



STATISTICS FOR DATA SCIENCE

Binomial Distribution

Prof. Uma D

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Prof. Uma D

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**.



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.)

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** .

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



Binomial Random Variable = Sum of IID Bernoulli Random Variables

Binomial or Not?

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

$p = P(\text{Alzheimer's gene}) = 0.1 \longrightarrow$ Binomial

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

$p = P(\text{defective}) = 2/20$

$p = P(\text{defective}) = 1/19$ Not Binomial

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

Rule of thumb:

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

- say $n/N \leq .05$, 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 ?

STATISTICS FOR DATA SCIENCE

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.

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

STATISTICS FOR DATA SCIENCE

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

$$P(\text{heads}) \times P(\text{heads}) \times P(\text{heads}) \times P(\text{tails}) \times P(\text{tails}) = (1/2)^3 \times (1/2)^2$$

Another way to get exactly 3 heads: THHHT

$$\text{Probability of this exact outcome} = (1/2)^1 \times (1/2)^3 \times (1/2)^1 = (1/2)^3 \times (1/2)^2$$

$(1/2)^3 \times (1/2)^2$ 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:

$$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + \dots$$

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

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Outcome	Probability
---------	-------------

THHHT	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HHHTT	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

TTHHH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HTTHH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HHTTH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HTHHT	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

THTHH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HTHTH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

HHTHT	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

THHTH	$(1/2)^3 \times (1/2)^2$
-------	--------------------------

10 arrangements	$\times (1/2)^3 \times (1/2)^2$
-----------------	---------------------------------

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

ways to
arrange
3 heads
in 5
trials

$$\binom{5}{3}$$

$${}_5C_3 = \frac{5!}{3!2!} = 10$$

$$P(x=3) = {}_5C_3 (0.5)^3 (0.5)^2$$

Factorial review: $n! = n(n-1)(n-2)\dots$

STATISTICS FOR DATA SCIENCE

Binomial Distribution

Probability Mass Function



STATISTICS FOR DATA SCIENCE

Binomial Distribution

Mean and Variance of Binomial Distribution



STATISTICS FOR DATA SCIENCE

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?

STATISTICS FOR DATA SCIENCE

Binomial Distribution



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

STATISTICS FOR DATA SCIENCE

Binomial Distribution



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

STATISTICS FOR DATA SCIENCE

Binomial Distribution: Problem



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

Example

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

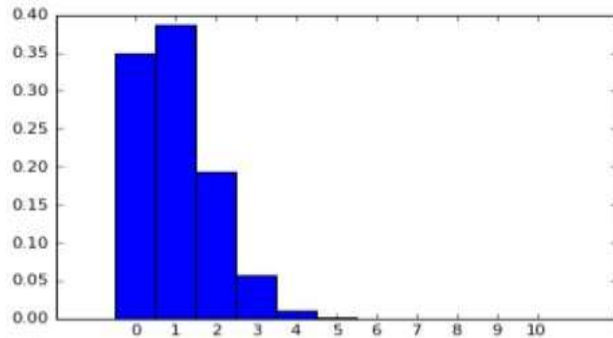
a) $n = 10, p = 0.10$

b) $n = 10, p = 0.5$

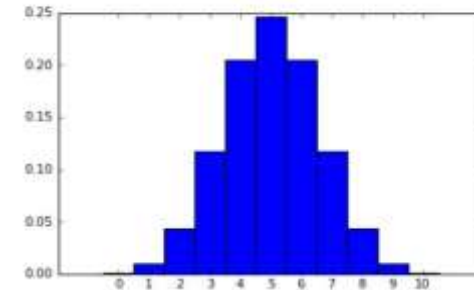
c) $n = 10, p = 0.90$

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

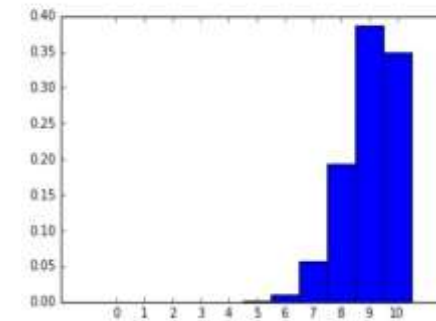
$n = 10, p = 0.1$



$n = 10, p = 0.5$



$n = 10, p = 0.9$



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 .

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.

STATISTICS FOR DATA SCIENCE

Binomial Distribution

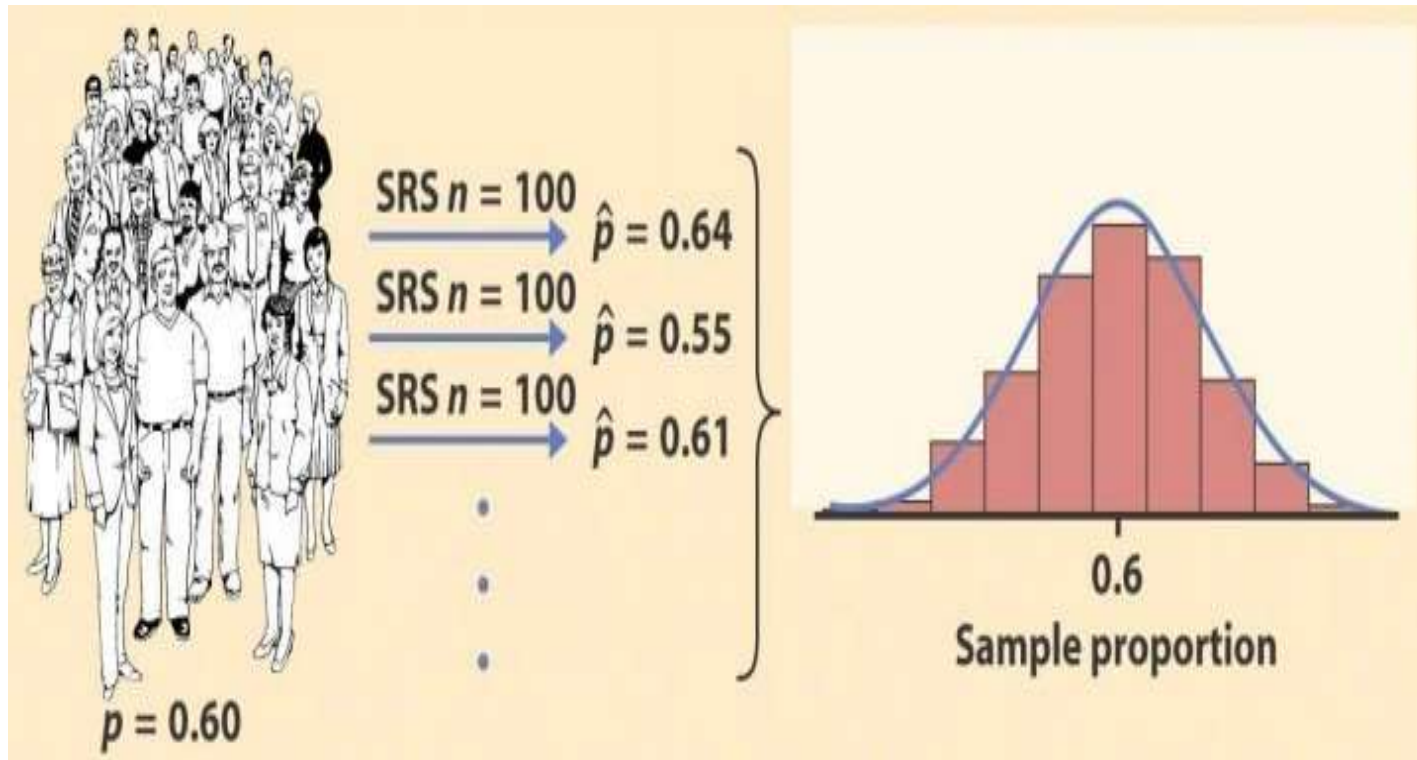


Computing uncertainty of \hat{p}

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 .



STATISTICS FOR DATA SCIENCE

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.
2. Find the uncertainty in the estimate.
3. Estimate the number of rods that must be sampled to reduce the uncertainty to 1%

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.

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:

- Less than 2 smokers?
- More than 5?
- What are the expected value and variance of the number of smokers?

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.

STATISTICS FOR DATA SCIENCE

Binomial Distribution



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. $n=10, p=0.9$

STATISTICS FOR DATA SCIENCE

Binomial Distribution



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. $n=100, p=0.75$



THANK YOU

Prof. Uma D

Department of Computer Science and Engineering