Essentials of Data Science With R Software - 1

Probability Theory and Statistical Inference

Univariate Random Variables

:::

Lecture 22

Random Variables - Discrete and Continuous

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Need of Random Variable:

Some concepts are required to draw statistical conclusions from a sample of data about a population of interest.

For example, suppose we know the starting salary of a sample of 100 students graduating in statistics. We can use this knowledge to draw conclusions about the expected salary for the population of all students graduating in statistics.

Need of Random Variable:

If a newly developed drug is given to a sample of selected Covid patients, then some patients may show improvement and some patients may not, but we are interested in the consequences for the entire population of patients.

Random variables are the foundation of the theoretical concepts required for making such conclusions.

They form the basis for statistical tests and inference

Need of Random Variable:

Suppose we toss a coin and it’s outcomes are Head and Tail.

So the sample space is Ω = {ω: where ω is Head or ω is Tail}.

It is difficult to handle the elements if they are not the numbers.

Need of Random Variable:

We can solve this issue as follows:

Let *X* be a function such that

$$X(\omega) = \begin{cases} 1 & \text{if}\omega \text{ is Head} \\ 0 & \text{if}\omega \text{ is Tail} \end{cases}$$

Thus *X* is a real valued function defined on Ω which takes us from Ω to a space of real numbers {0, 1}.

***X* is called a random variable.**

So the head is denoted by 0 and tail is denoted by 1.

Sample space (Ω) = {0, 1}

Need of Random Variable:

This can be simulated in R by the `sample` command by drawing one observation between 0 and 1 by simple random sampling with replacement.

```
sample(c(0,1), size=1, replace = T)
```

Need of Random Variable:

The outcome of this experiment is observed as follows:

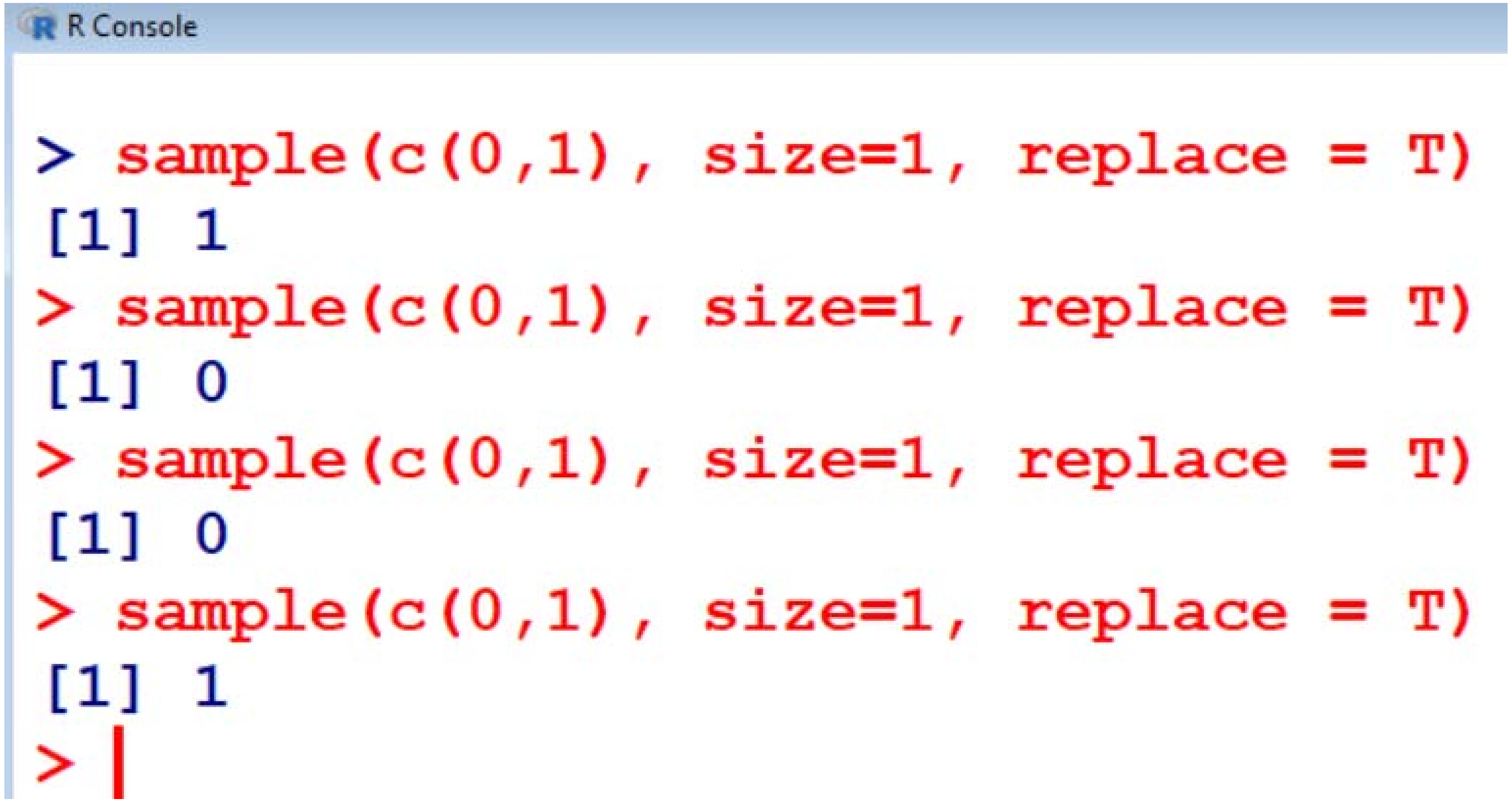
```
> sample(c(0,1), size=1, replace = T)
[1] 1

> sample(c(0,1), size=1, replace = T)
[1] 0

> sample(c(0,1), size=1, replace = T)
[1] 0

> sample(c(0,1), size=1, replace = T)
[1] 1
```

Need of Random Variable:



```
> sample(c(0,1), size=1, replace = T)
[1] 1
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 1
> |
```

Need of Random Variable:

In any random experiment, we are interested in the value of some numerical quantity determined by the result.

We are not interested in all the details of the experiments.

These quantities of interest that are determined by the result of the experiment are known as *random variables*.

Need of Random Variable:

Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities of its possible values.

For example

$$P(X = 0) = P(\text{Head}) = 1/2$$

$$P(X = 1) = P(\text{Tail}) = 1/2$$

We may therefore view *X* as a random variable which collects the possible outcomes of a random experiment and captures the uncertainty associated with them.

Need of Random Variable:

For example, suppose we toss two dice and suppose we want to study about the sum of the points on the upper face to be 7.

Our interest will not be in the individual data points (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) or (6, 1).

These quantities of interest that are determined by the result of the experiment are known as *random variables*.

Need of Random Variable:

$P\{X = 2) = P\{(1, 1)\} = 1/36$

$P(X = 3) = P\{(1, 2), (2, 1)\} = 2/36$

$P(X = 4) = P\{(1, 3), (2, 2), (3, 1)\} = 3/36$

$P(X = 5) = P\{(1, 4), (2, 3), (3, 2), (4, 1)\} = 4/36$

$P(X = 6) = P\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = 5/36$

$P(X = 7) = P\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} = 6/36$

$P(X = 8) = P\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} = 5/36$

$P(X = 9) = P\{(3, 6), (4, 5), (5, 4), (6, 3)\} = 4/36$

$P(X = 10) = P\{(4, 6), (5, 5), (6, 4)\} = 3/36$

$P(X = 11) = P\{(5, 6), (6, 5)\} = 2/36$

$P(X = 12) = P\{(6, 6)\} = 1/36$

Need of Random Variable:

In other words, the random variable *X* can take on any integral value between 2 and 12 and the probability that it takes on each value is obtained.

Let Ω represent the sample space of a random experiment, and let *R* be the set of real numbers.

Then the random variable *X* is assigning one and only one number to each element $\omega \in \Omega$, $X(\omega) = x$, $x \in R$, i.e. $X : \Omega \rightarrow R$.

For example, in $P(X = 2) = P\{(1, 1)\} = 1/36$, the random variable X assigns $X = 2$ to element $\omega = (1, 1)$.

Need of Random Variable:

In other words, the random variable *X* can take on any integral value between 2 and 12 and the probability that it takes on each value is obtained.

Since *X* must take on some value, we must have

$$1 = \boldsymbol{P}\left(\Omega\right) = \boldsymbol{P}\left(\bigcup_{\boldsymbol{i}=2}^{12}\left\{\boldsymbol{X} = \boldsymbol{i}\right\}\right) = \sum_{\boldsymbol{i}=2}^{12}\boldsymbol{P}\left\{\boldsymbol{X} = \boldsymbol{i}\right\}$$

Random Variable:

Let Ω represent the sample space of a random experiment, and let R be the set of real numbers.

A random variable is a function X which assigns to each element $\omega \in \Omega$ one and only one number

$$X(\omega) = x,\ x \in R, \text{ i.e. } X : \Omega \rightarrow R.$$

Random Variable:

It is a convention to denote random variables by capital letters (e.g. X) and their values by small letters (e.g. x).

If X is height of students, then x = 168 Centimetre is the value of X.

Similarly, x_1 = 170 centimetre indicates the first value of X and

x_2 = 180 centimetre indicates the second value of X.

Discrete Random Variable:

Random variables whose set of possible values can be written either as a finite sequence x_1, x_2, . . . ,x_n, or as an infinite sequence x_1, x_2, . . . are said to be *discrete*.

A sample space is discrete if it consists of a finite or countable infinite set of outcomes.

For instance, a random variable whose set of possible values is the set of nonnegative integers is a discrete random variable.

Discrete Random Variables:

Example: **A customer care phone contains 30 external lines.**

At a particular time, the system is observed, and some of the lines are being used.

Let the random variable *X* denote the number of lines in use.

Then, *X* can assume any of the integer values 0 through 30.

When the system is observed, if 5 lines are in use, $x = 5$.

Discrete Random Variables:

Example:

A batch of 1000 machined parts contains 20 that do not conform to customer requirements.

The random variable is the number of parts in a sample of five parts that do not conform to customer requirements.

Continuous Random Variable:

Suppose a dimensional length is measured such as vibrations, temperature fluctuations, calibrations, cutting tool wear, bearing wear, and raw material changes.

In practice, there can be small variations in the measurements due to many causes.

In an experiment like this, the measurement is represented as a random variable *X* and it is reasonable to model the range of possible values of *X* with an interval of real numbers.

The model provides for any precision in length measurements.

Continuous Random Variable:

There also exist random variables that take on a continuum of possible values.

These are known as *continuous* random variables.

One example is the random variable denoting the lifetime of a bulb, when the bulb's lifetime is assumed to take on any value in some interval (*a, b*), *a* > 0, *b* > 0.

Random Variable:

It is evident from the coin toss or the rolling of dice experiment that we need to know $P(X = x)$ to describe the respective random variable.

We assume up to now that we have this knowledge.

However, a sample of data can be used to estimate unknown probabilities and other quantities given a prespecified uncertainty level.

Random Variable:

A sample space is continuous if it contains an interval (either finite or infinite) of real numbers.

A sample space is discrete if it consists of a finite or countable infinite set of outcomes.

Random Variable:

More generally, we can say that it is mandatory to know $P(X \in A)$ for all possible A which are subsets of R.

If we choose $A = (-\infty, x]$, $x \in R$, we have

$$
\begin{aligned}
P(X \in A) &= P(X \in (-\infty, x]) \\
&= P(-\infty < X \leq x) \\
&= P(X \leq x).
\end{aligned}
$$

This consideration gives rise to the definition of the cumulative distribution function.