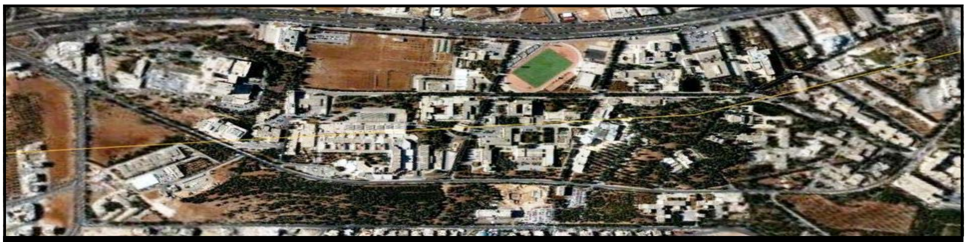
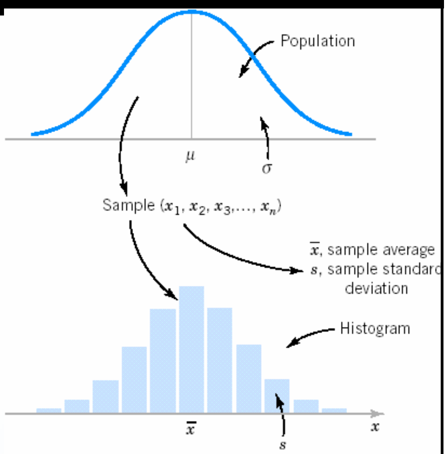


Process Modeling by Statistical Methods (0905331)
06- Point Estimation

Dr. Ali Khalaf Al-matar
Chemical Engineering Department
University of Jordan
banihaniali@yahoo.com

Population and Sample

- A **population** consists of the totality of the observations with which we are concerned.
- A **sample** is a subset of observations selected from a population.



Random Sampling: "Statistic"

- The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size n if (a) the X_i's are independent random variables, and (b) every X_i has the same probability distribution.
- A **statistic** is any function of the observations in a random sample
  - Sample mean and standard deviation are statistics.
- Since a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a **sampling distribution**.

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

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

as $n \to \infty$, is the standard normal distribution.

Sampling Distributions of Means

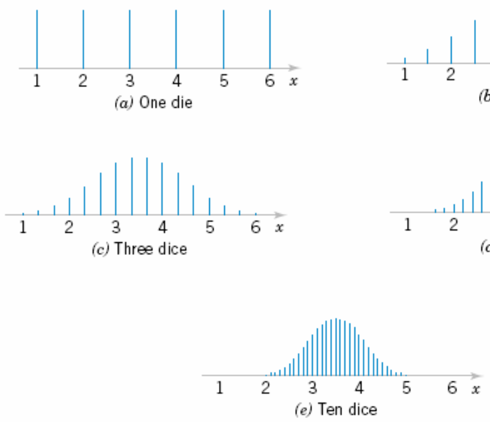


Figure 7-1 Distributions of average scores from throwing dice.

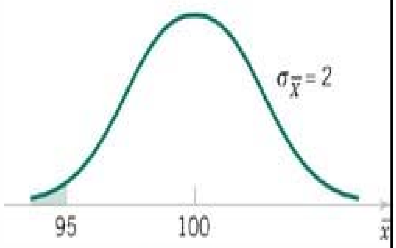
If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

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

is approximately standard normal, if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

Example 7-1: Resistors

- An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than 95 ohms.



Note that the sampling distribution of $\overline{X}$ *is normal, with mean = 100 ohms and a standard deviation*

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$z = \frac{95 - 100}{2} = -2.5$$

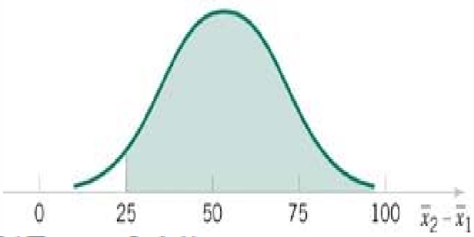
$$P(\overline{X} < 95) = P(Z < -2.5) = 0.0062$$

Example 7-3: Aircraft Engine Life

- The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of $n_1 = 16$ components is selected from the "old" process and a random sample of $n_2 = 25$ components is selected from the "improved" process. What is the probability that the difference in the two sample means $\overline{X}_2 - \overline{X}_1$ is at least 25 hours? Assume that the old and improved processes can be regarded as independent populations.

- To solve this problem, we first note that the distribution of $X1$ is normal with mean $\mu_1 = 5000$ hours and standard deviation $\sigma_1 / \sqrt{n_1} = 40 / \sqrt{16} = 10$ hours, and the distribution of X_2 is normal with mean $\mu_2 = 5050$ hours and $\sigma_2 / \sqrt{n_2} = 30 / \sqrt{25} = 6$ hours.
- Now the distribution of $X_2 - X_1$ is normal with mean $\mu_2 - \mu_1 = 5050 - 5000 = 50$ hours and variance $\sigma_2^2 / n_2 + \sigma_1^2 / n_1 = (6)^2 + (10)^2 = 136$ hours². This sampling distribution is shown in Fig. 7-4. The probability that $\bar{X}_2 - \bar{X}_1 \geq 25$ is the shaded portion of the normal distribution in this figure.

$$z = \frac{25 - 50}{\sqrt{136}} = -2.14$$



$$P(\bar{X}_2 - \bar{X}_1 \geq 25) = P(Z \geq -2.14)$$
$$= 0.9838$$

Point Estimates

- A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$.
  - The statistic $\hat{\Theta}$ is called the **point estimator**.

Required for Population	Estimate	
The mean μ of a single population	$\hat{\mu} = \bar{x}$	the sample mean
The variance σ^2 (or σ) of a single population	$\hat{\sigma}^2 = s^2$	the sample variance
The proportion p of items in a population that belong to a class of interest	$\hat{p} = x/n$	the sample proportion, where x is the number of items in a random sample of size n that belong to the class of interest.
The difference in means of two populations, $\mu_1 - \mu_2$	$\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$	the difference between the sample means of two independent random samples.
The difference in two population proportions, $p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	difference between two sample proportions computed from two independent random samples.

Unbiased Estimators

- The point estimator $\hat{\Theta}$ is an **unbiased estimator** for the parameter θ if

$$E(\hat{\Theta}) = \theta$$
$$E(\hat{\Theta}) - \theta = 0$$

- If the estimator is not unbiased, then the **bias** of the estimator is defined as: $E(\hat{\Theta}) - \theta$

Sample mean is an unbiased estimator for population mean
Sample variance is an unbiased estimator for population variance
Sample standard deviation is a biased estimator for population σ.

Minimum Variance Unbiased Estimator

- If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).
  - MVUE is most likely among all unbiased estimators to produce an estimate $\hat{\theta}$ that is close to the true value of θ.
  - If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ², the sample mean X is the MVUE for μ.

Standard Error of an Estimator

- The **standard error** of an estimator $\hat{\Theta}$ is its standard deviation, given by $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. If the standard error involves unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\Theta}}$ produces an **estimated standard error,** denoted by $\hat{\sigma}_{\hat{\Theta}}$.
  - Suppose we are sampling from a normal distribution with mean μ and variance σ^2. Now the distribution of X is normal with mean μ and variance σ^2/n, so the standard error of $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

sample standard deviation →

$$\hat{\sigma}_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Example 7-5: Thermal Conductivity

- An article in the *Journal of Heat Transfer* (Trans. ASME, Sec. C, 96, 1974, p. 59) described a new method of measuring the thermal conductivity of Armco iron. Using a temperature of 100°F and a power input of 550 watts, the following 10 measurements of thermal conductivity (in Btu/hr-ft-°F) were obtained: 41.60, 41.48, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 4.04.
- A point estimate of the mean thermal conductivity at 100°F and 550 watts is the sample mean or $\bar{x}$=41.924 Btu/hr-ft-°F.

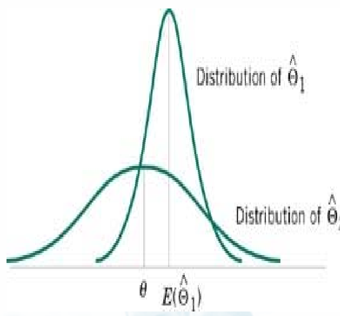
$$\hat{\sigma}_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

Mean Square Error of an Estimator

- The **mean square error** of an estimator $\hat{\Theta}$ of the parameter θ is

$$MSE(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 = E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2$$
$$= V(\hat{\Theta}) + \text{Bias}^2$$

  - important criterion for comparing two estimators



$$\text{Relative efficiency} = \frac{MSE(\hat{\Theta}_1)}{MSE(\hat{\Theta}_2)} \begin{cases} < 1, & \hat{\Theta}_1 \text{ is more efficient estimator} \\ > 1, & \hat{\Theta}_2 \text{ is more efficient estimator} \end{cases}$$

Point Estimates & Confidence Intervals

- Point estimate
  - the value computed from sample information, that is used to estimate the population parameter,
- Confidence interval
  - a range of values constructed from sample data so the parameter occurs within that range at a specified probability. The specified probability is called the level of confidence.

Standard Error of the Sample Mean

- Standard error of the mean, population standard deviation known

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Standard error of the mean based on the sample standard deviation

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

σ: population standard deviation

s: sample standard deviation

n: sample size

Confidence Interval for a Mean

- Use the standard values for the standard normal distribution values (z). Confidence interval for a mean

$$\bar{X} \pm z \frac{s}{\sqrt{n}}$$

Example

An experiment involves selecting a random sample of 256 managers. One item of interest is annual income. The sample mean is $45420, and the sample standard deviation is $2050

1. What is the estimated mean income of all middle managers?
2. What is the 95% confidence interval for the population mean?
3. What are the 95% confidence limits for the population mean?
4. What degree of confidence is being used?
5. Interpret the findings.

Solution

The point estimate of the population mean is $45420

The confidence interval can be found by

$$\bar{X} \pm z\frac{s}{\sqrt{n}} = \$45420 \pm 1.96\frac{\$2050}{\sqrt{256}}$$

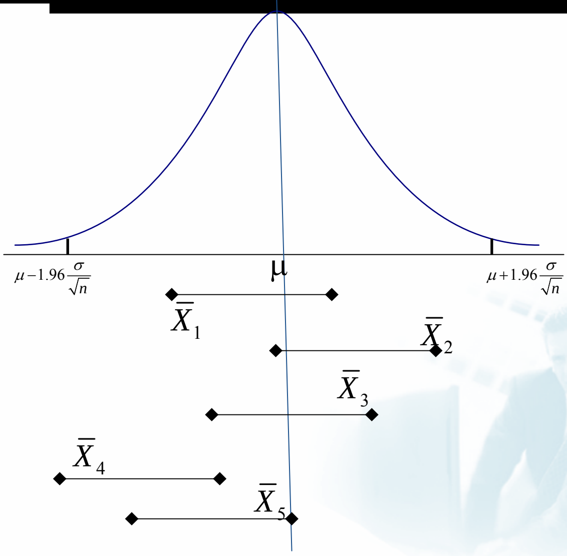
$$= \$45420 \pm 251.125$$

Which means the confidence interval is $45169 and $45671

The end points of the confidence interval are called the confidence limits ($45169 and $45671)

The measure of confidence interval is referred to as the degree of confidence or the level of confidence. In this case it is 0.95.

Interpretation: either an interval contains the population mean or not. About 5 of 100 confidence intervals would not contain the population mean.



Confidence Interval for a Population Proportion

- A proportion is based on a count of the number of successes relative to the total number sampled
- Confidence interval using a population proportion $p \pm z\,\sigma_p$
- Standard error of the sample proportion

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

- Confidence interval for a sample proportion

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

Example

- Suppose 1600 of 2000 union members samples said they plan to vote for the proposal to merge with another union. Using 0.95 level of confidence, what is the interval estimate for the population proportion? Based on the confidence interval what conclusions can be drawn?

Solution

- The proportion is 1600/2000 = 0.8
- Use the formula to estimate the interval

$$p \pm z\sqrt{\frac{p(1-p)}{n}} = 0.80 \pm 1.96\sqrt{\frac{0.80(1-0.80)}{2000}}$$

$$= 0.782 \text{ and } 0.818$$

- If the resolution requires 75% of the votes to approve it. Based on the sample results, when all union members vote, the proposal will probably pass because the 0.75 lies below the interval 0.782 and 0.818.

Finite Population Correction

- The populations sampled before were very large or assumed infinite. What happens when the sampled population is finite?
- For a finite population, where the total number of objects is N and the size of the sample is n, the following adjustments is made to the standard errors of the sample means and the proportion
- Apply correction for
  - Population is finite, and
  - Sample to population ratio is greater than 5%.

Finite Population Correction Factors

- Standard error of the sample mean corrected for finite population

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- Standard error for the sample proportions corrected for finite population

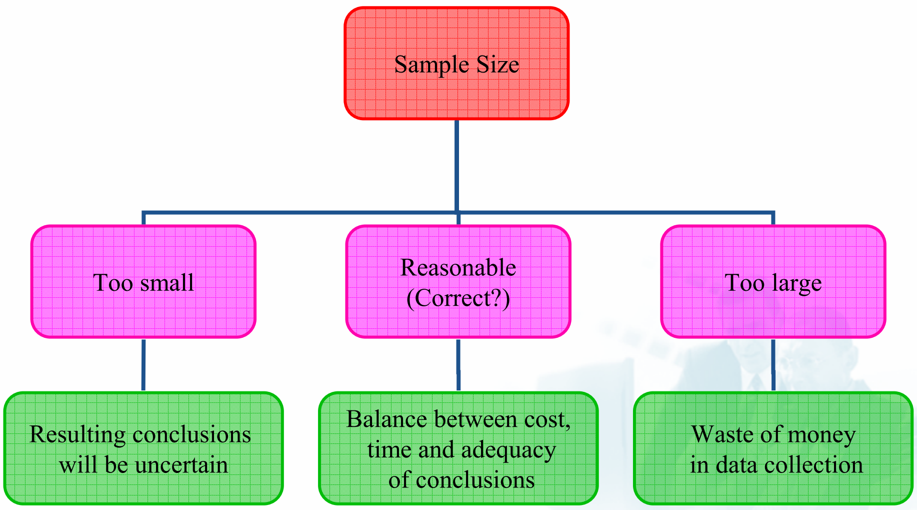
$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Example

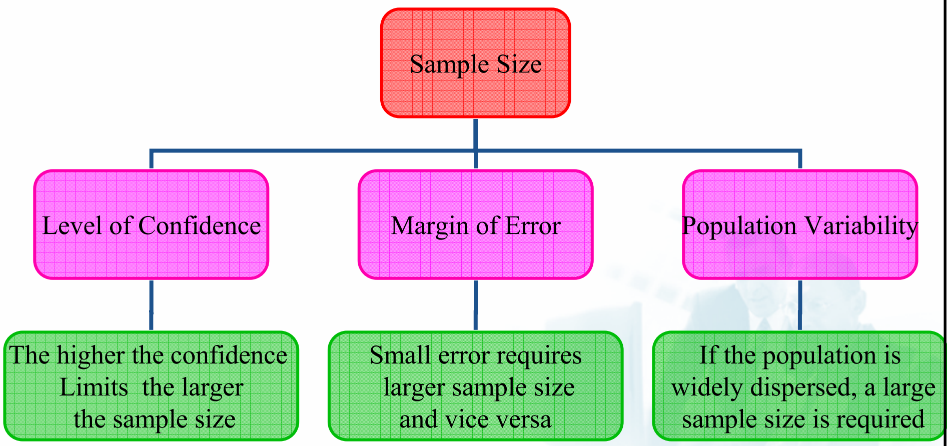
- There are 250 families in the small town of Scandia. A poll of 40 families revealed the mean annual charity contributions is \$450 with a standard deviation of \$75. Construct a 95% confidence interval for the mean annual contribution.
  - Note that the population is finite and that the sample to population ratio is greater than 5% (40/250=16%).
  - Use the finite population correction factor

$$\bar{X} \pm z \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \$450 \pm 1.96\frac{\$75}{\sqrt{40}}\sqrt{\frac{250-40}{250-1}}$$
$$= \$450 \pm 23.243\sqrt{0.8434} = \$450 \pm 21.35$$
$$= [\$428.65, \$471.35]$$

Choosing Sample Size



Factors Affecting Sample Size



Finding Sample Size

- We can express the interaction among the three factors as

$$E = z\frac{s}{\sqrt{n}}$$

- The required sample size to estimate a mean is given by

$$n = \left(\frac{z \cdot s}{E}\right)^2$$

n: sample size
z: standard normal value corresponding to the desired level of confidence
s: estimate of the population standard deviation
E: maximum allowable error

Example

- A student in the public administration wants to determine the mean amount members of the city council earn. The error in estimating the mean is to be less than \$100 with a 95% level of confidence. The student found a report by the Department of Labor that estimated the standard deviation to be \$1000. What is the required sample size? If the level of confidence was changed to 99%, what is the required sample size?

E = \$100

z = 1.96 (CORRESPONDING TO 95%)

s = \$1000

$$n = \left(\frac{z \cdot s}{E}\right)^2 = \left(\frac{1.96 \cdot \$1000}{\$100}\right)^2 = 384.16 = 385$$

z = 2.58 (Corresponding to 99%)

$$n = \left(\frac{z \cdot s}{E}\right)^2 = \left(\frac{2.58 \cdot \$1000}{\$100}\right)^2 = 665.64 = 666$$

Finding Sample Size of a Proportion

- The same criteria can be used to find the sample size for a proportion

$$n = p(1-p)\left(\frac{z}{E}\right)^2$$

n: sample size

z: standard normal value corresponding to the desired level of confidence

p: estimate of the population proportion

E: maximum allowable error

Example

- The study in the last example also estimates the proportion of cities that have private refuse collectors. The student wants the estimate to be within 0.10 of the population proportion, the desired level of confidence is 90%, and no estimate is available for the population proportion. What is the required sample size?

$E = 0.1$

$z = 1.65$ (corresponding to 90%)

Proportion estimate is not given, assume it to be 0.5 !?!

$$n = p(1-p)\left(\frac{z}{E}\right)^2 = (0.5)(0.5)\left(\frac{1.65}{0.1}\right)^2 = 68.06 = 69$$