Basic Elements of Statistics

Statistics

- **What is Statistics?**

Statistics

- What is Statistics?

- Concerned with collection of data, its description and analysis

Statistics

- **What is Statistics?**
- **Concerned with collection of data, its description and analysis**
- **Leading to drawing of conclusions from the data**

Statistics

- What is Statistics?
- Concerned with collection of data, its description and analysis
- Leading to drawing of conclusions from the data
- Art of learning from the data

Statistics

- **What is Statistics?**
- **Concerned with collection of data, its description and analysis**
- **Leading to drawing of conclusions from the data**
- **Art of learning from the data**
- **Fancy buzzwords: Data mining, Big data**

Descriptive Statistics

- Statistical analysis begins with a given set of data:

- For instance, the government regularly collects and publicizes data concerning yearly precipitation totals, earthquake occurrences, the unemployment rate, the gross domestic product, and the rate of inflation.

- Statistics can be used to describe, summarize, and analyse these data.

Descriptive Statistics

- In other situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data.

- The experiment chosen should depend on the use that one wants to make of the data.

- Example case: suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective.

Descriptive Statistics

- In other situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data.

- The experiment chosen should depend on the use that one wants to make of the data.

- Example case: suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective.

- This part of statistics, concerned with the description and summarization of data, is called descriptive statistics.

Descriptive vs Inferential Statistics

- ➢ Once an experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion.

- ➢ This part of statistics, concerned with the drawing of conclusions, is called inferential statistics.

- ➢ To be able to draw a conclusion from the data, we must take into account the possibility of chance.

- ➢ For instance, in previous example, suppose that the average score of members of the first group is quite a bit higher than that of the second. Can we conclude that this increase is due to the teaching method used?

- ➢ Or is it possible that the teaching method was not responsible for the increased scores but rather that the higher scores of the first group were just a chance occurrence?

Descriptive vs Inferential Statistics

- To be able to draw logical conclusions from data, we usually make some assumptions about the chances (or probabilities) of obtaining the different data values.

- The totality of these assumptions is referred to as a probability model for the data.

- Because the basis of statistical inference is the formulation of a probability model to describe the data, an understanding of statistical inference requires some knowledge of the theory of probability.

- In other words, statistical inference starts with the assumption that important aspects of the phenomenon under study can be described in terms of probabilities; it then draws conclusions by using data to make inferences about these probabilities.

Descriptive vs Inferential Statistics

- In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the population.

- The population is often too large for us to examine each of its members.

- In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements.

- This subgroup of a population is called a sample.

- Choosing a right sample is very important.

Brief History of Statistics

History of Statistics

- A systematic collection of data on the population and the economy was begun in the Italian city states of Venice and Florence during the Renaissance.

- The term statistics, derived from the word state, was used to refer to a collection of facts of interest to the state.

- Population statistics : Western Europe.

History of Statistics

- A systematic collection of data on the population and the economy was begun in the Italian city states of Venice and Florence during the Renaissance.

- The term statistics, derived from the word state, was used to refer to a collection of facts of interest to the state.

- Population statistics : Western Europe.

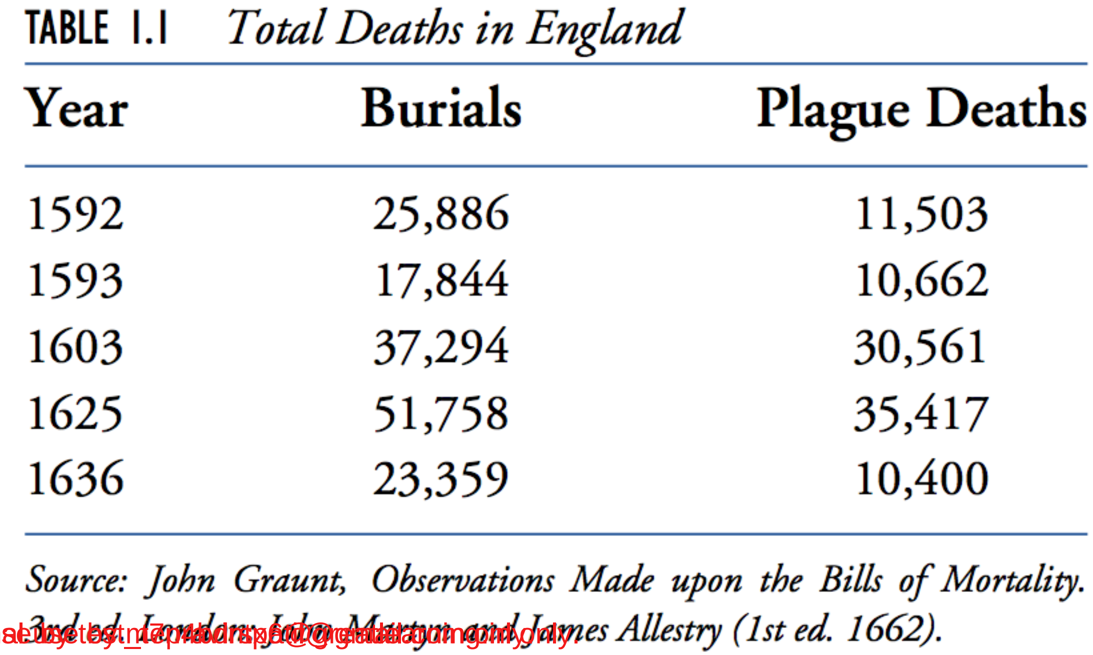

TABLE 1.1 *Total Deaths in England*

Year	Burials	Plague Deaths
1592	25,886	11,503
1593	17,844	10,662
1603	37,294	30,561
1625	51,758	35,417
1636	23,359	10,400

Source: *John Graunt, Observations Made upon the Bills of Mortality. 3rd ed. London: John Martyn and James Allestry (1st ed. 1662).*

A Few Applications

- Engineering : Manufacturing
- Engineering : Design
- Medicine : Clinical trials
- Sports : Baseball : Sabermetrics
- Sports : Cricket : Duckworth-Lewis

A Few Applications

- Engineering : Manufacturing
- Engineering : Design
- Medicine : Clinical trials
- Sports : Baseball : Sabermetrics
- Sports : Cricket : Duckworth-Lewis

Duckworth, FC & Lewis, AJ "A fair method for resetting the target in interrupted one-day cricket matches" Journal of the Operational Research Society, (Mar 1998) Volume 49, No. 3, pp 220-227

Introduction to Statistics

Types of Statistics

1. **<u>DESCRIPTIVE</u>**

2. **<u>INFERENTIAL</u>**

Types of Statistics

1. **DESCRIPTIVE**

- Provides a visual, tabular or numeric summary of large amounts of data to explain its key characteristics.

Types of Statistics

1. **DESCRIPTIVE**

- Provides a visual, tabular or numeric summary of large amounts of data to explain its key characteristics.
  a) identify patterns in large amounts of data
  b) data set is assumed as stand-alone

Types of Statistics

1. **<u>DESCRIPTIVE</u>**

- Provides a visual, tabular or numeric summary of large amounts of data to explain its key characteristics.
  a) identify patterns in large amounts of data
  b) data set is assumed as stand-alone

2. **<u>INFERENTIAL</u>**

- Uses a sample of data to make inferences about the general population.

Types of Statistics

1. **<u>DESCRIPTIVE</u>**

- Provides a visual, tabular or numeric summary of large amounts of data to explain its key characteristics.
  a) identify patterns in large amounts of data
  b) data set is assumed as stand-alone

2. **<u>INFERENTIAL</u>**

- Uses a sample of data to make inferences about the general population.
  a) assumes that sample is representative of a larger population
  b) draws conclusions about population based on the smaller sample

Statistical Inference

WHAT KIND OF INFERENCES CAN WE MAKE IN STATISTICS?

- We can estimate unknown population parameters based on properties of a sample

Statistical Inference

WHAT KIND OF INFERENCES CAN WE MAKE IN STATISTICS?

- We can estimate unknown population parameters based on properties of a sample

- We can test hypothesis about a population based on sample parameters

Samples and Population

We will introduce the concepts of sample and population

Samples and Population

We will introduce the concepts of sample and population

- **Examples of Populations**

Samples and Population

We will introduce the concepts of sample and population

- **Examples of Populations**
    - ➢ All applications received for credit cards from Bank XYZ

Samples and Population

We will introduce the concepts of sample and population

- **Examples of Populations**
  - ➢ All applications received for credit cards from Bank XYZ
  - ➢ All consumers of Product X

Samples and Population

We will introduce the concepts of sample and population

- **Examples of Populations**
  - All applications received for credit cards from Bank XYZ
  - All consumers of Product X
- **Examples of Samples**

Samples and Population

We will introduce the concepts of sample and population

- **Examples of Populations**
  - All applications received for credit cards from Bank XYZ
  - All consumers of Product X
- **Examples of Samples**
  - All applications received in the last 3 months
  - Women consumers over the age of 45 that have bought Product Y last 6 months

Sample vs Population

Why do we need to separate the two?

Sample vs Population

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyse data on the population.

Sample vs Population

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyse data on the population.

- It is easier to take a subset of the population, analyse the subset, and then make inferences about the population.

Sample vs Population

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyse data on the population.

- It is easier to take a subset of the population, analyse the subset, and then make inferences about the population.

The second point depends on a fundamental assumption –

Sample vs Population

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyse data on the population.
- It is easier to take a subset of the population, analyse the subset, and then make inferences about the population.

The second point depends on a fundamental assumption –

Representativeness

Sample vs Population

Why do we need to separate the two?

- Population (or the Universe) tends to be very large, making it difficult (or impossible) to collect and analyse data on the population.

- It is easier to take a subset of the population, analyse the subset, and then make inferences about the population.

The second point depends on a fundamental assumption –

Representativeness

We have to find a sample that is representative of the population that it belongs to

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

- **How many samples are possible?**

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

- **How many samples are possible?**

20?

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

- **How many samples are possible?**

20?

500?

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

- **How many samples are possible?**

20?

500?

10000 C 500?

Clearly, many samples are possible

Sample vs Population

Imagine we have a population of 10,000 respondents to a survey, and we want to take a sample of 500.

- **How many samples are possible?**

20?

500?

10000 C 500?

Clearly, many samples are possible

- **Will all samples be “good” or representative samples?**

Sample vs Population

We can choose many samples from a population, however a good sample is that which is chosen:

1. **Without bias**

2. **Full coverage**

3. **Nonresponse inclusive**

Sample vs Population

We can choose many samples from a population, however a good sample is that which is chosen:

1. **Without bias:** e.g. not choosing only high income respondents

2. **Full coverage**

3. **Nonresponse inclusive**

Sample vs Population

We can choose many samples from a population, however a good sample is that which is chosen:

1. **Without bias:** e.g. not choosing only high income respondents

2. **Full coverage:** all segments in population are correctly represented

3. **Nonresponse inclusive**

Sample vs Population

We can choose many samples from a population, however a good sample is that which is chosen:

1. **Without bias:** e.g. not choosing only high income respondents

2. **Full coverage:** all segments in population are correctly represented

3. **Nonresponse inclusive:** if 20% of your population are defaulters, your sample ideally should also have 20% defaulters

Sample vs Population

We can choose many samples from a population, however a good sample is that which is chosen:

1. **Without bias:** e.g. not choosing only high income respondents

2. **Full coverage:** all segments in population are correctly represented

3. **Nonresponse inclusive:** if 20% of your population are defaulters, your sample ideally should also have 20% defaulters

Behaviour of the sample should be like the behaviour of the population

Sample vs Population

Behavior of the sample should be like the behavior of the population

Sample vs Population

Behavior of the sample should be like the behavior of the population

- i.e. the descriptive statistics of sample data point should be equal to or close to population statistics

Sample vs Population

<u>Behavior of the sample should be like the behavior of the population</u>

- i.e. the descriptive statistics of sample data point should be equal to or close to population statistics

<u>Example:</u>

In the survey respondents example, say we take income as an attribute. Average income is $50,000.

Sample vs Population

<u>Behavior of the sample should be like the behavior of the population</u>

- i.e. the descriptive statistics of sample data point should be equal to or close to population statistics

<u>Example:</u>

In the survey respondents example, say we take income as an attribute. Average income is $50,000. Sample of 500 respondents should show average income of $50,000, or close.

Sample vs Population

Behavior of the sample should be like the behavior of the population

- i.e. the descriptive statistics of sample data point should be equal to or close to population statistics

Example:

In the survey respondents example, say we take income as an attribute. Average income is $50,000. Sample of 500 respondents should show average income of $50,000, or close.

How do we pick a representative sample from all the possible samples?

Choosing a Representative Sample

The best way to pick a representative sample is:

By picking sample elements at random!

Choosing a Representative Sample

The best way to pick a representative sample is:

By picking sample elements at random!

Imagine that in our survey respondents population of 10,000:

- ○ 60% of respondents can be categorized as having medium levels of income

Choosing a Representative Sample

The best way to pick a representative sample is:

By picking sample elements at random!

Imagine that in our survey respondents population of 10,000:

- 60% of respondents can be categorized as having medium levels of income
- 30% as low income

Choosing a Representative Sample

The best way to pick a representative sample is:

By picking sample elements at random!

Imagine that in our survey respondents population of 10,000:

- ○ 60% of respondents can be categorized as having medium levels of income
- ○ 30% as low income
- ○ 10% as high income

Choosing a Representative Sample

A Thought Experiment

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category:

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category:

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category: 30%

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category: 30%

High income category:

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category: 30%

High income category: 10%

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category: 30%

High income category: 10%

As we keep picking (at random), will we likely be creating a representative sample?

Choosing a Representative Sample

The chances of picking one respondent, completely at random, and finding that they belong to the Medium Income category: 60%

Low income category: 30%

High income category: 10%

As we keep picking (at random), will we likely be creating a representative sample?

We have introduced another concept here: **Likelihood**

How to Choose a Sample?

So how do we select a sample? Essentially, at random

How to Choose a Sample?

So how do we select a sample? Essentially, at random

Simple random sample

Stratified random sampling

How to Choose a Sample?

So how do we select a sample? Essentially, at random

Simple random sample

- All members in a population have an equal chance of being selected in the sample

Stratified random sampling

How to Choose a Sample?

So how do we select a sample? Essentially, at random

Simple random sample

- All members in a population have an equal chance of being selected in the sample

Stratified random sampling

- Population members are first divided into groups or strata based on meaningfulness. Then random samples are taken from each strata

How to Choose a Sample?

So how do we select a sample? Essentially, at random

Simple random sample

- All members in a population have an equal chance of being selected in the sample

Stratified random sampling

- Population members are first divided into groups or strata based on meaningfulness. Then random samples are taken from each strata

While building a sample remember representativeness. We assume that the sample represents the population, so any inferences we draw about the sample will be true of the population