MATH 353

PROBABILITY AND STATISTICS

UNIT 1

Alexander Boateng, PhD

Department of Statistics and Actuarial Sciences, KNUST

UNIT 1: OUTLINE

1. Introduction to Statistics
2. Measures of Central Tendency
3. Measures of Dispersion
4. Measures of Position
5. Exploratory Data Analysis
6. Summary Statistics

UNIT OBJECTIVES:

After completing this unit, you should be able to:

- Summarize data, using measures of central tendency such as **Mean**, **Median**, **Mode and Mid-range**.
- Describe data, using measures of dispersion or variation such as **Range**, **Variance and Standard Deviation.**
- Identify the position of values in a dataset, using various measures of position, such as **Percentiles, Deciles and Quartiles.**

INTRODUCTION TO STATISTICS

Definition 1:

Statistics is a science that helps to make decisions and draw conclusions in the presence of variability.

For example: Civil engineers working in the transportation industry.

Definition 2:

Statistics is the science concerned with developing and studying methods of collecting, analyzing, interpreting and presenting empirical data.

ROLE OF STATISTICS

<u>The roles of statistics are to:</u>

1. Collect, summarize, and present data:

   Descriptive Statistics → Record Keeping

2. Analyze, and make an inference about the population that produced the data:

   Inferential Statistics → Prediction with probability.

DEFINITION OF TERMINOLOGIES

- **Population:-** Is the entire collection of objects or outcomes about which data is collected.
- The entire members of outcomes or objects of one's interest can be either **finite or infinite** to analyst at the start of analysis. **Example:**
    - Stocks data (finite)
    - Manufacturing of automobiles (infinite)
- **Sample:-** Is a subset of the population containing the observed objects or the outcomes and the resulting data. **Example:** 10 stocks chosen from 100 stocks.

DEFINITION OF TERMINOLOGIES cont'd

- **Parameter:-** Is the numerical measure of a population characteristic. → the population mean (= average), the population variance, the population deviation.

  It is denoted by: μ , σ^2, σ , β etc.

- **Statistic:-** Is the numerical measure of a sample characteristic. → the sample mean (= average), the sample variance, the sample deviation.

  It is denoted by: $\bar{X}, s^2, s$ etc.

PARAMETRIC STATISTICS vs. NONPARAMETRIC STATISTICS

The study of statistics can be divided into two (2).

That is ***Parametric Statistics*** and ***Nonparametric Statistics***

- **Parametric Statistics:-** Is a branch of statistics which assumes that the sample data comes from a population that can be adequately modeled by a probability distribution that has fixed set of parameters.

- **Nonparametric Statistics:-** Refers to statistical method in which the data are not assume to come from prescribed models that are determined by a small number of parameters. It is based on distribution-free.

STATISTICAL DESCRIPTIVE MEASURES

Statistical Measures can be grouped into two namely:

- <u>**Measures of Central Tendency**</u>
- <u>**Measures of Dispersion (Variation)**</u>

Measures of Central Tendency	Measures of Dispersion (Variation)
1. Mean	1. Range
2. Median	2. Variance
3. Mode	3. Standard Deviation
4. Weighted Mean	4. Skewness
5. Harmonic Mean	6. Kurtosis
	7. Quartiles
	8. Percentiles
	9. Mean Deviation

ARITHMETIC MEAN

- $Mean = Expected\,Value = Average\;(\bar{X})$
- $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

NB: $\sum = sigma = total$

- Mean from Frequency Distribution Tables is given as:
- $\bar{X} = \frac{\sum fx}{\sum f}$

X	x_1	x_2	x_3	$\cdots$	x_n
f	f_1	f_2	f_3	$\cdots$	f_n

Example 1.0

- Find the mean of the following set of numbers: $148, 153, 156, 157, 160.$

Solution 1.0

- $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- $\bar{X} = \frac{148+153+156+157+160}{5}$
- $\bar{X} = 154.8$

Example 1.1

- If the average of the values, $9, 6, 7, 8, 5$ and x is 8. Find the value of x.

Solution 1.1

- $\bar{X} = 8 = \frac{9+6+7+8+5+x}{6}$
- $\bar{X} = 8 = \frac{35+x}{6}$

$8 \times 6 = 35 + x$

$48 = 35 + x$

$x = 48 - 35$

$\therefore x = 13.$

Example 1.2

- A sample of 100 boxes of matches was taken and a record made of the number of matches per box. The results were as follows.

Number of Matches / box (x)	47	48	49	50	51
Frequency (f)	4	20	35	24	17

Solution 1.2

x	f	fx
47	4	188
48	20	960
49	35	1715
50	24	1200
51	17	867
	$\sum f = 100$	$\sum fx = 4930$

- $\bar{X} = \dfrac{\sum fx}{\sum f}$

  $= \dfrac{4930}{100}$

  $= 49.30$

MEAN FROM GROUPED DATA
Example 1.3

- Telephone calls arriving at switchboard are answered by the telephonist. The following table shows the time, to the nearest second, recorded as being taken by the telephonsit to answer the calls received during one day.

Time to Answer (to nearest second)	Number of Calls
10-19	20
20-24	20
25-29	15
30	14
31-34	16
35-39	10
40-59	10

- Calculate an estimate of the mean time taken to answer the calls.

Solution 1.3

- Class Midpoint $= \frac{Lower\ limit + Upper\ limit}{2}$
- Class Midpoints:
- $\frac{10+19}{2}, \frac{20+24}{2}, \frac{25+29}{2}, 30, \frac{31+34}{2}, \frac{35+39}{2}, \frac{40+59}{2}$
- $mean\ (\bar{X}) = \frac{\sum fx}{\sum f}$

$$= \frac{2940}{105}$$

$$= 28$$

Class Midpoint (x)	Number of Calls (x)	fx
14.5	20	290
22	20	440
27	15	405
30	14	420
32.5	16	520
37	10	370
49.5	10	495
	$\sum f = 105$	$\sum fx = 2940$

$\therefore$ The mean time is 28 seconds.

Example 1.4

- The marks obtained by 40 students in an examination is shown below.

63	76	87	61	78	85	77	87	74	77
80	77	74	88	72	78	79	89	85	90
77	70	81	69	75	78	73	86	83	91
69	96	65	88	84	74	84	81	83	75

Example 1.4 cont'd.

Class Boundaries	Tally	Frequency (f)	Class Midpoint
59.5-64.5			
64.5-69.5			
69.5-74.5			
74.5-79.5			
79.5-84.5			
84.5-89.5			
89.5-94.5			
94.5-99.5			

- Copy and complete the table above.
- Find the mean mark of the students.

Solution 1.4

- $mean\ (\bar{X}) = \frac{\sum fx}{\sum f}$

$= \frac{3155}{40}$

$= 78.875$

Class Boundaries	Tally	Frequency (f)	Class Midpoint (x)	fx
59.5-64.5	||	2	62	124
64.5-69.5	|||	3	67	201
69.5-74.5	~~||||~~ |	6	72	432
74.5-79.5	~~||||~~ ~~||||~~ |	11	77	847
79.5-84.5	~~||||~~ ||	7	82	574
84.5-89.5	~~||||~~ |||	8	87	696
89.5-94.5	||	2	92	184
94.5-99.5	|	1	97	97
		$\sum f = 40$		$\sum fx = 3155$

WEIGHTED MEAN

- In some occasions, calculating the ordinary mean is not suitable.
- In these occasions, you wish to place greater emphasis on some values of the data.
- In general, if $x_1, x_2, x_3, \cdots, x_n$ are given with respective weighting $w_1, w_2, w_3, \cdots, w_n$, then
- $\text{Weighted Mean} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}$

$$= \frac{\sum w_i x_i}{\sum w_i}$$

Example 1.5

- Find the weighted mean of the numbers 8 and 12 if they are given the weights 2 and 3 respectively.

Solution 1.5

- Weighted Mean $= \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2}$

$$= \frac{8 \times 2 + 12 \times 3}{2+3}$$

$$= \frac{16+36}{5}$$

$$= 10.4$$

Example 1.6

- The weighted mean of the two numbers 30 and 15 is 20. If the weightings are 2 and x respectively, find the value of x.

Solution 1.6

- Weighted Mean $= \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2}$

$20 = \frac{30 \times 2 + 15 \times x}{2 + x}$

$20(2 + x) = 60 + 15x$

$20x - 15x = 60 - 40$

$5x = 20$

$x = \frac{20}{5}$

$x = 4$

HARMONIC MEAN

- **Harmonic Mean:-** Is calculated by dividing the number of values in the set by sum of the inverse of the values in the set. That is:
- $\text{HM} = \dfrac{Number\ of\ Values}{\left(\frac{1}{Value_1}\right)+\left(\frac{1}{Value_2}\right)+\cdots+\left(\frac{1}{Value_n}\right)}$
- $\text{Weighted Harmonic Mean} \ = \dfrac{\sum w}{\sum \frac{w}{a}}$
- where w is the weight.

Example 1.7

- Calculate the harmonic mean of x where $x = \{4, 15, 17, 5, 22\}$

Solution 1.7

- The number steps of values is 6. This is numerator.
- The sum of the inverses is $\frac{1}{4} + \frac{1}{15} + \frac{1}{17} + \frac{1}{5} + \frac{1}{22} = 0.621.$

  This is denominator.
- The harmonic mean is $\frac{6}{0.621} = 9.66$

Example 1.8

- Calculate the harmonic mean for the following:

x	1	3	5	7	9	11
f	2	4	6	8	10	12

Solution 1.8

- Calculation of Harmonic Mean is shown below:

x	f	$\frac{1}{x}$	$\frac{f}{x}$
1	2	1.000	2.000
3	4	0.333	1.332
5	6	0.200	1.200
7	8	0.143	1.144
9	10	0.111	1.111
11	12	0.091	1.092
	$\sum f = 42$		$\sum \frac{f}{x} = 7.897$

- $\text{HM} = \dfrac{\Sigma f}{\Sigma \frac{f}{x}}$

$$= \frac{42}{7.897}$$

$$= 5.331$$

∴ Harmonic mean is 5.331.

Example 1.9

- Kofi decided to take a quick tour around town on her bike, if the information below describes the speeds he travelled over each equal-length segment, what was his average speed for the trip?

Segment #1	8mph	3miles
Segment #2	12mph	3miles
Segment #3	14mph	3miles
Segment #4	7mph	3miles

Solution 1.9

- Sum distance = 3miles + 3miles + 3miles + 3miles = 12 miles
- Calculation of times:
- Segment 1: $\frac{3}{8} = 0.375hr$
- Segment 2: $\frac{3}{12} = 0.25hr$
- Segment 3: $\frac{3}{14} = 0.214hr$
- Segment 4: $\frac{3}{7} = 0.429hr$
- Total time = 0.375 + 0.25 + 0.214 + 0.429 = 1.268 hrs
- Average rate $= \frac{Total\ distance}{Total\ time} = \frac{12}{1.268} = 9.46mph$

GEOMETRIC MEAN

- Geometric Mean:- Is a method of finding the “middle” value in a set that contains some values that are intrinsically more influential than others. It is calculated by raising the product of a series of numbers to the inverse of the total length of the series.
- Computation:
- It is the nth root of the product of n numbers. That is for a set of numbers: $x_1, x_2, \cdots, x_n$.
- $GM = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}} = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} = (x_1 \cdot x_2 \cdots x_n)^{\frac{1}{n}}$
- $GM = Antilog\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right)$

Example 1.10

- What is the geometric mean of 2, 3, and 6?

Solution 1.10

- $GM = (x_1 \cdot x_2 \cdots x_n)^{\frac{1}{n}}$
  $= (2 \times 3 \times 6)^{\frac{1}{3}}$
  $= 3.30$

Example 1.11

- Ama's monthly salary jumped from GHC2,500 to GHC5,000 over the course of ten years.

Calculate average yearly increase using geometric mean.

Solution 1.11

- Find the geometric mean = $[2500 \times 5000]^{\frac{1}{2}} = 3{,}535.533$
- Average increase over ten years = $\frac{3535.533}{10} = 353.55$

Example 1.12

- Calculate the geometric mean of the annual percentage growth rate of profit in a business corporate from the year 2010 to 2015 if profits for these periods are:

50, **72**, **54**, **82**, **93.**

Solution 1.12

x_i	50	72	54	82	93	Total
$\log x_i$	1.6990	1.8573	1.7324	1.9138	1.9685	9.1710

Solution 1.12 cont'd.

- $GM = Antilog\left(\frac{\sum_{i=1}^{n} \log x_i}{n}\right)$

  $= Antilog\left(\frac{9.1710}{5}\right)$

  $= Antilog(1.8342)$

  $= 68.26$

- Geometrical mean of the annual percentage growth rate of profit is 68.26.

Alternatively:

- $GM = \sqrt[5]{50 \times 72 \times 54 \times 82 \times 93} = 68.265$

Example 1.13

- Find the geometric mean for the following data, which gives the defective screws obtained in a factory.

Diameter (cm)	5	15	25	35
Number of Defective Screws	5	8	3	4

Solution 1.13

- $GM = Antilog\left(\frac{\sum_{i=1}^{n} f_i \log x_i}{N}\right)$

$$= Antilog\left(\frac{23.2739}{20}\right)$$

$$= 14.58$$

x_i	f_i	$\log x_i$	$f_i \log x_i$
5	5	0.6990	3.4950
15	8	1.1761	9.4088
25	3	1.3979	4.1937
35	4	1.5441	6.1764
	$\sum f_i = N = 20$		$\sum f_i \log x_i = 23.2739$

WEIGHTED GEOMETRIC MEAN

- If the various observations $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$ are not of equal importance in the data, weighted geometric mean is calculated.

- Weighted GN of the observations $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}$ with respective weights as $\boldsymbol{w_1}, \boldsymbol{w_2}, \cdots, \boldsymbol{w_n}$ is given by:

- $GM = antilog\left[\frac{\sum w_i \log x_i}{\sum w_i}\right]$

- Weighted geometric mean observations is equal to the antilog of weighted arithmetic mean of their logarithms.

Example 1.14

- Calculate weighted geometric mean of the following data.

Diameter (cm)	5	15	25	35
Number of Defective Screws	5	8	3	4

Solution 1.14

- $GM = antilog\left[\frac{\sum w_i \log x_i}{\sum w_i}\right]$

$= antilog\left(\frac{27.1554}{25}\right)$

$= antilog(1.0862)$

$= 12.20$

x_i	w_i	$\log x_i$	$w_i \log x_i$
5	10	0.6990	6.9900
8	9	0.9031	8.1278
44	3	1.6435	4.9304
160	2	2.2041	4.4082
500	1	2.6990	2.6990
	$\sum w_i = 25$		$\sum w_i \log x_i = 27.1554$

MEDIAN

Example 1.15

- Find the Median of $\mathbf{8, 5, 7, 10, 15, 21.}$

Solution 1.15

- Arrange the values in the ascending order: $\mathbf{5, 7, 8, 10, 15, 21}$.
- The Median falls into 8 and 10.
- $\therefore$ The median = $\frac{8+10}{2} = 9$.

Example 1.16

- Find the median value from the set of values: $\mathbf{17, 15, 9, 13, 21, 7, 32}$.

Solution 1.16

- Arrange the values in the ascending order: $\mathbf{7, 9, 13, 15, 17, 21, 32.}$
- The median is 15.

Example 1.17

- The table gives the heights, measured to the nearest metre of 300 trees.

Height	2	3	4	5	6	7	8	9
Number of Trees	14	21	42	83	118	12	7	3

- Find the Median Height.

Solution 1.17

- Total Frequency (N) = $\sum f = 300$
- Median is $\left(\frac{\sum f}{2}\right)^{th}$ value

  = $\left(\frac{300}{2}\right)^{th}$ value = 150th value.
- 150th value occurs in the cumulative frequency 160.
- This CF = 160 corresponds to Height = 5.

∴ The Median value is 5.

Height (m)	Number of Trees	Cumulative Frequency (CF)
2	14	14
3	21	$14 + 21 = 35$
4	42	$35 + 42 = 77$
5	83	$77 + 83 = 160$
6	118	$160 + 118 = 278$
7	12	$278 + 12 = 290$
8	7	$290 + 7 = 297$
9	3	$297 + 3 = 300$

MEDIAN FOR GROUPED DATA FORMULA

$$\text{Median} = L_m + \left(\frac{\frac{N}{2} - F}{f_m} \right) \times i$$

where:

- N = the total frequency
- F = the cumulative frequency before the median class
- f_m = the frequency of the median class
- L_m = the lower class boundary of the median class
- i = the class width or size

Example 1.18

- Calculate the median for the distribution data below.

Class	50-59	60-69	70-79	80-89	90-99
Frequency	6	9	12	15	8

Solution 1.18

- $\text{Median} = L_m + \left(\frac{\frac{N}{2} - F}{f_m}\right) \times i$
- N = 50, $\frac{N}{2} = \frac{50}{2} = (25)^{th}$ value falls between 69.5 – 79.5
- F = 15, $f_m = 12$, $L_m = 69.5$
- $i = 10 = 79.5 - 69.5$
- $\text{Median} = 69.5 + \left(\frac{\frac{50}{2} - 15}{12}\right) \times 10$

$$= 69.5 + \left(\frac{25-15}{12}\right) \times 10$$

$$= 69.5 + \left(\frac{10}{12}\right) \times 10$$

$$= 69.5 + \frac{100}{12}$$

$$= 69.5 + 8.33$$

$$= 77.8$$

Class Boundaries	Frequency f	Cumulative Frequency CF
49.5-59.5	6	6
59.5-69.5	9	$6 + 9 = 15$
69.5-79.5	12	$15 + 12 = 27$
79.5-89.5	15	$27 + 15 = 42$
89.5-99.5	8	$42 + 8 = 50$
	N=50	

End of Slides

Thank You