

IAM 530
ELEMENTS OF PROBABILITY AND
STATISTICS

LECTURE 1-INTRODUCTION TO
STATISTICS

WHAT IS STATISTICS?

Statistics: The science of collecting, describing, analysing and interpreting data. It concerns what can be learned from data.

In general statistical techniques are employed either to describe what did happen or to predict what might happen.

Its methods are used in different areas of science
e.g.

- Biology
- Medicine
- Psychology
- Social Science
- Engineering
- Actuaries
- Finance

BASIC DEFINITIONS

- **Variable:** is a characteristic about each individual element of a population or sample.
- **Data:** are the different values associated with a variable.
- **Population:** consists of all the items or individuals about which you want to draw a conclusion.
- **Sample:** is the portion of a population selected for analysis.
- **Parameter:** is a numerical measure that describes a characteristic of a population.
- **Statistic:** is a numerical measure that describes a characteristic of a sample.
- **Experiment:** is a planned activity whose results yield a set of data.

Example: An insurance company is interested in learning about the average premiums of their customers.

The **population** is all customers.

A **sample** is any subset of that population.

The **variable** is the “premium” of each customer.

The **data** would be the set of values in the sample.

The **experiment** would be the method used to select the premiums.

The **parameter** of interest is the “average” premiums of all customers.

The **statistic** is the “average” premiums for all customers in the sample.

AREAS OF STATISTICS

Descriptive Statistics: collection, presentation, and description of sample data.

Inferential Statistics: making decisions and drawing conclusions about populations.

Modelling : developing prediction equations to understand a complex system.

Descriptive Statistics

- Collect data
 - e.g., Survey
- Present data
 - e.g., Tables and graphs
- Characterize data
 - e.g., Sample mean

Descriptive Statistics

Graphical Methods

- Frequency and relative frequency distributions
- Bar Charts
- Histograms
- Line Charts
- Pie-Charts

Descriptive Statistics

Numerical Methods

Measures of Central Tendency

- Sample Mean
- Sample Median
- Sample Mode

Measures of Dispersion(Variability)

- Range
- Mean Absolute Deviation (MAD)
- Sample Variance
- Sample Standard Deviation

Inferential Statistics

- Estimation of parameters (Point and Confidence Estimation).
 - e.g., Estimate the population mean
- Hypothesis testing
 - e.g., Test the claim that the population mean is equal to a value

Modelling

- Building a Model
 - e.g., Building a simple linear regression equation to model the relation between two or more variables by using a sample data.
- Forecasting
 - e.g., Using the regression model in prediction.

VARIABLES

- They may be classified as:
 - **Quantitative** (numerical) variables have values that represent quantities.
 - **Continuous**
 - **Discrete**
 - **Categorical** variables have values that can only be placed into categories.
 - **Nominal** values are the arbitrary numbers that represent categories.
 - **Ordinal** (ranked) Often ordinal variables are re-coded to be quantitative.

DATA COLLECTION

- First problem a statistician faces: how to obtain the data. It is important to obtain **good**, or **representative**, data.
- **Target Population**
- **Sampled Population**

METHODS USED TO COLLECT DATA

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

SAMPLING

Sampling Frame: A list of the elements belonging to the population from which the sample will be drawn. It is important that the sampling frame be representative of the population.

Sample Design: The process of selecting sample elements from the sampling frame.

SAMPLING DESIGN METHODS

- **Simple Random Sample** : All possible samples with the same number of observations are equally likely to be selected.
- **Stratified Sampling**: Population is separated into mutually exclusive sets (strata) and then sample is drawn using simple random samples from each strata.
- **Convenience Sample**: It is obtained by selecting individuals or objects without systematic randomization.
- **Systematic Sample**: A sample in which every k th item of the sampling frame is selected, starting from the first element which is randomly selected from the first k elements.

DATA TYPES

- **Cross sectional data** is collected at a certain point in time
 - Test score in a statistics course
 - Starting salaries of an MBA program graduates
- **Time series data** is collected over successive points in time
 - Weekly closing price of gold
 - Amount of crude oil imported monthly
- **Panel data** is collected over successive points in time as well

COMPARISON OF PROBABILITY AND STATISTICS

Probability: Properties of the population are assumed known. Answer questions about the sample based on these properties.

Statistics: Use information in the sample to draw a conclusion about the population.