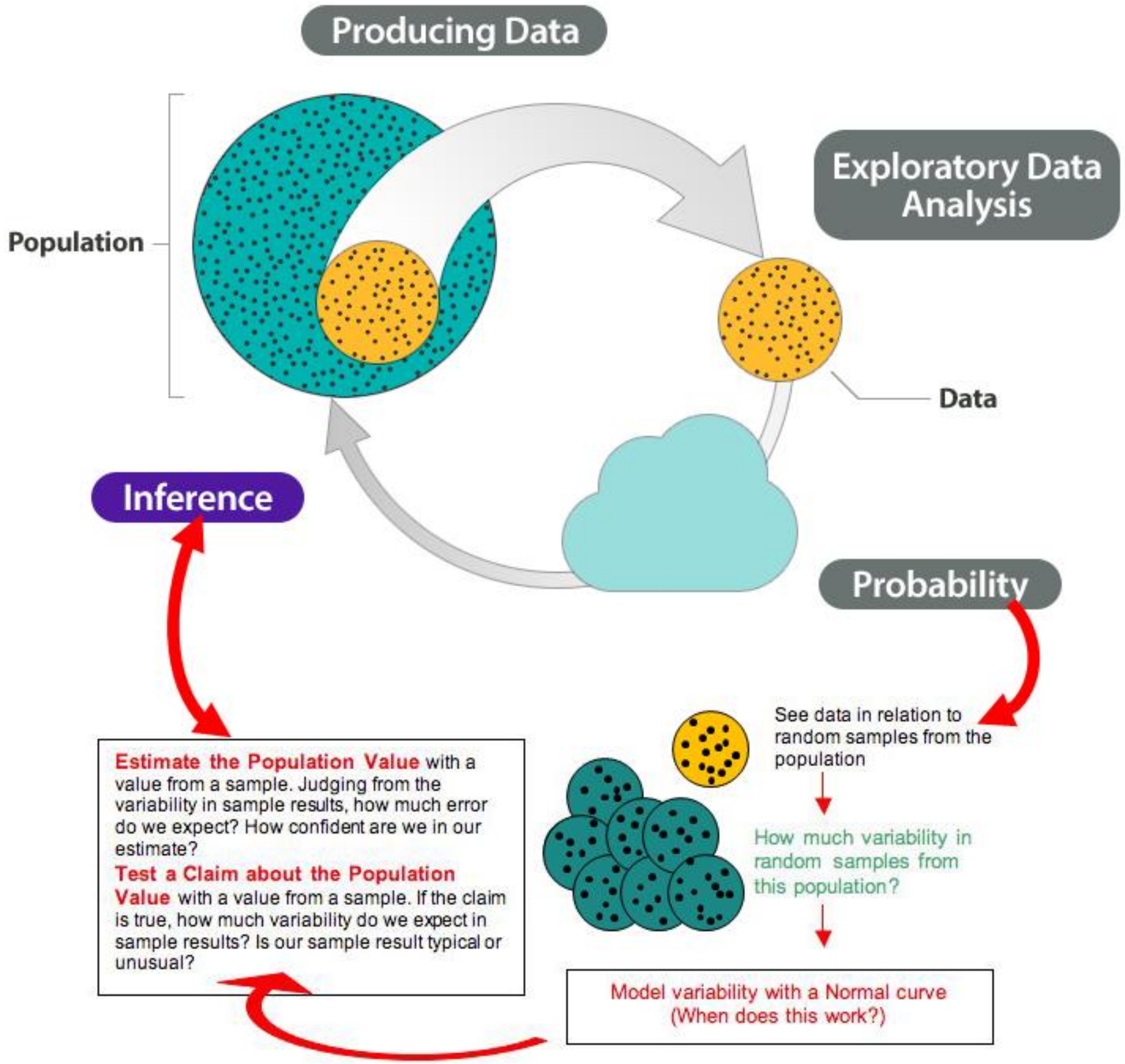
Statistical Inference

Statistical inference (1)

- Inferences are *generalizations* about a population that are made on the *basis of a sample* collected from the population.

Statistical inference deals with drawing conclusions about population parameters from an analysis of the sample data.

- To be meaningful, a statistical inference must include a *specification of the uncertainty* that is determined using the idea of *probability* and *sampling distribution* of statistic.



Statistical inference (2)

- The two most important *types of inferences* are:
  - *a)* *Estimation* of parameter(s)
  - *b)* *Testing* of statistical hypotheses

- More specifically, depending on the purpose of the study, we may wish to do one, two, or all three of the following:
  - *a)* *Estimate a single* value for the unknown μ (point estimation).
  - *b)* *Determine an interval* of plausible values for μ (interval estimation).
  - c) Decide whether or not the mean time μ is the specified value(testing statistical hypotheses)

- **Point estimation of a population mean**
- Confidence interval for a population mean
- Testing hypotheses about a population mean

Point estimation of a population mean (1)

- The aim of point estimation is to calculate, from the sample data, **a *single number*** that is likely to be *close* to the unknown value of the parameter.

> A statistic intended for estimating a parameter is called a **point estimator,** or simply an **estimator.** The standard deviation of an estimator is called its **standard error:** S.E.

- Recall, the properties of sample mean $\bar{X}$,

1. $E(\bar{X}) = \mu$
2. $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ so, S.E.$(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
3. With large n, $\bar{X}$ is nearly normally distributed

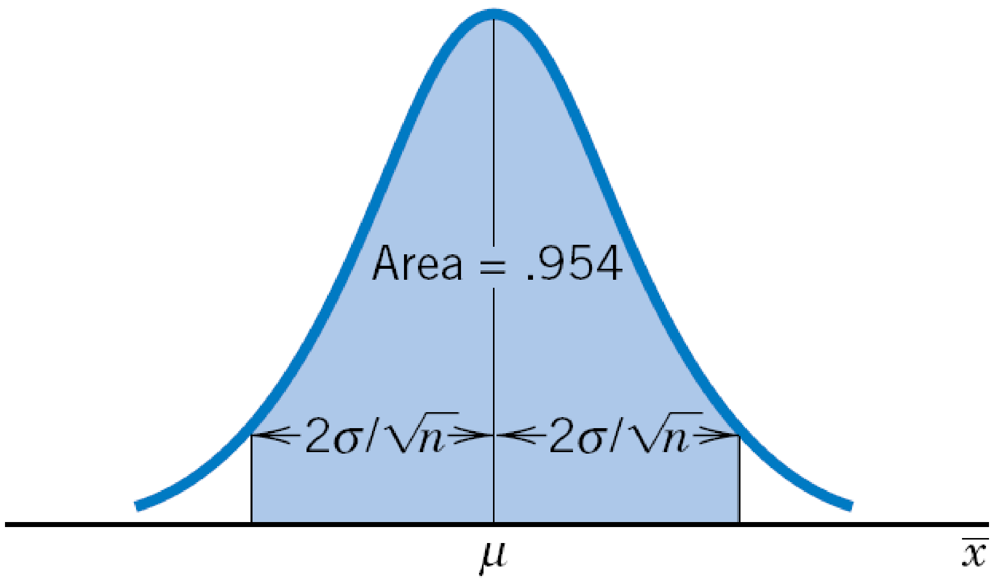
   with mean μ and standard Deviation $\frac{\sigma}{\sqrt{n}}$

Point estimation of a population mean (2)

- In a normal distribution, the interval running *two standard deviations* on either side of the mean contains probability 0.954.

- Prior to sampling, the probability is 0.954 that the estimator $\bar{X}$ will be at most a distance $\frac{2\sigma}{\sqrt{n}}$ from the true parameter value μ.

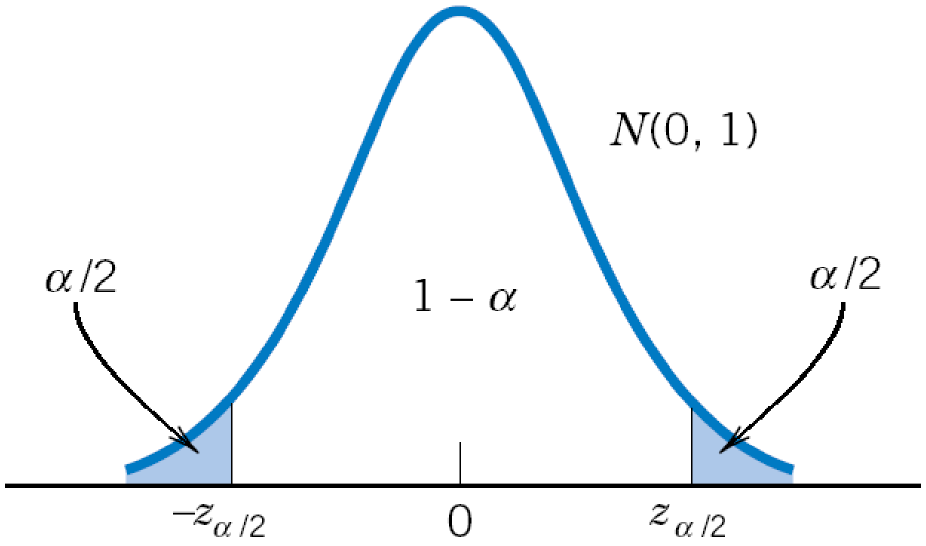
- This probability statement can be rephrased by saying that ***when we are estimating*** μ by the 95.4% ***error margin*** is $\frac{2\sigma}{\sqrt{n}}$.



Point estimation of a population mean (3)

Notation

$\boldsymbol{z_{\alpha/2}} = \boldsymbol{Upper\ \ \alpha/2}$ point of standard normal distribution. That is, the area to the right of $\boldsymbol{z_{\alpha/2}}$ is $\boldsymbol{\alpha/2}$ and the area between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $\mathbf{1-\alpha}$ (see Figure on the right).



Values of $z_{\alpha/2}$

$1-\alpha$	.80	.85	.90	.95	.99
$z_{\alpha/2}$	1.28	1.44	1.645	1.96	2.58

Point estimation of a population mean (4)

Point Estimation of the Mean

Parameter: Population mean μ.

Data: $X_1, \ldots, X_n$ (a random sample of size n)

Estimator: $\overline{X}$ (sample mean)

$$\text{S.E.}(\overline{X}) = \frac{\sigma}{\sqrt{n}} \qquad \text{Estimated S.E.}(\overline{X}) = \frac{S}{\sqrt{n}}$$

For large n, the $100(1 - \alpha)\%$ error margin is $z_{\alpha/2}\,\sigma/\sqrt{n}$.
(If σ is unknown, use S in place of σ.)

Point estimation of a population mean (5)

Example:

The following table consisting of 40 measurements of the time devoted to community service the past month, **give a point estimate** of the population mean amount of time and **state a 95% error margin**.

0	0	0	0	0	0	0	1	1	1
2	2	2	2	2	3	3	3	3	4
4	4	4	5	5	5	5	5	5	5
5	5	6	6	6	8	10	15	20	25

Answer:

$$\bar{\mathrm{x}} = \frac{\sum \mathrm{x_i}}{40} = 4.55,$$

$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{39}} = \sqrt{26.715} = 5.17.$ To calculate the 95% error margin, we set $1 - \alpha$ so that $\alpha/2=0.025$

and $z_{\alpha/2} = 1.96$. Therefore, the 95% error margin is

$$\frac{1.96\, s}{\sqrt{n}} = \frac{1.96\ x\ 5.17}{\sqrt{40}} = \mathbf{1.60} \text{ hours}$$

We do *not* expect the population mean to be *exactly 4.55* and we attach a *95% error margin* of *plus* and *minus* 1.60 hours.

- **Point estimation of a population mean**
- **Confidence interval for a population mean**
- Testing hypotheses about a population mean

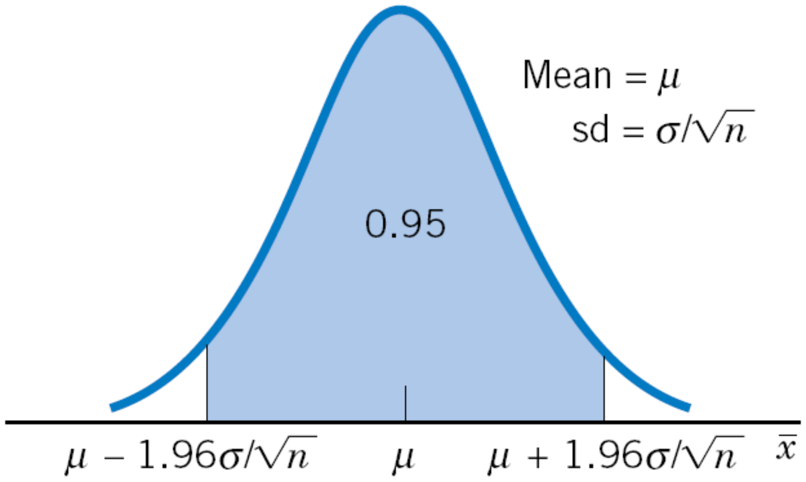
Confidence interval for a population mean (1)

- Confidence interval estimator for population parameter is a rule for determining (based on sample) an interval that is *likely to include the parameter*. The corresponding estimate is called ***confidence interval estimate***.
- A probability statement about $\bar{x}$ based on the normal distribution provides the cornerstone for the development of a confidence interval.
- The normal table shows that the probability is 0.95 that a normal random variable will lie within 1.96 standard deviations from its mean. For $\bar{x}$, then we have,

$$P\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x}$ is the same as $\mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$

$\bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}$ is the same as $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu$



……..next page

Confidence interval for a population mean (2)

…cont'ed

By transposing from one side of an inequality to the other.

$$\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{\text{x}} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] \text{ is equivalent to } \left[\bar{\text{x}} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{\text{x}} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Both events state that the difference lies between 1.96 and thus, the probability statement

$$P\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{\text{x}} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95 \quad \text{can be expressed as}$$

$$P\left[\bar{\text{x}} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{\text{x}} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Because σ is assumed to be known, both the upper and lower endpoints can be computed as soon as the sample data are available.

A **95% confidence interval for μ** when the population is normal and σ known is;

$$\boldsymbol{\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)}$$

Confidence interval for a population mean (3)

Example:

The daily carbon monoxide (CO) emission from a large production plant will be measured on 25 randomly selected weekdays. The production process is always being modified and the current mean value of daily CO emissions μ is unknown. Data collected over several years confirm that, for each year, the distribution of CO emission is normal with a standard deviation of 0.8 ton. Suppose the sample mean is found to be $\bar{x} = 2.7$ tons. Construct a 95% confidence interval for the current daily mean emission μ.

Answer:

$$\left(2.7 - 1.96\frac{0.8}{\sqrt{25}},\ \ 2.7 + 1.96\frac{0.8}{\sqrt{25}}\right) = (\mathbf{2.39},\ \mathbf{3.01})\ \textbf{tons} \text{ is a } 95\% \text{ confidence interval for } \mu.$$

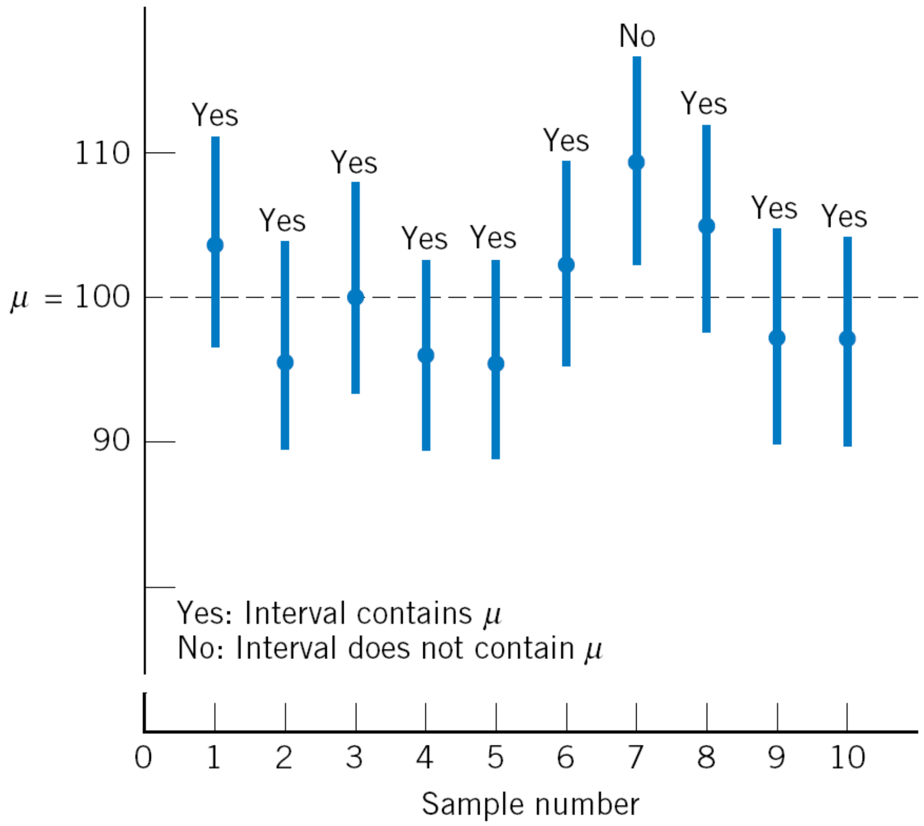
Since μ is unknown, we cannot determine whether or not μ lies in this interval.

- *In summary*, when the population is normal and σ is known, a $100(1-a)\%$ confidence interval for μ is given by

$$\left(\bar{\boldsymbol{x}} - \boldsymbol{z}_{\boldsymbol{\alpha}/2}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}},\ \ \bar{\boldsymbol{x}} + \boldsymbol{z}_{\boldsymbol{\alpha}/2}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

Interpretations of confidence intervals (1)

- Let $\mu = 100$ and $\sigma = 10$. Ten samples of size 7 are selected, and a 95% confidence interval $\bar{x} \pm 1.96\ x\ 10/\sqrt{7}$ is computed from each.
- Each vertical line (from the graph) segment represents one confidence interval.
- The midpoint of a line is the observed value of $\bar{x}$ for that particular sample.
- Note that all the intervals are of the same length $2x1.96\ x\ \sigma/\sqrt{n} = 14.8$



Interpretations of confidence intervals (2)

- Stated in terms of a 95% confidence interval for μ,

1. Before we sample, a confidence interval $\left(\bar{\mathrm{x}} - 1.96\frac{\sigma}{\sqrt{n}},\quad \bar{\mathrm{x}} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is a random interval that attempts to cover the true value of the parameter μ.

2. The probability $P\left[\bar{\mathrm{x}} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{\mathrm{x}} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$ interpreted as the long-run relative frequency over many repetitions of sampling asserts that about 95% of the intervals will cover μ.

3. Once $\bar{\mathrm{x}}$ is calculated from an observed sample, the interval $\left(\bar{\mathrm{x}} - 1.96\frac{\sigma}{\sqrt{n}},\quad \bar{\mathrm{x}} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is presented as a 95% confidence interval for μ.

4. In any application we never know if the 95% confidence interval covers the unknown mean μ, we adopt the terminology confidence once the interval is calculated.

- **Point estimation of a population mean**
- **Confidence interval for a population mean**
- **Testing hypotheses about a population mean**

Testing hypotheses about a population mean (1)

- The goal of testing statistical hypotheses is to *determine* if a claim or conjecture about some feature of the population, a parameter, is *strongly supported* by the information obtained from the sample data.

- While formulating the hypotheses,
  - The *claim* or the research hypothesis that *we wish* to establish is called the ***alternative hypothesis*** $\boldsymbol{H_1}$.
  - The opposite statement, *one that nullifies* the research hypothesis, is called the ***null hypothesis*** $\boldsymbol{H_0}$

Formulation of H_0 and H_1

When our goal is to establish an assertion with substantive support obtained from the sample, the negation of the assertion is taken to be the null hypothesis H_0 and the assertion itself is taken to be the alternative hypothesis H_1 .

Testing hypotheses about a population mean (2)

- *Rejection of H_0* amounts to saying that H_1 is substantiated, whereas *nonrejection* or retention of H_0 means that **H_1 fails to be substantiated**.
- A key point is that a decision to reject H_0 *must be based on strong evidence*. Otherwise, the claim H_1 could not be established beyond a reasonable doubt.

> The random variable $\overline{X}$ whose value serves to determine the action is called the **test statistic.**
>
> A **test of the null hypothesis** is a course of action specifying the set of values of a test statistic $\overline{X}$, for which H_0 is to be rejected.
>
> This set is called the **rejection region** of the test.

- A test is completely specified by a *test statistic* and the *rejection region*.

Testing hypotheses about a population mean (3)

Two types of error and their probabilities

- When our sample-based decision is to reject H_0, either we have a correct decision (if H_1 is true) or we commit a *type I error* (if H_0 is true). On the other hand, a decision to retain H_0 either constitutes a correct decision (if H_0 is true) or leads to a *type II error*.

Two Types of Error

Type I error: Rejection of H_0 when H_0 is true

Type II error: Nonrejection of H_0 when H_1 is true

α = Probability of making a Type I error (also called the **level of significance**)

β = Probability of making a Type II error

Testing hypotheses about a population mean (4)

Example:

For the upgraded teller machine, let μ be the population mean transaction time. Because μ is claimed to be lower than 270 seconds, we formulate the hypothesis as:

$$H_0: \mu = 270 \quad \text{versus} \quad H_1: \mu < 270 \qquad n = 38 \qquad \alpha = 0.05 \; \sigma = 24$$

Decision Based on Sample	Unknown True Situation: H_0 True $\mu = 270$	Unknown True Situation: H_1 True $\mu < 270$
Reject H_0	Wrong rejection of H_0 (Type I error)	Correct decision
Retain H_0	Correct decision	Wrong retention of H_0 (Type II error)

Testing hypotheses about a population mean (5)

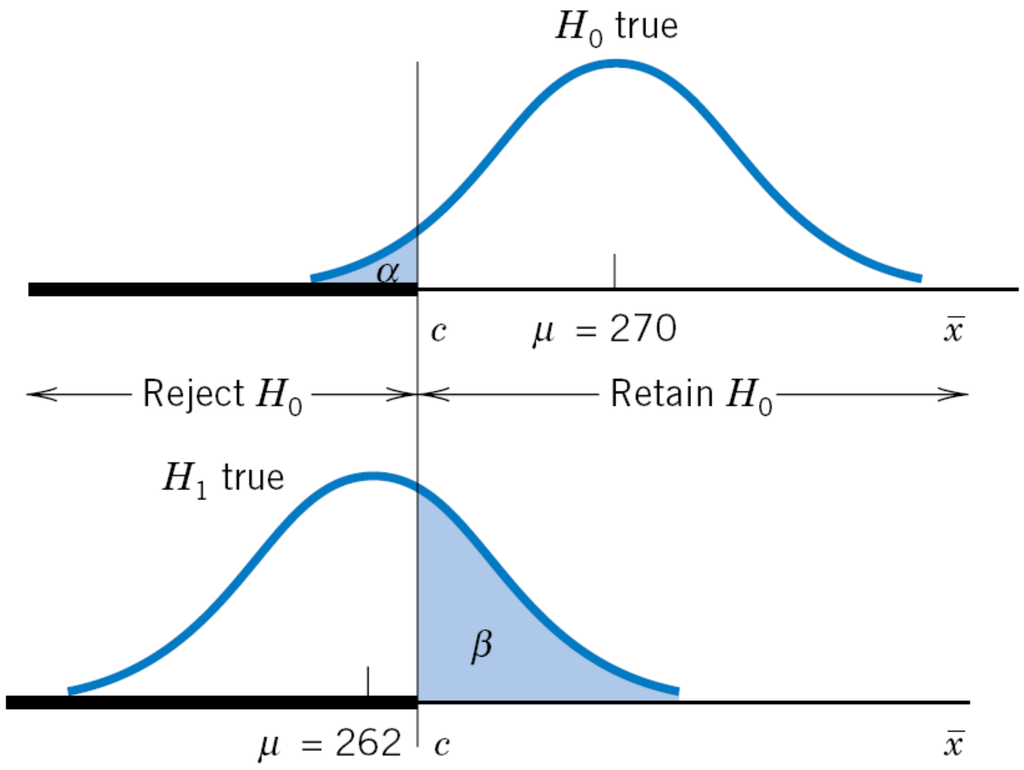
Example…cont'ed

Rejection region of the form $R: \bar{X} \leq c$; so that,

$\alpha = P[\bar{X} \leq c]$ when $\mu = 270$ $(H_0$ true)

$\beta = P[\bar{X} > c]$ when $\mu < 270$ $(H_1$ true)

- The probability β depends on the numerical value of μ that prevails under H_1.
- Figure on the right shows that:
  - The type I error probability α as the shaded area under the normal curve that has $\mu = 270$
  - The type II error probability β as the shaded area under the normal curve that has $\mu = 262$, a case of H_1being true.



Testing hypotheses about a population mean (6)

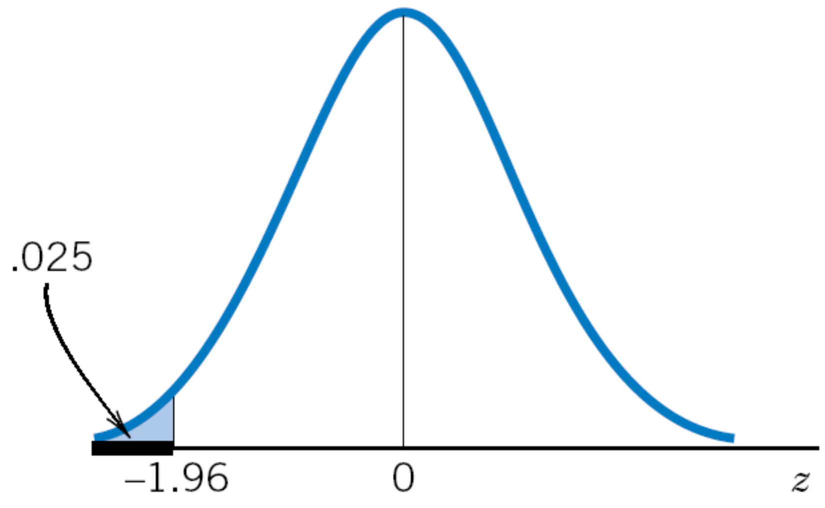
Example…cont'ed

- Suppose, the sample mean and standard deviation are found to be 261 and 22 seconds, respectively. Test the null hypothesis $H_0 : \mu = 270$ versus $H_1 : \mu < 270$ using a 2.5% level of significance and state whether or not the claim $\mu < 270$ is substantiated.

Answer:

- Employ the test statistic, $Z = \frac{\bar{X} - 270}{s/\sqrt{38}} = \frac{261 - 270}{22/\sqrt{38}} = -2.52$. The rejection region is $R : \bar{X} \leq -1.96$ for $\alpha = 0.025$ given that with the level of significance α, we set the rejection region $R : Z \leq -z_\alpha$ (**large sample normal test** or a **Z test**).

- The observed z is in R, the null hypothesis is rejected at the level of significance a $\alpha = 0.025$. We conclude that the claim of a reduction in the mean transaction time is *strongly supported* by the data.



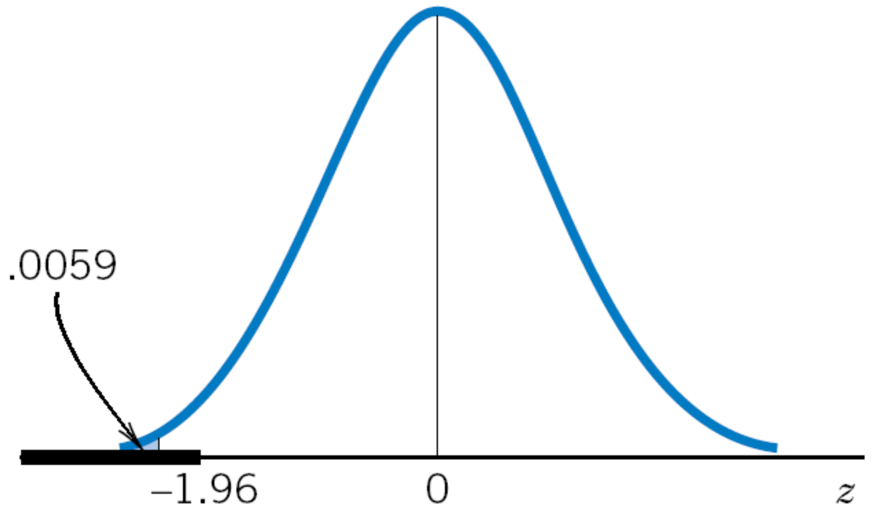
Testing hypotheses about a population mean (7)

Example…cont'ed

How small an α could we use and still arrive at the conclusion of rejecting H_0?

Answer:

- Consider the observed $z = 2.52$ itself as the cutoff point (critical value) and calculate the rejection probability $P[\bar{X} \leq -2.52] = 0.0059$.
- The smallest possible α that would permit rejection of H_0 , on the basis of the observed $z = 2.52$ is therefore 0.0059 (called **significance probability** or **P–value** of the observed z.
- The P–value serves as a measure of the *strength of evidence* against H_0.



Testing hypotheses about a population mean (8)

***P-value*:- how strong is a rejection of H_0?**

- A ***small P–value*** means that the null hypothesis is ***strongly rejected*** or the result is ***highly statistically significant***.

> The **P–value** is the probability, calculated under H_0, that the test statistic takes a value equal to or more extreme than the value actually observed.

- The general procedures of hypotheses testing:

The Steps for Testing Hypotheses

1. Formulate the null hypothesis H_0 and the alternative hypothesis H_1.
2. Test criterion: State the test statistic and the form of the rejection region.
3. With a specified α, determine the rejection region.
4. Calculate the test statistic from the data.
5. Draw a conclusion: State whether or not H_0 is rejected at the specified α and interpret the conclusion in the context of the problem. Also, it is a good statistical practice to calculate the P–value and strengthen the conclusion.

Testing hypotheses about a population mean (9)

Example:

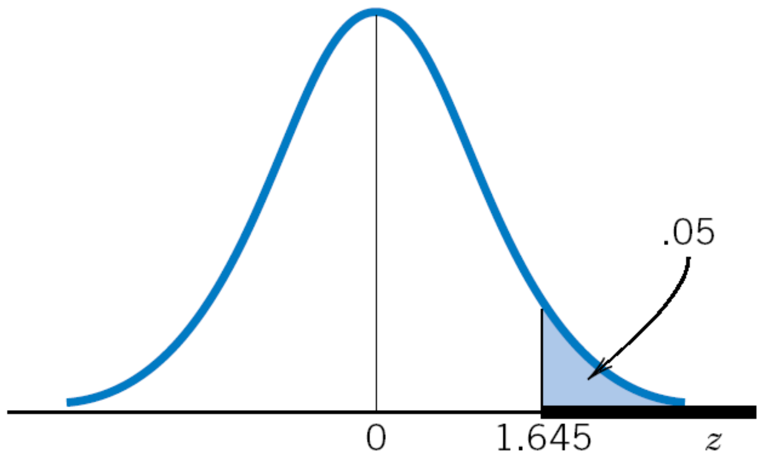
A brochure inviting subscriptions for a new diet program states that the participants are expected to lose over 22 pounds in five weeks. Suppose that, from the data of the five-week weight losses of 56 participants, the sample mean and standard deviation are found to be 23.5 and 10.2 pounds, respectively. Could the statement in the brochure be substantiated on the basis of these findings? Test with $\alpha = 0.05$. Also calculate the P-value and interpret the result.

Answer:

- Because our aim is to substantiate the assertion that $\mu > 22$ pounds, we formulate the hypotheses $H_0: \mu = 22$ versus $H_1: \mu > 22$, $n = 56$, $\bar{X}$ denote the sample mean weight loss of the 56 participants

$$z = \frac{\bar{X} - 22}{s/\sqrt{56}} = \frac{23.5 - 22}{10.2/\sqrt{56}} = 1.10$$

- Since H_1 *is right-sided*, the rejection region should be of the form $R: Z \geq c$. Because $z_{0.05} = 1.645$ the test with level of significance 0.05 has the rejection region $R: Z \geq 1.645$.



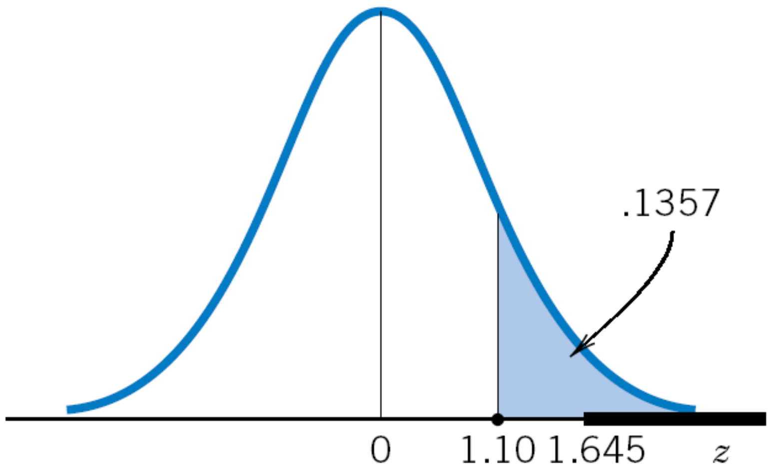
Testing hypotheses about a population mean (10)

Answer…cont'ed

- Since 1.10 is not in R, we do not reject the null hypothesis. We conclude that, with level of significance $\alpha = 0.05$, the stated claim that $\mu > 22$ is not substantiated.

- The significance probability of this result is P-value= $P[Z \geq 1.10] = 0.1357$ is the smallest α at which H_0 could be rejected.

- So we conclude that the data do not provide a strong basis for rejection of H_0.



Testing hypotheses about a population mean (11)

Large Sample Tests for μ

When the sample size is large, a **Z test** concerning μ is based on the normal test statistic

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

The rejection region is one- or two-sided depending on the alternative hypothesis. Specifically,

$$\begin{array}{lcl} H_1: \mu > \mu_0 & \text{requires} & R: Z \geq z_\alpha \\ H_1: \mu < \mu_0 & & R: Z \leq -z_\alpha \\ H_1: \mu \neq \mu_0 & & R: |Z| \geq z_{\alpha/2} \end{array}$$

Testing hypotheses about a population mean (12)

<u>One-sided</u> hypotheses versus **<u>two-sided</u>** hypothesis

$$H_0: \mu = \mu_0 \quad \text{versus} \quad H_1: \mu \neq \mu_0$$

- H_0 is to be rejected if $\bar{X}$ is too far away from μ_0 in either direction i.e. if Z is too small or too large.
- For a level α test, we divide the rejection probability α equally between the two tails and construct the rejection region

$R: Z \leq z_{\alpha/2}$ or $Z \geq z_{\alpha/2}$ which can be expressed in the more compact notation

$$R: |Z| \geq z_{\alpha/2}$$

Example: Do these data indicate that the population mean time is different from 2.6 hours?

0	0	0	0	0	0	0	1	1	1
2	2	2	2	2	3	3	3	3	4
4	4	4	5	5	5	5	5	5	5
5	5	6	6	6	8	10	15	20	25

Testing hypotheses about a population mean (13)

Answer:

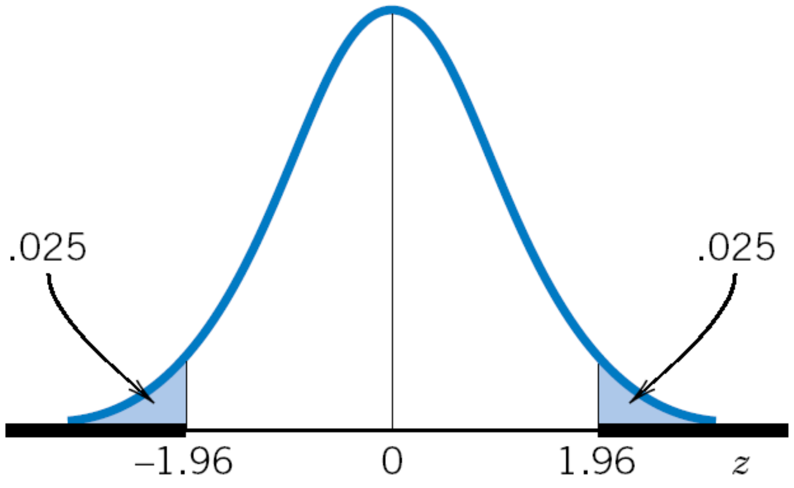
- We are seeking evidence in support of $\mu \neq \mu_0$ so the hypotheses should be formulated as

$$H_0: \mu = 2.6 \quad \text{versus} \quad H_1: \mu \neq 2.6$$

- The sample size $n = 40$ being large, we will employ the test statistic $z = \frac{\bar{X} - 2.6}{s/\sqrt{40}}$

- The two-sided form of H_1 dictates that the rejection region must also be two-sided.

- Let us choose $\alpha = 0.05$, then $\alpha/2 = 0.025$ and $z_{\alpha/2} = 1.96$. Consequently, for $\alpha = 0.05$, the rejection region is $R: |Z| \geq 1.96$



Testing hypotheses about a population mean (14)

Answer…cont'ed

- From the data, $\bar{x} = 4.55$ and $s = 5.17$, so the observed value of the test statistic is

$$z = \frac{\bar{x} - 2.6}{s/\sqrt{40}} = \frac{4.55 - 2.6}{5.17/\sqrt{40}} = 2.39$$

- Since $|Z| = 2.39$ is larger than 1.96, we reject H_0 at $\alpha = 0.05$.

- The significance probability or p-value is precisely;

$$\text{P-value} = P[|Z| \geq 2.39\,]$$
$$= P[|Z| \leq -2.39\,] + P[|Z| \geq 2.39\,]$$
$$= 2x0.0084 = \mathbf{0.0168}$$

- With α as small as 0.0168, H_0 is still rejected. This very small P–value gives strong support for H_1.