STATISTICS FOR DATA SCIENCE

Binomial Distribution

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

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Binomial Distribution

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

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Assume that you are tossing a coin 10 times.
You will get a number of heads between 0 and 10.

You may then carry out another 10 trials, in which you will also have a number of heads between 0 and 10.

By doing this many times, you will have a data set which has the **shape of the binomial distribution**.

Source: www.citoolkit.com

STATISTICS FOR DATA SCIENCE

Binomial Distribution

If a total of n Bernoulli trials are conducted:

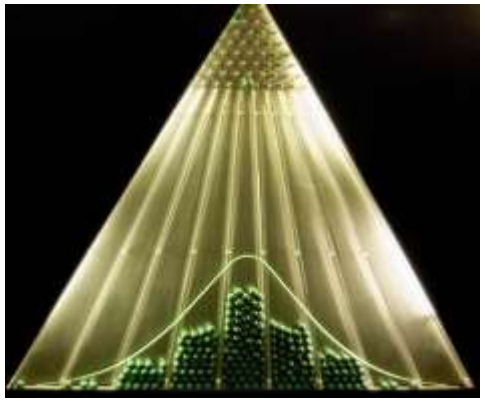
- Trials are independent. (fixed number of trials)
- Each trial has only two possible outcomes – its a Bernoulli trial
- Probability of success remains the same for each trial.
- X – represents the number of successes in n independent and identically distributed Bernoulli trials then,

X has the binomial distribution with parameters n and p $X \sim Bin(n, p)$

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Binomial Distribution is a **discrete** probability distribution.



A total of n Bernoulli trials are conducted.
Let **X** represents the **number of successes** in n independent and identically distributed Bernoulli trials.

Then X has the binomial distribution with **parameters n and p.**

$$X \sim Bin(n, p)$$

Source: www.mathstackexchange.com

Binomial Distribution

Binomial Random Variable = Sum of IID Bernoulli Random Variables

Total of n Bernoulli trials are conducted each with success probability 1.

Y, Y, ... Y - represent n Bernoulli Random Variables.

Hence for i = 1, 2 . . . n , $Y \sim Bernoulli(p)$

$Y_i = 1$ if the i^{th} trial is a success $Y_i = 0$ if the i^{th} trail is a failure

Let X represent no of successes among n trails.

$X = Y_1 + Y_2 + ... + Y_n$, $X \sim Bin(n, p)$.

This shows binomial random variable can be expressed as sum of Bernoulli random variables

Binomial or Not?

Select three people from a population and suppose that 10% of the population has the Alzhemier's gene. We select randomly 5 people. Is this a Binomial Experiment or not?

p=P(Alzhemier's gene)=0.1 ⟶ Binomial

2 out of 20 Laptops are defective. We randomly select 3 for testing. Is this a Binomial Experiment or not?

p=P(defective)=2/20
p=P(defective)=1/19 Not Binomial

Note : The independence is a key assumption that often violated in real life applications.

Examples

Which of the following are binomial experiments?

1. Telephone surveying a group of 200 people to ask if they voted for George Bush.
2. You take a survey of 50 traffic lights in a certain city, at 3 p.m., recording whether the light was red, green, or yellow at that time.

   (No of outcomes > 2)
3. Asking 100 people if they have ever been to Paris.

Examples

A coin is flipped 10 times. Let X be the no.of.heads that appear.

What is the probability distribution of X?

Solution

There are 10 independent Bernoulli trials, each with success probability $p = 0.5$. The random variable X is equal to the number of successes in the 10 trials. Therefore $X \sim \text{Bin}(10, 0.5)$.

Binomial or Not?

Consider 2 out of 20 PCs are defective. We randomly select 3 for testing. Is this a binomial experiment?

1. The experiment consists of n=3 identical trials
2. Each trial result in one of two outcomes
3. The probability of success (finding the defective) is 2/20 and remains the same
4. The trials are not independent.
5. For example, P(success on the 2nd trial | success on the 1st trial) = 1/19, not 2/20

Good Rule of Thumb

Rule of thumb:

If the sample size n is relatively large to the population size N,

- say n/N <= .05, the resulting experiment can use binomial distribution.
- say n/N > .05, the resulting experiment would not be binomial.

Example

A lot contains several thousand components. 10% of which are defective. Seven components are sampled from the lot.

Let X represent the no.of.defective components in the sample.

What is the distribution of X?

Example

Solution

Since the sample size is small compared to the population (i.e., less than 5%), the number of successes in the sample approximately follows a binomial distribution. Therefore we model X with the Bin(7, 0.1) distribution.

Probability Mass Function of a Binomial Random Variable

A biased coin has probability 0.6 of coming up heads. The coin is tossed three times.

Let X be the number of heads. Then $X \sim \text{Bin}(3, 0.6)$.

We will compute $P(X = 2)$.

Probability Mass Function of a Binomial Random Variable

- There are three arrangements of two heads in three tosses of coin,

  HHT,

  HTH,

  THH.

  We first compute the probability of HHT. The event HHT is a sequence of independent events: H on the first toss, H on the second toss, T on the third toss.

Probability Mass Function of a Binomial Random Variable

- We know the probabilities of each of these events separately:

$$P(\text{H on the first toss}) = 0.6,\ P(\text{H on the second toss}) = 0.6,\ P(\text{T on the third toss}) = 0.4$$

Since the events are independent, the probability that they all occur is equal to the product of their probabilities

Probability Mass Function of a Binomial Random Variable

$$P(\text{HHT}) = (0.6)(0.6)(0.4) = (0.6)^2(0.4)^1$$

Similarly, $P(\text{HTH}) = (0.6)(0.4)(0.6) = (0.6)^2(0.4)^1$, and $P(\text{THH}) = (0.4)(0.6)(0.6) = (0.6)^2(0.4)^1$. It is easy to see that all the different arrangements of two heads and one tail have the same probability. Now

$$\begin{aligned} P(X = 2) &= P(\text{HHT or HTH or THH}) \\ &= P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) \\ &= (0.6)^2(0.4)^1 + (0.6)^2(0.4)^1 + (0.6)^2(0.4)^1 \\ &= 3(0.6)^2(0.4)^1 \end{aligned}$$

Probability Mass Function of a Binomial Random Variable

- We see that the number 3 represents the number of arrangements of two successes (heads) and one failure (tails),
- 0.6 is the success probability p,
- the exponent 2 is the number of successes,
- 0.4 is the failure probability 1 − p,
- and the exponent 1 is the number of failures.
- We can now generalize this result to produce a formula for the probability of x successes in n independent Bernoulli trials with success probability p, in terms of x,n, and p. In other words, we can compute
- $P(X = x)$ where $X \sim Bin(n, p)$

Probability Mass Function of a Binomial Random Variable

$$P(X = x) = (\text{number of arrangements of } x \text{ successes in } n \text{ trials}) \cdot p^x (1-p)^{n-x}$$

If $X \sim \text{Bin}(n,\ p)$, the probability mass function of X is

$$p(x) = P(X = x) = \begin{cases} \dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \qquad (4.4)$$

Text-Book-Chapter 4.2- Page.No:206

Example

Find the probability mass function of the random variable X if

$X \sim Bin(10, 0.4)$.

Find $P(X = 5)$.

Solution

We use Equation(4.4) with $n = 10$ and $p = 0.4$. The probability mass function is

Example

$$P(X = x) = (\text{number of arrangements of } x \text{ successes in } n \text{ trials}) \cdot p^x (1-p)^{n-x}$$

If $X \sim \text{Bin}(n,\ p)$, the probability mass function of X is

$$p(x) = P(X = x) = \begin{cases} \dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \qquad (4.4)$$

$$p(x) = \begin{cases} \dfrac{10!}{x!(10-x)!} (0.4)^x (0.6)^{10-x} & x = 0, 1, \ldots, 10 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X = 5) = p(5) = \frac{10!}{5!(10-5)!} (0.4)^5 (0.6)^{10-5}$$

$$= 0.2007$$

Text-Book-Chapter 4.2- Page.No:207

Mean and Variance of Binomial Distribution

$$\text{Mean} = n * (0\,(1 - p) + 1 * p) = np$$

$$\text{Variance} = n* ((0 - p)^2\,(1 - p) + (1 - p)^2 \; *(p)\,)$$
$$= n\,p(1 - p)$$

Example

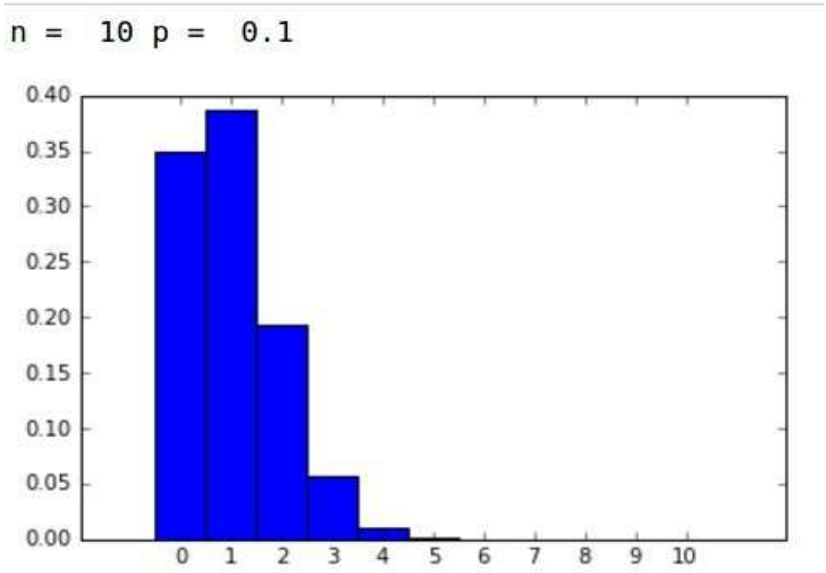
Find the effect of changing p when n is fixed and is small.

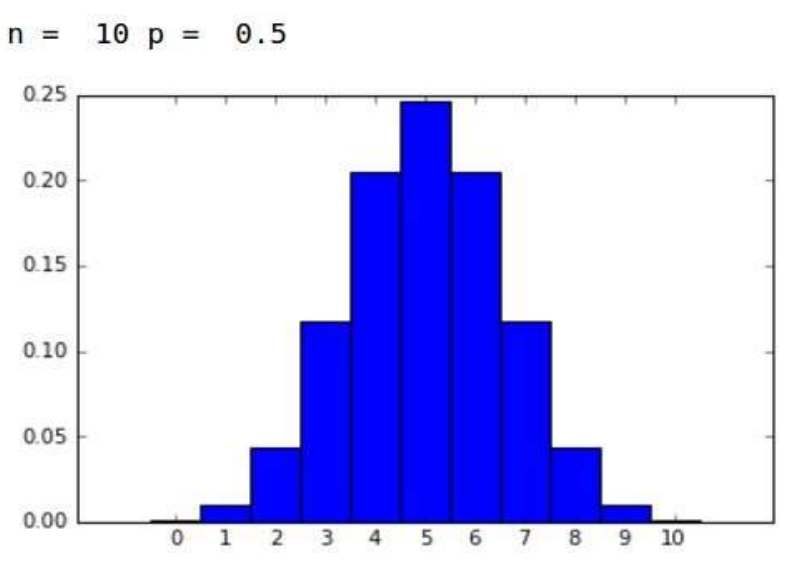
a) n = 10, p = 0.10

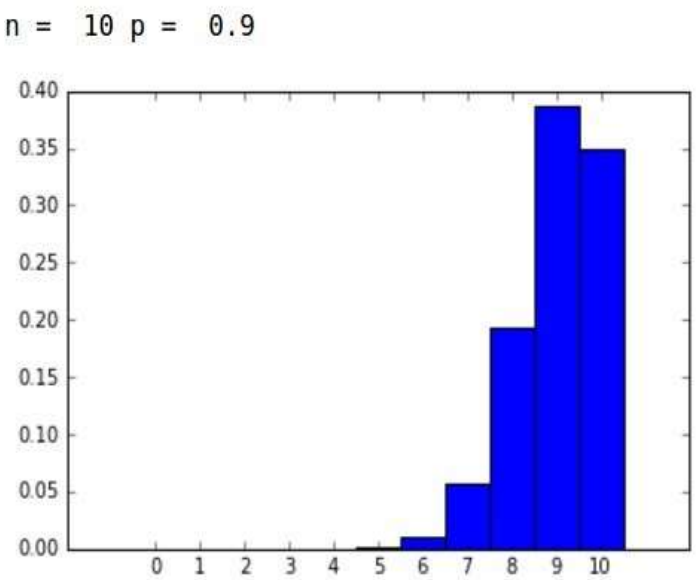
b) n = 10, p = 0.5

c) n = 10, p = 0.90

For small samples, binomial distributions are skewed when p is different from 0.5.







Using a Sample Proportion to Estimate a Success Probability

- Conduct n independent Bernoulli Trials.
- Count X – no of successes.
- Sample proportion – denotes estimated value of p

$$\hat{p} = \frac{\text{number of successes}}{\text{number of trials}} = \frac{X}{n}$$

Sample proportion is just an estimate of p and is not equal to p.

If we take another sample, the value of sample proportion might come out differently. Hence, there might be some uncertainty in the estimated value

Example

A quality engineer is testing the calibration of a machine that packs ice cream into containers. In a sample of 20 containers, 3 are underfilled. Estimate the probability p that the machine underfills a container.

Solution

The sample proportion of underfilled containers is

$p = 3/20 = 0.15$.

We estimate that the probability p that the machine underfills a container is 0.15 as well.

Uncertainty in the Sample Proportion

- It is important to realize that the sample proportion $\hat{p}$ is just an estimate of the success probability p, and in general, is not equal to p.
- If another sample were taken, the value of $\hat{p}$ would probably come out differently.
- In other words, there is uncertainty in $\hat{p}$
- For $\hat{p}$ to be useful, we must compute its bias and its uncertainty

Example

A quality engineer takes a random sample of 100 steel rods from a days production, and finds that 92 of them meet specifications.

1. Estimate the proportion of the days production that meets specifications.
2. Find the uncertainty in the estimate.
3. Estimate the no.of.rods that must be sampled to reduce the uncertainty to 1%?

Example

Solution:

1) Sample proportion = 92/100

2) Uncertainty = sqrt(p (1 – p) / n)
= sqrt(0.92 * 0.08 / 100)
= 0.027

3) Given, Uncertainty = 0.01 P =0.92 n = ?
n = p (1- p) / square(uncertainty)
= 0.92 * 0.08 / square(0.01)
= 736

Computing Bias

Bias – is intentional or unintentional favoring of one outcome over the other in the population.

In statistics, Bias of an estimator is the difference between estimator's expected value and true value of parameter being estimated.

μ_p - p

$$\mu_{\hat{p}} = \mu_{X/n} = \frac{\mu_X}{n} = \frac{np}{n} = p$$

Since $\mu_{\hat{p}} = p$, $\hat{p}$ is unbiased; in other words, its bias is 0.

Computing Uncertainty

Uncertainty – is the standard deviation of sample proportion.

The uncertainty is the standard deviation $\sigma_{\hat{p}}$.

deviation of X is $\sigma_X = \sqrt{np(1-p)}$. Since $\hat{p} = X/n$

$$\sigma_{\hat{p}} = \sigma_{X/n} = \frac{\sigma_X}{n}$$

$$= \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

In practice, when computing the uncertainty in $\hat{p}$, we don't know the success probability p, so we approximate it with $\hat{p}$.

Text-Book-Chapter 4.2- Page.No:210

Example

The safety commissioner in a large city wants to estimate the proportion of buildings in the city that are in violation of fire codes. A random sample of 40 buildings is chosen for inspection, and 4 of them are found to have fire code violations. Estimate the proportion of buildings in the city that have fire code violations, and find the uncertainty in the estimate.

Example

Solution

Let p denote the proportion of buildings in the city that have fire code violations. The sample size (number of trials) is n = 40. The number of buildings with violations (successes) is X = 4.

We estimate p with the sample proportion:

$$\widehat{p} = \frac{X}{n} = \frac{4}{40} = 0.10$$

Example

the uncertainty in $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Substituting $\hat{p} = 0.1$ for p and 40 for n, we obtain

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.10)(0.90)}{40}}$$

$$= 0.047$$

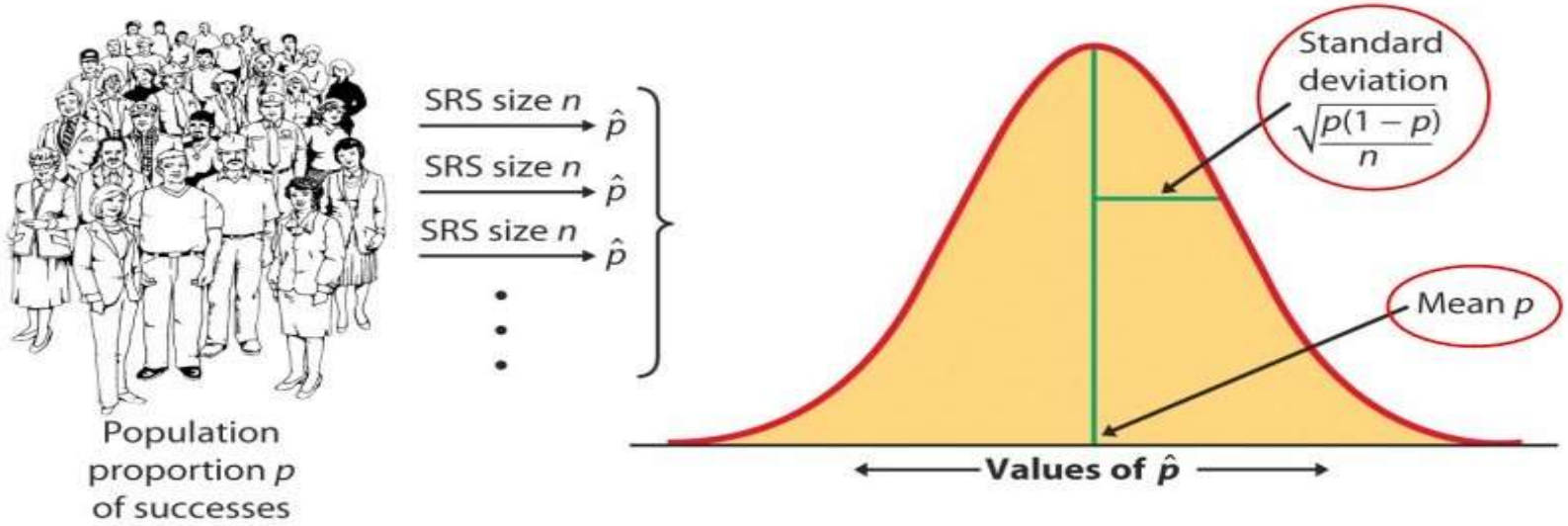
It turned out that the uncertainty in the sample proportion was rather large. We can reduce the uncertainty by increasing the sample size.

Sampling distribution of the Sample Proportion

Sampling distribution of the sample proportion

The sampling distribution of $\hat{p}$ is never exactly normal. But as the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately normal.

The normal approximation is most accurate for any fixed n when p is close to 0.5, and least accurate when p is near 0 or near 1.

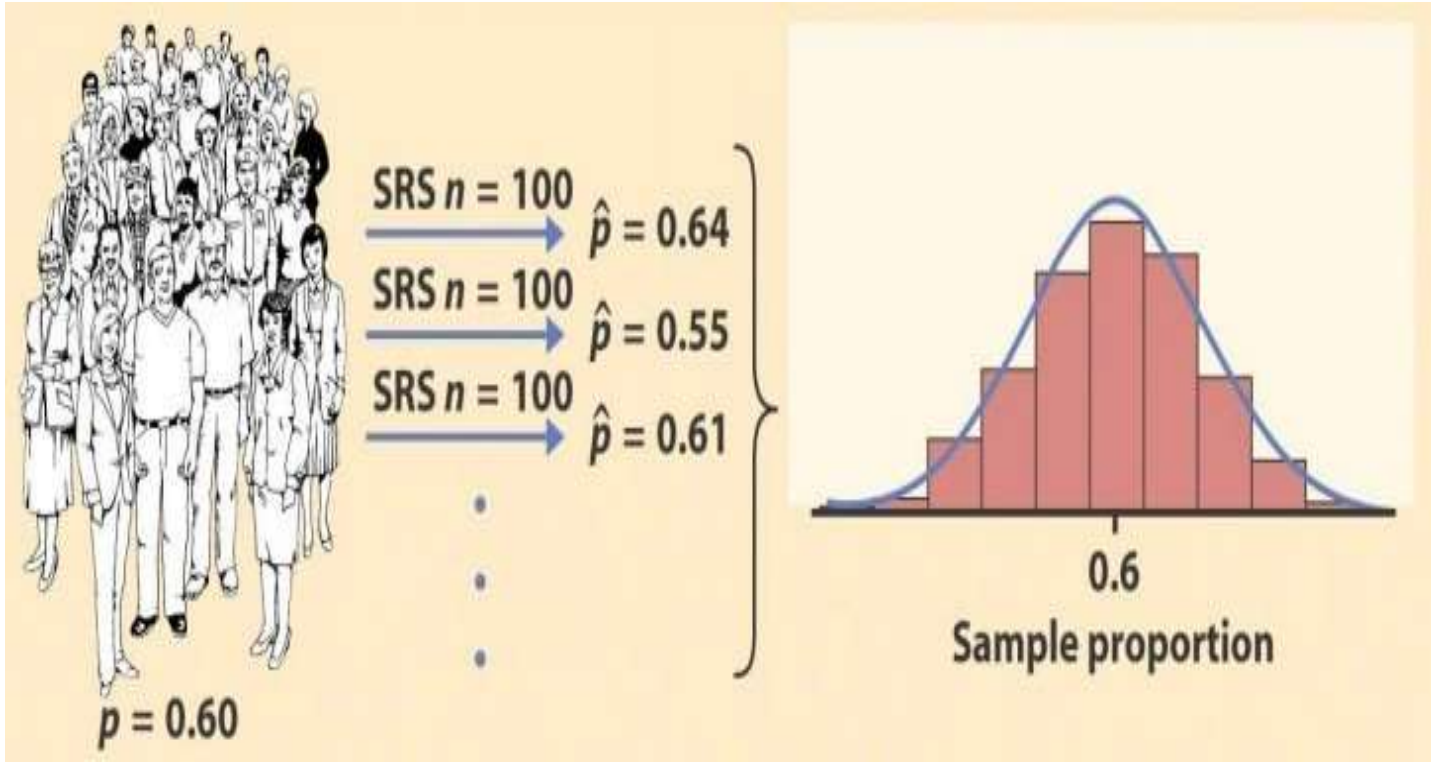


Source:https://cobalt.rocky.edu/~ulrich.hoensch/FS_2012/MAT210/LectureNotes/Lecture10_Inference_Ethics_Randomness

Uncertainty in the Sample Proportion

- Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. This is called sampling variability.
- If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the sampling distribution—will follow a predictable pattern.
- The variability decreases as the sample size increases. So larger samples usually give closer estimates of the population proportion p.

Uncertainty in the Sample Proportion



Source:https://cobalt.rocky.edu/~ulrich.hoensch/FS_2012/MAT210/LectureNotes/Lecture10_Inference_Ethics_Randomness

Applications

Interested about the occurrence of an event, not its magnitude.

1)Whether a smoker quit smoking altogether, rather than evaluate daily reductions in the number of cigarettes smoked.

2)In a clinical trial, a patient's condition may improve or not. We study the number of patients who improved, not how much better they feel.

3)Is a person ambitious or not? The binomial distribution describes the number of ambitious persons, not how ambitious they are.

4)In quality control we assess the number of defective items in a lot of goods, irrespective of the type of defect.

THANK YOU

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

Department of Computer Science and Engineering