1. Introduction To Statistics

1.1 What is Statistics ?

“A branch of mathematics dealing with the collection, analysis, interpretation and presentation of masses of numerical data”– Webster’s New Collegiate Dictionary

“The branch of the scientific method which deals with the data obtained by counting or measuring the properties of populations”– Fraser (1958)

“The entire science of decision making in the face of uncertainty” – Freund and Walpole (1987)

“The technology of the scientific method concerned with (1) the design of experiments and investigations, (2) statistical inference”– Mood, Graybill and Boes (1974)

All definitions imply that

“Statistics is a theory of information with inference making as its objective”.

2. Population

The large body of data that is the target of our interest.

3. Sample

The subset selected from the population.

4. Objective of Statistics

“to make an inference about a population based on information contained in a sample and to provide an associated measure of goodness for the inference”.

Examples

1. A medical scientist wants to estimate the average length of time until the recurrence of a certain disease.

Population: times until recurrence for all people who have had a particular disease.

Objective: to estimate the average time until recurrence.

2. A city engineer wants to estimate the average weekly water consumption for single-family dwelling units in the city.

Population: weekly water consumption for all single-family dwelling units in the city.

Objective: to estimate the average weekly water consumption for all families.

1.2 Characterizing a Set of Measurements: Graphical Methods

Frequency distributions provide a graphic and very useful method for describing (summarizing) a set of numbers (data).

An individual population or any set of measurements can be characterized by a relative frequency distribution which can be represented by a relative frequency (rf)- histogram.

To make a rf- histogram:

- Subdivide the axis of measurements into intervals of equal width.

- Construct rectangle over each interval such that the height of the rectangle is proportional to to the fraction of the total number of measurements falling (rf) in each cell.

Example: To characterize 10 measurements

2.1, 2.4, 2.2, 2.3, 2.7, 2.5, 2.4, 2.6, 2.6 and 2.9
we divide these measurements into intervals of equal width (say, 0.2 unit) starting from 2.05 and calculate the rf for each interval:

Interval	2.05-2.25	2.25-2.45	2.45-2.65	2.65-2.85	2.85-3.05
Relative Frequency	0.2	0.3	0.3	0.1	0.1

The rf-distribution is obtained by drawing a curve passing through the mid points of the width of the rectangles and extended on both ends to meet the x-axis.

Probabilistic interpretation derived from the histogram:

If a measurement is selected at random from the original data set, the probability that it will fall in a given interval is proportional to the area under the histogram lying over that interval.

Example:

The probability that a randomly selected measurement falls in the interval 2.05-2.45
= 0.2 + 0.3 = 0.5 (half measurements fall in this interval)
= area of the rectangle on 2.05-2.45

1.3 Characterizing a Set of Measurements: Numerical Methods

Numerical quantities measuring the sample information are useful when we wish to make an inference and measure the goodness of that inference.

We can mathematically derive certain properties of these sample quantities and make probability statements regarding the goodness of our inferences.

Two types of Numerical Descriptive Measures of a Data Set:

1. Measures of Central Tendency:

The arithmetic mean (or simply mean) of a sample of n measured responses $y_i, i = 1, \ldots, n$ is:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The corresponding population mean is denoted by μ .
The mean of a set of measurements only locates the center of the distribution of data.
Two sets of data could have widely different frequency distributions but equal means.
To describe data adequately we must define measures of data variability.

2. Measures of Dispersion or Variation:

The variance of a sample of n measured responses $y_i, i = 1, \ldots, n$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and the corresponding population variance is denoted by σ^2.
The standard deviation of a sample of measurements is:

$$s = \sqrt{s^2}$$

and the corresponding population standard deviation by $\sigma = \sqrt{\sigma^2}$.

Empirical Rule:

Many data sets in real life have approximately the mound (or bell) - shaped frequency distribution, known as the normal distribution. For such a distribution, it follows that the interval with endpoints

$\mu \pm \sigma$	contains approximately 68% of the measurements
$\mu \pm 2\sigma$	contains approximately 95% of the measurements
$\mu \pm 3\sigma$	contains almost all of the measurements

Exercise 1.7. We calculate: $\bar{y} = 1252.44$ and $s = 393.75$.

Supposing the distribution of measurements (y) is approximately normal, we have the following based on the empirical rule:

k	$\bar{y} \pm ks$	Interval Boundaries	Frequency	Expected
			Actual	Empirical Rule
1	1252.44 ± 393.75	858.69-1646.19	45	34
2	1252.44 ± 787.50	464.94-2039.94	48	47.5
3	1252.44 ± 1181.25	71.19-2433.69	49	50

It can be deduced that:

1. approximately 68% of the income data are between
   $ 858.69-$1646.19.
2. approximately 95% of the income data are between
   $ 464.94-$2039.94.
3. almost all of the income data are between
   $ 71.19-$2433.69.

Thus, knowledge of the mean and the standard deviation gives us a pretty well picture of the frequency distribution of the data on federal government income from taxes.

4. Suppose a state is randomly selected out of 50 states.

What is the probability that the selected state’s contribution to federal income is between $ 858.69-$1646.19?

Based on the empirical rule, we find the answer = 0.68.