate normal distribution.

Solution:

$n=7$ $\qquad \overline{X} = \sum_{i=1}^{n} X_i / n = 10.0$

$$S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)} = 0.283$$

First, a point estimate for μ is $\overline{X} = \sum_{i=1}^{n} X_i / n = 10.0$

Now, we need to find a confidence interval for μ.

α = ??

95%=(1−α)100% ⇔ 0. 95=(1−α) ⇔ α=0.05 ⇔

$\alpha/2$=0.025

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.447$ (with $\nu = n - 1 = 6$ degrees of freedom)

	0.025
	↓
6	→ $t_{0.025} = 2.447$

A 95% C.I. for μ is

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$\Leftrightarrow \overline{X} - t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$\Leftrightarrow 10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right)$$

$$\Leftrightarrow 10.0 - 0.262 < \mu < 10.0 + 0.262$$

$$\Leftrightarrow 9.74 < \mu < 10.26$$

$$\Leftrightarrow \mu \in (9.74, 10.26)$$

We are 95% confident that $\mu \in (9.74, 10.26)$.

Standard Error of a Point Estimate

Notes:

- Note: a $(1-\alpha)100\%$ C.I. for μ, when σ^2 is known, is

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = \overline{X} \pm Z_{\frac{\alpha}{2}} s.e(\overline{X}).$$

- Note: a $(1-\alpha)100\%$ C.I. for μ, when σ^2 is unknown and the distribution is normal, is

$$\overline{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} = \overline{X} \pm t_{\frac{\alpha}{2}} \hat{s}.e(\overline{X}). \quad (\nu = n-1 \text{ df})$$

Two Samples: Estimating the Difference between Two Means $(\mu_1 - \mu_2)$

Recall: For two independent samples:

- $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = E(\bar{X}_1 - \bar{X}_2)$
- $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = Var(\bar{X}_1 - \bar{X}_2)$
- $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Point Estimation of $(\mu_1 - \mu_2)$

$\bar{X}_1 - \bar{X}_2$ is a "good" point estimate for $\mu_1 - \mu_2$.

Confidence Interval of $(\mu_1 - \mu_2)$

(i) First Case: σ_1^2 and σ_2^2 are known:

- $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Result:

a (1−α)100% confidence interval for $(\mu_1 - \mu_2)$is :

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or $$(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

or $$\left((\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X}_1 - \bar{X}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

(ii) Second Case: $\sigma_1^2=\sigma_2^2=\sigma^2$ is unknown:

- If σ_1^2 and σ_2^2 are unknown but $\sigma_1^2=\sigma_2^2=\sigma^2$, then the pooled estimate of σ^2 is

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

where S_1^2 is the variance of the 1-st sample and S_2^2 is the variance of the 2-nd sample. The degrees of freedom of S_p^2 is $\nu=n_1+n_2-2$.

Result:

a $(1-\alpha)100\%$ confidence interval for $\mu_1-\mu_2$ is :

$$(\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}<\mu_1-\mu_2<(\bar{X}_1-\bar{X}_2)+t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}$$

or

$$(\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2<(\bar{X}_1-\bar{X}_2)+t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

or $(\bar{X}_1-\bar{X}_2)\pm t_{\frac{\alpha}{2}}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

or $\left((\bar{X}_1-\bar{X}_2)-t_{\frac{\alpha}{2}}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},(\bar{X}_1-\bar{X}_2)+t_{\frac{\alpha}{2}}\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$

where $t_{\frac{\alpha}{2}}$ is the t-value with $\nu=n_1+n_2-2$ degrees of freedom.

Example (First Case)

An experiment was conducted in which two types of engines, *A* and *B*, were compared. Gas mileage in miles per gallon was measured. 50 experiments were conducted using engine type *A* and 75 experiments were done for engine type *B*. The gasoline used and other conditions were held constant. The average gas mileage for engine *A* was 36 miles per gallon and the average for engine *B* was 42 miles per gallon. Find 96% confidence interval for $\mu_B - \mu_A$, where μ_A and μ_B are population mean gas mileage for engines *A* and *B*, respectively. Assume that the population standard deviations are 6 and 8 for engines *A* and *B*, respectively.

Solution:

Engine A	Engine B
$n_A=50$	$n_B=75$
$\bar{X}_A=36$	$\bar{X}_B=42$
$\sigma_A=6$	$\sigma_B=8$

A point estimate for $\mu_B - \mu_A$ is

$$\bar{X}_B - \bar{X}_A = 42 - 36 = 6$$

Confidence interval:

$\alpha = ??$

$96\% = (1-\alpha)100\% \Leftrightarrow 0.96 = (1-\alpha) \Leftrightarrow \alpha = 0.04 \Leftrightarrow \alpha/2 = 0.02$

$Z_{\frac{\alpha}{2}} = Z_{0.02} = 2.05$

A 96% C.I. for $\mu_B - \mu_A$ is

$$(\bar{X}_B - \bar{X}_A) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}} < \mu_B - \mu_A < (\bar{X}_B - \bar{X}_A) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}$$

$$(\overline{X}_B - \overline{X}_A) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}$$

$$(42 - 36) \pm Z_{0.02} \sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$6 \pm (2.05) \sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$6 \pm 2.571$$

$$3.43 < \mu_B - \mu_A < 8.57$$

We are 96% confident that $\mu_B - \mu_A \in (3.43, 8.57)$.

Example: (Second Case)

To compare the resistance of wire *A* with that of wire *B*, an experiment shows the following results based on two independent samples (original data multiplied by 1000):

Wire *A*: 140, 138, 143, 142, 144, 137

Wire *B*: 135, 140, 136, 142, 138, 140

Assuming equal variances, find 95% confidence interval for $\mu_A - \mu_B$ Where $\mu_A(\mu_B)$ is the mean resistance of wire A(B).

Solution:

Wire A	Wire B
$n_A=6$	$n_B=6$
$\bar{X}_A=140.67$	$\bar{X}_B=138.50$
$S^2_A=7.86690$	$S^2_B=7.10009$

A point estimate for $\mu_A - \mu_B$ is

$$\bar{X}_A - \bar{X}_B = 140.67-138.50 = 2.17$$

Confidence interval:

$$95\% = (1-\alpha)100\% \Leftrightarrow 0.95 = (1-\alpha) \Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025$$

$$\nu = df = n_A + n_B - 2 = 10$$

$$t_{\frac{\alpha}{2}} = t_{0.025} = 2.228$$

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

$$= \frac{(6-1)(7.86690) + (6-1)(7.10009)}{6+6-2} = 7.4835$$

$$S_p = \sqrt{S_p^2} = \sqrt{7.4835} = 2.7356$$

A 95% C.I. for $\mu_A - \mu_B$ is

$$(\bar{X}_A - \bar{X}_B) - t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} < \mu_A - \mu_B < (\bar{X}_A - \bar{X}_B) + t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

or $$(\bar{X}_A - \bar{X}_B) \pm t_{\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

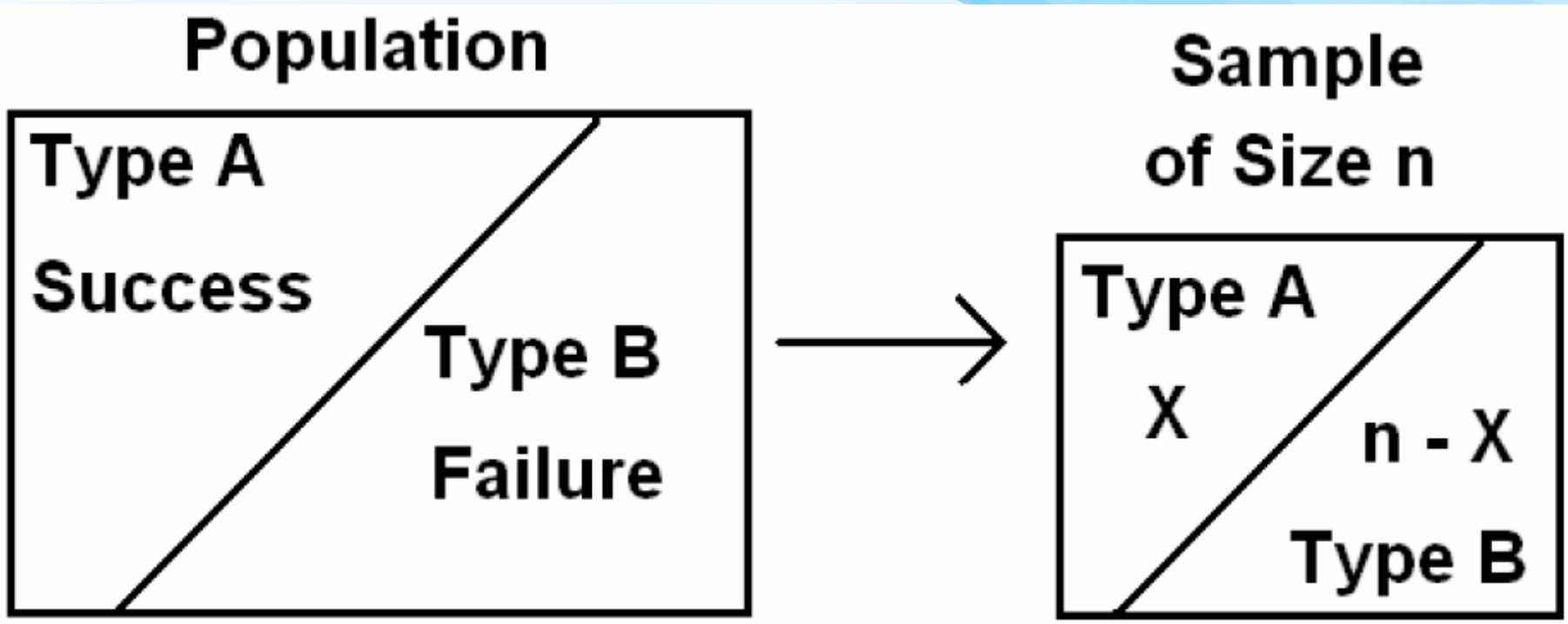
$$(140.67 - 138.50) \pm (2.228)(2.7356)\sqrt{\frac{1}{6} + \frac{1}{6}}$$

$$2.17 \pm 3.51890$$

$$-1.35 < \mu_A - \mu_B < 5.69$$

We are 95% confident that $\mu_A - \mu_B \in (-1.35, 5.69)$

Single Sample: Estimating of a Proportion



$.p$ = Population proportion of successes (elements of Type A) in the population

$$= \frac{A}{A+B} = \frac{\text{no. of elements of type } A \text{ in the population}}{\text{Total no. of elements}}$$

$.n$ = sample size

X = no. of elements of type A in the sample of size n.

$\hat{p}$ = Sample proportion of successes (elements of Type A) in the sample

$$= \frac{X}{n}$$

Recall that:

(1) $X \sim \text{Binomial}(n, p)$

(2) $E(\hat{p}) = E\left(\frac{X}{n}\right) = p$

(3) $\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{pq}{n}$; $q = 1 - p$

(4) For large n, we have

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \quad \text{(Approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \quad \text{(Approximately)}$$

Point Estimation for p

A good point estimator for the population proportion p is given by the statistic (sample proportion):

$$\hat{p} = \frac{X}{n}$$

Confidence Interval for p

Result:

For large n, an approximate $(1-\alpha)100\%$ confidence interval for p is :

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad ; \quad \hat{q} = 1 - \hat{p}$$

$$\left(\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \; , \quad \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

$$\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Example

In a random sample of $n = 500$ families owing television sets in the city of Hamilton, Canada, it was found that $x = 340$ subscribed to HBO. Find 95% confidence interval for the actual proportion of families in this city who subscribe to HBO.

Solution:

p = proportion of families in this city who subscribe to HBO.

n = sample size = 500

X = no. of families in the sample who subscribe to HBO. = 340

$\hat{p}$= proportion of families in the sample who subscribe to HBO $= \frac{X}{n} = \frac{340}{500} = 0.68$

$$\hat{q} = 1 - \hat{p} = 1 - 0.68 = 0.32$$

A point estimator for p is

$$\hat{p} = \frac{X}{n} = \frac{340}{500} = 0.68$$

Now,

95% = (1−α)100% ⇔ 0. 95 = (1−α) ⇔ α=0.05

⇔ α/2 = 0.025

$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$

A 95% confidence interval for p is:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \quad ; \quad \hat{q} = 1 - \hat{p}$$

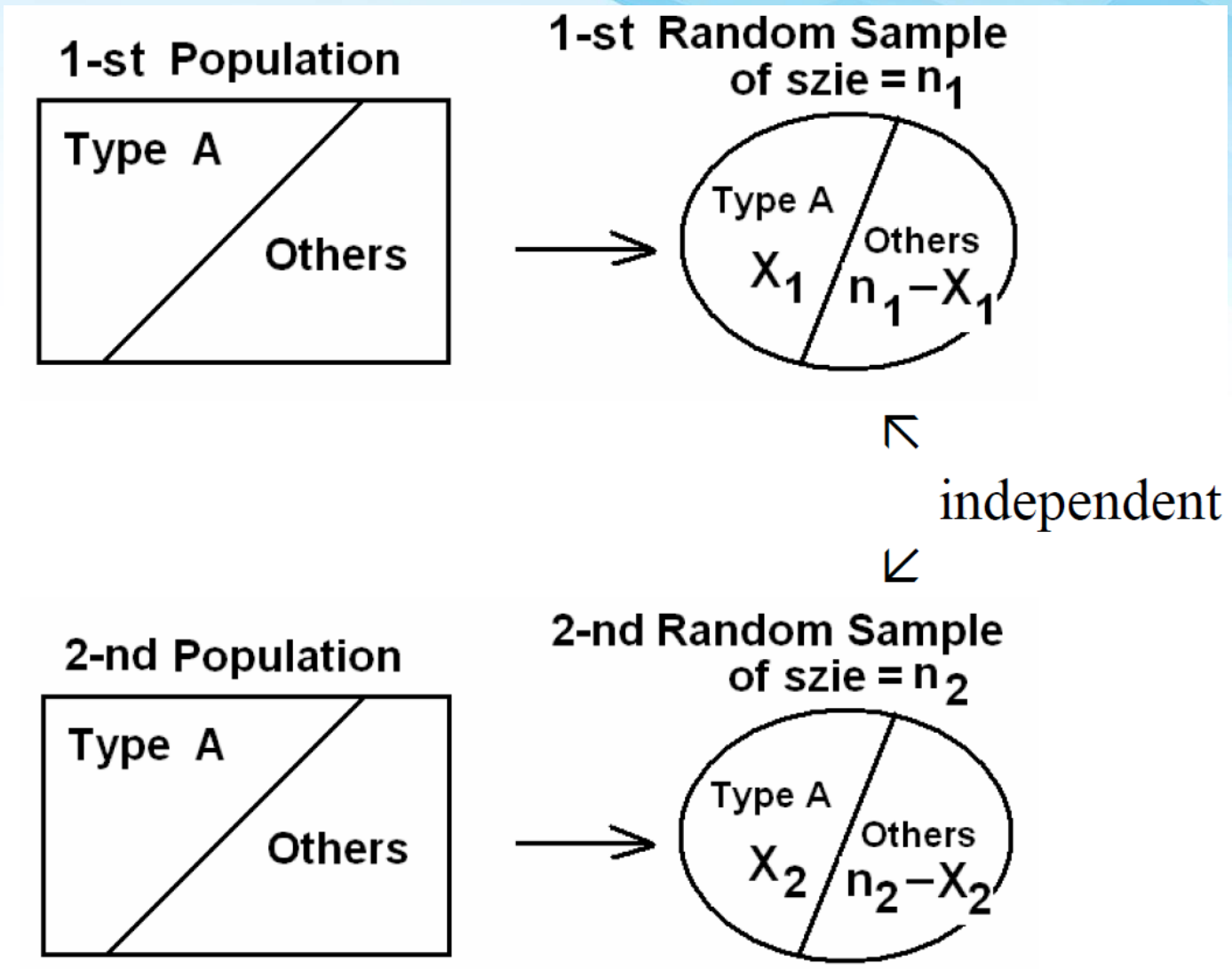
$$0.68 \pm 1.96 \sqrt{\frac{(0.68)(0.32)}{500}}$$

$$0.68 \pm 0.04$$

$$0.64 < p < 0.72$$

We are 95% confident that $p \in (0.64, 0.72)$.

Two Samples: Estimating the Difference between Two Proportions



Suppose that we have two populations:

- $.p_1$ = proportion of the 1-st population.
- $.p_2$ = proportion of the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type A in the 1-st sample.

$$X_1 \sim \text{Binomial}(n_1, p_1)$$

$$E(X_1) = n_1 p_1$$

$$\text{Var}(X_1) = n_1 p_1 q_1 \quad (q_1 = 1 - p_1)$$

- Let X_2 = no. of elements of type A in the 2-nd sample.

  $X_2 \sim \text{Binomial}(n_2, p_2)$

  $E(X_2) = n_2 p_2$

  $\text{Var}(X_2) = n_2 p_2 q_2 \quad (q_2 = 1 - p_2)$

- $\hat{p}_1 = \dfrac{X_1}{n_1}$ = proportion of the 1-st sample
- $\hat{p}_2 = \dfrac{X_2}{n_2}$ = proportion of the 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.

Results:

(1) $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

(2) $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$; $q_1 = 1 - p_1, q_2 = 1 - p_2$

(3) For large n_1 and n_2, we have

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right) \text{ (Approximately)}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1) \text{ (Approximately)}$$

Point Estimation for $p_1 - p_2$

A good point estimator for the difference between the two proportions, $p_1 - p_2$, is given by the statistic:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Confidence Interval for $p_1 - p_2$

Result:

For large $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$, an approximate $(1-\alpha)100\%$ confidence interval for $\boldsymbol{p_1 - p_2}$ is :

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

or

$$\left((\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \right)$$

or

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example

A certain change in a process for manufacture of component parts is being considered. Samples are taken using both existing and the new procedure to determine if the new process results in an improvement. If 75 of 1500 items from the existing procedure were found to be defective and 80 of 2000 items from the new procedure were found to be defective, find 90% confidence interval for the true difference in the fraction of defectives between the existing and the new process.

Solution

.p_1 = fraction (proportion) of defectives of the existing process

.p_2 = fraction (proportion) of defectives of the new process

$\hat{p}_1$ = sample fraction of defectives of the existing process

$\hat{p}_2$ = sample fraction of defectives of the new process

Existing Process	New Process
$n_1 = 1500$	$n_2 = 2000$
$X_1=75$	$X_2=80$
$\hat{p}_1 = \frac{X_1}{n_1} = \frac{75}{1500} = 0.05$	$\hat{p}_2 = \frac{X_2}{n_2} = \frac{80}{2000} = 0.04$
$\hat{q}_1 = 1 - 0.05 = 0.95$	$\hat{q}_2 = 1 - 0.04 = 0.96$

Point Estimation for $p_1 - p_2$:

A point estimator for the difference between the two proportions, $p_1 - p_2$, is:

$$\hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01$$

Confidence Interval for $p_1 - p_2$:

A 90% confidence interval for $p_1 - p_2$ is :

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{0.05} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$0.01 \pm 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}$$

$$0.01 \pm 0.01173$$

$$-0.0017 < p_1 - p_2 < 0.0217$$

We are 90% confident that $p_1 - p_2 \in (-0.0017, 0.0217)$.

Note:

Since $0 \in$ 90% confidence interval=$(-0.0017, 0.0217)$, there is no reason to believe that the new procedure produced a significant decrease in the proportion of defectives over the existing method ($p_1 - p_2 \approx 0 \Leftrightarrow p_1 \approx p_2$).