Nile University

Spring 2021

CIT651 – Introduction to Machine Learning and Statistical Data Analysis

Lec 3 - Statistics Basics

Mustafa Elattar

Statistical Analysis

- Textbook: “Applied Statistics and Probability for Engineers,” by Douglas C. Montgomery and George C. Runger, 6th edition, 2014

- Deals with Data
  - Collection
  - Presentation
  - Description
  - Analysis

- The objectives include:
  - Describe/understand a phenomenon
  - Predict outcomes

Statistical Analysis

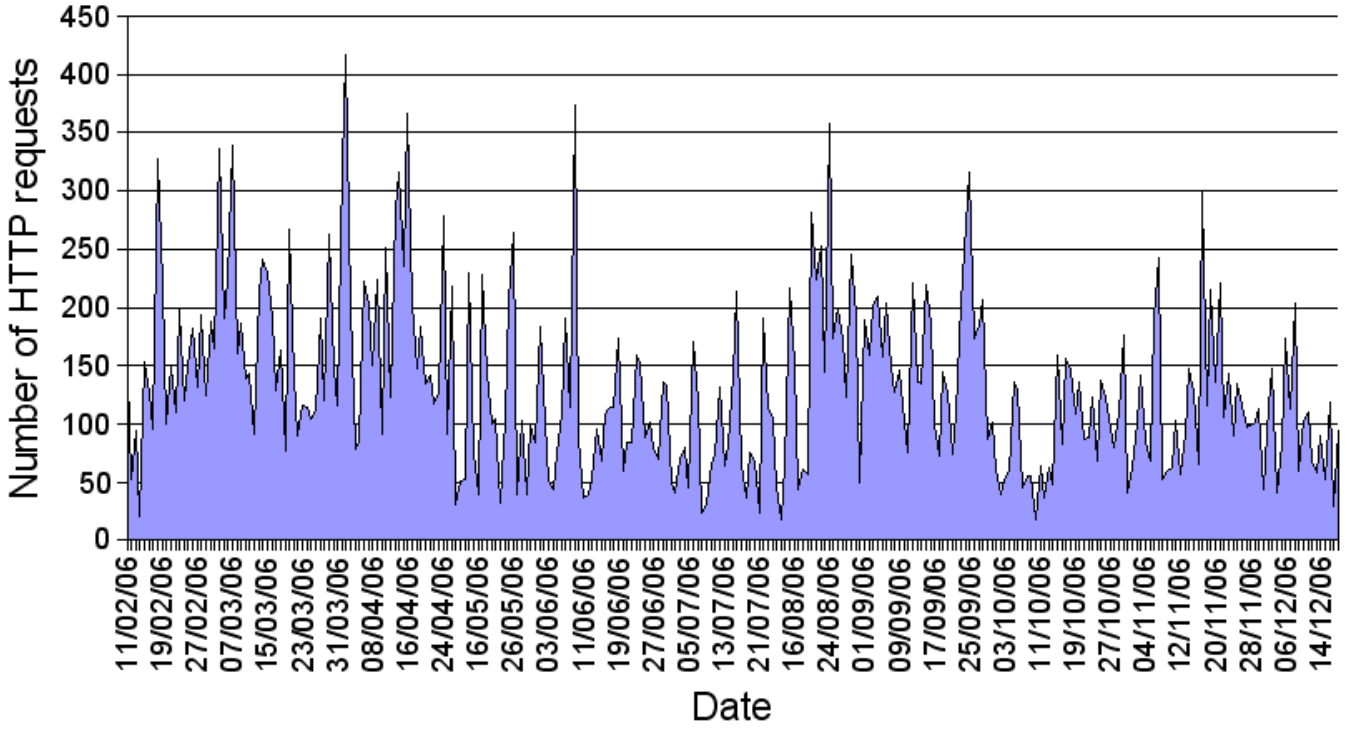
- A statistic: is a calculated numerical value that characterizes some aspect of a sample set of data

- Why statistics?. ... Randomness

- Applications:
  - Analyzing traffic patterns
  - Predicting stock market changes
  - Weather Forecasts/Analysis
  - Monitoring effectiveness of design, medication, solution, …
  - Assessment of the quality of a product, service, person, …

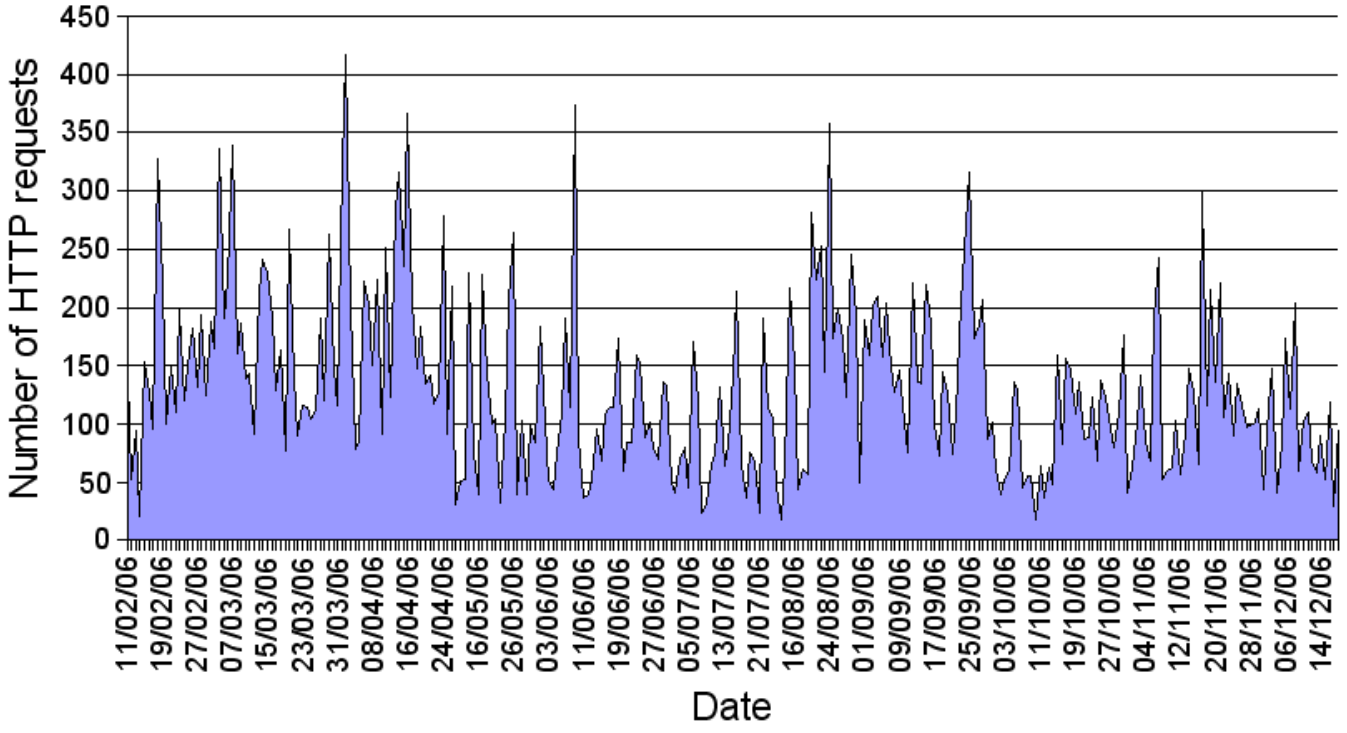
Statistical Analysis

- Given a website (www.nu.edu.eg), lets ask some direct questions:
  - What is the #visitors to the website?
  - What is the max #visitors? Min #visitors?



Statistical Analysis

- Now, let's ask more advanced questions:
  - In July, we started a marketing campaign, was it effective?
  - What is the expected number of visitors next month?
  - Compared to www.AUC.edu, which site has more visitors?
  - Which day of the week has the largest number of visitors?



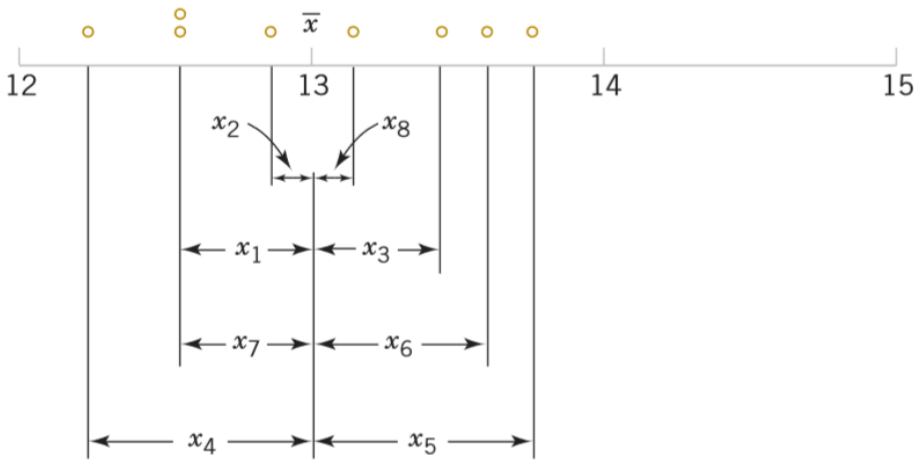
Statistical Analysis

- Statistical Data Analysis includes 3 topics:
  - Basic Probability
  - Data Collection and Description
  - Data Presentation:
    - Stem-and-Leaf Diagrams
    - Histograms
    - Box Plots
    - Time Series plots
    - Multivariate Data
  - Inference and Hypothesis Testing

Stem-and-Leaf Diagrams

- Dot Diagram: Represent each sample with a dot
- Example: Find the dot diagram for the variable x

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	–0.4	0.16
2	12.9	–0.1	0.01
3	13.4	0.4	0.16
4	12.3	–0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	–0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60



Stem-and-Leaf Diagrams

- The dot diagram is a useful data display for small samples up to about 20 observations

- However, when the number of observations is large, other graphical displays may be more useful such as Stem-and-Leaf Diagrams

- To construct Stem-and-Leaf Diagram:

  (1) Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
  (2) List the stem values in a vertical column.
  (3) Record the leaf for each observation beside its stem.
  (4) Write the units for stems and leaves on the display.

TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171				
199	151	142	163				
160	175	149	87				
196	201	200	176				

for the data

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Histograms

- The histogram is a visual display of the frequency distribution constructed as follows:

  (1) Label the bin (class interval) boundaries on a horizontal scale.
  (2) Mark and label the vertical scale with the frequencies or the relative frequencies.
  (3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

- Example:

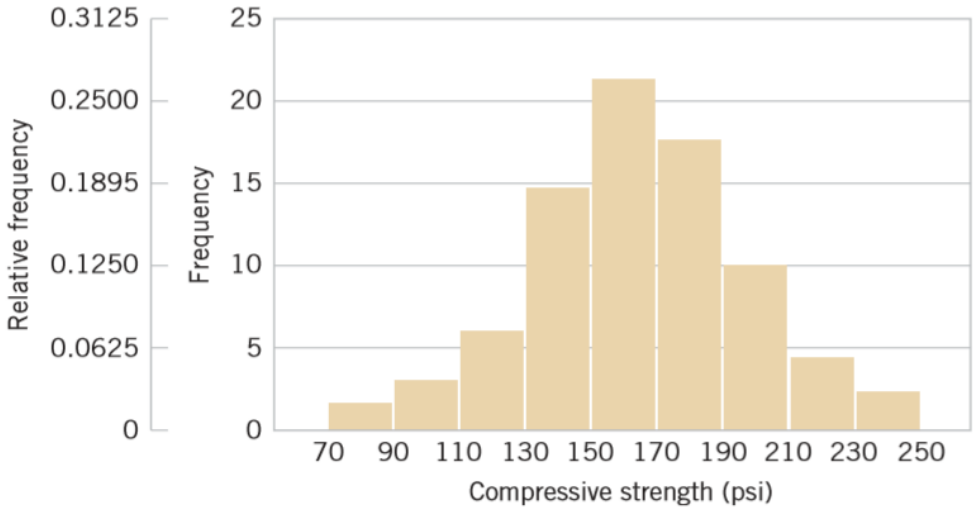
TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

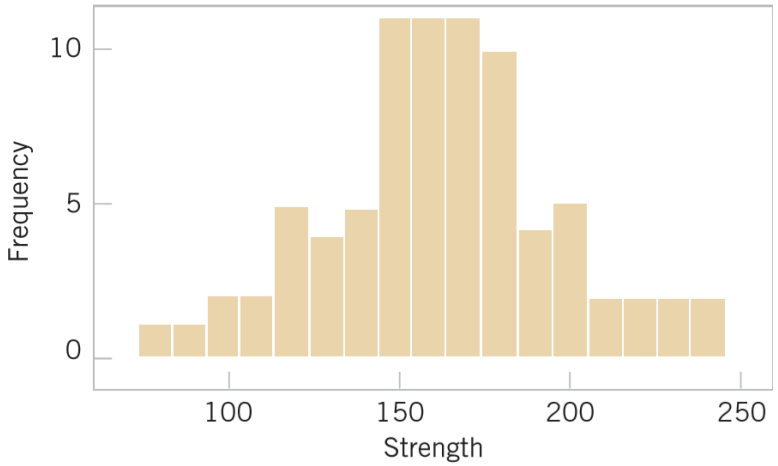
Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Histograms

- A histogram with 9 bins



- A histogram of the same data with 17 bins



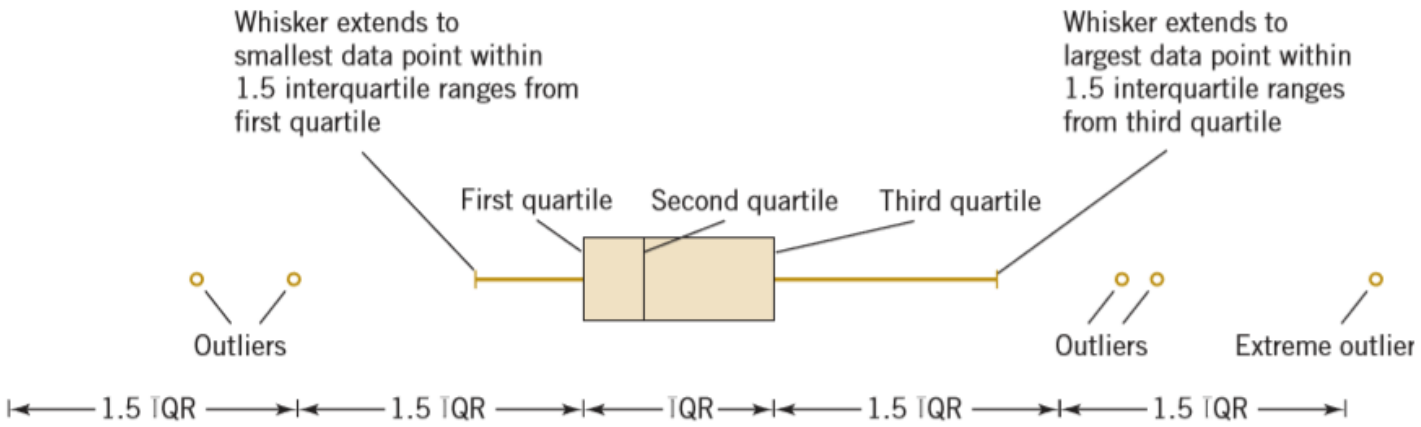
Box Plots

- The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of outliers

- First, we define the concept of dividing the data into quartiles. When an ordered set of data is divided into four equal parts, the division points are called quartiles:
  - The first or lower quartile, q_1, is a value that has approximately 25% of the observations below it and approximately 75% of the observations above
  - The second quartile, q_2, has approximately 50% of the observations below its value. The second quartile is exactly equal to the median
  - The third or upper quartile, q_3, has approximately 75% of the observations below its value

Box Plots

- A sample box plot:



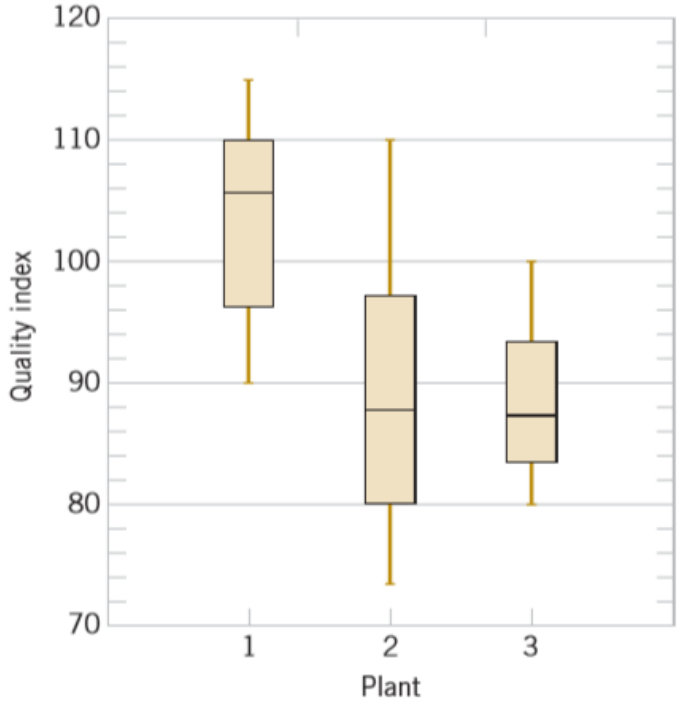
- A point beyond a whisker, but less than three interquartile ranges from the box edge, is called an outlier

- A point more than three interquartile ranges from the box edge is called an extreme outlier

Box Plots

- Box plots are very useful in graphical comparisons among data sets because they have high visual impact and are easy to understand

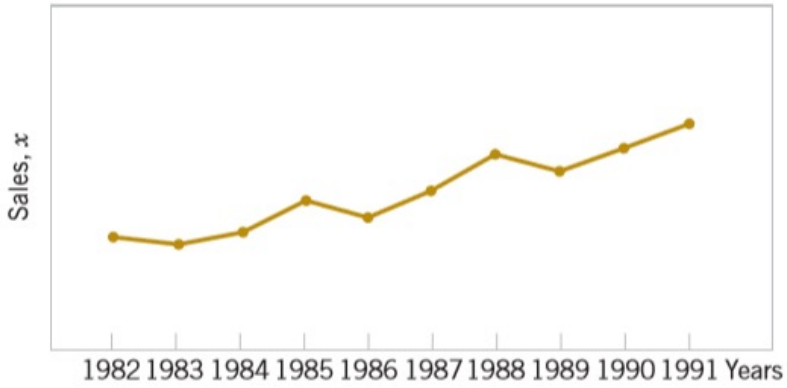
- Example:

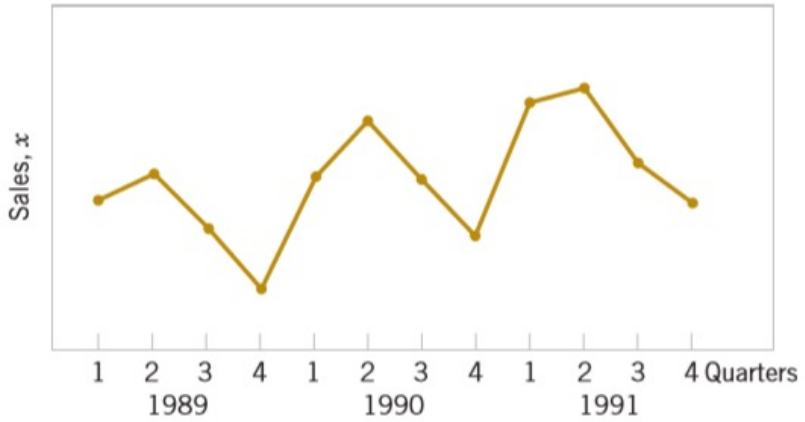


Time Sequence Plots

- A time series or time sequence is a data set in which the observations are recorded in the order in which they occur

- Examples:





Scatter Diagrams

- A scatter diagram is constructed by plotting each pair of observations with one measurement in the pair on the vertical axis of the graph and the other measurement in the pair on the horizontal axis

- Examples:

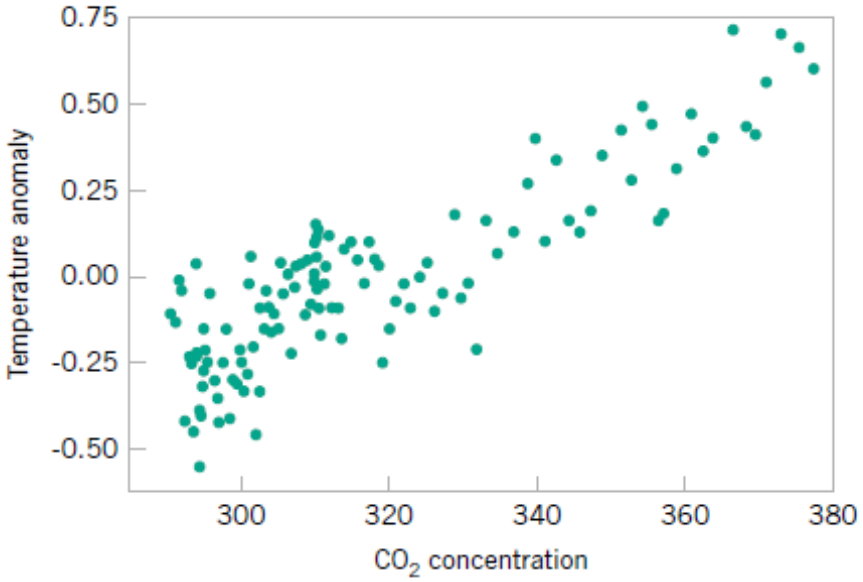
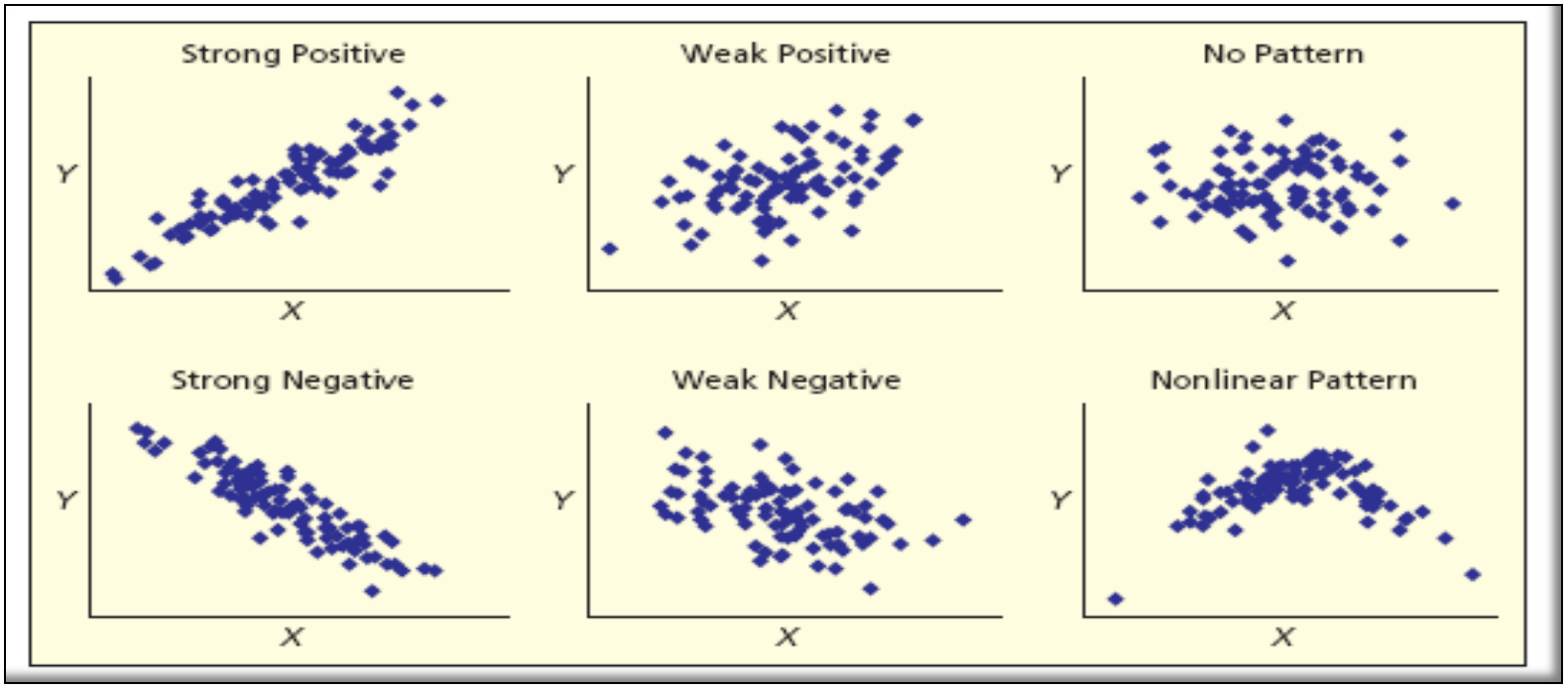


Figure 2-21 Scatter diagram of global mean air temperature anomaly versus global CO_2 concentration.

Scatter Diagrams

- Relationships between the variables can be determined from the scatter diagrams



Break

Point Estimation

- Suppose that we want to obtain a point estimate (a reasonable value) of a population parameter.

- Before the data are collected, the observations are considered to be random variables, say $X_1, X_2, \ldots, X_n$

- Therefore, any function of the observation, or any **statistic**, is also a random variable.

- Example:

  The sample mean $\bar{X}$ and the sample variance S^2 are statistics and random variables

Point Estimation

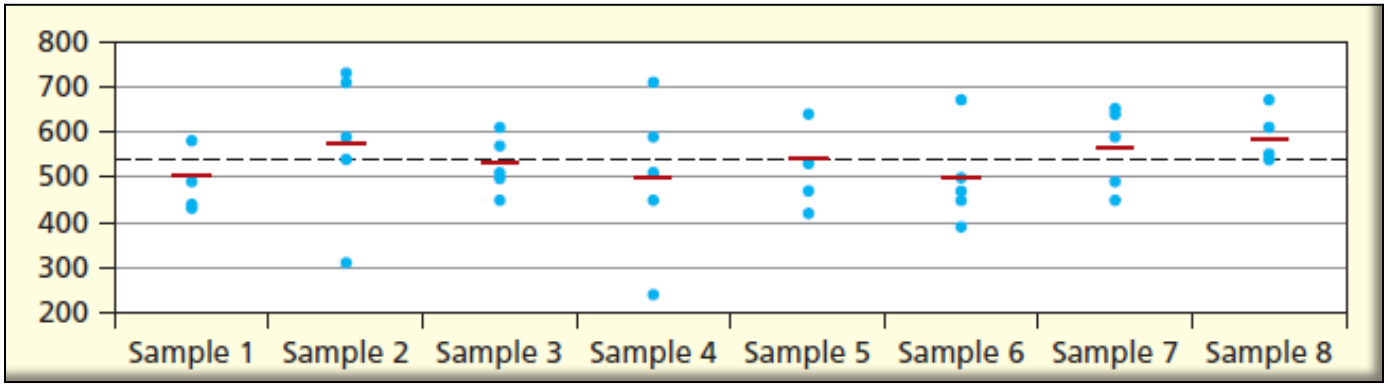
- Example: Consider eight random samples of size n = 5 from a large population (2,632) of GMAT scores for MBA applicants

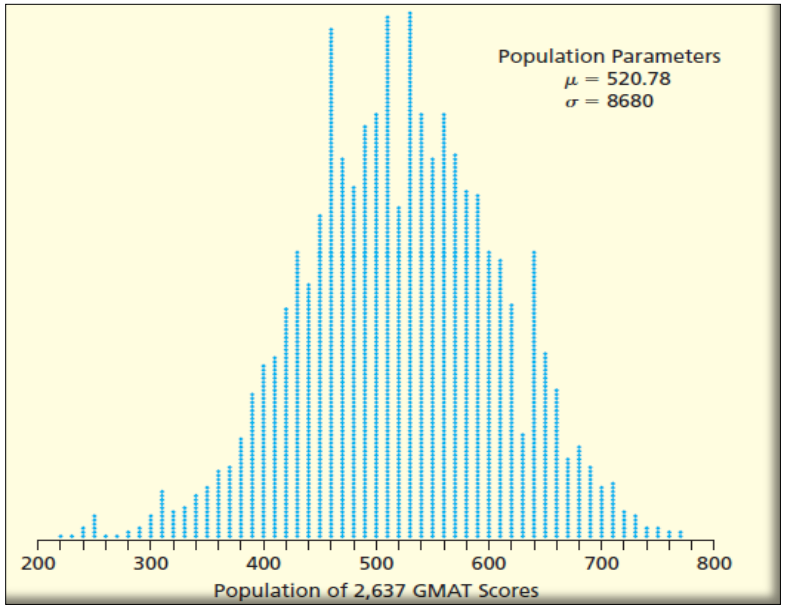
Random Samples from the GMAT Score Population							
Sample 1	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*	*Sample 6*	*Sample 7*	*Sample 8*
490	310	500	450	420	450	490	670
580	590	450	590	640	670	450	610
440	730	510	710	470	390	590	550
580	710	570	240	530	500	640	540
430	540	610	510	640	470	650	540
$\bar{x}_1 = 504.0$	$\bar{x}_2 = 576.0$	$\bar{x}_3 = 528.0$	$\bar{x}_4 = 500.0$	$\bar{x}_5 = 540.0$	$\bar{x}_6 = 496.0$	$\bar{x}_7 = 564.0$	$\bar{x}_8 = 582.0$

- Sample means tend to be close to the population mean (μ = 520.78)

Point Estimation

- The dot plot shows that the samples' means have much less variation than the individual sample items





Point Estimation

- Because a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a **sampling distribution**

- Example: Suppose that the random variable X is normally distributed with an unknown mean μ. The sample mean is a point estimator of the unknown population mean μ. Therefore, $\hat{\mu} = \bar{X}$

- After the sample has been selected, the numerical value $\bar{x}$ is the point estimate of μ. Thus, if x_1 = 25, x_2 = 30, x_3 = 29 and x_4 = 31, the point estimate of μ is

$$\bar{x} = \frac{25 + 30 + 29 + 31}{4} = 28.75$$

Central Limit Theorem

- Consider determining the sampling distribution of the sample mean $\bar{X}$. Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2

- Each observation in this sample, say, $X_1, X_2, \ldots, X_n$ is thus a normally and independently distributed random variable with mean μ and variance σ^2

- Because linear functions of independent normally distributed random variables are also normally distributed, we conclude that the sample mean

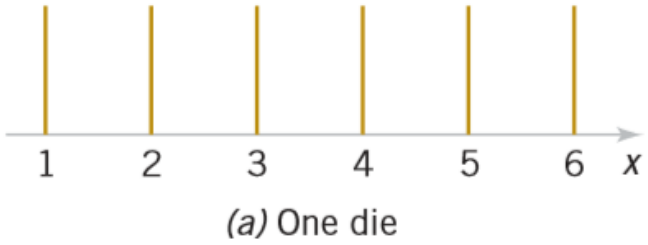
$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

has normal distribution with mean $\mu_{\bar{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$ and variance

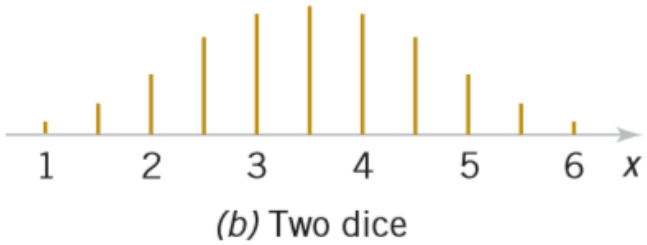
$$\sigma^2_{\bar{X}} = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n} = \sigma^2$$

Central Limit Theorem

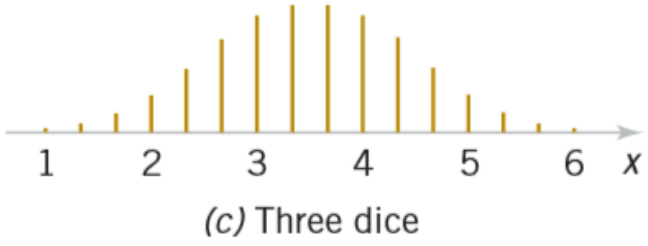
- Example: Although the distribution of rolling one die is far from normal, the distribution of average score of multiple rolls is approximated reasonably well by a normal distribution for sample sizes as small as five



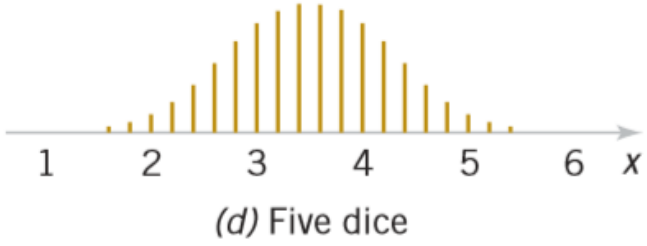
(a) One die



(b) Two dice



(c) Three dice



(d) Five dice

(e) Ten dice

Central Limit Theorem

- If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean μ and variance σ^2/n if the sample size n is large

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

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad (7\text{-}1)$$

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

- The central limit theorem is the reason why many of the random variables encountered in engineering and science are normally distributed

Gaussian Distribution Standardization

- Any Gaussian random variable can be standardized as follows

Standardizing a Normal Random Variable

If X is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma} \qquad (4\text{-}10)$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, Z is a standard normal random variable.

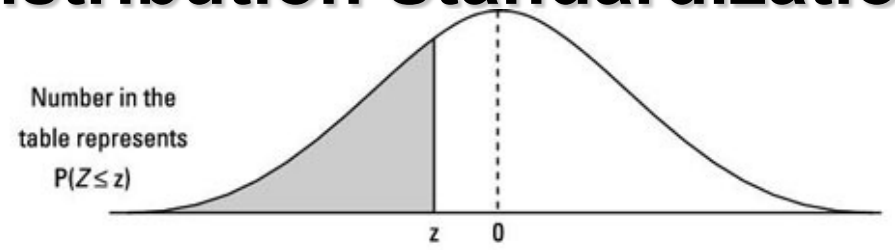
- Proof:
  - $E(Z)$ = 0 → obvious
  - $Var(Z)$ = 1:
    - Consider a random variable $Y = aX + b$. First, we prove that $Var(Y) = a^2 Var(X)$

Gaussian Distribution Standardization

$$
\begin{aligned}
Var\left(aX+b\right) &= E\left[\left\{\left(aX+b\right)-\left(a\mu+b\right)\right\}^2\right] \\
&= E\left[a^2\left(X-\mu\right)^2\right] \\
&= a^2E\left[X^2-2X\mu+\mu^2\right] \\
&= a^2\left(E\left(X^2\right)-\mu^2\right) \\
&= a^2Var\left(X\right)
\end{aligned}
$$

- If $Z=\frac{X-\mu}{\sigma}$ is compared to $Z=aX+b$, then $a=\frac{1}{\sigma}, b=\frac{-\mu}{\sigma}$
- Therefore, $Var(Z) = a^2Var(X) = 1$

Gaussian Distribution Standardization



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985

Break

Hypothesis Testing

- Many problems in engineering require that we decide which of two competing claims or statements about some parameter is true

- The statements are called hypotheses, and the decision-making procedure is called hypothesis testing

- Example: Suppose that our interest focuses on the mean height of a group of people (a parameter of the height distribution). We are interested in deciding whether or not the mean height is 160 centimeters. We may express this formally as

H_0: μ = 160 cm $\qquad\qquad$ H_1: $\mu \neq$ 160cm

Hypothesis Testing

- H_0 is the null hypothesis: This is a claim that is initially assumed to be true
- H_1 is the alternative hypothesis: It is a statement that contradicts the null hypothesis
- We cannot prove or accept a null hypothesis. We can only state that we fail to reject it or we don't have enough evidence to reject it
- A procedure leading to a decision about the null hypothesis is called a test of a hypothesis

Hypothesis Testing

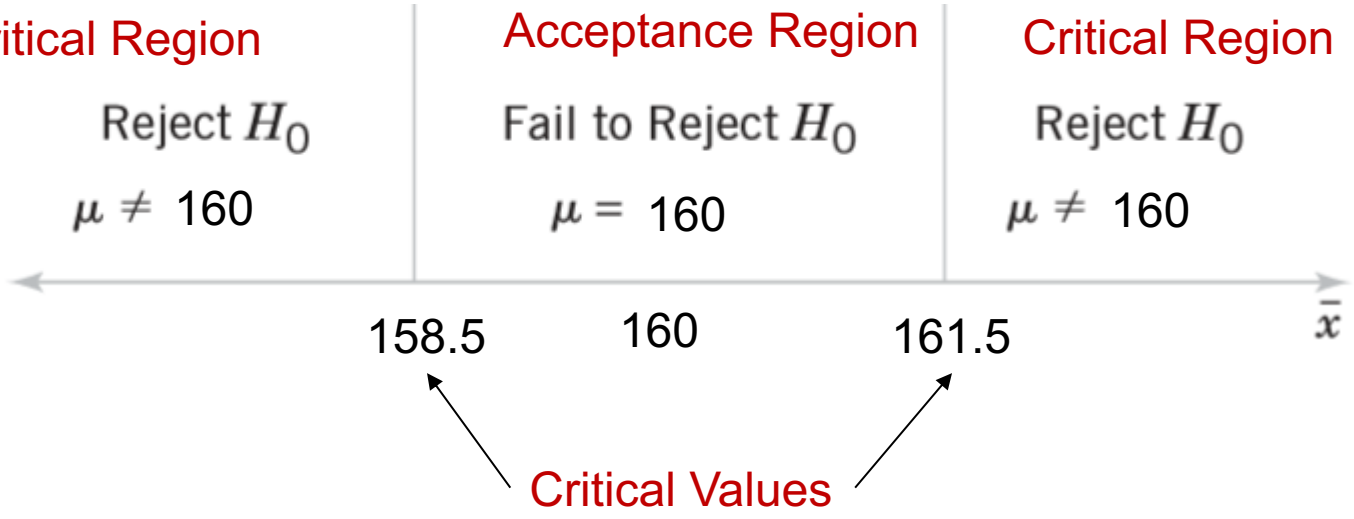
- Hypothesis testing has 7 steps:
  1. Parameter of interest: From the problem context, identify the parameter of interest
  2. Null hypothesis, H_0: State the null hypothesis, H_0
  3. Alternative hypothesis, H_1: Specify an appropriate alternative hypothesis, H_1
  4. Test statistic: Determine an appropriate test statistic
  5. Reject H_0 if: State the rejection criteria for the null hypothesis
  6. Computations: Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value
  7. Draw conclusions: Decide whether or not H_0 should be rejected and report that in the problem context

Hypothesis Testing

- Going back to the example of:

H_0: $\mu = 160$ cm H_1: $\mu \neq 160$cm

- Let $\bar{x}$ be the sample mean and assume that we will not reject the null hypothesis if $158.5 \leq \bar{x} \leq 161.5$



Hypothesis Testing Errors

- Two types of errors might occur:
  - Type I Error: Rejecting the null hypothesis H_0 when it is true
  - Type II Error: Failing to reject the null hypothesis H_0 when it is false
- Examples:
  - The true mean height is 160 cm but for the random sample, we observed $\bar{x}$ in the critical region → Type I Error
  - The true mean height is not 160 cm but for the random sample, we observed $\bar{x}$ in the acceptance region → Type II Error

Hypothesis Testing Errors

- In testing any statistical hypothesis, four different situations determine whether the final decision is correct or in error

TABLE • 9-1 Decisions in Hypothesis Testing

Decision	H_0 Is True	H_0 Is False
Fail to reject H_0	No error	Type II error
Reject H_0	Type I error	No error

- The probability of Type I error (significance level) is given by

Probability of Type I Error

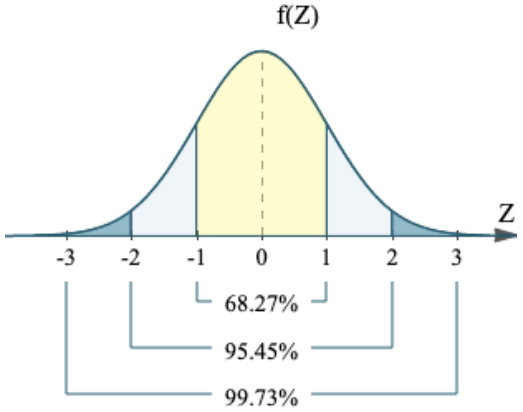
$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \qquad (9\text{-}3)$$

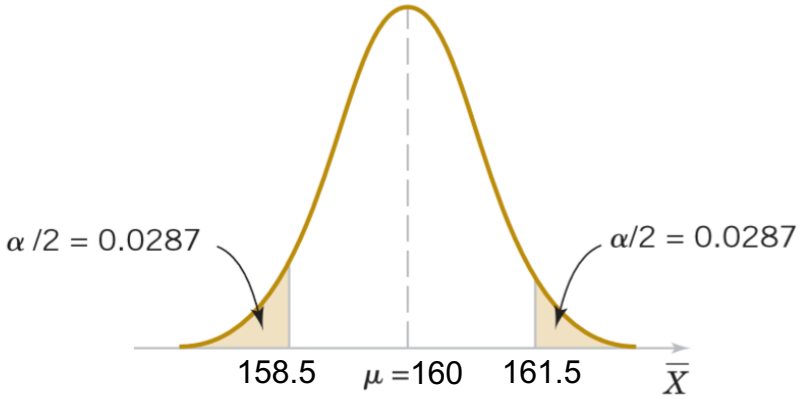
Hypothesis Testing Errors

- Example: Find α if the true mean of the height is 160 cm and the standard deviation of height is 2.5 cm and that the height has a distribution for which the conditions of the central limit theorem apply. Assume the sample size is 10

- Solution:
- Given that the conditions of central limit theorem apply, then the distribution of the sample mean is approximately normal with mean $\mu = 160$ cm and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{n}} = 0.79$

- The probability of making type I error can be found from

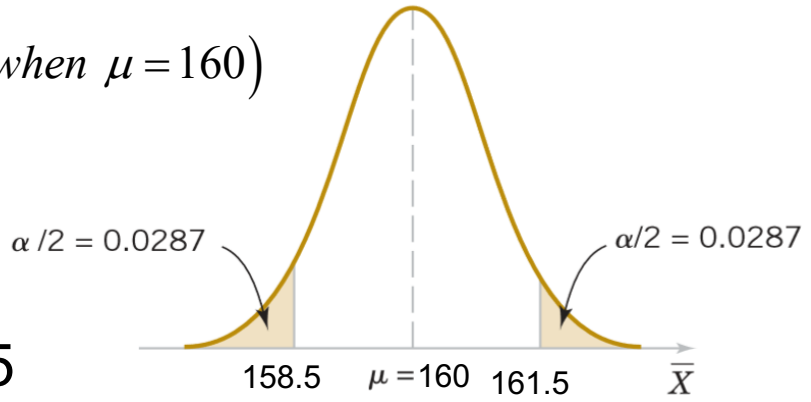




Hypothesis Testing Errors

- To find α

$$\alpha = P\left(\bar{X} < 158.5 \text{ when } \mu = 160\right) + P\left(\bar{X} > 161.5 \text{ when } \mu = 160\right)$$



- The z-values corresponding to the critical values 158.5 and 161.5 are

$$z_1 = \frac{158.5 - 160}{0.79} = -1.9 \text{ and } z_2 = \frac{161.5 - 160}{0.79} = 1.9$$

- From Gaussian distribution:

$$\alpha = P\left(z < -1.9\right) + P\left(z > 1.9\right) = 0.0287 + 0.0287 = 0.0574$$

- This implies that 5.74% of all random samples would lead to rejection of the null hypothesis

Hypothesis Testing Errors

- Because we can control the probability of making a type I error (or significance level), a logical question is what value should be used

- A widely used procedure in hypothesis testing is to use a significance level of $\alpha = 0.05$. This value has evolved through experience and may not be appropriate for all situations

Hypothesis Testing P-value

- One way to report the results of a hypothesis test is to state that the null hypothesis was or was not rejected at a specified α-value or level of significance. This is called fixed significance level testing

- Example: Knowing that H_0: $\mu = 160$ cm was rejected at the 0.05 significance level does not tell whether test statistic value was barely in or far into the critical region

- The P-value conveys much information about the weight of evidence against H_0

> The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

Hypothesis Testing P-value

- Example: Find the P-value if the observed sample mean $\overline{x}$ is 161.3 cm where the true mean of the height is 160 cm and the standard deviation of height is 2.5 cm and that the height has a distribution for which the conditions of the central limit theorem apply. Assume the sample size is 16

- Solution: Based on the definition of the P-value, the critical region for this test will be above $\overline{x}$ = 161.3 and the symmetric value 158.7

- The P-value will be

$$P-value = P\left(\bar{X} \le 158.7\right) + P\left(\bar{X} \ge 161.3\right)$$
$$= 1 - P\left(158.7 < \bar{X} < 161.3\right)$$
$$= 1 - P\left(\frac{158.7-160}{2.5/\sqrt{16}} < Z < \frac{161.3-160}{2.5/\sqrt{16}}\right) = 0.038$$

- The probability of obtaining a random sample whose mean is at least as far from 160 as 161.3 (or 158.7) is 0.038

Hypothesis Testing P-value

- Example:

Example 9-2 **Propellant Burning Rate** Air crew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 centimeters per second. We know that the standard deviation of burning rate is $\sigma = 2$ centimeters per second. The experimenter decides to specify a type I error probability or significance level of $\alpha = 0.05$ and selects a random sample of $n = 25$ and obtains a sample average burning rate of $\bar{x} = 51.3$ centimeters per second. What conclusions should be drawn?

We may solve this problem by following the seven-step procedure outlined in Section 9-1.6. This results in

1. **Parameter of interest:** The parameter of interest is μ, the mean burning rate.
2. **Null hypothesis:** H_0: $\mu = 50$ centimeters per second
3. **Alternative hypothesis:** H_1: $\mu \neq 50$ centimeters per second
4. **Test statistic:** The test statistic is

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

5. **Reject H_0 if:** Reject H_0 if the P-value is less than 0.05. To use a fixed significance level test, the boundaries of the critical region would be $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.
6. **Computations:** Because $\bar{x} = 51.3$ and $\sigma = 2$,

$$z_0 = \frac{51.3 - 50}{2 / \sqrt{25}} = 3.25$$

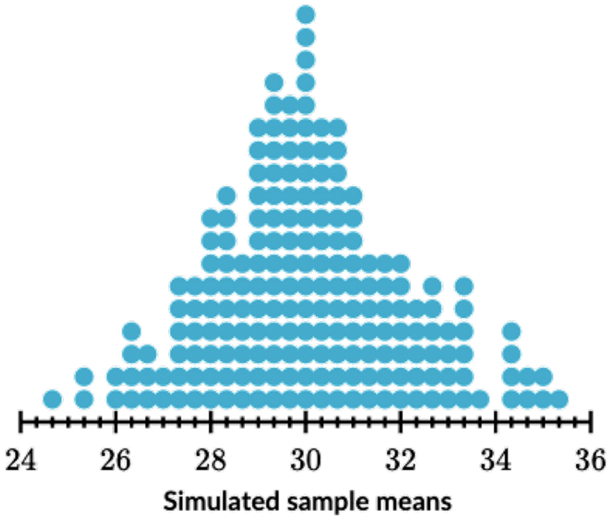
7. **Conclusion:** Because the P-value $= 2\left[1 - \Phi(3.25)\right] = 0.0012$ we reject H_0: $\mu = 50$ at the 0.05 level of significance.

Practical Interpretation: We conclude that the mean burning rate differs from 50 centimeters per second, based on a sample of 25 measurements. In fact, there is strong evidence that the mean burning rate exceeds 50 centimeters per second.

Example

The ages of workers in a certain industry are approximately normally distributed with a mean of 30 years and a standard deviation of 3.5 years. A recruiter wondered if that held true for workers in a certain state. The recruiter took a random sample of $n = 3$ of these workers from the state, and the mean age of the workers in the sample was $\bar{x} = 26$ years.

To see how likely a sample like theirs was to occur by random chance alone, the recruiter performed a simulation. They simulated 200 samples of $n = 3$ ages from a normal population with a mean of 30 years and standard deviation of 3.5 years. They recorded the mean of the ages in each sample. Here are the sample means from their 200 samples:



They want to test $H_0 : \mu = 30$ years vs. $H_a : \mu \neq 30$ where μ is the true mean age of these workers in this state.

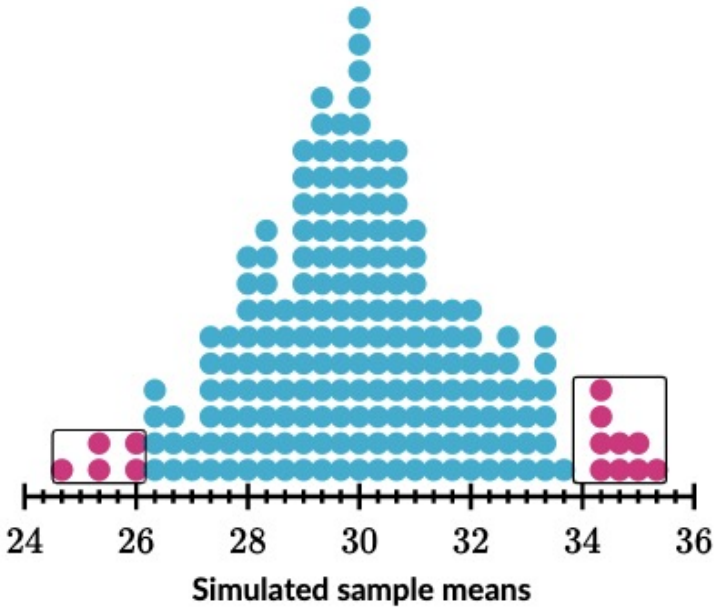
Based on these simulated results, what is the approximate p-value of the test?

Note: The sample result was $\bar{x} = 26$ years.

The $n = 3$ workers in the sample had a mean age of $\bar{x} = 26$ years.

Since the alternative hypothesis is $H_a : \mu \neq 30$ years, we can find the approximate p-value of this result by looking at how often a sample mean *as far or farther than* 26 years occurred in the simulation. We need to look for sample means this far *above or below* the hypothesized mean.

The sample mean $\bar{x} = 26$ years is 4 years below the hypothesized mean of 30 years. The simulation produced a sample mean as far or farther than this distance in 14 out of 200 samples:



$$p\text{-value} \approx \frac{14}{200} \approx 0.07$$

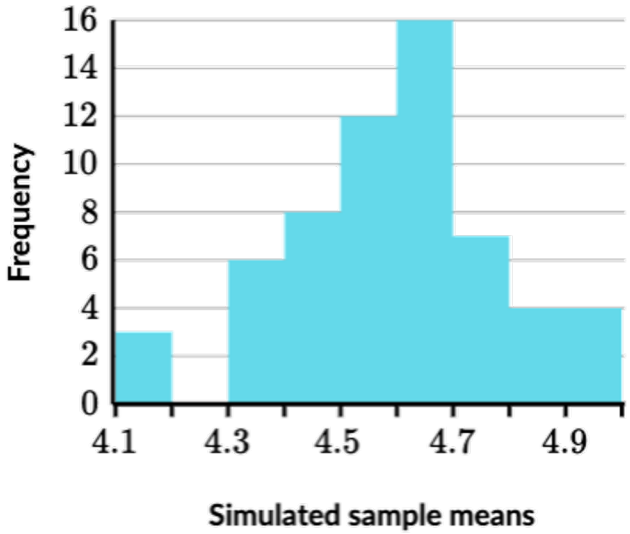
Example

A nutritionist suspected that her company's clients had below average cholesterol. They obtained a random sample of 8 clients of the same age and gender. These clients had a mean cholesterol level of $\bar{x} = 4.28\ \text{mmol/L}$ (millimoles per liter).

To see how likely a sample like this was to happen by random chance alone, the nutritionist performed a simulation. They simulated 60 samples of $n = 8$ cholesterol levels from a normal population with a mean of $4.6\ \text{mmol/L}$ and a standard deviation of $0.5\ \text{mmol/L}$ (these are generally accepted values for people with the same age and gender of those in the sample). They recorded the mean of the cholesterol levels in each sample. Here are the sample means from their 60 samples:

They want to test $H_0 : \mu = 4.6\ \text{mmol/L}$ vs. $H_a : \mu < 4.6\ \text{mmol/L}$ where μ is the mean cholesterol level for all clients like those sampled.

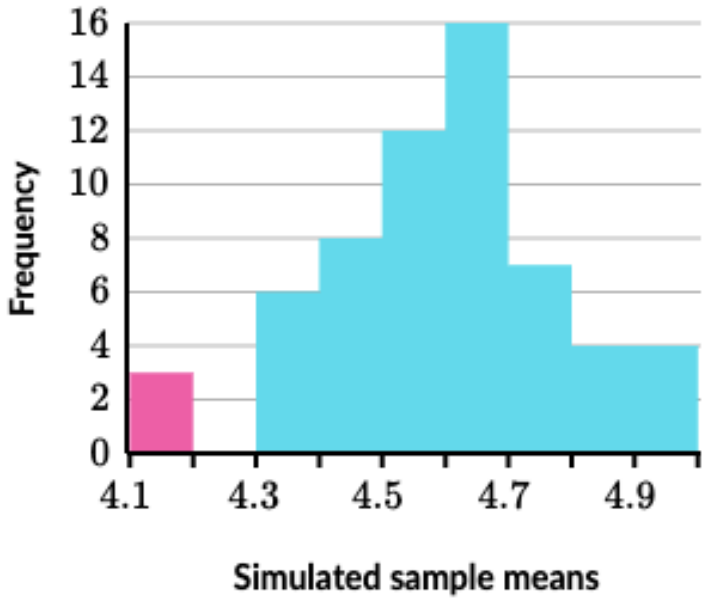
Based on these simulated results, what is the approximate p-value of the test?



The $n = 8$ clients had a mean cholesterol level of $\bar{x} = 4.28 \text{ mmol/L}$.

Since the alternative hypothesis is $H_a : \mu < 4.6 \text{ mmol/L}$, we can find the approximate p-value of this result by looking at how often a sample result *as low or lower than* $\bar{x} = 4.28 \text{ mmol/L}$ occurred in the simulation.

The simulation produced a sample mean at or below $\bar{x} = 4.28 \text{ mmol/L}$ in 3 out of 60 samples:



$$p\text{-value} \approx \frac{3}{60} \approx 0.05$$