STATISTICS FOR DATA SCIENCE

Random Variables

Prof. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Random Variables

Prof. Uma D

Topics to be covered...

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

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

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)

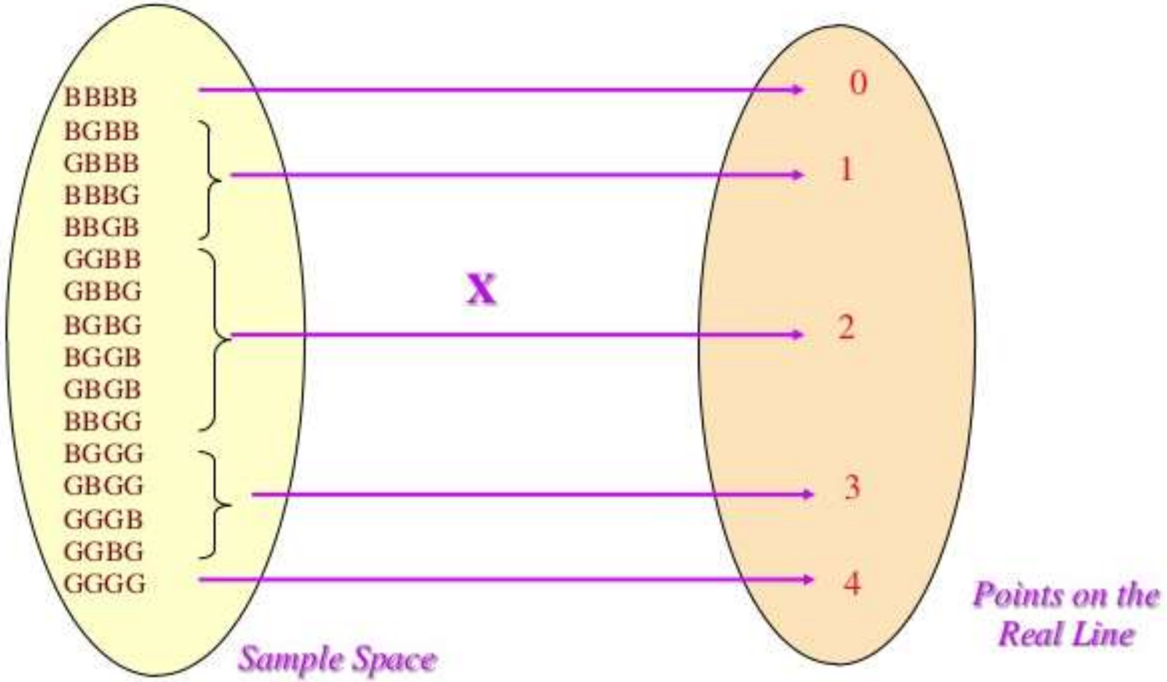
Note:
Each possible outcome is assigned a single numeric value.
All outcomes are assigned a numeric value.
The value assigned varies over the outcome.

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

Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Random Variables



Source: slideshare.net

Random Variables

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

(OR)

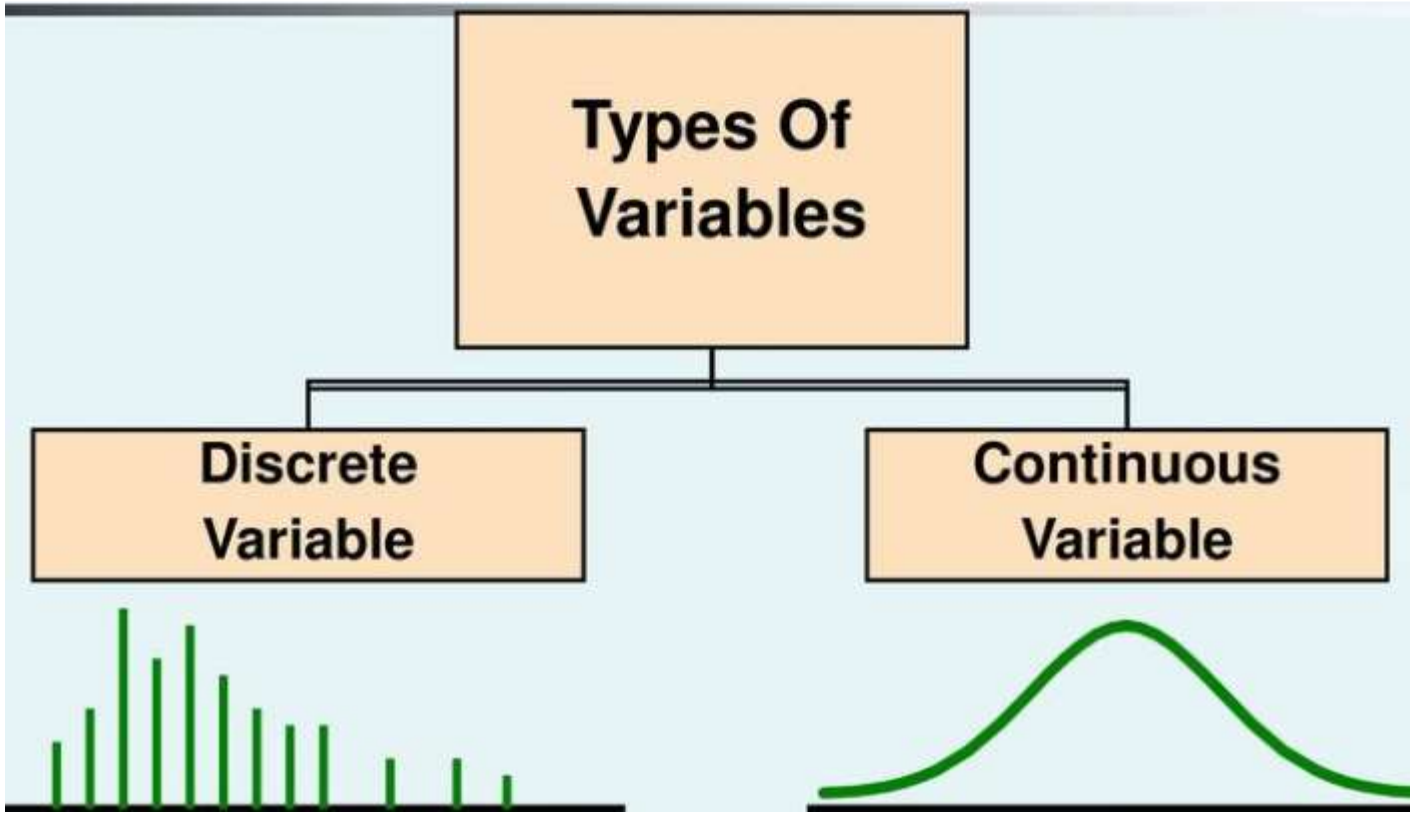
A **random variable** is a **quantitative variable** whose **value depends on a chance in some way.**

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/

Types of Random Variables



Types Of Variables

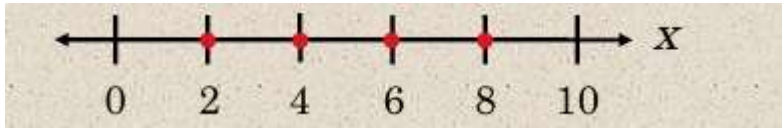
Discrete Variable

Continuous Variable

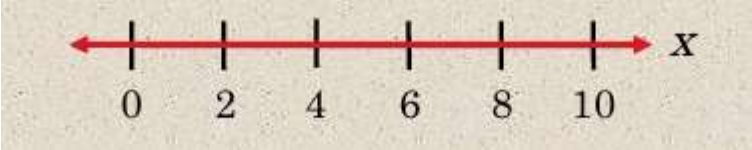
Source: www.slideshare.net

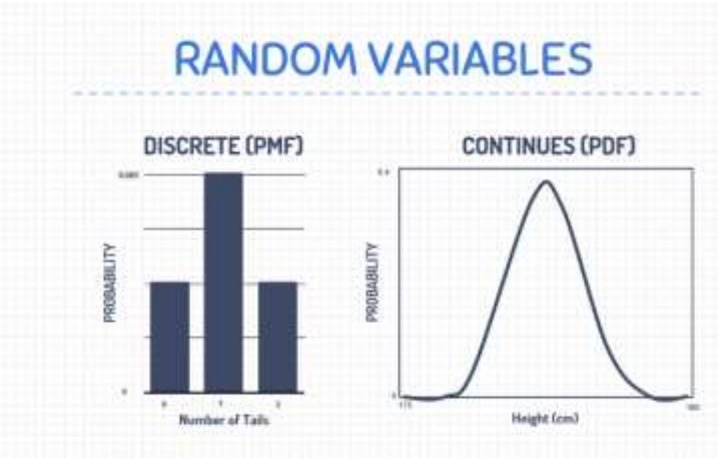
Random Variables

A random variable is **discrete** if it has a finite or countable number of possible outcomes that can be listed.



A random variable is **continuous** if it has an uncountable number of possible outcomes, represented by the intervals on a number line.





0

Source: slideshare.net

Types of Random Variables

A **discrete random variable** is one which can take on a **countable number** of distinct values such as 1, 2, 3, 4 …….100….

The number associated with discrete random variables is the set of **counting** numbers {1,2,3,4,…….} including {0}.

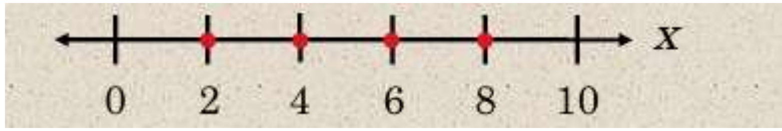
Discrete random variables have a **countable number of outcomes.**

Examples: **Dead/alive**, **dice, counts,** etc.

Types of Random Variables

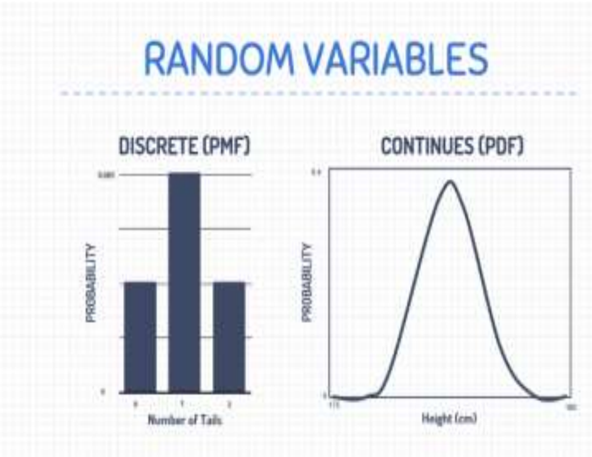
Discrete Random Variables:

- Comes from a discrete set.
- Whole numbers.
- Takes a countable number of possible values. Ex.: [1, 2]
- Has discrete jumps(or gaps) between successive values
- Has measurable probability associated with individual values counts.



Source Image: medium.com,slideshare.net

Types of Random Variables



Continuous Random Variables

Can take on an **infinite number of possible values**, corresponding to **every value** in an **interval.**

Example: [4.3,6.2]

Continuous random variables are usually **measurements** such as weight, height, temperature, income, etc.

Examples: **blood pressure**, **weight,** the **speed of a car**, the **real numbers** from **1 to 6**.

Source Image: medium. com

Probability Distribution

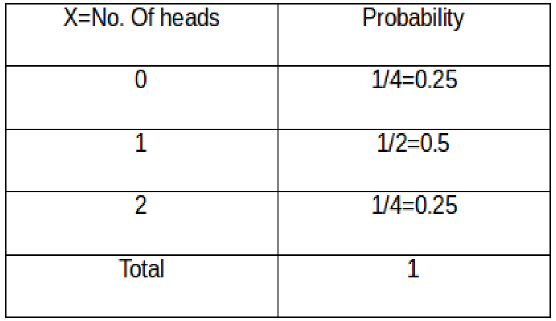
A **probability distribution** lists each possible value the random variable can assume, together with its probability.

Helps in finding all the possible values a random variable can take between the minimum and maximum possible values.

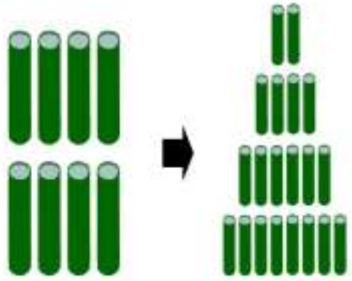
It is a way to shape the sample data to make predictions and draw conclusions about an entire population.

It refers to the frequency at which some events or experiments occur.

Used to **model real-life events** for which the outcome is uncertain.



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

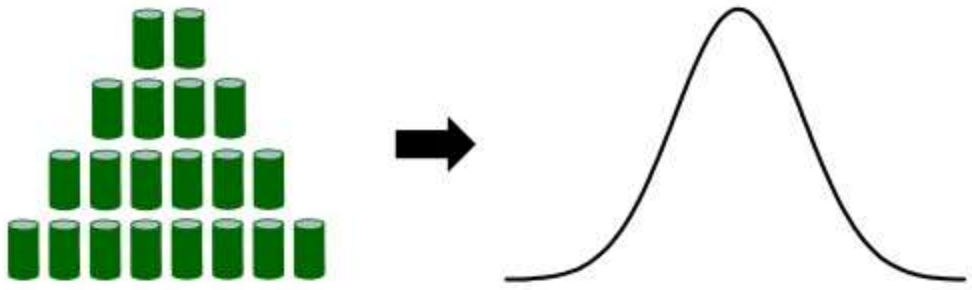


Source: www.citoolkit.com

Probability Distribution:

You might be certain if you examine the whole population.

To draw conclusions from sample data, you should compare **values obtained** from the **sample with** the **theoretical values** obtained from the **probability distribution.**

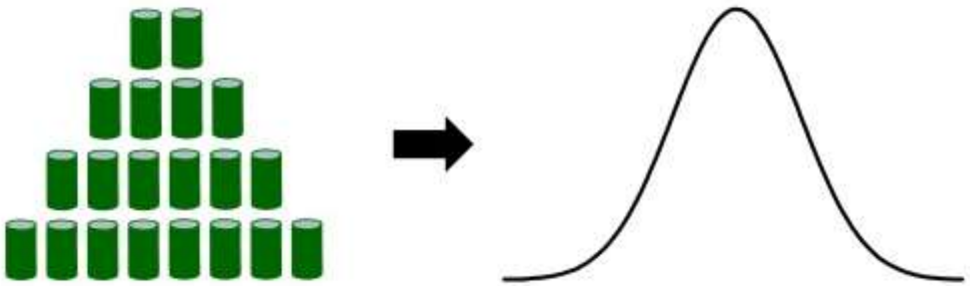


Source: www.citoolkit.com

Probability Distribution:

There will always be a **risk of drawing false conclusions** or making **false predictions**.

We need to be **sufficiently confident** before taking any decision by setting confidence levels(90 or 95 or 98).



Source: www.citoolkit.com

Use of Probability Distribution:

Insurance managers may use them to forecast the uncertain future claims.

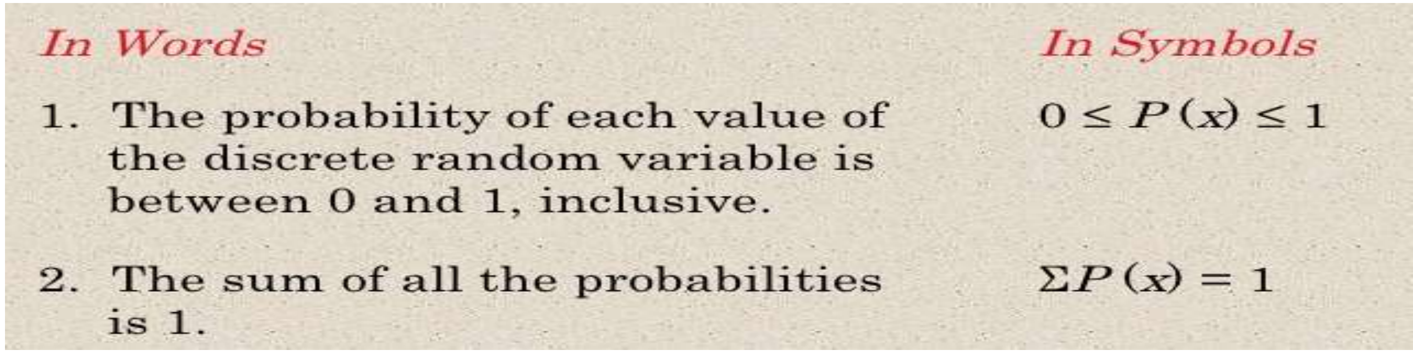
Fund managers may use them to determine the possible returns a stock may earn in the future.

Restaurant managers may use them to resolve future customer complaints.

Source: www.citoolkit.com

Probability Functions of a Discrete Random Variable

A probability distribution function must satisfy the following conditions.



In Words	*In Symbols*
1. The probability of each value of the discrete random variable is between 0 and 1, inclusive.	$0 \leq P(x) \leq 1$
2. The sum of all the probabilities is 1.	$\Sigma P(x) = 1$

Probability Mass Function /Probability Distribution p(x)= P(X=x)

Cumulative Distribution Function F(x)=P(X<=x)

Source: slideshare.net

Probability Mass Function

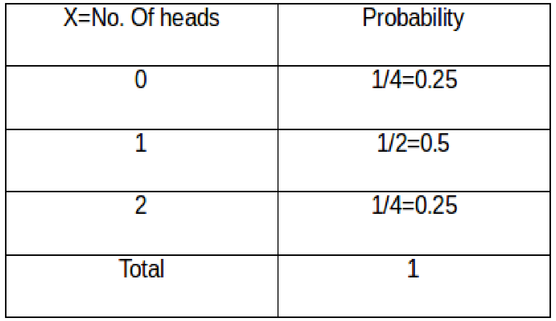
A **probability function** maps the possible values of x against their respective **probabilities of occurrence, p(x) .** p(x) is a number from 0 to 1.0

The description of the **possible values of X** and the **probability** of each value.

The probability mass function specifies the probability that a **random variable** is equal to **a given value**. i.e. **P(X=x)**.

The area under a probability function is always 1.

.

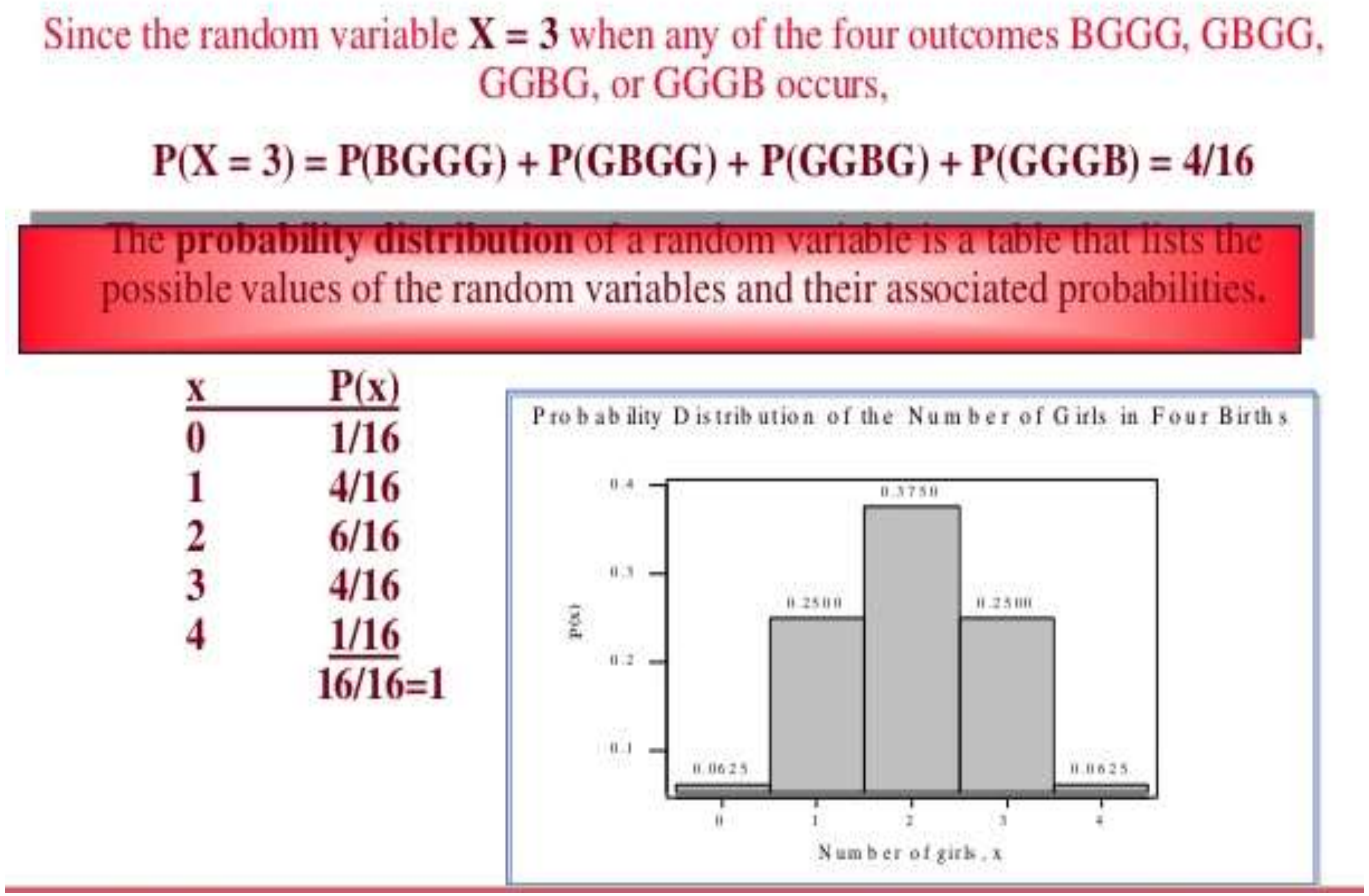


X=No. Of heads	Probability
0	1/4=0.25
1	1/2=0.5
2	1/4=0.25
Total	1

Source: slideshare.net

STATISTICS FOR DATA SCIENCE

Random Variables



Since the random variable **X = 3** when any of the four outcomes BGGG, GBGG, GGBG, or GGGB occurs,

$$P(X = 3) = P(BGGG) + P(GBGG) + P(GGBG) + P(GGGB) = 4/16$$

The **probability distribution** of a random variable is a table that lists the possible values of the random variables and their associated probabilities.

x	P(x)
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16
	16/16=1

Probability Distribution of the Number of Girls in Four Births
P(x)
0.4
0.3
0.2
0.1
0.0625
0.2500
0.3750
0.2500
0.0625
0
1
2
3
4
Number of girls, x

Source: slideshare.net

Probability Functions of a Discrete Random Variable

Cumulative Distribution Function

A cumulative distribution function specifies the probability that a random variable is **less than or equal** to a **given value**. i.e. **P(X<=x).**

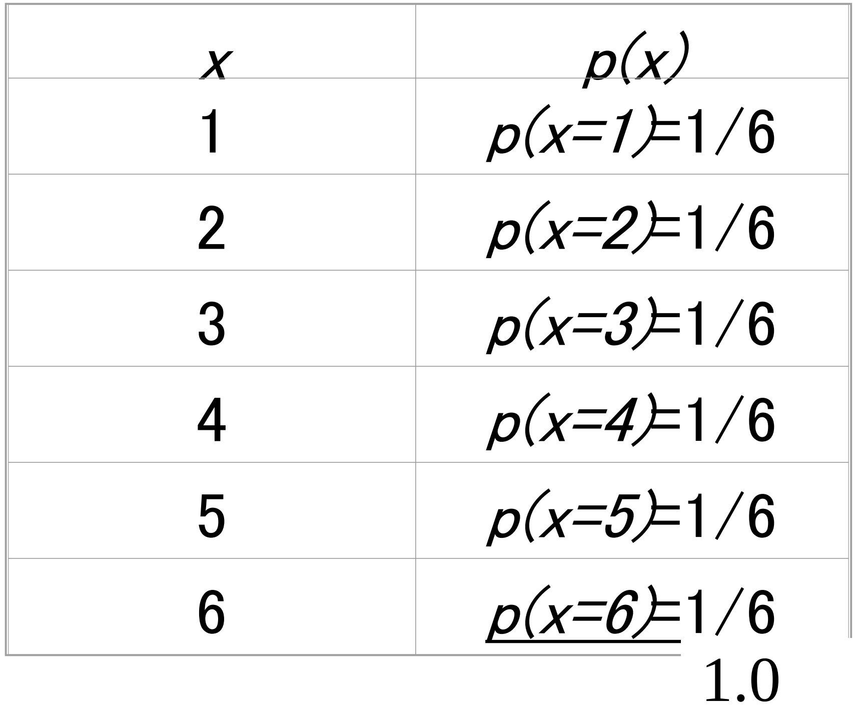
The cdf of a r.v. is the function F(x) = P(X<=x)

Source: slideshare.net

Probability Mass vs. Cumulative Distribution Functions

Probability Mass Function

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

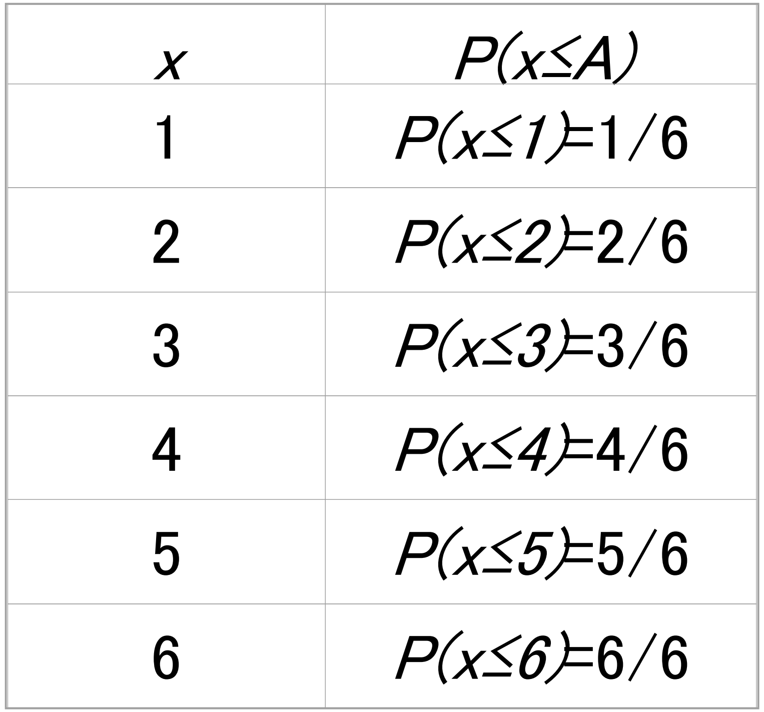


x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$

1.0

Cumulative Distribution Function

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



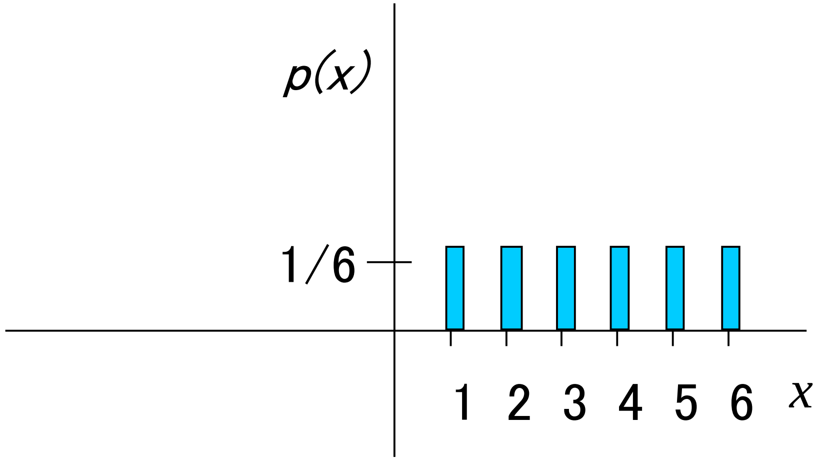
x	$P(x \leq A)$
1	$P(x \leq 1)=1/6$
2	$P(x \leq 2)=2/6$
3	$P(x \leq 3)=3/6$
4	$P(x \leq 4)=4/6$
5	$P(x \leq 5)=5/6$
6	$P(x \leq 6)=6/6$

Source: slideshare.net

Probability Mass vs. Cumulative Distribution Functions

Probability Mass Function

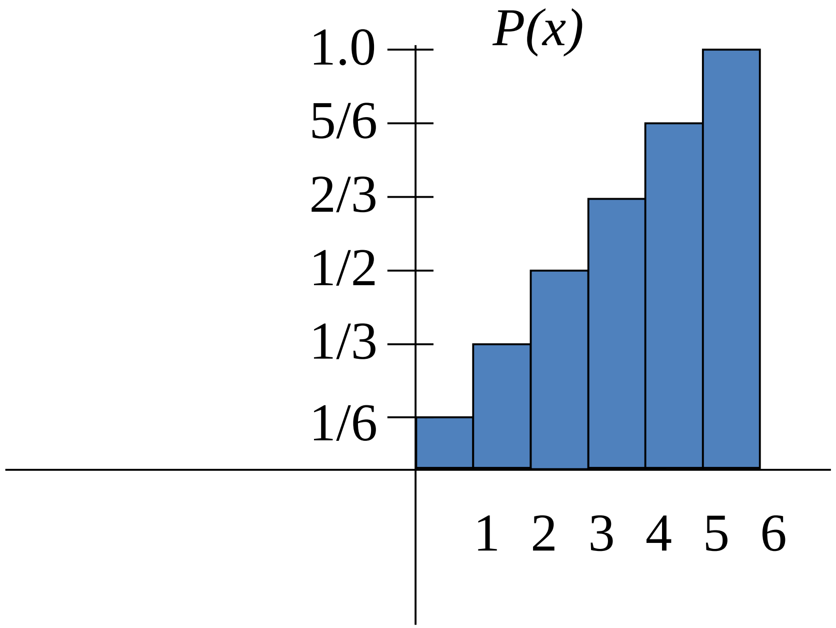
$p(x) = P(X=x)$



$$\sum_{\text{all } x} P(x) = 1$$

Cumulative Distribution Function

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



Source: slideshare.net

Problems

The number of patients seen in the ER in any given hour is a random variable represented by *X*. The probability distribution for *x* is:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14)= .1$

b. At least 12 patients arrive $p(x\geq 12)= (.2 + .1 +.1) = .4$

c. At most 11 patients arrive $p(x\leq 11)= (.4 +.2) = .6$

Source: slideshare.net

Problems

Computer chips often contain surface imperfections. For a certain type of computer chip, 9% contain no imperfections, 22% contain 1 imperfection, 26% contain 2 imperfections, 20% contain 3 imperfections, 12% contain 4 imperfections, and the remaining 11% contain 5 imperfections.

Define a random variable and find the possible values for it.

Is RV discrete or continuous?

Problems

Let X = Number of imperfections in a randomly chosen chip.

X	:	0	1	2	3	4	5
P(X=x)	:	0.09	0.22	0.26	0.20	0.12	0.11

Expected Value and Variance

All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).

Expected value is an **extremely useful** concept for good decision-making!

Expected Value of a Random Variable

Expected value is just the average or mean (μ) of random variable *x*.

It's sometimes called a "weighted average" because more frequent values of X are weighted more highly in the average.

It's also how we expect X to behave on-average over the long run.

Expected Value

Discrete case:

$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

Continuous case:

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

E(X) = μ

these symbols are used interchangeably

Recall the following probability distribution of ER arrivals:

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1

$$\boxed{\sum_{i=1}^{5} x_i\, p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3}$$

Variance

σ^2=Var(x) =E[(x-μ)^2]

"The expected (or average) squared distance (or deviation) from the mean"

$$\boxed{\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)}$$

Var(X)= σ^2

SD(X) = σ

Variance

Discrete Random Variable:

$$Var(X) = \sum_{\text{all } x} (x_i - \mu)^2 \, p(x_i)$$

Continuous Random Variable:

$$Var(X) = \int_{\text{all } x} (x_i - \mu)^2 \, p(x_i) dx$$

Source: www.slideshare.net

Discrete Random Variables: Problems

Example: The lottery

The Lottery (also known as a tax on people who are bad at math…)

A certain lottery works by picking 6 numbers from 1 to 49. It costs $1.00 to play the lottery, and if you win, you win $2 million after taxes.

If you play the lottery once, what are your expected winnings or losses?

Source: www.slideshare.net

Discrete Random Variables: Problems

$$\frac{1}{\binom{49}{6}} = \frac{1}{\frac{49!}{43!6!}} = \frac{1}{13{,}983{,}816} = 7.2 \text{ x } 10^{-8}$$

“49 choose 6”

Out of 49 numbers, this is the number of distinct combinations of 6.

x$	*p(x)*
-1	.999999928
+ 2 million	*7.2 x 10*$^{-8}$

E(X) = P(win)*$2,000,000 + P(lose)*-$1.00

= 2.0 *x 10*^{6} ** 7.2 x 10*$^{-8}$*+ .999999928 (-1)*

= *.144 - .999999928 = -$.86*

Negative expected value is never good! You shouldn’t play if you expect to lose money!

Source: www.slideshare.net

Problems

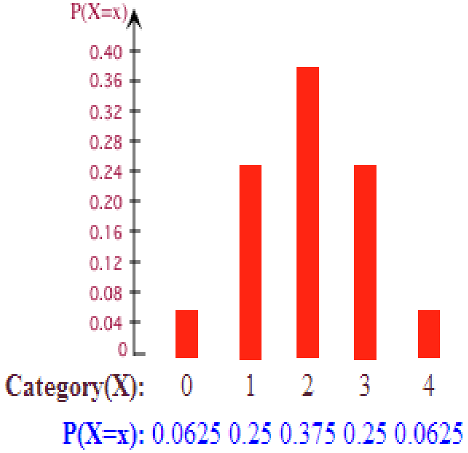
If you play the lottery every week for 10 years, what are your expected winnings or losses?

520 x (-.86) = -$447.20

Source: www.slideshare.net

Probability Histogram

When the possible values of a discrete r. v. are evenly spaced, the probability mass function can be represented by a histogram with rectangles centred at the possible values of a r. v.



The area of a rectangle centred at a value x is P(X=x).

It is often useful to display the data collected in an experiment in the form of a histogram.

Having a graphical representation is helpful because it allows the researcher to visualize what shape the distribution takes.

Source: www.slideshare.net

Problems

Do It Yourself !!!

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.

Do It Yourself !!!

Approximately 60% of full-term newborn babies develop jaundice. 2 full-term babies are randomly sampled. What is the probability distribution of X, if X represents the number that develops jaundice?

1. Define a random variable and find the possible values for it.
2. Is RV discrete or continuous?
3. Write the probability distribution of X , where X represents number of flaws in the wire.
4. Write cumulative distribution of X.
5. Find mean and variance of X.
6. Draw probability histogram for X.

Problems

Do It Yourself !!!

How casinos can afford to give so many free drinks… (Gambling)?

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet $1 that an odd number comes up, you win or lose $1 according to whether or not that event occurs. If random variable X denotes your net gain, X=1 with probability 18/38 and X= -1 with probability 20/38.

Is it good to play this game?

THANK YOU

Prof. Uma D

Department of Computer Science and Engineering