STATISTICS FOR DATA SCIENCE

Binomial Distribution

Prof. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Prof. Uma D

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

Conditions for Binomial Distribution:

There are **only two possible outcomes** to each trial(success and failure).

The **number of trials** are **fixed**.

The **probability of success** is **identical** for all trials.

The **trials** are **independent**(i.e. Carrying out one trial has no effect on any other trials.)

Source: www.citoolkit.com

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.**

$$X \sim \text{Bin}(n, p)$$

Source: www.mathstackexchange.com

Binomial Random Variable = Sum of IID Bernoulli Random Variables

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

Let $Y_1, Y_2, Y_3, \ldots Y_n$ represent n Bernoulli r.v.s.

For $i = 1$ to n, $Y \sim \text{Bernoulli}(p)$

$Y_i = 1$ if the ith trial is success

$Y_i = 0$ if the ith trial is failure

let $X = Y_1 + Y_2 + \ldots + Y_n$

$X \sim \text{Bin}(n, p)$

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.

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?

$$X \sim Bin(7, 0.1)$$

Binomial Distribution: Applications

Asking customers if they will shop again in the next 12 months.

Taking 10 samples from a large batch which is 3 percent defective(as past history shows).

Counting the number of individuals who own more than one car.

Counting the number of correct answers in a multi-choice exam.

Source: www.citoolkit.com

Examples for Binomial Distribution:

- A fixed number of observations (trials), n
  - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
  - e.g., head or tail in each toss of a coin; disease or no disease
  - Generally called “success” and “failure”
  - Probability of success is p, probability of failure is 1 – p
- Constant probability for each observation
  - e.g., Probability of getting a tail is the same each time we toss the coin

Source: www.slideshare.net

Binomial Distribution

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Solution:

One way to get exactly 3 heads: HHHTT

P(heads)xP(heads) xP(heads)xP(tails)xP(tails) =$(1/2)^3 x (1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1 x (1/2)^3 x (1/2)^1$ =$(1/2)^3 x (1/2)^2$

$(1/2)^3 x (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$$(1/2)^3 x (1/2)^2 + (1/2)^3 x (1/2)^2 + (1/2)^3 x (1/2)^2 + \ldots..$$

for as many unique arrangements as there are—but how many are there??

Source: www.slideshare.net

Outcome	Probability
THHHT	$(1/2)^3$ *x* $(1/2)^2$
HHHTT	$(1/2)^3$ *x* $(1/2)^2$
TTHHH	$(1/2)^3$ *x* $(1/2)^2$
HTTHH	$(1/2)^3$ *x* $(1/2)^2$
HHTTH	$(1/2)^3$ *x* $(1/2)^2$
HTHHT	$(1/2)^3$ *x* $(1/2)^2$
THTHH	$(1/2)^3$ *x* $(1/2)^2$
HTHTH	$(1/2)^3$ *x* $(1/2)^2$
HHTHT	$(1/2)^3$ *x* $(1/2)^2$
THHTH	$(1/2)^3$ *x* $(1/2)^2$

10 arrangements *x* $(1/2)^3$ *x* $(1/2)^2$

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

$${}_5C_3 = 5!/3!2! = 10$$

The probability of each unique outcome (note: they are all equal)

$P(x=3)=5C_3\ (0.5)^3\ (0.5)^2$

Factorial review: n! = n(n-1)(n-2)…

Source: www.slideshare.net

Probability Mass Function

Let X be a r. v. with $X \sim Bin(n, p)$

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

$$p(x) = \begin{cases} \frac{n!}{x!\,(n-x)!} p^{x} (1-p)^{n-x} & , \end{cases}$$

X : No. of successes in n trials

Mean and Variance of Binomial Distribution

$X \sim Bin(n, p)$

$\mu_x = n * p$

Variance $\sigma_x^2 = npq$ or $\sigma_x^2 = np(1-p)$

Standard Deviation $\sigma_x = \sqrt{\text{Variance}} = \sqrt{np(1-p)}$

Binomial Distribution

1. You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample 500 exposed people, how many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

2. What's the probability that **at most** 10 exposed people develop the disease?

Source: www.slideshare.net

/ 1. How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

$$X \sim Bin(500, .05)$$

$$\mu_x = E(x) = n * p = 500 \times .05 = 25$$

$$Var(x) = \sigma_x^2 = npq = 500 \times .05 \times .95 = 23.75$$

$$SD(x) = \sigma_x = \sqrt{npq} = \sqrt{23.75} = 4.87$$

$$\mu_x = 25 \pm 4.87$$

Source: www.slideshare.net

STATISTICS FOR DATA SCIENCE

Binomial Distribution

2. What’s the probability that **at most** 10 exposed subjects develop the disease?

$$P(X \le 10) = P(X=0) + P(X=1) + P(X=2) + \cdots + P(X=10)$$

$$= \binom{500}{0}(0.05)^{0}(0.95)^{500} + \binom{500}{1}(0.05)^{1}(0.95)^{499} + \cdots$$

$$+ \binom{500}{10}(0.05)^{10}(0.95)^{490}$$

$$P(X \le 10) < 0.01$$

Source: www.slideshare.net

Binomial Distribution:Problem

A coin is flipped 3 times.
What is the probability head turns up exactly 2 times.

n = No. of trials = 3 $X \sim Bin(3, 0.5)$

$x = 2$

$$P(X=2) = \binom{3}{2}(0.5)^{2}(0.5)^{3-2}$$

$$= 0.375$$

$P($Exactly 2 heads$) = P(HHT) + P(HTH) + P(THH)$

Source: www.slideshare.net

Example

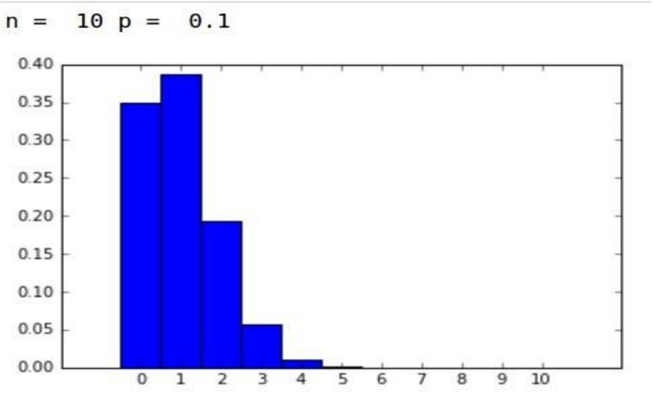
Find the effect of changing p when n is fixed and is small.

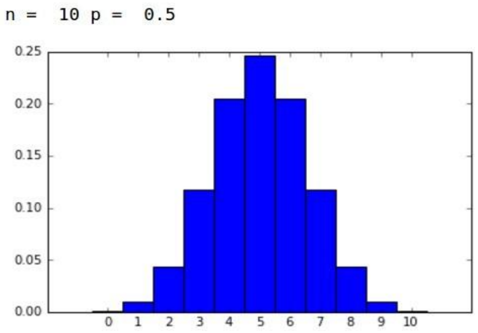
a) n = 10, p = 0.10

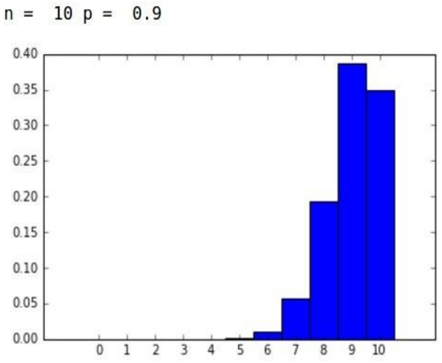
b) n = 10, p = 0.5

c) n = 10, p = 0.90

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







Sample Proportion to Estimate a Success Probability

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

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

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

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.

Computing Bias of $\hat{p}$

Bias : An 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.

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

$$\begin{aligned} \text{Bias} &= \mu_{\hat{p}} - p \\ &= p - p \\ &= 0 \end{aligned}$$

$\therefore \mu_{\hat{p}}$ is unbiased.

Source: www.slideshare.net

Computing uncertainty of $\hat{p}$

Uncertainty = SD

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

Since p is not known

$$\hat{p} = \frac{X}{n}$$

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

Source: www.slideshare.net

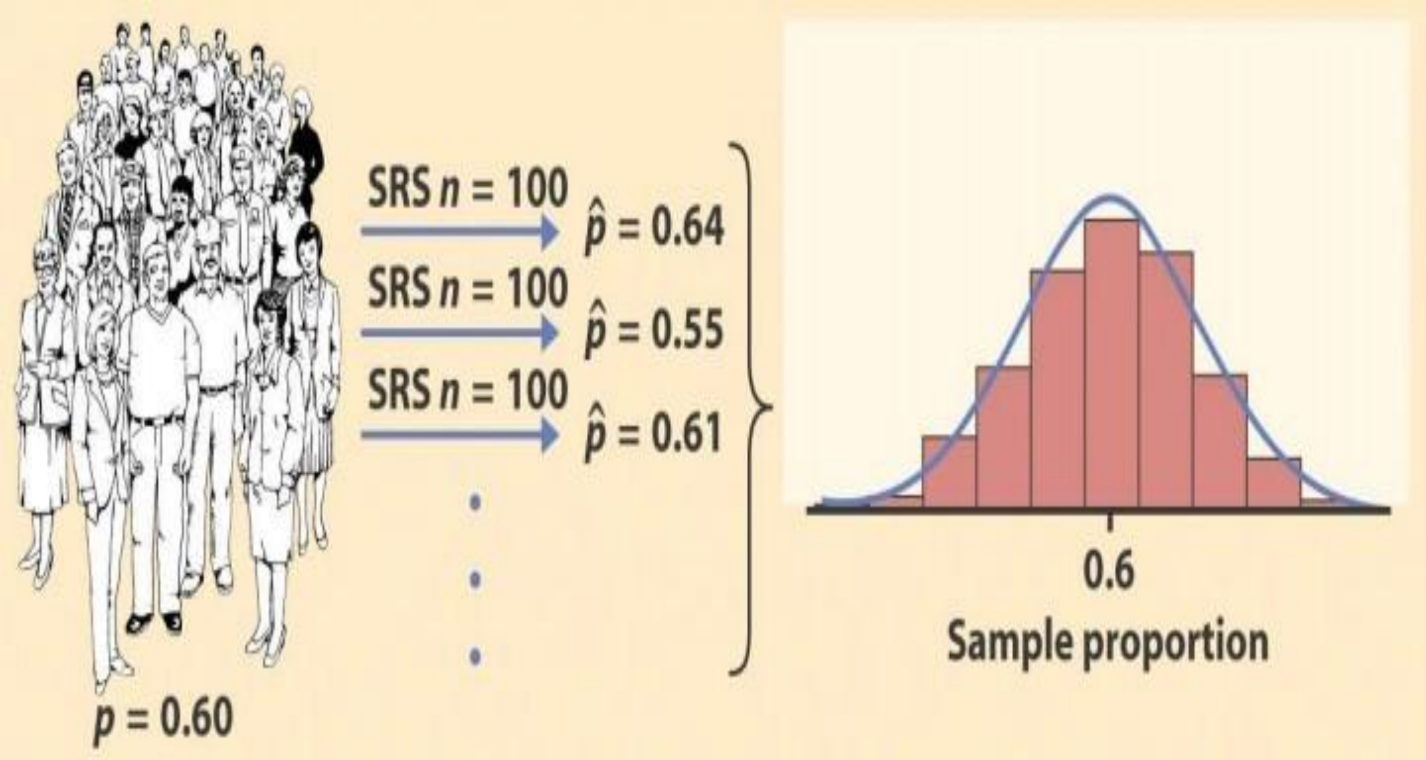
Sampling Distribution

1. 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.

2. 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.

3.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

Binomial Distribution: Problem

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 day's production that meets specifications. $\hat{p} = \frac{92}{100} = .92$
2. Find the uncertainty in the estimate. $\sigma_x = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.92 \times 0.08}{100}} = 0.027$
3. Estimate the number of rods that must be sampled to reduce the uncertainty to 1%

Uncertainty $= 0.01$

$\hat{p} = 0.92 \quad n = ?$

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

$\sigma^2 = \frac{p(1-p)}{n}$

$\Rightarrow n = \frac{p(1-p)}{\sigma^2} = \frac{0.92 \times 0.08}{(0.01)^2} = 736$

$\therefore n = 736$

Source: www.slideshare.net

Do It Yourself !!!

1. Telephone surveying a group of 200 people to ask if they voted for Trump.

2. You take a survey of 40 traffic lights in a certain city, at 2 p.m., recording whether the light was yellow, red, or green at that time.

3. Asking 50 people if they have ever been to Taj Mahal.

Source: www.slideshare.net

Do It Yourself !!!

You are conducting a case-control study of smoking and lung cancer. If the probability of being a smoker among lung cancer cases is .6, what's the probability that in a group of 8 cases you have:

a. Less than 2 smokers?
b. More than 5?
c. What are the expected value and variance of the number of smokers?

Source: www.slideshare.net

Binomial Distribution: Problem

Do It Yourself !!!

The quality manager of a fortune cookie company believes that a larger than acceptable proportion of paper fortunes being used are blank.

Suppose she takes a sample of 320 fortune cookies from the production line and 15 of the paper fortunes are blank.

1. Calculate the estimate of the proportion.
2. Calculate uncertainty in the estimate.

Source: www.slideshare.net

Do It Yourself !!!

Find the effect of changing p when n is fixed and small. Write your observations. (Coding Assignment)

1. n=10, p=0.05
2. n=10, p=0.1
3. n=10, p=0.5
4. n=10, p=0.9

Source: www.slideshare.net

Do It Yourself !!!

Find the effect of changing p when n is fixed and large. Write your observations. (Coding Assignment)

1. n=100, p=0.1
2. n=100, p=0.25
3. n=100, p=0.5
4. n=100, p=0.75

Source: www.slideshare.net

THANK YOU

Prof. Uma D

Department of Computer Science and Engineering