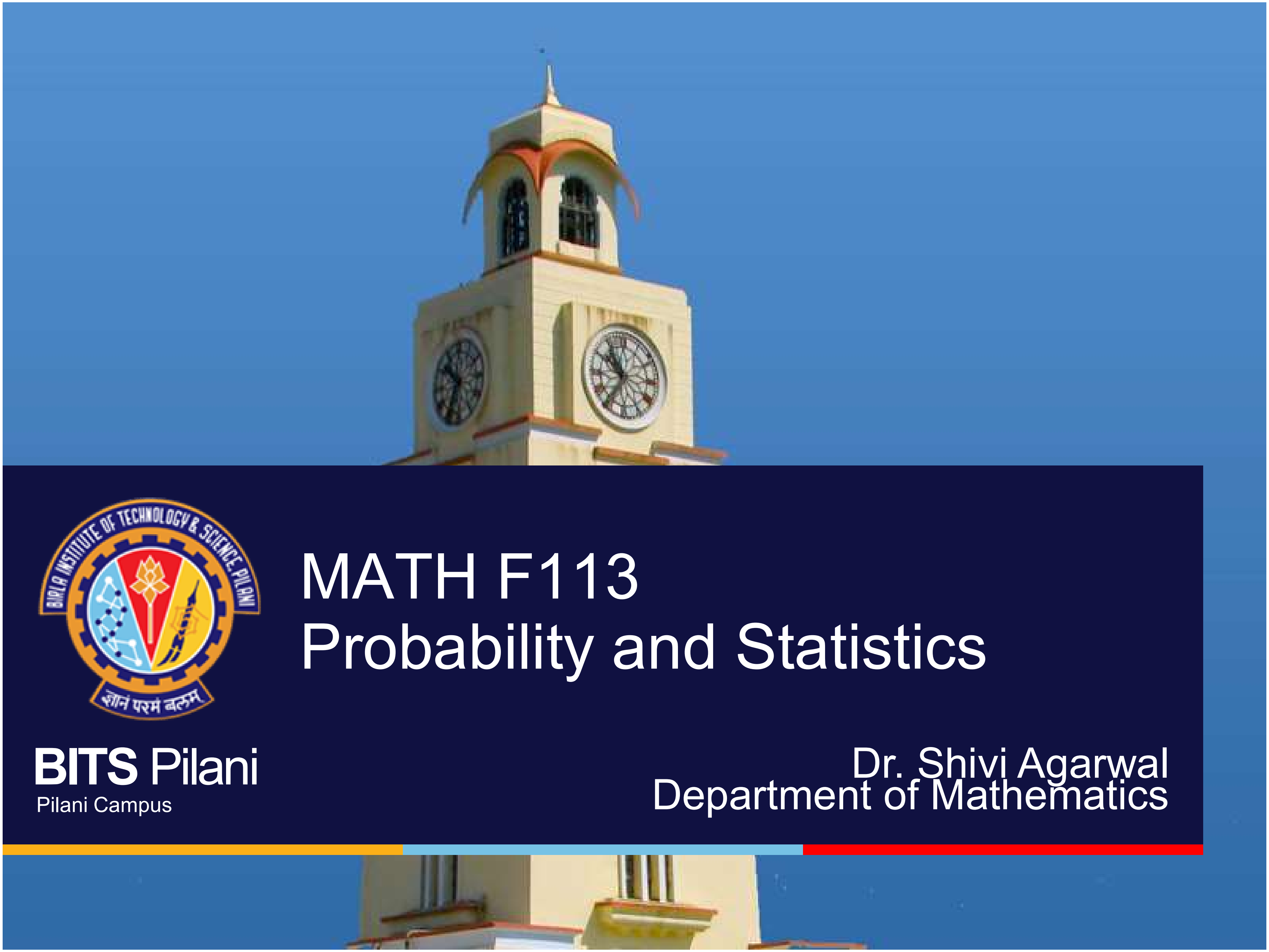

MATH F113
Probability and Statistics

Dr. Shivi Agarwal
Department of Mathematics

BITS Pilani
Pilani Campus

CHAPTER-6

DESCRIPTIVE STATISTICS

Statistics

- In Statistics, we want to study properties of a (large) group of objects, generally termed as *population*.
- Methods of statistics study small subsets of population. This is called *sample*. The science developed for this purpose is called descriptive statistics.
- Study of samples is used to infer the properties of the entire population. The science developed for this purpose is called statistical inference.

- Population of interest : values of a random variable.
- Its properties of interest : parameters of its distribution.

Example

- If the population has normal distribution, the location and shape are described by μ and σ.
- For a binomial distribution consisting of n trials, the shape is determined by p

- Often the values of parameters that specify the exact form of a distribution are unknown.
- You must rely on the sample to make inference about these parameters.

Need of Sampling

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

- To estimate the proportion of tires manufactured by Dunlop tires which are sturdy, the quality control dept of the company needs to apply a certain amount of stress on tires and see if they survive it. Rather than applying stress to each and every tire, it is practical to do it on an appropriate sample and use it to infer for the whole set of tires.

- » If you want that the sample should provide reliable information about the population, you must select your sample in a certain way!

Population

- Many times populations are described by the distribution of the observations.
- We refer to population in terms of the *corresponding probability distribution $f(x)$ of the random variable X.*

Random Sampling

- Characteristics of a statistical Problem:

1. Associated with the problem is a large group of objects about which inferences are to be made. This group of objects is called the ***population*.**

2. There is at least one random variable whose behavior is to be studied relative to the population.

- The population is too large to study in its entirety, or techniques used in the study are destructive in nature. In either case we must draw conclusions about the population based on observing only a portion or “sample” of objects drawn from the population

Definition : A set of observations $X_1, \ldots,$ X_n constitutes a **random sample** of size n from the **infinite population** $f(x)$ if

1. Each X_i is a random variable whose distribution is given by $f(x)$,
2. These n random variables are *independent.*

A random sample of size n from the distribution of X is a collection of n independent random variables, each with the same distribution.

Remarks

(1) Here the sample can be thought as ordered and with repetition allowed, i.e., an n-tuple (x_1, …, x_n) represents a sample.

(2) Let r.v. X with density $f_X(x)$ denote the population. The *random sample* can be thought as the n-dimensional random variable (X_1, …, X_n) whose joint density is

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_X(x_1) \ldots f_X(x_n)$$

3. If we want to interpret sample obtained by choosing values of X, we can interpret as choosing n values of X randomly and independently with replacement and in order, whether population is finite or infinite. For infinite populations or large populations compared to sample (say 20 times larger than sample size), even if each choice is done randomly and replacement is not allowed, we more or less have these assumptions satisfied.

4. When population is finite, replacement is not done and ordering is not given importance, sampling theory is slightly different. We won’t discuss it and assume the hypotheses in definition.

5. The random sample and its characteristics will be denoted by capital letters and a particular sample and its characteristics by corresponding small letters. This is in tune with the practice of denoting random variables by capital letters and its values by corresponding small letters. Thus a particular sample is a value $(x_1, \ldots, x_n)$ of the n-dimensional r.v. $(X_1, \ldots, X_n)$, or equivalently, a point in n-dimensional space. The values x_1, …, x_n are also called observed values of X.

- Sec 6.2 (graphical representation and analysis of data) not in syllabus.

A random sample of size n from the distribution of X is a collection of n independent random variables, each with the same distribution as X.

NOTE: Once a random sample has been drawn, we commonly use the data gathered to evaluate pertinent *statistics*.

6.3. Statistic

Definition: A statistic of a random sample $(X_1, \ldots, X_n)$ from population X is a function $H(X_1, \ldots, X_n)$ of the n-dimensional r.v. $(X_1, \ldots, X_n)$. It itself is a r.v. whose values are values of the statistic on particular samples, i.e., the values of the statistic are determined by the sample considered.

Ex. 1)The component X_1 of $(X_1, \ldots, X_n)$.

2) $\Sigma_{1 \leq i \leq n} X_i$.

Use of Statistic

- To estimate population parameters, like mean, variance, etc, suitable statistic is used.
- Statistic itself is different from the population parameter. As such, value of statistic depends on the sample used, where as population parameter has nothing to do with sample and is fixed.

Ex 6.1.1

Consider the random variable X, the number of times per hour that a television signal is interrupted by random interference. Assume that this random variable has a Poisson distribution with unknown μ and unknown variance σ^2. To approximate the value of each of these parameters, we intend to observe the signal for ten randomly selected non overlapping one-hour periods over a week's time. Let X_i ,

(i = 1,2…10) denote the number of interruptions that occur during the *i*th observation period. The random variables X_1, X_2, , X_3 ,………., X_{10} constitute a random sample of size 10 from a Poisson distribution with unknown mean μ and unknown variance σ^2. When experiment is conducted, these data result:

$$x_1 = 1\ ;\ x_2 = 0\ ; x_3 = 0\ ;\ x_4 = 2\ ;\ x_5 = 1\ ;$$

$$x_6 = 1\ ;\ x_7 = 0\ ; x_8 = 0\ ;\ x_9 = 3\ ;\ x_{10} = 0$$

The observed value of the statistics $\sum X_i$, $\sum X_i^2$, $\sum X_i/n$, $\max_i \{X_i\}$ and $\min_i \{X_i\}$ based on this sample are 8, 16, 0.8, 3 and 0, respectively. Note that the random variable $X_1 - \mu$ is ***not*** a statistic. Since μ is unknown , we cannot determine its numerical value from a random sample.

Useful Statistic

Location Statistic: The mean or theoretical average value of X is our primary measure of the center of location of X.

Definition: (Sample Mean). Let X_1, $X_{2,}$, X_3 ,…., X_n be a random sample from the distribution of X. The statistic $\overline{X}$ is called the sample mean and is denoted by

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Median Location

Let $x_1, x_{2,}, \ldots, x_n$ be a sample of observations arranged in order from the smallest to the largest. The sample median is the middle observation if ‘n’ is odd. It is the average of the two middle observations if ‘n’ is even. We denote the median of a sample by :

$$\tilde{X} = \left(\begin{matrix} X^{\left(\frac{n+1}{2}\right)} & ;\, n\tilde{i}s\, odd \\ \dfrac{X^{\left(\frac{n}{2}\right)} + X^{\left(\frac{n}{2}+1\right)}}{2} & ;\, if \;\; n\, is\, even \end{matrix} \right)$$

Measures of Variability

The variance of a random variable, given by $\sigma^2 = E[(X - \mu)^2]$ measures the 'variability' of X about the mean.

Sample variance and sample standard deviation :

Let $X_1, X_{2,}, X_3, \ldots\ldots\ldots\ldots, X_n$ be a random sample of size 'n' from the distribution of X. Then the statistic

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$$

S^2 is called the ‘sample variance’. Furthermore, the statistic ‘S’ is called the ‘sample standard deviation’. Which is the positive square root of S^2.

Computational formula :

Let X_1, $X_{2,}$, X_3 ,………., X_n be a random sample of size ‘n’ from the distribution of X. The sample variance is given by :

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Definition: (Sample range). The sample range is defined to be the difference between the largest and the smallest observations with subtraction in the order largest minus smallest.

$$\text{Sample mean } \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\text{Sample minimum} = \min_i \{X_i\}$$

$$\text{Sample maximum} = \max_i \{X_i\}$$

$$\text{Sample Range} = \max_i \{X_i\} - \min_i \{X_i\}$$

$$\text{Sample Variance } S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}.$$

Example: A random sample of size 9 yields the following observations on the random variable X, the coal consumption in millions of tons by electric utilities for a given year:

406 395 400 450 390 410 415 401 408

The observed value of the sample mean for these data is

$$\overline{X} = \sum_{i=1}^{n} x_i /n = (406+395+400+\ldots\ldots.+408)$$

$$= 3675/9 = 408.3 \text{ million tons}$$

The average value for X for this sample is 408.3 million tons.

Example: These data constitute a sample of observations on X, the coal consumption in millions of tons by electric utilities for a given year:

390 400 406 410 450 395 401 408 415

To compute the sample variance, we must evaluate the statistics $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ for this sample.

The observed values are $\sum_{i=1}^{9} x_i = 3675$ $\sum_{i=1}^{9} x_i^2 = 1{,}503{,}051$

The observed value of S^2 is

$$S^2=\left(9\sum_{i=1}^{9} x_i^2 - \left(\sum_{i=1}^{9} x_i\right)^2\right)/9(8)=(9(1503051)-3675^2)/9(8)$$

$$= 303.25$$

Remember that variance is usually considered to be unitless because the physical unit attached to it is often meaningless.

The observed value of S is

$$s = \sqrt{s^2} = \sqrt{303.25} = 17.4 \text{ million tons}$$

Notice that the physical measurement unit associated with s matches that of the original data and that 17.4 million tons is the standard deviation for this sample.

Using a particular sample,

Sample mean $\bar{x}$ *is used to* approximate the population mean μ.

Sample variance s^2 is used to approximate the population variance σ^2, while sample s.d s is used to approximate population s.d. σ.

Ex. 17 : Consider these data sets :

1 3 2 1 2 4 1

Ist 2 5 4 2nd 2 5 2 5

4 3 3 1 5 5 3

(a) Find the sample mean and sample median for each data set. (3; 3 for 1st, 3, 2.5 for 2nd)

(b) Find the sample range for each data set.

(4,4)

(c) Find sample variance and sample s.d. for each data set.(1.5,1.2 for 1st, 2.91,1.7 for 2nd)

(d) Would you be surprised to hear someone claim that these data were drawn from the same population?

Q. 25 (Approx. of σ by range) :

If X is normal, $P[-2\sigma < X-\mu < 2\sigma] = 0.95 \sim 1$

So we can assume $4\sigma \sim$ range, or

Approximately σ=(estimated range)/4

If X is not normal, by Chebyshev inequality :

$P[-3\sigma < X-\mu < 3\sigma] \geq 0.89$,

So we can approx take

σ=(estimated range)/6.

Q 6.3.18: The observed values of the statistics $\sum X_i$ and

$\sum_{i=1}^{50} X_i^2$ are 63,707 and 154,924,261 respectively.

(a) Would you be surprised to hear someone claim that the mean life span of the lithium batteries used in this model calculator is 1270 hours? Explain.

(b) Find the sample variance and sample standard deviation for these data.

Sec 6.4 (Boxplots) not in syllabus.