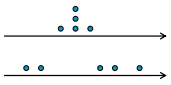
Data Description

Describe Distribution with Numbers

Example: Birth weights (in lb) of 5 babies born from two groups of women under different care programs.

Group 1: 7, 6, 8, 7, 7

Group 2: 3, 4, 8, 9, 11



Measure of Central Tendency

Describing Center

Measure of Central Tendency

Mean: the average value of the data.

If the values of a sample of n observations are denoted by $x_1, x_2, ..., x_n$, their **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

* If the data were for the whole population then the result from this calculation would be called the **population mean**, and the notation for it is μ.

Example: Birth weights (in lb) of 5 babies born from a group of women under certain diet.

7, 6, 8, 7, 7

Sol:

mean = (7 + 6 + 8 +7 + 7) / 5 = 35/5 = 7

[near the center of the data set]

Example: (number of hysterectomies performed by 15 male doctors)

27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

=> mean = **41.33**

Median: of a data set is

- the data value exactly in the middle of its ordered list if the number of pieces of data is odd,
- the mean of the two middle data values in its ordered list if the number of pieces of data is even.

[median is not influenced by outliers and is best for nonsymmetric distribution]

Example: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, 25, 25, 27, 28, 31, 33, **34**, 36, 37, 44, 50, 59, 85, 86
median = **34**

Example: (Birth weights for 6 infants.)

5, 7, 6, 8, 5, 9

ordered list => 5, 5, 6, 7, 8, 9

median = (6+7) / 2 = ***6.5***

Mode: of a data set is the observation that occurs most frequently.

Example 1: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, **25, 25**, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

Mode = **25**

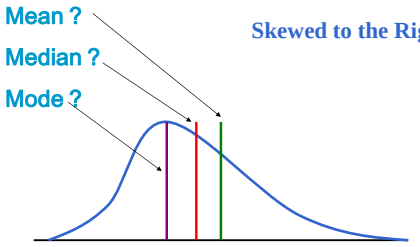
Example 2: (Blood type of 15 students)
A, B, A, A, O, AB, A, A, B, B, O, O, A, A, A

Mode = A

A – 8
B – 3
O – 3
AB – 1

What is a Modal class?

Class	Frequency	Relative Freq.	Cumulative R.F.
90< - 110	2	2/22 = .091	2/22
110< - 130	2	2/22 = .091	4/22
130< - 150	4	4/22 = .182	8/22
150< - 170	2	2/22 = .091	10/22
170< - 190	**7**	**7/22 = .318**	**17/22**
190< - 210	3	3/22 = .136	20/22
210< - 230	1	1/22 = .045	21/22
230< - 250	0	0/22 = .000	21/22
250< - 270	0	0/22 = .000	21/22
270< - 290	1	1/22 = .045	22/22
Total	**22**	**1.000**	



Midrange

- The average of the lowest and the highest values in the data set.

$$\text{Midrange} = \frac{\text{Lowest Value} + \text{Highest Value}}{2}$$

Example: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

Lowest value = 20
Highest value = 86

Midrange = (20 + 86) / 2 = **53**

Weighted Mean

Example: (Grade point average)
A student received 3 A's, 5 B's, 2 C's.

Class (grade point, x)	Frequency (weight, w)
4	3
3	5
2	2

$$\text{Average grade point} = \frac{3 \times 4 + 5 \times 3 + 2 \times 2}{3 + 5 + 2} = \frac{31}{10} = 3.1$$

weight

Weighted Mean

$$\text{Weighted mean} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_k \cdot x_k}{w_1 + w_2 + \ldots + w_k} = \frac{\Sigma\, w \cdot x}{\Sigma\, w}$$

Where $w_1, w_2, \ldots$ are the weights and $x_1, x_2, \ldots$ are the values (or class midpoint or class mark).

Mean Estimation

Class	Frequency (w)	Class Mark (x)	w · x
90 - < 110	1	100	1x100
110 - < 130	2	120	2x120
130 - <150	3	140	3x140
150 - < 170	1	160	1x160
Total	**7**		**920**

$$\text{Estimated mean} = \frac{920}{7} = \mathbf{131.43}$$

Measure of Variation

Describing Spread

Measure of Spread:

Range = largest data value – smallest data value

Sample from group I (diet program I):
7, 6, 8, 7, 7
=> mean = (7 + 6 + 8 +7 + 7) / 5 = 35/5 = 7

Sample from group II (diet program II):
3, 4, 8, 9, 11
=> mean = (3 + 4 + 8 + 9 + 11) / 5 = 35/5 = 7

Does the mother's diet program affect the birth weights of babies?

Is there any difference between the two samples?

range of sample I = 8 - 6 = **2**

range of sample II = 11 - 3 = **8**

Variance and Standard Deviation

Measure the spread of the data around the center of the data.

Example: Birth weights (in lb) of 5 babies born from a group of women under diet program II.

3, 4, 8, 9, 11 ⇒ mean = $\bar{x}$ = 7

Data Value x_i	Deviation from mean $x_i - \bar{x}$	Squared Dev. $(x_i - \bar{x})^2$
3	3 – 7 = – 4	16
4	4 – 7 = – 3	9
8	8 – 7 = 1	1
9	9 – 7 = 2	4
11	11 – 7 = 4	16
Total	0	46

Sample Variance = 46/4 = 11.5 lb.

Sample Standard Deviation = $\sqrt{46/4}$ = 3.39 lb.

If n observations are denoted by $x_1, x_2, \ldots, x_n$, their variance and standard deviation are

Sample Variance: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

(unbiased estimator for variance of an infinite population.)

Sample Standard Deviation: $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Sample Mean: $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A Short Cut formula:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{291 - \frac{35^2}{5}}{4} = 11.5$$

Data, x	x^2
3	9
4	16
8	64
9	81
11	121
35	**291**

Chapter 3 - 25

What is the sample standard deviation of the weights of babies from the sample of mothers who received diet program I?

Diet program I Data: 7, 6, 8, 7, 7

Diet I: mean = 7, s = 0.71
Diet II: mean = 7, s = 3.39

Does the mother's diet program affect the birth weights of babies?

Chapter 3 - 26

About *s* (sample standard deviation) :

- *s* measures the spread around the mean.
- the larger *s* is, the more spread out the data are.
- if *s* = 0, then all the observations must be equal.
- *s* is strongly influenced by outliers.

Chapter 3 - 27

Population Parameters

If N observations are denoted by $x_1, x_2, \ldots, x_N$, are all the observation in a finite population, their mean, μ, variance σ^2, and standard deviation, σ, are

Population Mean: $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

Population Standard Deviation: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

Chapter 3 - 28

Notation:

If for any population which their mean and variance exist, the notations for these measures are usually defined as

Population Mean: μ

Population Variance: σ^2

Population Standard Deviation: σ

These are ideal numbers. In practice, usually we don't exactly know these values and wish to estimate them.

Chapter 3 - 29

The Use of Mean and Standard Deviation

- Describe distribution
- Understand the center and the spread of the distribution

Chapter 3 - 30

Actual Length of 12 foot 2x4

	$\overline{X}$	S
My Lowe's	12.51	0.12
Homeowner Depot	12.02	0.03
Wood Lot	12.01	0.91

Profit Margin (1972-1981)

	$\overline{X}$	S
American Water Works	7.61	.68
Brown & Sharpe	7.62	7.39
Campbell Soup	13.65	1.05
McDonald's	20.04	1.02
Pam American	–.98	14.18

Measure of Relative Variability

Which of the following data has relatively lower variability?

Analyst A: (Slide A)
123, 124, 128, 133, 126, 122, 129

Analyst B: (Slide B)
9, 10, 13, 10, 12, 13, 11

Coefficient of Variation (C.V.): is the standard deviation expressed as a percentage of the mean. It is a unit-free measure of dispersion. It provides a measurement for comparing relative variability of data sets from different scales.

$$\text{C.V.} = \frac{S}{\overline{X}} \cdot 100\%$$

Example: One wishes to compare the quality of works from two blood cell count analysts. The average from repeated counts on slide A used by analyst A was **126.43** lb with a s.d.= **3.87,** and average from analyst B for slide B is **11.14** with a s.d.= **1.57**.

C.V. (Analyst A) = (3.87/126.43)x100% = **3.06%**
C.V. (Analyst B) = (1.57/11.14) x100% = **14.12%**

Analyst A has lower variability!

Chebychev's inequality

There is at least $1 - (1/k^2)$ of the data in a data set lie within k **standard deviation** of their **mean**.

Example: Heart rates for asthmatic patients in a state of respiratory arrest has a **mean of 140** beats per minute and a **standard deviation of 35.5** beats per minute. What percentage of the population of this type of patients have heart rates lie between two standard deviations of the mean in a state of respiratory arrest?

(i.e., 140-2x35.5 = **69** & 140+2x35.5 = **211**)

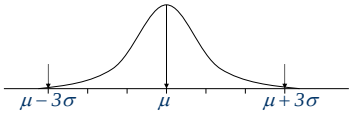
It will be at least 75%,
because, 1 – (1/**2**^{2}) = ¾ = 75%.

Chapter 3 - 37

Empirical Rule:

Properties of a symmetric and Normal distribution

- the distribution is symmetric about it mean (μ),
- 68% of the area is between $\mu - \sigma$ and $\mu + \sigma$,
- 95% of the area is between $\mu - 2\sigma$ and $\mu + 2\sigma$,
- 99.7% of the area is between $\mu - 3\sigma$ and $\mu + 3\sigma$.



Chapter 3 - 38

Approximation with E.R.

Assume that the heart rate for a particular population has a mean of 70 per minutes and standard deviation of 5.

If the heart rate for this population is bell-shaped normally distributed, what percentage of the population have heart rate between 60 to 80?

About **95%**, because it is between 2 σ's.

Chapter 3 - 39

Measure of Position

Standard Score, Quartile, Percentile

Chapter 3.a - 40

Z-score (Standard Score)

If x is an observation from a distribution that has mean μ, and standard deviation σ, the standardized value of x is,

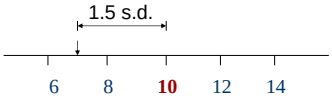
z-score of x : $$z = \frac{x - \mu}{\sigma} = \frac{x - mean}{standard\ deviation}$$

"$\mu + 3\sigma$" has a z-score 3, since it is 3 s.d. from mean.

Chapter 3 - 41

If a distribution has a mean 10 and a s.d. 2, the **value 7** has a z-score –1.5.

z-score = (7 – 10)/2 = – 1.5.



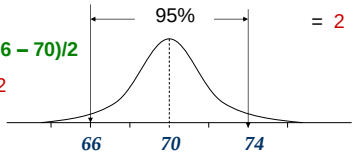
Chapter 3 - 42

Heart rates for a certain population at a certain condition follow a **bell shape symmetric** distribution with ***mean 70*** and ***standard deviation 2***.

What is the standard scores of the value **74** and the value 66?

Z_{74} = (74 – 70)/2 = 2

Z_{66} = (66 – 70)/2 = -2



Sample z-score

$$z = \frac{x - \bar{x}}{s}$$

Example: If the mean of a random sample is **5** and the standard deviation is **2**, what would be the sample z-score of the value **6**?

$\bar{x} = 5, \quad s = 2, \quad x = 6$

$$z = \frac{6-5}{2} = \frac{1}{2} = 0.5$$

Percentile

- The **percentile corresponding to a given value** X is computed by using the following formula.

$$percentile = \frac{(number\ of\ values\ below\ X) + 0.5}{total\ number\ of\ values} \cdot 100\%$$

Example: A sample of number of times visited class website by 15 students is the following: 27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28. Find the percentile of the data value 31 in this sample.

- Sol: X = 31

Ordered data:

20, 25, 25, 27, 28, **31**, 33, 34, 36, 37, 44, 50, 59, 85, 86

$$percentile = \frac{5 + 0.5}{15} \cdot 100\% = 36.67 = \mathbf{37}$$

(round to the nearest integer)

The value 31 is the 37-th percentile.

Find a Data Value Corresponding to a Given Percentile

Step 1: Sort the data.

Step 2: Compute position index c

$c = n \cdot p / 100$

n = total number of values

p = percentile (*If for 90th percentile, p = 90.*)

Step 3 (find position):

1) If c is not whole number, round **up** c to the next whole number.
2) If c is a whole number, the percentile is at the position that is halfway between c and $c + 1$.

Example: A sample of number of times visited class website by 15 students is the following: 27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28. Find the 90th percentile of the data in this sample.

- Sol: n = 15, p = 90.

Ordered data:

20, 25, 25, 27, 28, **31**, 33, 34, 36, 37, 44, 50, 59, 85, 86

$c = np/100$ = 15 x 90 / 100 = 13.5

Round c to 14. The 14th number in the ordered list is the 90th percentile and that is **85**.

Quartiles

- The first quartile, Q_1, or 25th percentile, is the median of the lower half of the list of ordered observations below the median of the data set.
- The third quartile, Q_3, or 75th percentile, is the median of the upper half of the list of ordered observations above the median of the data set.

Example: [even number of data]

6, 60,61,63,64,64,65,65,65,66,67,69,71,71,71,72,72,72,72,73,74,75

Q_1 = **64** **Median** = **68** Q_3 = **72**

Measure of spread:

Interquartile range (IQR) = $Q_3 - Q_1$

IQR = 72 - 64 = 8

Example: [odd number of data values]

60,61,63,64,64,65,65,65,66,67,**69**,71,71,71,72,72,72,72,73,74,75

Q_1 = **64.5** **Median** = **69** Q_3 = **72**

Measure of spread:

Interquartile range (IQR) = $Q_3 - Q_1$

IQR = 72 – 64.5 = 7.5

Exploratory Data Analysis

Stemplot and Boxplot

Stemplots (or Stem-and-leaf plots)

- -- leading digits are called stems
- -- final digits are called leaves

Example:
(number of hysterectomies performed by 15 male doctors)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

→

2 | 05578
3 | 13467
4 | 4
5 | 09
6 |
7 |
8 | 56

Example:
Number of hysterectomies performed by **15 male doctors**:
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

by **10 female doctors**, the numbers are:
5, 7, 10, 14, 18, 19, 25, 29, 31, 33

	(Male)	(Female)
	2 | 05578	0 | 57
	3 | 13467	1 | 0489
→	4 | 4	2 | 59
	5 | 09	3 | 13
	6 |	
	7 |	
	8 | 56	

Back-to-back stem-plot

(Female)	(Male)
75 |	0 |
9840 |	1 |
95 |	2 | 05578
31 |	3 | 13467
|	4 | 4
|	5 | 09
|	6 |
|	7 |
|	8 | 56

Example: (Height data with gender)
Female: 60, 63, 64, 65, 65, 65, 66, 67
Male: 61, 64, 69, 71, 71, 71, 72, 72, 72, 72, 73, 74, 75
(See data sheet)

Back-to-back

Female		Male
555430	6 |	149
	7 |	1112222345

Split-back-to-back

Female		Male	
430	6* |	14	* => 0 - 4
76555	6# |	9	# => 5 - 9
	7* |	111222234	
	7# |	5	

The five-number summary

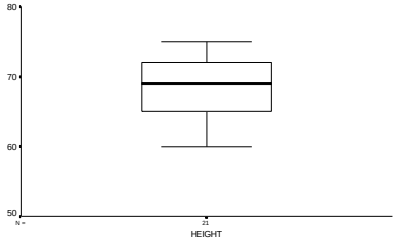
- .Minimum value
- .Q_1
- .Median
- .Q_3
- .Maximum value

Boxplot

Example: (data sheet without outlier "6")

60,61,63,64,64,65,65,65,66,67,**69**,71,71,71,72,72,72,72,73,74,75

Min = 60, Q_1 = 64.5, Median = 69, Q_3 = 72, Max = 75.



Outliers

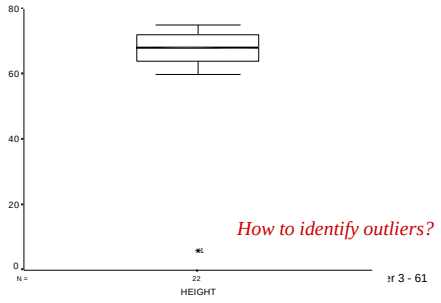
- The extremely high or extremely low data value when compared with the rest of the data values.

With 6 in the data:

6, 60,61,63,64,64,65,65,65,66,67,69,71,71,71,72,72,72,72,73,74,75

Q_1 = 64 **Median = 68** **Q_3 = 72**

IQR = 72 - 64 = 8



How to identify outliers?

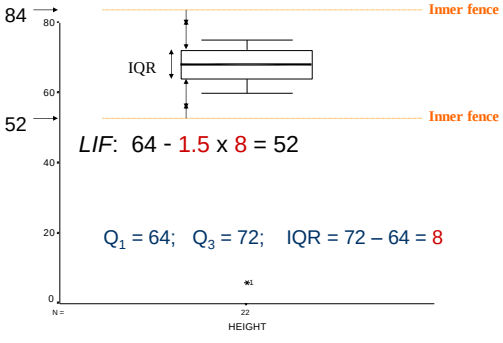
Inner and outer fences for outliers

- The inner fences are located at
  a distance of 1.5 **IQR** below Q_1
  (*lower* inner fence = Q_1 - 1.5 x IQR)
  and at a distance of 1.5 **IQR** above Q_3
  (*upper* inner fence = Q_3 + 1.5 x IQR).
- The outer fences are located at
  a distance of 3 **IQR** below Q_1
  (*lower* outer fence = Q_1 – 3 x IQR)
  and at a distance of 3 **IQR** above Q_3
  (*upper* outer fence = Q_3 + 3 x IQR) .

IQR = 72 – 64 = 8; Q_1 = 64; Q_3 = 72

- The inner fences are located at
  a distance of 1.5 **IQR** below Q_1
  (*lower* inner fence = 64 - 1.5 x 8 = 52)
  and at a distance of 1.5 **IQR** above Q_3
  (*upper* inner fence = 72 + 1.5 x 8 = 84).
- The outer fences are located at
  a distance of 3 **IQR** below Q_1
  (*lower* outer fence = 64 – 3 x 8 = 40)
  and at a distance of 3 **IQR** above Q_3
  (*upper* outer fence = 72 + 3 x 8 = 96) .

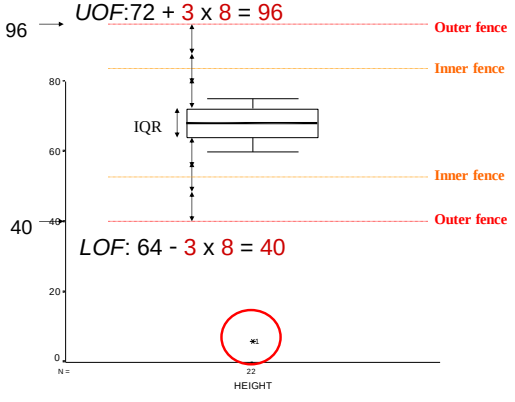
UIF: 72 + 1.5 x 8 = 84



LIF: 64 - 1.5 x 8 = 52

Q_1 = 64; Q_3 = 72; IQR = 72 – 64 = 8

UOF:72 + 3 x 8 = 96



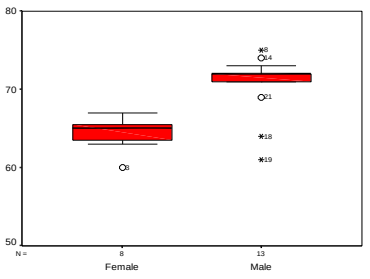
LOF: 64 - 3 x 8 = 40

Mild and Extreme outliers

- Data values falling between the inner and outer fences are considered **mild outliers**.
- Data values falling outside the outer fences are considered **extreme outliers**.

When outliers exist, the whisker extended to the smallest and largest data values *within the inner fence.*

Side-by-side Box Plot



**

3 - 67

Remarks:

- If the distribution of the data is symmetric, then the ***mean*** and ***median*** will be about the same.
- The five-number summary and boxplot are best for nonsymmetric data.
- The ***median*** and ***quartiles*** are not influenced by outliers.
- The ***mean*** and ***standard deviation*** are most appropriate to use only if the data are symmetric because both of these measures are easily influenced by outliers.

Chapter 3 - 68

Boxplot

For the following data:

13 72 78 40 50 56 50 52 57
69 130 142 51 52

Find the **five-number-summary & IRQ**

Make a **boxplot.**

Chapter 3 - 69