STATISTICS FOR DATA SCIENCE

Random Variables

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Random Variables

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- Random Variables
- Types of Random Variables
- Discrete Random Variables

STATISTICS FOR DATA SCIENCE

Random Variables

- A random variable is the outcome of an experiment (i.e. a random process) expressed as a number.
- Assigns numerical value to each outcome.

Looking Ahead: In Inference, we'll want to draw conclusions about population proportion or mean, based on sample proportion or mean. To accomplish this, we will explore how sample proportion or mean behave in repeated samples. If the samples are random, sample proportion or sample mean are **random variables.**

STATISTICS FOR DATA SCIENCE

Random Variables

Consider 4 sequential births.

S={BBBB,BBBG,BBGB,BBGG,BGBB,BGBG,BGGB,BGGG,
GBBB,GBBG,GBGB,GBGG,GGBB,GGBG,GGGB,GGGG}

Probability of each outcome is 1/16.

Now, count the number of girls in each set of four sequential births and assign a number based on number of girls.

BBBB(0),BBBG(1),BBGB(1),BBGG(2),BGBB(1),BGBG(2),BGGB(2)
BGGG(3),GBBB(1),GBBG(2),GBGB(2),GBGG(3),GGBB(2),GGBG(3)
GGGB(3),GGGG(4)

Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Random Variables

BBBB(0),BBBG(1),BBGB(1),BBGG(2),BGBB(1),BGBG(2),BGGB(2)
BGGG(3),GBBB(1),GBBG(2),GBGB(2),GBGG(3),GGBB(2),GGBG(3)
GGGB(3),GGGG(4)

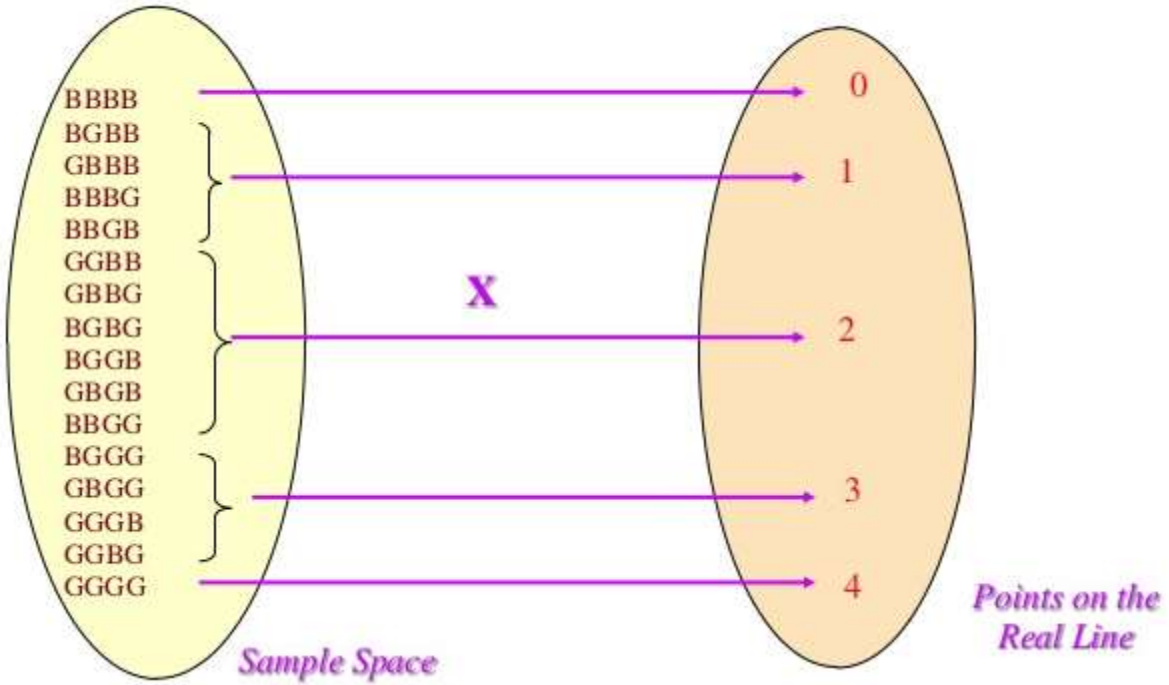
Note:
Each possible outcome is assigned a single numeric value.
All outcomes are assigned a numeric value.
The value assigned varies over the outcome.

The **count of the number of girls** is a **random variable.**

Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Random Variables



Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Random Variables

A **random variable** is a **variable** whose **value is determined by chance.**

(OR)

A **random variable** is a **quantitative variable** whose **value depends on a chance in some way.**

A random variable is the outcome of an experiment (i.e. a random process) expressed as a number.

A **random variable** assigns a numerical value to each outcome in a sample space.

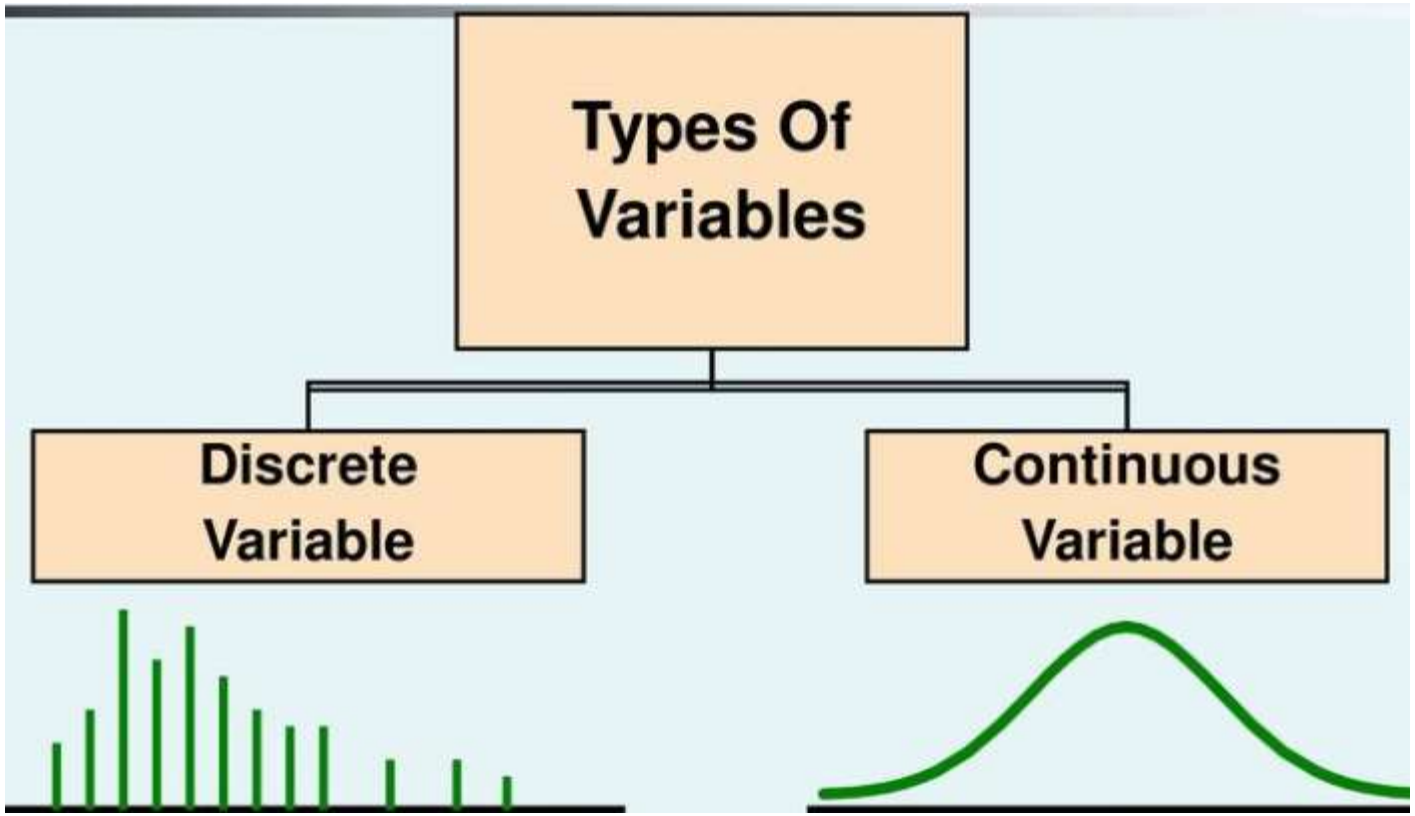
Source: http://clipart-library.com/

Example – Random variables

- Refer Text-Book-Chapter 2.4- Page.No:90 & 91

STATISTICS FOR DATA SCIENCE

Types of Random Variables



Text-Book-Chapter 2.4- Page.No:91

Source: www.slideshare.net

Types of Random Variables

Discrete Random Variables:

- Comes from a discreteset.
- Whole numbers.
- Takes a countable number ofpossible values.
- Ex-[1,2]

Continuous Random Variables:

- Can take on an infinite number of possible values, corresponding to every value in an interval.
- Ex-[2.5,3.6]

STATISTICS FOR DATA SCIENCE

Types of Random Variables

Discrete random variable

- Sample space is finite or countably many elements
- The probability function f(x) Is often tabulated
- Calculation of probabilities

$$P(a < X < b) = \sum_{a<t<b} f(t)$$

Continuous random variable

- Sample space has infinitely many elements
- The density function f(x) is a continuous function
- Calculation of probabilities

$$P(a < X < b) = \int_a^b f(t)\,dt$$

Source: www.slideshare.net

Problems to be solved

Identify the type of random variables:

1. No. of. customers who visits a bank every day.
2. Volume of milk produced by cow.
3. Time between lightening strikes ina thunderstorm.
4. Number of free throws an NBA player makes in his next 20 attempts.

STATISTICS FOR DATA SCIENCE

Discrete Random variables

- A random variable is discrete if its possible values form a discrete set.
- This means if the possible values are arranged in order, there is a gap between each value and the next one.
- The set of possible values may be infinite.

  Example:
- Set of all integers
- Set of all positive integers.

Example

Tossing two coins simultaneously. Findthe probability of getting an head?

Solution:

X=No. Of heads	Probability
0	1/4=0.25
1	1/2=0.5
2	1/4=0.25
Total	1

Probability Mass Function

$p(x) = P(X = x)$

Where,

X is the random variable

x is the value of the random variable.

1) $0 <= p(x) <= 1$

2) $\sum_{x} p(x) = 1$

Probability mass function is sometimes called the

probability distribution.

Examples

The number of flaws in a 1-inch length of copper wire varies from wire to wire. Overall, 48% of the wires produced have no flaws.39% have one flaw. 12%have two flaws and 1% have three flaws.

Write the Probability distribution of X, where X represents the no. of flaws in thewire.

Examples

Solution:

Let X be the no.of.flaws in a randomly selected piece of wire.

X	P(X)
0	0.48
1	0.39
2	0.12
3	0.01

Note:

The probability mass function over all the corresponding random variables is always equal to 1.

STATISTICS FOR DATA SCIENCE

Cumulative Distribution Function

- A function called the **cumulative distribution function** (cdf) specifies the probability that a random variable is less than or equal to a given value.
- The cumulative distribution function of the random variable X is the function $F(x) = P(X \leq x)$.

Example

Specifies the probability that a random variable is less than or equal to a given value.

$$F(x) = P(X <= x)$$

For previous example,

calculate

F(1)=P(X<=1)?

P(X<=1)= P(0)+P(1)

X	P(X)
0	0.48
1	0.39
2	0.12
3	0.01

Example

Solution:

P(X<=1)= P(0)+P(1)

X	P(X)
0	0.48
1	0.39
2	0.12
3	0.01

$$F(x) = P(X \leq x) = \sum_{t \leq x} p(t)$$

F(1)=P(X<=1)= P(0)+P(1)

= 0.48+0.39

= 0.87

Mean for Discrete Random Variable

- Let X be a discrete random variable with probability mass function . $p(x) = P(X = x)$
- The **mean** of X is given by,

$$\mu_X = \sum_x xP(X = x)$$

where the sum is over all possible values of X.

- The mean of X is sometimes called the expectation, or expected value, of X and may also be denoted by $E(X)$ or by μ.

STATISTICS FOR DATA SCIENCE

Example - Mean for Discrete Random Variable

A certain industrial process is brought down for recalibration whenever the quality of the items produced falls below specifications. Let X represent the number of times the process is recalibrated during a week, and assume that X has the following probability mass function.

x	0	1	2	3	4
$p(x)$	0.35	0.25	0.20	0.15	0.05

Find the mean of X.

Example - Mean for Discrete Random Variable

Solution:

The mean of X is given by:

$$\mu_X = \sum_x xP(X = x)$$

$$\mu_X = 0(0.35) + 1(0.25) + 2(0.20) + 3(0.15) + 4(0.05) = 1.30$$

Variance for Discrete Random Variables

- Let X be a discrete random variable with probability mass function $p(x) = P(X = x)$
- The **variance** of X is given by

$$\sigma_x^2 = \sum_x (x - \mu_x)^2 P(X = x)$$

$$\sum_x x^2 P(X = x) - \mu_x^2.$$

- The variance of X may also be denoted by $V(X)$ or by $\sigma 2$.
- The standard deviation is the square root of the variance.

$$\sigma^2 = \sqrt{\sigma_x^2}$$

Example - Variance for Discrete Random Variable

For the earlier example, find variance and standard deviation for the random variable X.

Solution:

We compute the variance by using

$$\sigma_X^2 = \sum_x (x - \mu_X)^2 P(X = x)$$

$$\begin{aligned}\sigma_X^2 &= (0 - 1.30)^2 P(X = 0) + (1 - 1.30)^2 P(X = 1) + (2 - 1.30)^2 P(X = 2) \\ &\quad + (3 - 1.30)^2 P(X = 3) + (4 - 1.30)^2 P(X = 4) \\ &= (1.69)(0.35) + (0.09)(0.25) + (0.49)(0.20) + (2.89)(0.15) + (7.29)(0.05) \\ &= 1.51\end{aligned}$$

The standard deviation is $\sigma_X = \sqrt{1.51} = 1.23$.

Example

A resistor in a certain circuit is specified to have a resistance in the range 99 Ω–101 Ω. An engineer obtains two resistors. The probability that both of them meet the specification is 0.36, the probability that exactly one of them meets the specification is 0.48, and the probability that neither of them meets the specification is 0.16. Let X represent the number of resistors that meet the specification. Find the probability mass function, and the mean, variance, and standard deviation of X.

Solution

The probability mass function is $P(X = 0) = 0.16$, $P(X = 1) = 0.48$, $P(X = 2) = 0.36$, and $P(X = x) = 0$ for $x \neq 0$, 1, or 2. The mean is

$$\mu_X = (0)(0.16) + (1)(0.48) + (2)(0.36)$$

$$= 1.200$$

Example

The variance is

$$\sigma_X^2 = (0 - 1.200)^2(0.16) + (1 - 1.200)^2(0.48) + (2 - 1.200)^2(0.36)$$

$$= 0.4800$$

The standard deviation is $\sigma_X = \sqrt{0.4800} = 0.693$.

Probability Histogram

- When the possible values of a discrete random variable are evenly spaced, the probability mass function can be represented by a histogram, with rectangles centered at the possible values of the random variable.
- The area of the rectangle centered at a value x is equal to

  $P(X = x)$.

- Such a histogram is called a probability histogram, because the areas represent probabilities.

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering