Probability and Statistics

Descriptive Statistics for grouped and ungrouped data

Lecture 4.
Class 1.
Time: 8:30- 10:30
Department: Bit

Review and Introduction

What is Statistics?

- Science of gathering (تجميع), analyzing, interpreting, and presenting data

Population Vs Sample

- **Population** — the whole: a collection of persons, objects, or items under study
- **Sample** — a portion of the whole: a subset of the population

Descriptive Statistics

Descriptive vs. Inferential Statistics

- **Descriptive Statistics** — statistics gathered on a group to describe or reach conclusions about that same group only.

- **Inferential Statistics** — statistics gathered on sample data to reach conclusions about the population from which the sample was taken.

Descriptive Statistics

Descriptive Statistics

- Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data.

Example:

The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed below:

Descriptive Statistics

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Descriptive Statistics

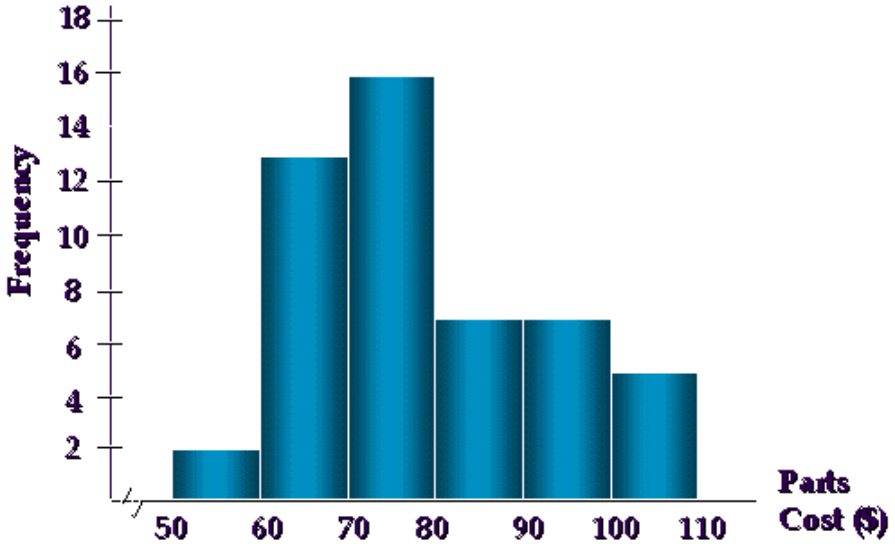
Tabular Summary

Cost ($)	Parts Frequency		Percent Frequency
50-59	2	4	
60-69	13	26	
70-79	16	32	
80-89	7	14	
90-99	7	14	
100-109	5	10	
Total	**50**		**100**

التكرار النسبي = (تكرار الفئة / التكرار الكلي) * 100

Descriptive Statistics

Graphical Summary (Histogram)



Numerical Descriptive Statistics

The most common numerical descriptive statistic is the average (or mean). Others: mode, median, variance, standard deviation, etc.

Statistical Inferences

Statistical Inference

- The process of using data obtained from a small group of elements (the sample) to make estimates and test hypotheses about the characteristics of a larger group of elements (the population).

Parameter vs. Statistic

Parameter معلمة — descriptive measure of the population

Usually represented by Greek letters

Statistic احصاءة —descriptive measure of a sample

Usually represented by Roman letters

Statistical Inferences

Population Parameters:

μ : denotes population mean (expected value)

σ^2: denotes population variance.

σ : denotes population standard deviation

Sample Parameters:

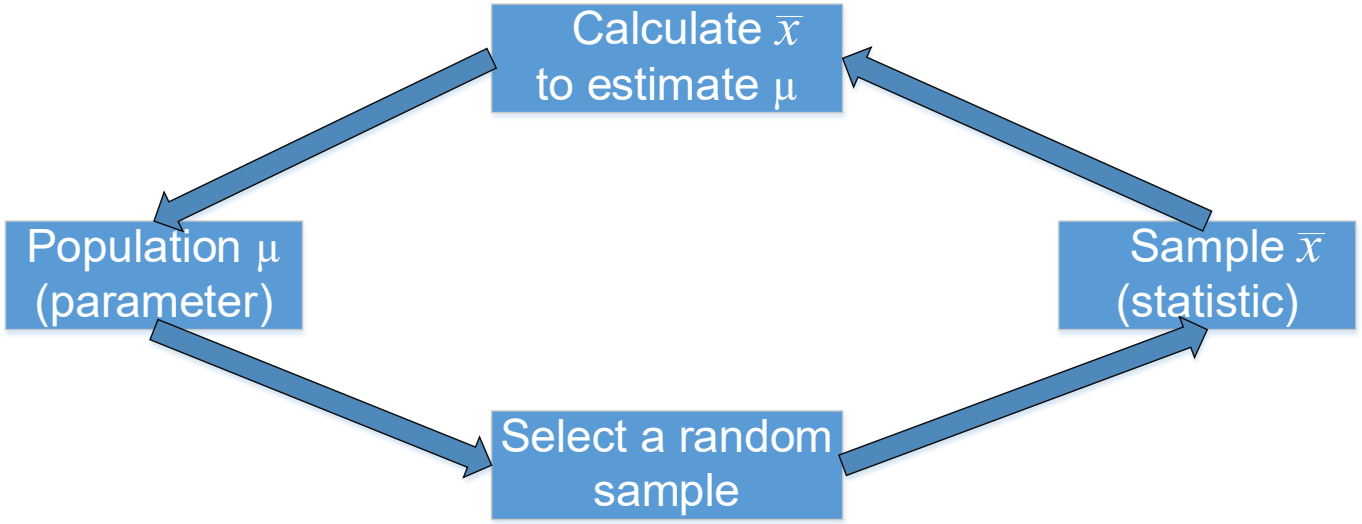
$\bar{x}$: denotes sample mean

s^2 : denotes sample variance.

s : denotes sample standard deviation

Statistical Inferences

Process of Inferential Statistics



Data Measurement

Example:

- Students of a university are classified by the school using a nonnumeric label*: Business, Humanities, Education*, and so on.

Alternatively, a numeric code could be used (e.g. 1 denotes Business, 2 denotes Humanities, 3 denotes Education, …).

- Employment Classification: 1 for Educator; 2 for Construction Worker; 3 for Manufacturing Worker

Data Measurement

Nonparametric statistics:

- A class of statistical techniques that make few assumptions about the population
- Used with nominal and ordinal level data

Parametric statistics:

- A class of statistical techniques that contain assumptions about the population
- Used only with interval & ratio level data

Ungrouped vs. Grouped Data

Ungrouped data

have not been summarized in any way
are also called raw data

Grouped data

have been organized into a frequency distribution

Ungrouped vs. Grouped Data

Example: Ages of a sample of managers

42	26	32	34	57
30	58	37	50	30
53	40	30	47	49
50	40	32	31	40
52	28	23**	35	25
30	36	32	26	50
55	30	58	64	52
49	33	43	46	32
61	31	30	40	60
74*	37	29	43	54

Ungrouped vs. Grouped Data

Frequency Distribution of Manager's Ages:

Class Interval	**Frequency**
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

Ungrouped vs. Grouped Data

Range and Class

Data Range: Range = Largest – Smallest

Ex: Range = 74 – 23 = 51

Number of Classes and Class Width

- The number of classes should be between 5 and 20.
  Fewer than 5 classes cause excessive مفرط summarization.
  More than 20 classes leave too much detail.

Ungrouped vs. Grouped Data

- Class Width

Divide the range by the number of classes for an approximate class width Round up to a convenient number

Ex: Approximate Class Width = 51/6 = 8.5

Class Width = 10

Ungrouped vs. Grouped Data

Relative Frequency

Class Interval	Frequency	Relative Frequency
20-under 30	6	.12
30-under 40	18	.36
40-under 50	11	.22
50-under 60	11	.22
60-under 70	3	.06
70-under 80	1	.02
Total	50	1.00

Ungrouped vs. Grouped Data

Cumulative Frequency

Class Interval	Frequency		Cumulative Frequency
20-under 30	6	6	
30-under 40	18	24	
40-under 50	11	35	
50-under 60	11	46	
60-under 70	3	49	
70-under 80		1	50
Total	50		

Measures of Central Tendency
Ungrouped Data

- Measures of central tendency yield تسـفر عـن information about “particular places or locations in a group of numbers.”

- Common Measures of Location

Mode, Median, Mean, Percentiles, Quartiles

Ungrouped vs. Grouped Data

Cumulative Relative Frequencies

Class Interval	Frequency	RF	Cu. Frequency	CRF
20-under 30	6	.12	6	.12
30-under 40	18	.36	24	.48
40-under 50	11	.22	35	.70
50-under 60	11	.22	46	.92
60-under 70	3	.06	49	.98
70-under 80	1	.02	50	1.00
Total	50	1.00		

Measures of Central Tendency
Ungrouped Data

Mode

- The most frequently occurring value in a data set

Bimodal -- Data sets that have two modes

Multimodal -- Data sets that contain more than two modes

Measures of Central Tendency
Ungrouped Data

Example:

35	37	37	39	40	40
41	41	43	43	43	43
44	44	44	44	44	45
45	46	46	46	46	48

Value 44 occurs 5 times

The mode is 44

•

Measures of Central Tendency
Ungrouped Data

Median

- Middle value in an ordered array of numbers.
- *Unaffected* by extremely large and extremely small values
- Median is determined without using all information from the data set.

Measures of Central Tendency
Ungrouped Data

Computational Procedure

- Arrange the observations in an ordered array.
- If there is an odd number of terms, the median is the middle term of the ordered array.
- If there is an even number of terms, the median is the average of the middle two terms.

Measures of Central Tendency
Ungrouped Data

Example:

- Ordered Array: 3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22
  - There are 17 terms in the ordered array.
  - Position of median = (n+1)/2 = (17+1)/2 = 9
  - The median is the 9th term, 15.

- Ordered Array: 3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21
  - There are 16 terms in the ordered array.
  - Position of median = (n+1)/2 = (16+1)/2 = 8.5
  - The median is between the 8th and 9th terms: 14.5.

Measures of Central Tendency
Ungrouped Data

Arithmetic Mean

- Commonly called ‘the mean’: the average of a group of numbers
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set

Measures of Central Tendency
Ungrouped Data

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Sample Mean


وقل ربي زدني علماً
UNIVERSITY OF INFORMATION TECHNOLOGY AND COMMUNICATIONS


جامعة تكنولوجيا المعلومات والاتصالات
University of Information Technology & Communications
2014
Businesses Informatics College
كلية معلوماتية الاعمال

THANK YOU