# Introduction to Data Science

Unit 2

## Random Variables

**Preet Kanwal**

**Assistant Professor**

**Department of CSE**

**PESU Bangalore**

# Agenda

- Discrete Random Variables
- Continuous Random Variables

# Random Variable

- Random variable is a quantitative variable whose value depends on a chance in some way.
- Outcome of an experiment
- Assigns numerical value to each outcome
- Types: Discrete and Continuous
- Can be thought of as a sample from a population

# Types of Random Variables

- Discrete
- Continuous

# Discrete Random Variables

- Comes from a discrete set.
- Whole numbers.
- Takes a countable number of possible values.

# Continuous Random Variables

Can take on an infinite number of possible values, corresponding to every value in an interval.
Example: [4,6]

# Identify Discrete or Continuous

1. Number of free throws an NBA player makes in his next 20 attempts.

2. Time between lightening strikes in a thunderstorm.

3. Velocity of next pitch in Major League Baseball.

4. Number of rolls of a die needed to roll a 3 for first time.

# Discrete Random Variables

Prof. Preet Kanwal

# Discrete Random Variables

Tossing two coins simultaneously. Find the probability of getting an head?

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

***Notation:***

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

Where,

X is the random variable

x is the value of the random variable.

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

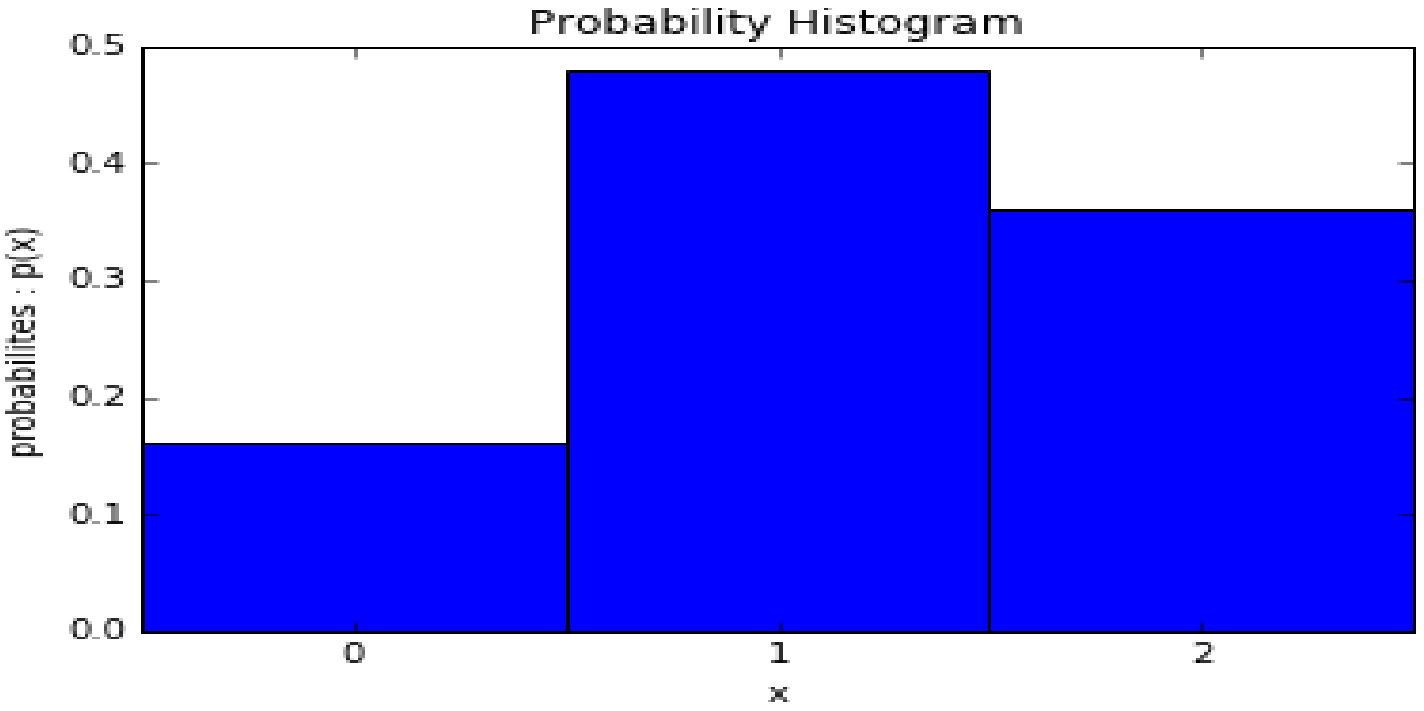
2) $\sum_{x} p(x) = 1$

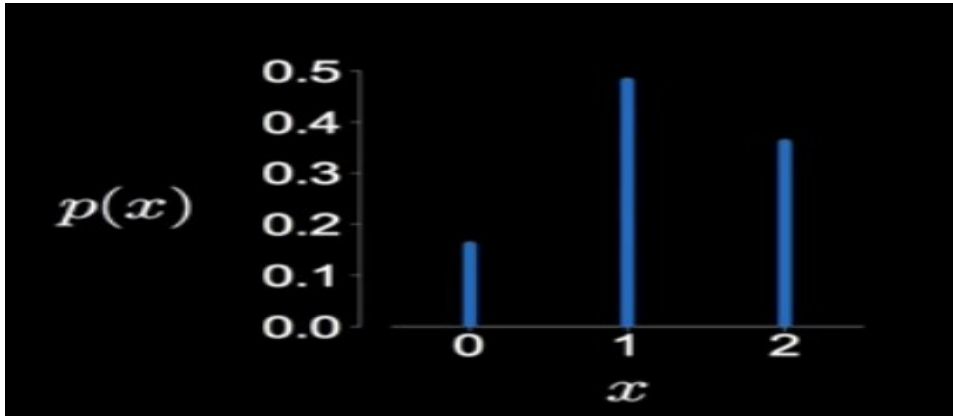
# Problem 1

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??

probability distribution of X

| x | p(x) |
|---|---|
| 0 | 0.16000 |
| 1 | 0.48000 |
| 2 | 0.36000 |



# Problem

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 the wire.

# Mean and Variance

- Mean (AKA Expected Value or Expectation)

$$E(X) = \mu_X = \sum_x xp(x)$$

- Variance

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

# Cumulative Distribution Function

- 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 P(X<=1)?

- P(X<=3)= P(0)+P(1)

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

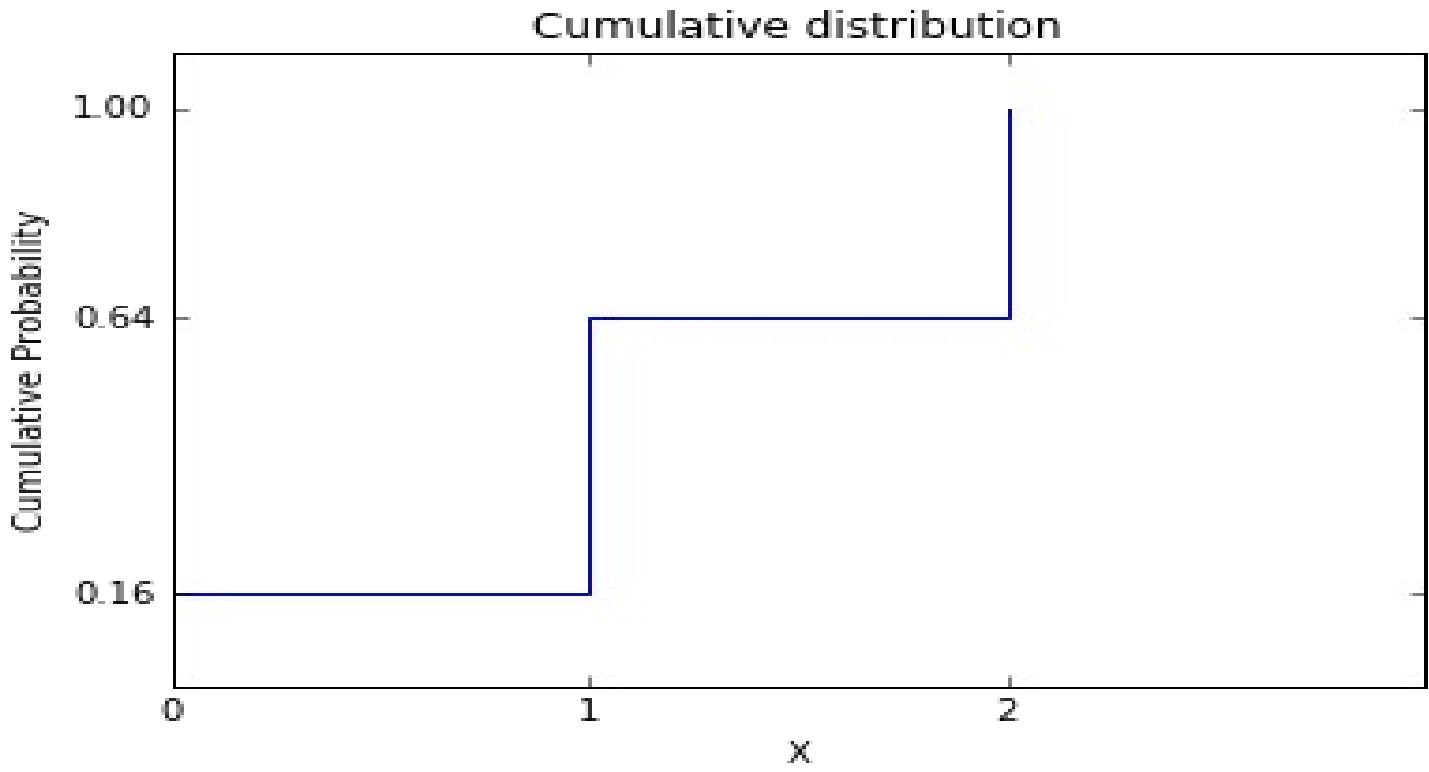
$$P(a \leq X \leq b) = F(b) - F(a)$$

$$F(-\infty) = 0$$

$$F(\infty) = 1$$

probability distribution of X

| x | p(x) | F(X) |
|---|---|---|
| 0 | 0.16000 | 0.16000 |
| 1 | 0.48000 | 0.64000 |
| 2 | 0.36000 | 1.00000 |



# Problem 2

Let X represent no of tiers with low air pressure on a randomly chosen car. Which of these functions is a possible Probability Mass function of X?

| | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|
| P1(x) | 0.2 | 0.2 | 0.3 | 0.1 | 0.1 |
| P2(x) | 0.1 | 0.3 | 0.3 | 0.2 | 0.2 |
| P3(X) | 0.1 | 0.2 | 0.4 | 0.2 | 0.1 |

1) Find mean and variance.
2) Find F(2).
3) Find P(1<= X <= 4).

# Problem 2 - Solution

p 3 (x) is the only probability mass function, because it is the only one whose probabilities sum to 1.

1) Mean = 2.0  and Variance = 1.2

2) F(2) = P(0) + P(1) + P(2)
= 0.1 + 0.2 + 0.4
= 0.7

3) P(1<= X <= 4) = F(4) – F(1)
= 1 – 0.3
= 0.7

# Problem 3

- On 100 different days a traffic engg counts the no of cars that pass through a certain intersection between 5 pm and 5.05 pm.
- The results are presented in the following table

| No of Cars | No of days | Proportion of days |
|---|---|---|
| 0 | 36 | 0.36 |
| 1 | 28 | 0.28 |
| 2 | 15 | 0.15 |
| 3 | 10 | 0.10 |
| 4 | 7 | 0.07 |
| 5 | 4 | 0.04 |

- X – No of cars passing through the intersection bet 5 and 5.05pmon a randomly chosen day.
- A) Someone suggests p1(X) = $(0.2)(0.8)^x$

B) Another person suggests p2(x) = $(0.4)(0.6)^x$

C) Compare PMF of A and B to the data in the table. Which PMF is better?

D) Someone says neither of it is a good model As its not agreeing with the data exactly!. Is the conclusion correct?

# Problem 3 - Solution

| x | p1(x) |
|---|---|
| 0 | 0.2 |
| 1 | 0.16 |
| 2 | 0.128 |
| 3 | 0.1024 |
| 4 | 0.0819 |
| 5 | 0.0655 |

| x | p2(x) |
|---|---|
| 0 | 0.4 |
| 1 | 0.24 |
| 2 | 0.144 |
| 3 | 0.0864 |
| 4 | 0.0518 |
| 5 | 0.0311 |

**C)** p 2 (x) appears to be the better model. Its probabilities are all fairly close to the proportions of days observed in
the data. In contrast, the probabilities of 0 and 1 for p 1 (x) are much smaller than the observed proportions.

**D)** No, this is not right. The data are a simple random sample, and the model represents the population. Simple
random samples generally do not reflect the population exactly.

# Problem 4

- 3 components are randomly sampled one at a time from a large lot.
- As each component is selected, its tested. If it passes the test a success occurs and If it fails a failure occurs.
- Assume 80% of components result in success
- Let X represent the no of successes among the 3 sampled components.

a) What are possible values for X?

b) Find P(X = 3).

c) FSS – event denotes 1st fails 2and 3 succeed. Find P(FSS).

d) Find P(SFS) and P(SSF).

e) P(X = 2).

f) Find Mean and Variance of X.

g) Y represents no of successes if 4 components are sampled. Find P(Y = 3).

# Problem 4 - Solution

a) X = {0, 1, 2, 3}

b) P(X = 3) = P(SSS) = (0.8) 3 = 0.512

c) P(FSS) = $(0.2)(0.8)^2$ = 0.128

d) P(SFS) = P(SSF) = $(0.8)^2$ (0.2) = 0.128

e) P(X = 2) = P(FSS) + P(SFS) + P(SSF) = 0.384

f) Mean = 2.4 and Variance = 0.48

g) P(Y = 3) =P(SSSF) + P(SSFS) + P(SFSS) + P(FSSS)

= $(0.8)^3$ (0.2) + $(0.8)^3$ (0.2) + $(0.8)^3$ (0.2) + $(0.8)^3$ (0.2)

= 0.4096

# Problem 5

- A certain type of component is packaged in lots of four.
- X – represents the no of properly functioning components in a randomly chosen lot.
- The probability mass function of X is given as:

$$p(x) = \begin{cases} cx & x = 1,2,3,4 \\ 0 & \text{otherwise} \end{cases}$$

where, c is a constant.

- Find c so that p(x) is a pmf.
- Find P(X =2).
- Find mean.
- Find Variance.
- Find Standard Deviation.

# Problem 5 - Solution

a) Find c so that p(x) is a pmf.

c(1 + 2 + 3 + 4 + 5) = 1, so c = 1/15.

b) Find P(X =2).

P(X = 2) = c(2) = 2/15 = 0.2

c) Mean = 11/3

d) Variance = 14/9

e) Standard Deviation = 1.2472

# Problem 6

- After manufacture, computer disks are tested for errors.
- X – No of errors detected on a randomly chosen disk.
- Following table represents cumulative distribution function F(x) of X:

| | |
|---|---|
| 0 | 0.41 |
| 1 | 0.72 |
| 2 | 0.83 |
| 3 | 0.95 |
| 4 | 1.00 |

- What is the Probab
- P(more than 3 erro
- P(Exactly one error is detected)
- What is the most probable no of errors to be detected?

# Problem 6 - Solution

1) What is the Probability that 2 or fewer errors occur?

$P(X \leq 2) = F(2) = 0.83$

2)P(more than 3 errors occur)

$P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - 0.95 = 0.05$

3) P(Exactly one error is detected)

$P(X = 1) = P(X \leq 1) - P(X \leq 0)$

$= F(1) - F(0)$

$= 0.72 - 0.41$

$= 0.31$

4) What is the most probable no of errors to be detected?

The value of x for which Probability is greatest is x = 0.

# Problem 7

Microprocessing chips are randomly sampled one by one from a large population and tested to determine if they are acceptable or not for a certain application.

90% of the chips in the population are acceptable.

a) Find the probability mass function of X, where X represents no. of chips that are tested up to and including the first acceptable chip.

b) Find P(X = 3)

# Problem 7 - Solution

a) $p(x) = (0.9)(0.1)^{x-1}$ $x = 1, 2, 3, ...$

0 otherwise

b) $P(X = 3) = P(UUA)$
$= P(U)P(U)P(A)$
$= (0.1)(0.1)(0.9)$
$= 0.009$

# Continuous Random Variables

Prof. Preet Kanwal

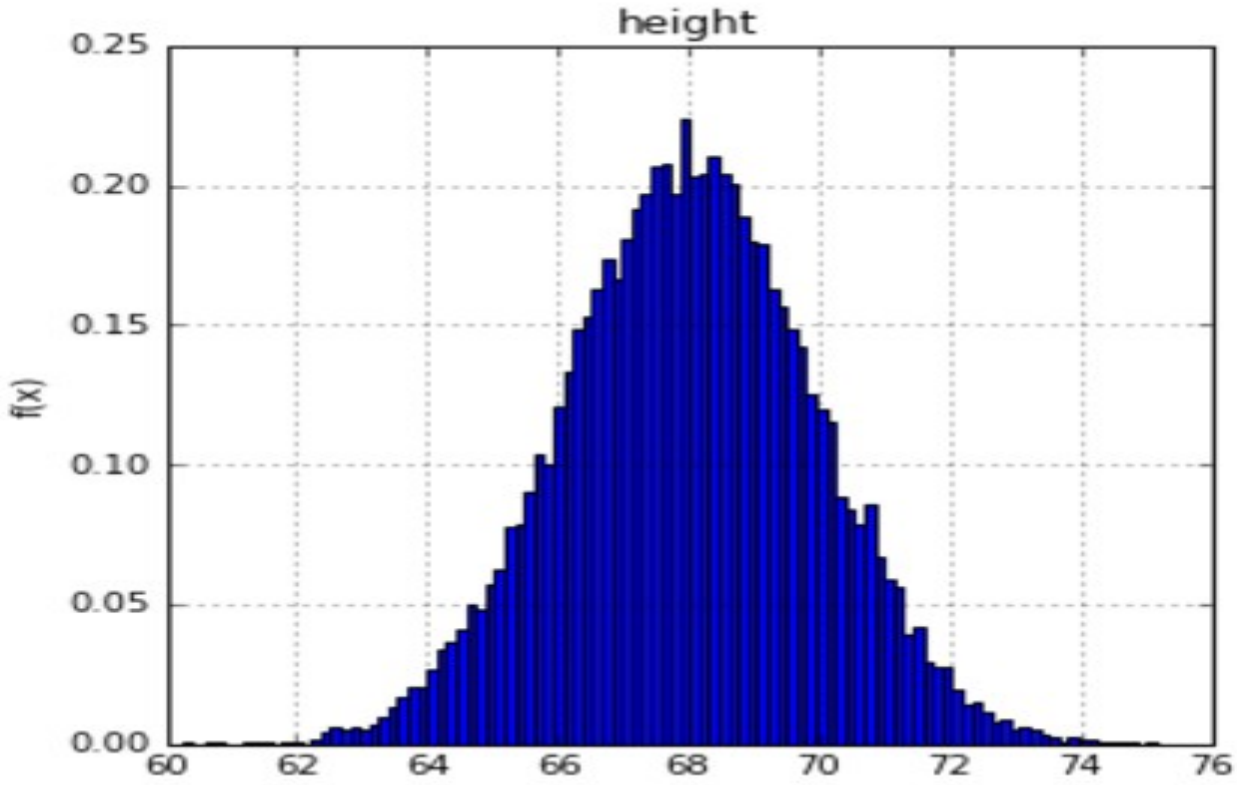
# Continuous Random variable

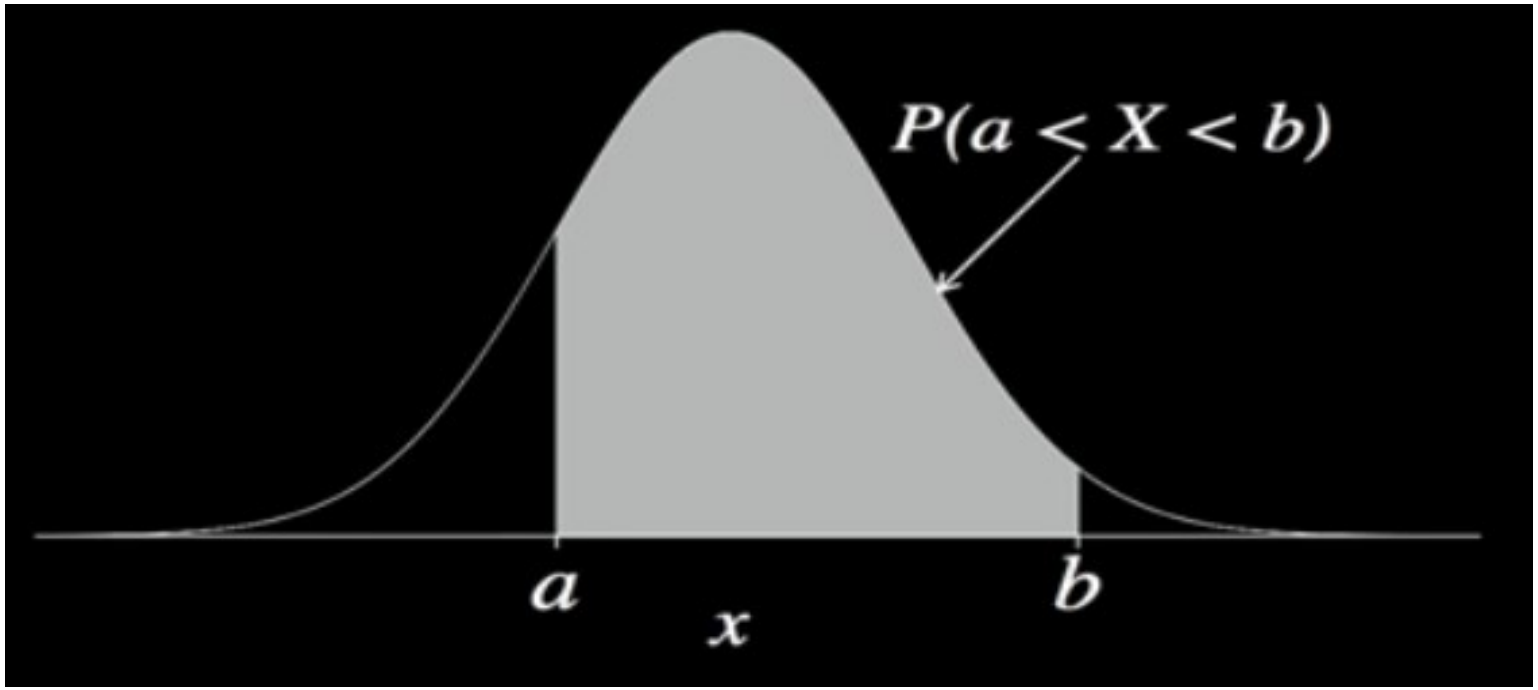
Looks like a smooth histogram.

Values where the curve is high are more likely to occur.

We model Continuous Random Variables with a a curve *f(x)* called a **probability density function(pdf).**

*f(x)* is function that represents the height of the curve at point x.






For Continuous Random Variables, Probabilities are areas under the curve – hence found using integration.

Probability of random variable X equal to any one of the specific value say a is 0. For any a, P(X = a ) = 0 .

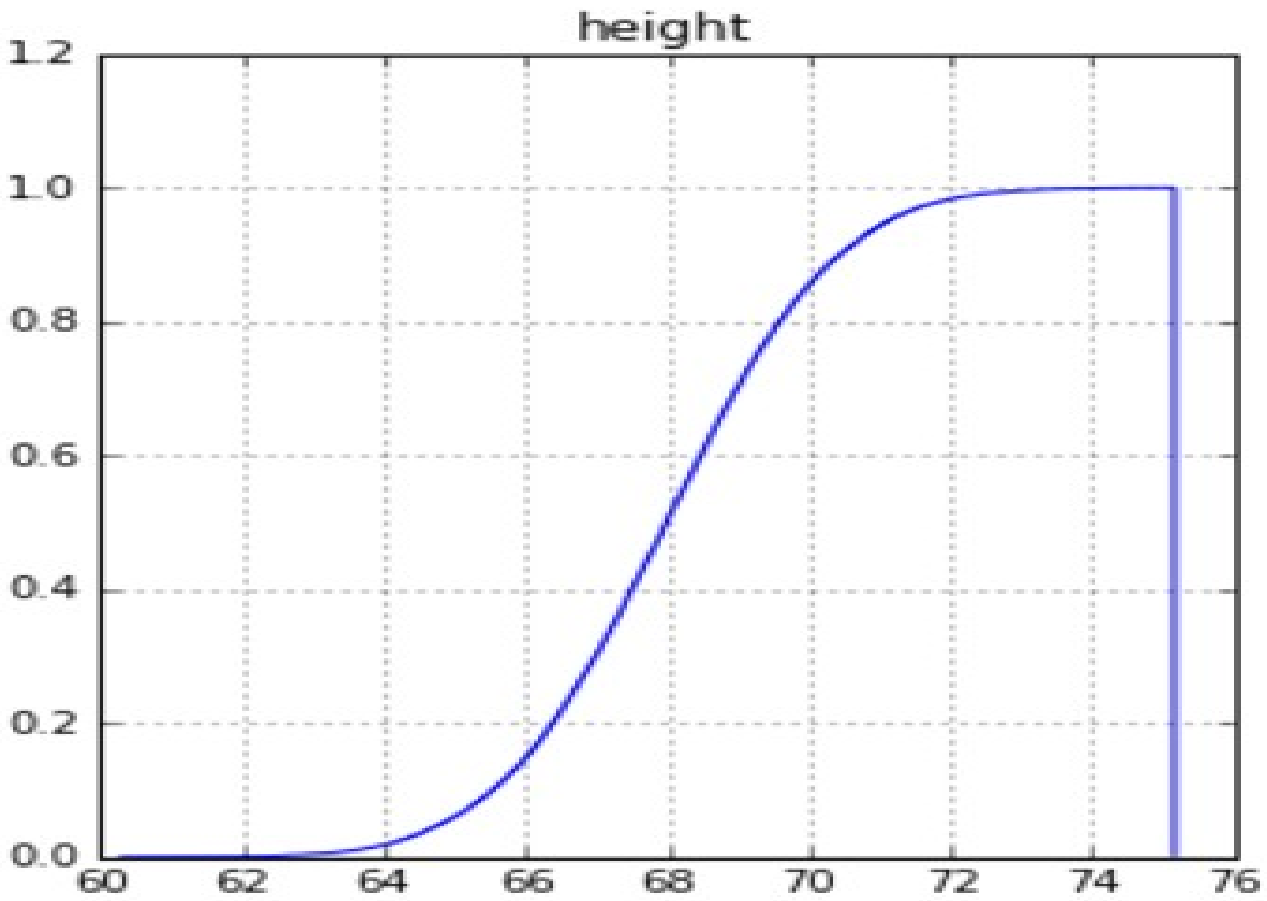
That's why P(a<= X <=b) = P(a< X <=b) =
P(a<= X <b) = P(a< X < b)

For any Continuous Probability Distribution,

$f(x) >= 0$ for all x

Area under the entire curve is equal to 1

# Plot of Cumulative Distribution Function (F(x)) :



# Problem 1

Suppose for a random variable X:

$$f(x) = cx^3$$

For 2 <= x <= 4 and 0 otherwise.

a) What value of c makes this a legitimate probability distribution?

b) What is P( X > 3) .

c) What is the median of this distribution?

d) What is the cumulative distribution function?

e) Find P(X <= 2.7).

f) Find mean and variance of this distribution.

# Problem 1 - Solution

a) c = 1/60

b) P(X> 3) = 0.729

c) Median = 3.415

d) CDF = 1/60 $(x^4 – 2^4)/ 240$

e) P(X <= 2.7) = 0.155

f) Mean = 248/75 = 3.3

Variance = 0.266

# Problem 2

Let X be a random variable with PDF given by

$$f(x)=\begin{cases} x/250 & 20 \le |x| \le 30 \\ 0 & \text{otherwise} \end{cases}$$

1) Find $P(X \ge 25)$ .
2) Find E(X) and Var(X).
3) Find CDF.
4) Find median.
5) Find $60^{th}$ percentile.

# Problem 2 - Solution

1) $P(X \geq 25) = 0.55$

2) $E(X) = 25.33$ and $Var(X) = 8.3911$

3) CDF.

$$F(x) = \begin{cases} 0 & x < 20 \\ (x^2 - 400)/500 & 20 \leq x \leq 30 \\ 1 & x > 30 \end{cases}$$

4) Median = 25.495

5) $60^{th}$ percentile = 26.458

# Problem 3

The main bearing clearance(in mm) in a certain type of engine is a random variable with probability density function:

$$f(x) = \begin{cases} 625x & 0 < x \leq 0.04 \\ 50 - 625x & 0.04 < x \leq 0.08 \\ 0 & \text{otherwise} \end{cases}$$

1)What is the probability the clearance id less than 0.02mm?
2)Find the mean clearance.
3)Find the standard deviation of the clearances.
4)Find cumulative distribution function of the clearances.
5)Find the median clearances.
6)The specification for clearance is 0.015 to 0.063mm. What is the probability that the specification is met?

# Problem 3 - Solution

1) $P(X < 0.02) = 0.125$

2) Mean clearance = 0.04

3) Variance = 0.0002667
   Standard deviation = 0.01633

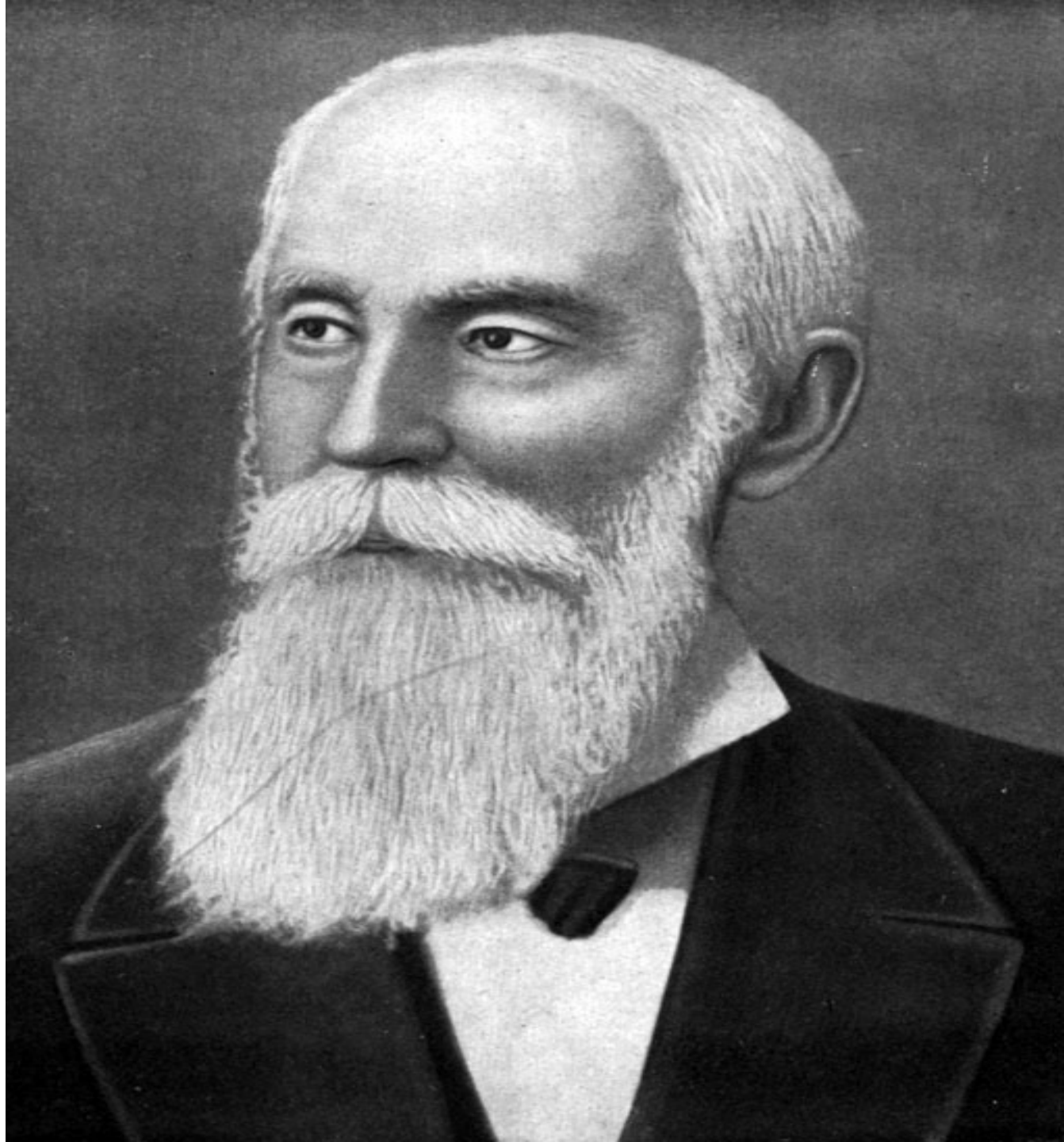
4) CDF:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 625\,x^2/2 & 0 < x \leq 0.04 \\ 50x - 312.5x^2 - 1 & 0.04 < x \leq 0.08 \end{cases}$$

5) Median = 0.04

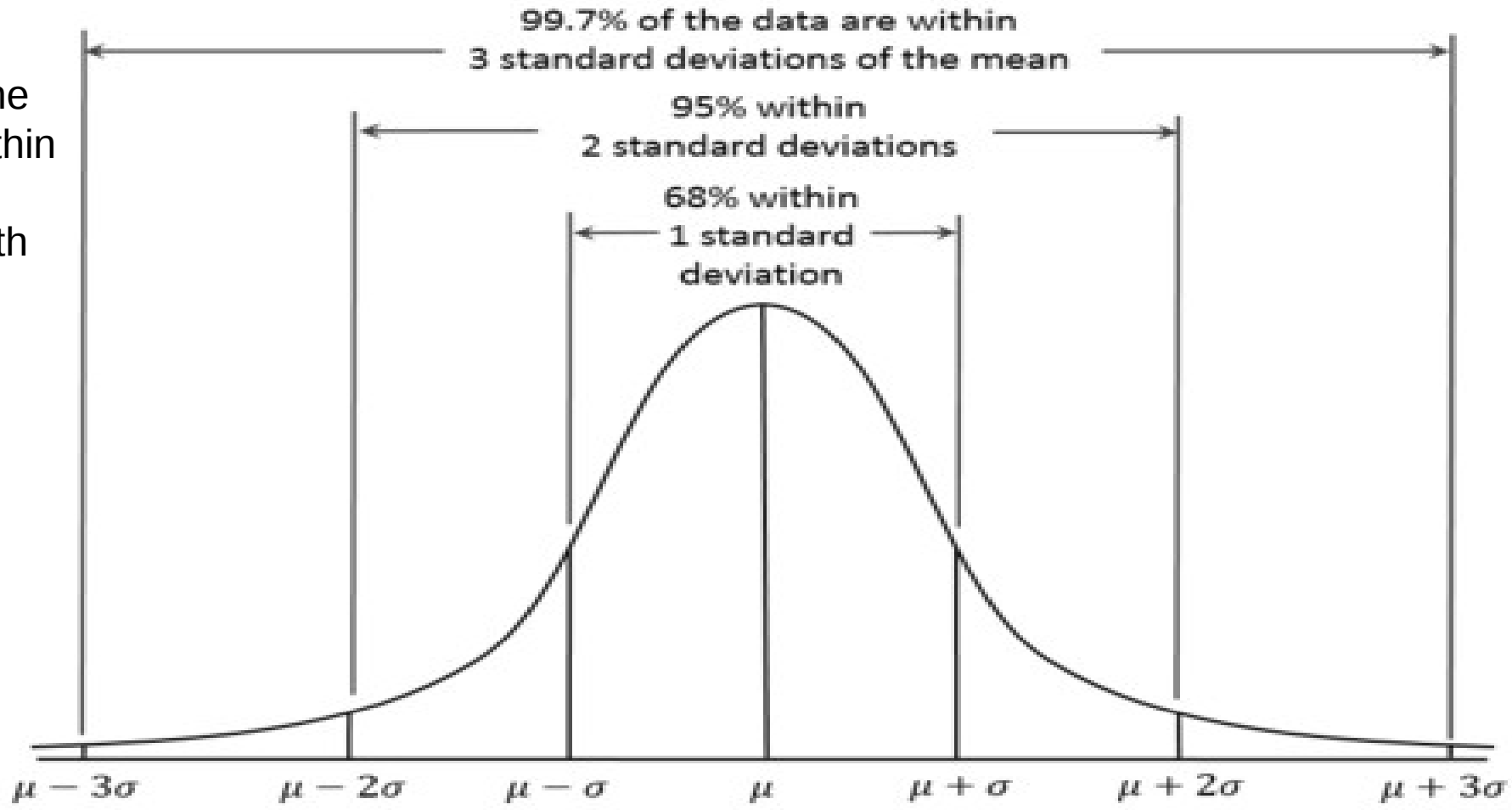
6) $P(0.015 < X < 0.063) = 0.9097$

# Chebyshev's Inequality



# 68–95–99.7 rule : when Data is distributed Normally

Shorthand used to remember the percentage of values that lie within a band around the mean in a **normal distribution** with a width of one, two and three standard deviations, respectively.



# 68–95–99.7 rule : when Data is distributed Normally

$$\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$$

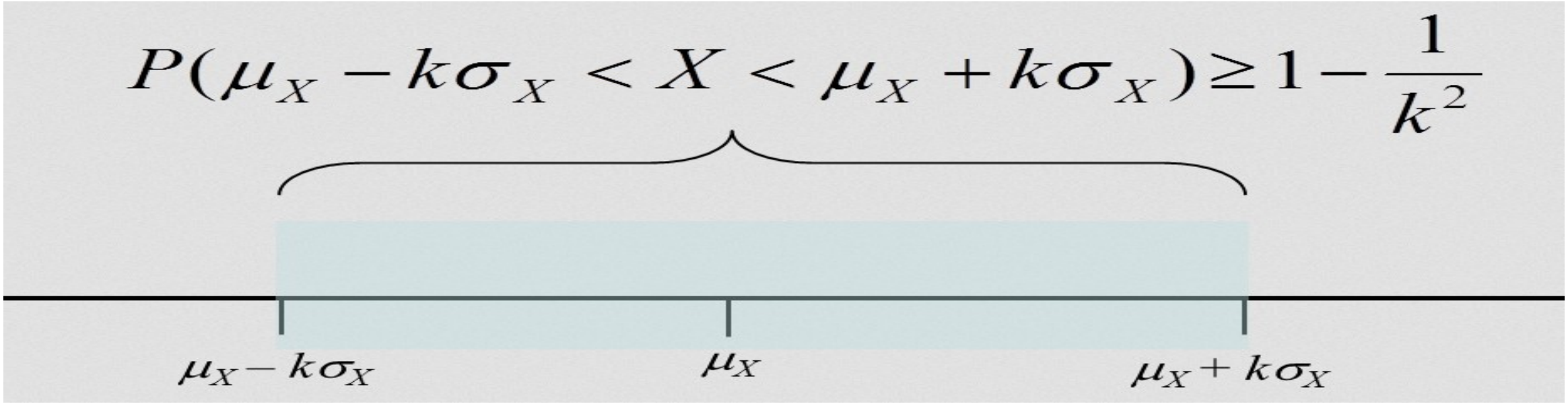
# When Data is not distributed Normally

But if the data set is not distributed normally, then a different amount could be within one standard deviation.

**Chebyshev's inequality** provides a way to know what fraction of data falls within **K standard deviations** from the mean for any data set.

# Statement of Chebyshev's Inequality

Chebyshev's inequality states that at least $1-1/K^2$ of data from a sample must fall within K standard deviations from the mean, where K is any positive real number greater than one.



# Illustration of the Inequality

To illustrate the inequality, we will look at it for a few values of K:

For K = 2 we have $1-1/K^2 = 1 - 1/4 = 3/4 = 75\%$. So Chebyshev's inequality says that at least 75% of the data values of any distribution must be within two standard deviations of the mean.

For K = 3 we have $1 – 1/K^2 = 1 - 1/9 = 8/9 = 89\%$. So Chebyshev's inequality says that at least 89% of the data values of any distribution must be within three standard deviations of the mean.

# Statement of Chebyshev's Inequality

**Chevyshev's Inequality can also be stated as follows:**

Chebyshev's inequality relates mean and standard deviation by providing a bound on the probablity that a Random Variable takes on a value that differs from its mean by K standard deviation or more is never greater than $1/k^2$

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Only the case k > 1 is useful.

When k ≤ 1 the right hand $1/k^2 \geq 1$ and the inequality is trivial as all probabilities are ≤ 1.

# Problem 1

Computers from a particular company are found to last on average for three years without any hardware malfunction, with standard deviation of two months. At least what percent of the computers last between 31 months and 41 months?

# Problem 1 – Solution

Mean lifetime = 3 years = 36 months.

Standard Deviation = 2 months

**To find % of the computers last from 31 months to 41 months.**

| 31 – mean | = | 31 – 36 | = 5 months

| 41 – mean | = | 41 – 36 | = 5 months

K = 5/ standard deviation = 5/2

=> K = 2.5

By Chebyshev's inequality,

at least $1 – 1/(2.5)^2$ = 84% of the computers last from 31 months to 41 months.

# Problem 2

Bacteria in a culture live for an average time of three hours with standard deviation of 10 minutes. At least what fraction of the bacteria live between two and four hours?

# Problem 2 – Solution

Mean = 3 hours = 180 minutes

Standard deviation = 10 minutes

**To find % bacteria live between two and four hours:**

| 2 – mean | = | 2 – 3 | = 1 hour = 60 minutes

| 4 – mean | = | 4 – 3 | = 1 hour = 60 minutes

K = 60 / standard dev = 60/ 10

=> K = 6

By Chebyshev's inequality,

So at least $1 – 1/6^2 = 35/36 = 97\%$ of the bacteria live between two and four hours.

# Problem 3

What is the smallest number of standard deviations from the mean that we must go if we want to ensure that we have at least 50% of the data of a distribution?

# Problem 3 – Solution

Here we use Chebyshev's inequality and work backward.

$1-1/K^2 = 0.50$

$1/K^2 = 0.50$

$K^2 = 1/0.5$

$K^2 = 2$

$K = \text{sqrt}(2)$

$K = 1.4$

By Chebyshev's inequality,

So at least 50% of the data is within approximately 1.4 standard deviations from the mean.

# Problem 4 (a)

The length of a metal pin manufactured by a certain process has mean 50 mm and standard deviation 0.45mm.

What is the largest possible value for the probability that the length of the metal pin is outside the interval [49.1 , 50.9] mm?

# Problem 4 (a) – Solution

Mean = 50 mm

Standard deviation = 0.45 mm

**To find P(X ≤49.1 or X ≥ 50.9) ≤ $1/K^2$**

Find K:

| 49.1 – mean | = | 49.1 – 50 | = 0.9

| 50.9 – mean | = | 50.9 – 50 | = 0.9

K = 0.9 / Standard deviation = 0.9 /0.45

K = 2

By Chebyshev's inequality,

P(X ≤ 49.1 or X ≥ 50.9) ≤ $1/K^2$ ≤ 1 /4 ≤ 0.25

# Problem 4 (b)

Assume theh PDF of X, the length of a metal pin is given as

$$f(x) = \begin{cases} [477 - 471(x - 50)^2] / 640 & 49 \le x \le 51 \\ 0 & \text{otherwise} \end{cases}$$

Compute the probability that the length of the metal pin is outside the interval [49.1 , 50.9] mm?

How close is this probability to Chebyshev bound?

# Problem 4 (b) – Solution

Mean = 50 mm

Standard deviation = 0.45 mm

**To find P(X ≤49.1 or X ≥ 50.9)**

$$P(X \leq 49.1 \text{ or } X \geq 50.9) = 1 - P(49.1 \leq X \leq 50.9)$$

$$= 0.01610$$

# Note:

Chebyshev's bound is generally much larger than the actual probability.

Hence should only be used when the distribution of the random variable is unknown.

# Transforming and Combining Random Variables

Prof. Preet Kanwal

# Linear Transformation on Random Variables

In this section:

1) We will learn how the mean and standard deviation are affected by transformations on Random Variables.
2) Study the effects of transformations on the shape, center and spread of the distribution.

# Why perform transformations??

- Change of Units
- To represent a different Situation.
- Example:

  The difference in blood pressure with and without taking a certain drug.

# Transformations to be Studied

- Addition – Adding a constant to each value of X.
- Subtraction – Subtracting a constant from each value of X.
- Multiplication – Multiplying each value of X by a constant.
- Division – Dividing each value of X by a constant.

where, X represents a Random Variable.

# Effect on a Random Variable of Multiplying (or Dividing) by a Constant

# Problem – Part A

- SRS travels offers a half-day trip in a tourist area.
- There must be at least 2 passengers for the visit to run.
- The vehicle provided by SRS travels can hold up to 6 passengers.
- Let X represent the No. Of passengers that turn up on a randomly selected day.

# Probability Distribution of X

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| p(x) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

1) Calculate average number of passengers that turn up for the visit.

2) Calculate SD.

# Probability Distribution of X

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| p(x) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

1) Mean = 3.75

2) SD = 1.090

# Problem – Part B

- SRS travels charges Rs. 150 per passenger.
- Let Y represent the amount SRS travels collects on a randomly selected day.
- Provide probability distribution of Y.
- Calculate mean and SD of Y.

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| p(x) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

Mean = 3.75

standard dev = 1.090

| Y | 300 | 450 | 600 | 750 | 900 |
|---|---|---|---|---|---|
| p(y) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

Mean = 562.50

standard dev = 163.50

# Effect on a Random Variable of Multiplying (or Dividing) by a Constant

- Multiplies (or divides) measures of center and location (mean, median, quartiles, percentiles) by same constant.
- Multiplies (or divides) measures of spread (range, IQR, standard deviation) by same constant.
- Does not change the shape of the distribution.
- Multiplying a random variable by a constant b multiplies the variance by $b^2$.

# Effect on a Random Variable of Adding (or Subtracting) a Constant

# Problem – Part C

- The amount spent on petrol and permit by SRS travels per trip is Rs. 100.
- Let Z represent the Profit made by SRS travels on a randomly selected day.
- Find the probability distribution of Z.
- Find mean and SD of Z.

| Y | 300 | 450 | 600 | 750 | 900 |
|---|---|---|---|---|---|
| p(y) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

Mean = 562.50

Standard dev = 163.50

| Z | 200 | 350 | 500 | 650 | 800 |
|---|---|---|---|---|---|
| p(z) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

Mean = 462.50

Standard dev = 163.50

# Effect on a Random Variable of Adding (or Subtracting) a Constant

- Adds (or subtracts) measures to center and location (mean, median, quartiles, percentiles) by same constant.
- Does not change measures of spread (range, IQR, standard deviation).
- Does not change the shape of the distribution.

# Note:

Whether we are dealing with data or random variables, the effects of a linear transformation are the same.

**Effect on a Linear Transformation on the Mean and Standard Deviation**

If $Y = a + bX$ is a linear transformation of the random variable $X$, then

- The probability distribution of $Y$ has the same shape as the probability distribution of $X$.
- $\mu_Y = a + b\mu_X$.
- $\sigma_Y = |b|\sigma_X$ (since $b$ could be a negative number).

# Combining Random Variables

# Combining Random Variables

- Many interesting statistics problems require us  to examine two or more random variables.
- Let's investigate result of adding and subtracting two random variables.

# Adding two Random Variables

# Problem – Find E(T) if T = X + Y

X – represents the number of passengers on a randomly selected tip with SRS travels.

| X | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| p(x) | 0.15 | 0.25 | 0.35 | 0.20 | 0.05 |

Mean = 3.75 and Standard Deviation = 1.090

Y – represents the number of passengers on a randomly selected tip with VRL logistics.

| Y | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| p(y) | 0.3 | 0.4 | 0.2 | 0.1 |

Mean = 3.10 and Standard Deviation = 0.943

# Distribution of T contains all possible combinations of X and Y

| $x_i$ | $p_i$ | $y_i$ | $p_i$ | $t_i = x_i + y_i$ | $p_i$ |
|---|---|---|---|---|---|
| 2 | 0.15 | 2 | 0.3 | 4 | (0.15)(0.3) = 0.045 |
| 2 | 0.15 | 3 | 0.4 | 5 | (0.15)(0.4) = 0.060 |
| 2 | 0.15 | 4 | 0.2 | 6 | (0.15)(0.2) = 0.030 |
| 2 | 0.15 | 5 | 0.1 | 7 | (0.15)(0.1) = 0.015 |
| 3 | 0.25 | 2 | 0.3 | 5 | (0.25)(0.3) = 0.075 |
| 3 | 0.25 | 3 | 0.4 | 6 | (0.25)(0.4) = 0.100 |
| 3 | 0.25 | 4 | 0.2 | 7 | (0.25)(0.2) = 0.050 |
| 3 | 0.25 | 5 | 0.1 | 8 | (0.25)(0.1) = 0.025 |
| 4 | 0.35 | 2 | 0.3 | 6 | (0.35)(0.3) = 0.105 |
| 4 | 0.35 | 3 | 0.4 | 7 | (0.35)(0.4) = 0.140 |
| 4 | 0.35 | 4 | 0.2 | 8 | (0.35)(0.2) = 0.070 |
| 4 | 0.35 | 5 | 0.1 | 9 | (0.35)(0.1) = 0.035 |
| 5 | 0.20 | 2 | 0.3 | 7 | (0.20)(0.3) = 0.060 |
| 5 | 0.20 | 3 | 0.4 | 8 | (0.20)(0.4) = 0.080 |
| 5 | 0.20 | 4 | 0.2 | 9 | (0.20)(0.2) = 0.040 |
| 5 | 0.20 | 5 | 0.1 | 10 | (0.20)(0.1) = 0.020 |
| 6 | 0.05 | 2 | 0.3 | 8 | (0.05)(0.3) = 0.015 |
| 6 | 0.05 | 3 | 0.4 | 9 | (0.05)(0.4) = 0.020 |
| 6 | 0.05 | 4 | 0.2 | 10 | (0.05)(0.2) = 0.010 |

# T – represents the total no of passengers that SRS and VRL can expect on a randomly selected day

E(X) = Mean of X= 3.75

E(Y) = Mean of Y = 3.10

E(T) = Mean of T = Mean of X + Mean of Y = 6.85

**Mean of the Sum of Random Variables**

For any two random variables $X$ and $Y$, if $T = X + Y$, then the expected value of $T$ is

$$E(T) = \mu_T = \mu_X + \mu_Y$$

# Problem – Find Var(T) where T = X + Y

**Variance of the Sum of Random Variables**

For any two *independent* random variables $X$ and $Y$, if $T = X + Y$, then the variance of $T$ is

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$$

In general, the variance of the sum of several independent random variables is the sum of their variances.

# Independent Random Variables

If knowing whether any event involving X alone has occured tells us nothing about the occurrence of any event involving Y alone, and vice versa, then,

X and Y are independent random variables.

Random variables describe outcomes that are unrelated to each other.

If they are **not** independent then a **covariance** term has to be introduced.

In our example we can assume X and Y to be independent since SRS and VRL and completely different organizations, operating independently.

# X + X not same as 2X

| | X | X +X | 2X |
|---|---|---|---|
| Values | 10, 18 | 20, 28<br>28, 36 | 20, 36 |
| mean | 14 | 28 | 28 |
| Variance | 16 | 32 | 64 |
| SD | 4 | 5.47 | 8 |
| Conclusion | | Adding values from the same distribution doubles mean and Variances. | Multiplying a Random Variable by a constant b mean is multiplied by b. Variances increase by square of that b. |

# Subtracting two Random Variables

## Mean of the Difference of Random Variables

For any two random variables $X$ and $Y$, if $D = X - Y$, then the expected value of $D$ is

$$E(D) = \mu_D = \mu_X - \mu_Y$$

In general, the mean of the difference of several random variables is the difference of their means. *The order of subtraction is important!*

## Variance of the Difference of Random Variables

For any two *independent* random variables $X$ and $Y$, if $D = X - Y$, then the variance of $D$ is

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$$

# Summary

# Linear Combination of Random Variables

If $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ are random variables

and

$c_1, c_2, c_3, \ldots\ldots\ldots\ldots\ldots c_n$ are constants

then,

$c_1X_1 + c_2X_2 + c_3X_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + c_nX_n$

is called linear combination of $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ .

# When performing transformation on a Random Variable:

- Multiplication and division affects mean and standard devaition.
- Addition and subtraction affects only mean.

# When combining two random variables

- Means combine very easily with addition or subtraction.
- You cannot add standard deviations.
- Variances can be added.
- Add Variances even if subtracting the random variable.

# Problem 1

If X and ,Y are independent random Variables such that

- E(X) = 9.5 and E(Y) = 6.8
- SD(X) = 0.4 and SD(Y) = 0.1
- Find Means and SD of the following:

1) 3X

2) Y – X

3) X + 4Y

# Problem 1 – Solution

| | Mean | SD |
|---|---|---|
| X | 9.5 | 0.4 |
| Y | 6.8 | 0.1 |
| 3X | 3 * 9.5 = 28.5 | 3 * 0.4 = 1.2 |
| Y – X | 6.8 – 9.5 = -2.7 | $\text{Sqrt}(0.1^2 + 0.4^2) = 0.4123$ |
| X + 4Y | 9.5 + 4 * 6.8 = 36.7 | $\text{Sqrt}(0.4^2 + (4 * 0.1)^2) = 0.5656$ |

# Problem 2

The lifetime of a certain lightbulb in a certain. application has mean 700 hours and standard deviation 20 hours.

As each bulb burns out, its replaced with a new bulb.

Find the mean and SD of the length of time that five bulbs will last.

# Problem 2 - Solution

Let X – represent the length of time the bulb lasts.

Hence for 5 bulbs we have a linear combination :

Let Y = X + X + X + X + X

E(X) = 700 hours SD(X) = 20 hour

**E(Y)** = E(X) + E(X) + E(X) + E(X) + E(X)

= 5 * 700 = **3500 hours**

**Var(Y)** = Var(X) + Var(X) + Var(X) + Var(X) + Var(X)

= 5 * $20^2$ = **2000 hours**

**SD(Y)** = sqrt(2000) = **44.72 hours**

# Problem 3

- Thickness X of a wooden shim(in mm)

  has pdf:

  $f(x) = 3/4 - (3(x - 5)^2)/4 \quad 4 \le x \le 6$

  $0 \quad$ otherwise

1) Find E(X) and Var(X)
2) Y -denotes thickness of shim in inches.

   1 mm = 0.0394 inches

   Find E(Y) and Var(Y)
3) If 3 shims are selected independently and stacked together. Find mean and variance of total thickness.

# Problem 3 - Solution

1) E(X) = 5 mm and Var(X)= 0.2 mm
2) Y -denotes thickness of shim in inches.

1 mm = 0.0394 inches

$E(Y) = 0.0394 * E(X) = 0.197$

$Var(Y) = 0.0394^2 * Var(X)$

3) If 3 shims are selected independently and stacked together. Find mean and variance of total thickness.

$Z = X_1 + X_2 + X_3$

$E(Z) = E(X1) + E(X2) + E(X3) = 3 * 5 = 15$ mm

$Var(Z) = Var(X1) + Var(X2) + Var(X3) = 3 * 0.2 = 0.6$ mm

# IID variables

If $X_1$, $X_2$, …, $X_n$ are independent random variables all with the same distribution.

Then they are called **independent and identically distributed (i.i.d)**

**Independent -** outcome of one obsv doesnt affect the outcome of other obsv

**identically distributed** – they have same mean and variance; possible outcomes will be same as the previous event with same probabilities.

In other words, the terms *random sample* and *IID* are basically one and the same.

# Example 1 : IID

- **Casino games :**
  - roulette wheel
  - roll of dice
  - deal of shuffled cards

result is independent of any other iteration

# Example 2 : IID

Coin toss repeated several times.

***independent*** since every time you flip a coin, the previous result doesn't influence your current result.

They are ***identically distributed***, since every time you flip a coin, the chances of getting head (or tail) are identical, no matter if its the 1st or the 100th toss (probability distribution is identical over time).

# Problem 4

- PDF of X:

  f(x) = 10 $9.95 < x < 10.05$

  0 otherwise

- PDF of Y:

  g(y) = 5 $4.9 < y < 5.1$

  0 otherwise

- Find P(X < 9.98 and Y > 5.01) if X and Y are independent.

# Problem 4 - solution

$P(X < 9.98 \text{ and } Y > 5.01) = 0.135$

# Independence and Simple Random Samples

**In practice**, we often do the **sampling without replacement,** that is, we do not allow one person to be chosen twice. [changes the distribution and probabilities]

However, **if the population is large**, then the probability of choosing one person twice is extremely low, and it can be shown that **the results obtained from sampling with replacement are very close to the results obtained using sampling without replacement.**

# Sample Mean – Most frequently encountered linear Combination.

Let's say we are looking at a sample of n random variables, $X_1$, $X_2$, …, $X_n$.

Let numberOfSamples = 40

Let the sample size be 100 (less than 5% of the population)

X1, X2,…, Xn designate the result of the 1st, 2nd, and nth sample from the population.

- Specifically, $X_1, X_2, \ldots, X_n$ is a simple random sample fom a population whose mean is $\mu$ and variance $\sigma^2$.
- Since, X1, X2,…, Xn are IID, they have the same distribution as that of the population.
- Hence, X1, X2,…, Xn has the same mean($\mu$), and variance $\sigma^2$ as that of the population.

# IID - Simulation

- Generate 3 samples($X_1$, $X_2$, $X_3$) of size 1000, 2000, 3000.
- Describe all three samples.
- How does the mean and standard dev of sample changes when compared to population mean and SD.

# Sampling Distribution of Sample mean

- It is the probability distribution of the mean.
- It is the distribution of the mean if we were to repeatedly draw samples from the population.
- In repeated sampling, the value of the sample mean would vary from sample to sample.
- The value of mean will be a random sample from the means sampling distribution.
- We will use mathematical arguments based on statistics sampling distribution to make statements about population parameters.
- Example: we are 95% confident that the sample mean lies within 22 units of population mean.

# Sampling distribution (histogram)- Using multiple samples, generate distribution of the sample mean

Write a function to generate a sampling distribution (histogram) of the mean height (or weight) of a sample.

Input to the function should be sample size and number of samples.

**Test your function using sample sizes of:**

10, 50, 100, and 1000 and

**number of samples of 100 and 1000.**

What do you observe?

# Facts

| No of Samples = 100 | | |
|---|---|---|
| Sample Size | Mean of Means | SD of means = Population SD / sqrt(noOfSamples) |
| 10 | Almost same | 0.55 > 0.19 |
| 50 | Almost same | 0.22 > 0.19 |
| 100 | Almost same | 0.19 > 0.19 |
| 1000 | Almost same | 0.057 > 0.19 |

| No of Samples = 1000 | | |
|---|---|---|
| Sample Size | Mean of Means | SD of means = Population SD / sqrt(noOfSamples) |
| 10 | Almost same | 0.60 > 0.060 |
| 50 | Almost same | 0.26 > 0.060 |
| 100 | Almost same | 0.185 > 0.060 |
| 1000 | Almost same | 0.059 > 0.060 |

# Problem 5

- A machine that fills cardboard boxes with cereal has a fill weight whose mean is 12.02 oz, with a standard deviation of 0.03 oz.
- A case consists of 12 boxes randomly sampled from the output of the machine.

1) Find the mean of the total weight of the cereal in the case.
2) Find the standard deviation.
3) Find the mean of the average weight per box of the cereal in the case.
4) Find the standard deviation of the average weight box.
5) How many boxes must be included in a case for the standard deviation of the average weight per box to be 0.005 oz?

# Problem 5 - Solution

- Mean weight of cereals in cardboard boxes = 12.02
- Standard deviation in weight of cereals in cardboard boxes = 0.03
- case consists of 12 boxes randomly sampled

$X_1, X_2, \ldots, X_{12}$

1) **Mean of the total weight of the cereal in the case:**

$X_1, X_2, \ldots, X_{12}$ are IID hence their mean = 12.02 and SD = 0.03

$Y = X_1 + X_2 + \ldots + X_{12}$

$E(Y) = E(X_1) + E(X_2) + \ldots + E(X_{12})$

$= 12 * 12.02 = 144.24$

2) **SD $= 12 * 0.03^2$**

# Problem 5 - Solution

**3) Mean of the average weight per box of the cereal in the case.**

$X_1, X_2, \ldots, X_{12}$ are IID hence their mean = 12.02 and SD = 0.03

$$\mu_{\bar{X}} = \mu$$

4) **SD of the average weight = 0.03 / sqrt(12)**

# Problem 5 - Solution

5) **standard deviation of the average weight per box to be 0.005 oz**

No of boxes to be included – n??

SD of the average weight = 0.03 / sqrt(n)

0.05 = 0.03 / sqrt(n)

# Problem 6

Two independent measurements are made of the lifetime of a charmed strange meson.

Each measurement has a SD of $7 * 10^{-15}$ seconds.

The lifetime of the meson is estimated by averaging the two measurements.

What is the standard deviation of this estimate?

# Problem 6 – Solution

Let the two measurements be X1, X2

SD(X1) = SD(X2) = $7 * 10^{-15}$ seconds.

The lifetime of the meson is estimated by averaging the two measurements.

SD of this estimate = $7 * 10^{-15}$ / sqrt(2)

= $4.95 * 10^{-15}$ seconds.

# Discrete Probability Distributions

Prof. Preet Kanwal

**A discrete probability distribution is a table (or a formula) listing all possible values that a discrete variable can take on, together with the associated probabilities.**

# Outline

1) Bernoulli Distribution
2) Binomial Distribution
3) Geometric Distribution
4) Negative Binomial Distribution
5) Poisson Distribution
6) HyperGeometric Distribution
7) Multinomial Distribution

# Bernoulli Distribution

Prof. Preet Kanwal

# Bernoulli Distribution

Jacob Bernoulli
( Swiss mathematician of the 17th century.)
(1654 – 1705)
Discovered constant e

A Bernoulli trial is a process that results in one of two possible outcomes.

Example:

- Tossing a coin
- Any yes or no question
- Success or failure
- Is the top card of a shuffled deck an ace?
- Was the newborn child a girl?

# Conditions of Bernoulli Distribution

- Its a Single Trial.

- The trial can result in one of the two possible outcomes, labelled success and failure.

- P(success) = p

The term "success" in this sense consists in the result meeting specified conditions

- P(failure) = 1 – p

More generally, given any probability space, for any event (set of outcomes), one can define a Bernoulli trial, corresponding to whether the event occurred or not.

# X ~ Bernoulli(p)

For any Bernoulli Trial,

A Random Variable X is defined as :

X = 1 if success occurs, where probability of success is denoted by p

X = 0 if Failure occurs, where probability of failure is (1 – p)

then X is said to have a Bernoulli distribution with probability p.

**Note: A Bernoulli Random Variable can only take values 0 and 1.**

# Probability distribution of X: : X ~ Bernoulli(p)

| x | f(x) |
|---|---|
| 0 | 1 – p |
| 1 | p |

Mean = $0 (1 – p) + 1 * p = p$

Variance = $(0 – p)^2 (1 – p) + (1 – p)^2 *(p) = p*(1 – p)$

# Problem 1

Approximately 1 in 200 American adults are lawyers.

One American adult is randomly selected.

What is the distribution of the number of lawyers?

# Problem 1 - Solution

X – represents the American adult is a lawyer.

X ~ Bernoulli (1 /200)

Probability Distribution of X

| X | p(x) |
|---|---|
| 0 | 199/200 |
| 1 | 1/200 |

# Problem 2

Suppose that a student takes a multiple choice test.

The test has 10 questions, each of which has 4 possible answers **(only one is correct).**

If the student blindly guesses the answer to each question, do the questions form a sequence of Bernoulli trials? If so, identify the trial outcomes and the parameter p.

# Problem 2 - Solution

For each question,
Either the answer chosen is correct or incorrect
P(Answer is correct) = ¼
P(Answer is incorrect) = ¾
Hence there are only 2 possible outcomes for each question.
Hence each question is a Bernoulli trial.

Since there are in total 10 questions, we have a sequence of Bernoulli trials.

# Problem 3

Candidate A is running  for office in a certain district.

Twenty persons are selected at random from the population of registered voters and asked if they prefer candidate A.

Do the responses form a sequence of Bernoulli trials?

If so identify the trial outcomes and the meaning of the parameter p.

# Problem 3 - Solution

Yes, the responses form a sequence of Bernoulli trials.

The outcomes are:
prefer A
and do not prefer A;

p is the proportion of voters in the entire district who prefer A

# Joining two Bernoulli Random variables

# Problem 4

At a certain fast-food restaurant,
25% of drink orders are for a small drink
35% for a medium drink
40% for a large drink
X = 1 if a randomly chosen order is for a small drink and 0 otherwise.
Y = 1 if a randomly chosen order is for a medium drink and 0 otherwise.
Z = 1 if a randomly chosen order is for a small or a medium drink and 0 otherwise.

1.Find Probability distribution of X , Y, Z.
2.Is it possible for both X and Y to be equal to 1?
3.Does $p_z = p_x + p_y$? Where $p_z$ , $p_x$ , $p_y$ denote success probabilities of Z, X, Y respectively.
4.Does Z = X + Y?

# Problem 4 -Solution

2. Is it possible for both X and Y to be equal to 1?
No. If the order is for small drink, it cannot also be for a medium drink.
Orders are independent.

3. Yes. $p_z = 0.60 = 0.25+0.35= p_x + p_y$

Mutually exclusive events. For one order if X occurs Y cannot occur and vice versa.

4. No Z ≠X +Y.

# Problem 5


A penny and a nickel are tossed. Both are fair coins.

X = 1 if the penny comes up heads, 0 otherwise.
Y = 1, if nickel comes up heads, 0 otherwise.
Z = 1, if both penny and nickel comes up heads, 0 otherwise.

1.Find PMF of X, Y, Z.
2.Are X and Y independent?
3.Does $p_z = p_x * p_y$?Where $p_z$ , $p_x$ , $p_y$ denote success probabilities of Z, X, Y respectively.
4.Does Z = X Y?

# Problem 5 - Solution

1.Are X and Y independent?

Yes.

Does $p_z = p_x * p_y$?

Yes. $P_z = 1/4 = 1/2 * 1/2 = p(x) * p(y)$

2.Does Z = X Y?

Yes.

- If both coins come up heads, then X = 1, Y = 1, and Z = 1, so Z = XY.
- If not, then Z = 0, and either X, Y, or both are equal to 0 as well, so again Z = XY.

# Problem 6

**Two dice are rolled**
X=1, if dice come up doubles.
X=0, otherwise.

Y=1, if sum is 6.
Y=0, otherwise.

Z=1, If dice come up both doubles and with sum of 6.
Z=0, otherwise.

1.Find PMF of X,Y,Z?
2.Are X and Y independent?
3.Does $p_z = p_x * p_y$?Where $p_z$ , $p_x$ , $p_y$ denote success probabilities of Z, X, Y respectively.
4.Does Z = X Y?

# Problem 6 - Solution

| X | P(x) |
|---|---|
| 0 | 30/36 |
| 1 | 6/36 |

| Y | P(y) |
|---|---|
| 0 | 31/36 |
| 1 | 5/36 |

| Z | P(z) |
|---|---|
| 0 | 35/36 |
| 1 | 1/36 |

1.X and Y are not independent.

2. Does $p_z = p_x * p_y$

No, $P(Z = 1) \neq P(X = 1)P(Y = 1)$.

3.Does Z = X Y? No

# Bernoulli Distribution Applications

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes, such as:
(a) The state of a telephone at a given time that can be either free or busy.
(b) A person who can be either healthy or sick with a certain disease.
(c) The preference of a person who can be either for or against a certain political candidate.

# Other common Discrete Probability Distributions based on the assumption of **independent Bernoulli trials:**

## Binomial Distribution

Binomial distribution is the distribution of number of successes in n independent Bernoulli trials

## Geometric Distribution

Geometric distribution is the distribution of number of trials to get the first success in n independent Bernoulli trials

## Negative Binomial Distribution

Negative binomial distribution is the distribution of the number of trials required to get $r^{th}$ success in n independent Bernoulli trails

# Binomial Distribution

Prof. Preet Kanwal

# Binomial Distribution properties

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

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

# 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 ,

$$Y_i \sim \text{Bernoulli}(p)$$

$Y_i = 1$ 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.

$$X = Y_1 + Y_2 + .... + Y_n$$

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

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

# 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.

3. Asking 100 people if they have ever been to Paris.

# 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.

# Problem 1

A coin is flipped 3 times.

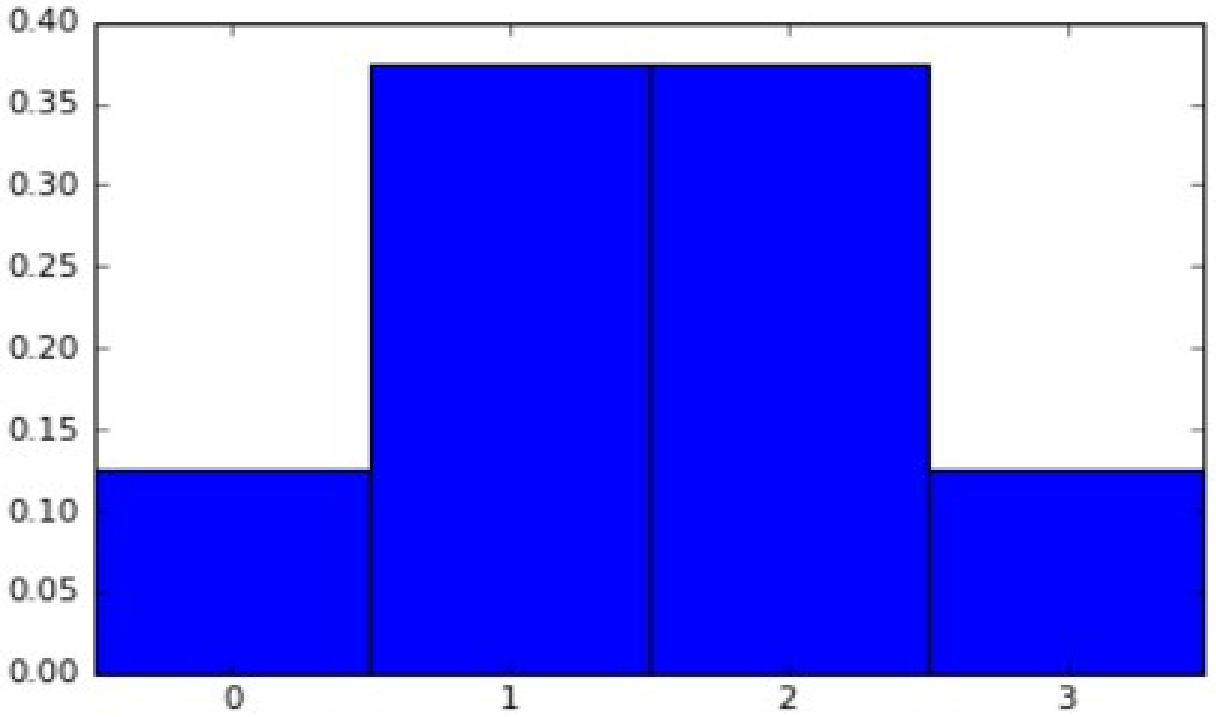
What is the probability head turns up

Exactly 2 times.

# Problem 1 - Solution

Probability mass function

| X | P(x) |
|---|---|
| 0 | 0.125 |
| 1 | 0.375 |
| 2 | 0.375 |
| 3 | 0.125 |



# Probability mass function of a Binomial Random Variable

number of trials = 3

P(success) = 0.5 , success – getting heads

Exactly 2 heads: P(HHT) + P(HTH) + P(THH)

$= 3 * p^2 * (1-p)^1$

$= 3C2 * p^2 * (1-p)^{3-1}$

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

$$\mathbf{P(X = x) = {}^{n}C_{x} * p^{x} * (1 - p)^{n-x}}$$

# Mean & Variance of Binomial Distribution??

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

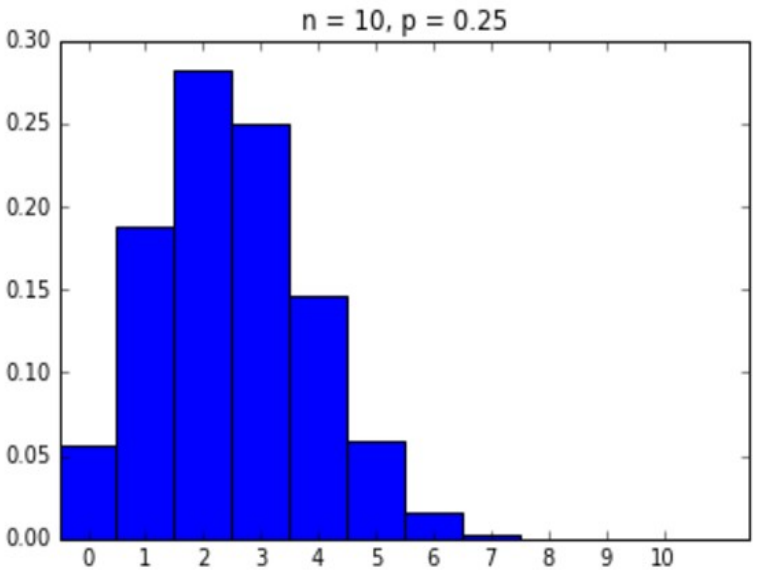
$$\text{Variance} = n* ((0 - p)^2\,(1 - p) + (1 - p)^2 \ *(p)\ )$$
$$= n\,p(1 - p)$$

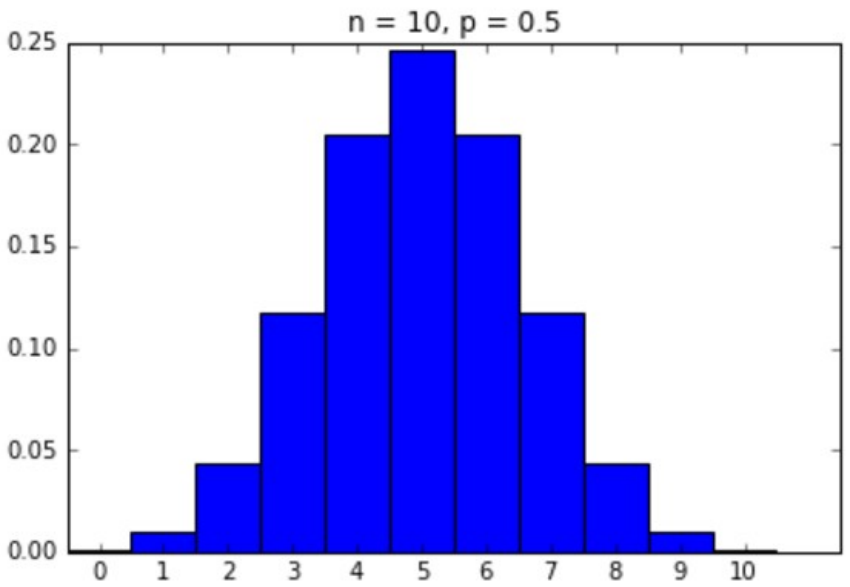
# CodingAssignment – Part A

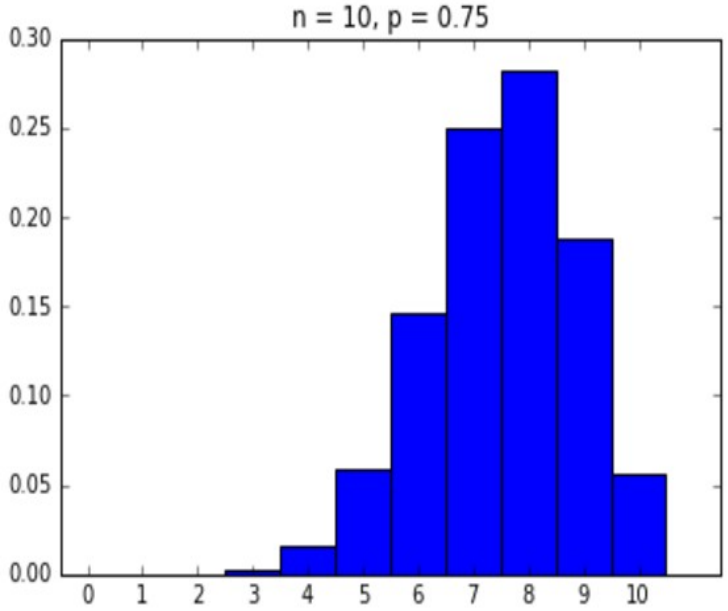
Find the effect of changing p when n is fixed.
a) n = 10, p = 0.10
b) n = 10, p = 0.5
c) n = 10, p = 0.90

# Coding Assignment - observation

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



# CodingAssignment – Part B

Find the effect of changing p when n is fixed.
a) n = 100, p = 0.25
b) n = 100, p = 0.5
c) n = 100, p = 0.75

# Problem 2

A balanced six sided die is rolled 3 times.

What is the probability a 5 comes up exactly twice?

# Problem 2 - Solution

A balanced six sided die is rolled 3 times.

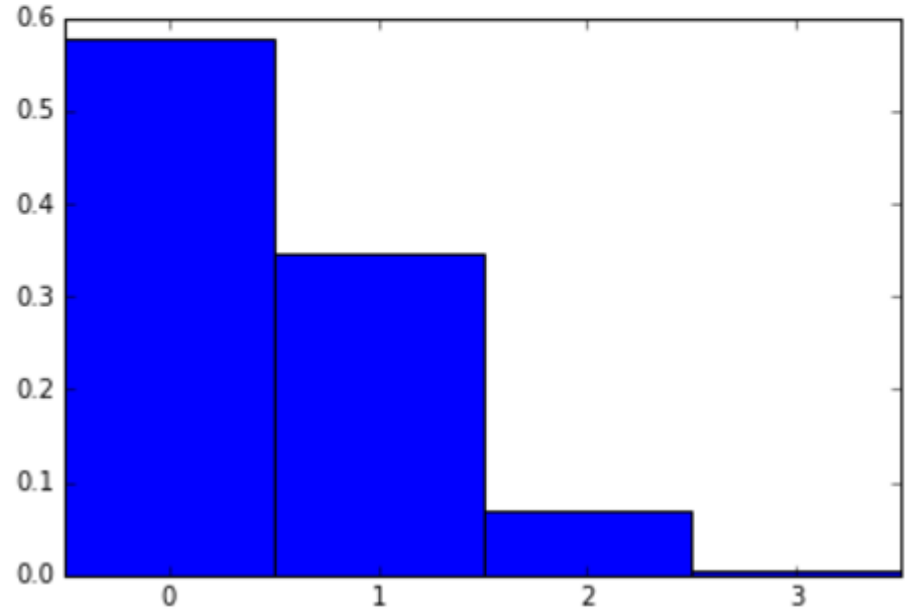
Let X represent number of 5's in 3 rolls.

$X \sim Bin(3, 1/6)$

$P(X = 2) = {}^{3}C_{2} * (1/6)^{2} * (5/6)^{3-2}$

$= 0.0694$

| X | p(x) |
|---|---|
| 0 | 0.578703009259537 |
| 1 | 0.3472226388883889 |
| 2 | 0.06944469444461111 |
| 3 | 0.004629657407462964 |
| total | 1.0 |



# Problem 3

A coin is flipped 100 times.

What is the probability head turns up

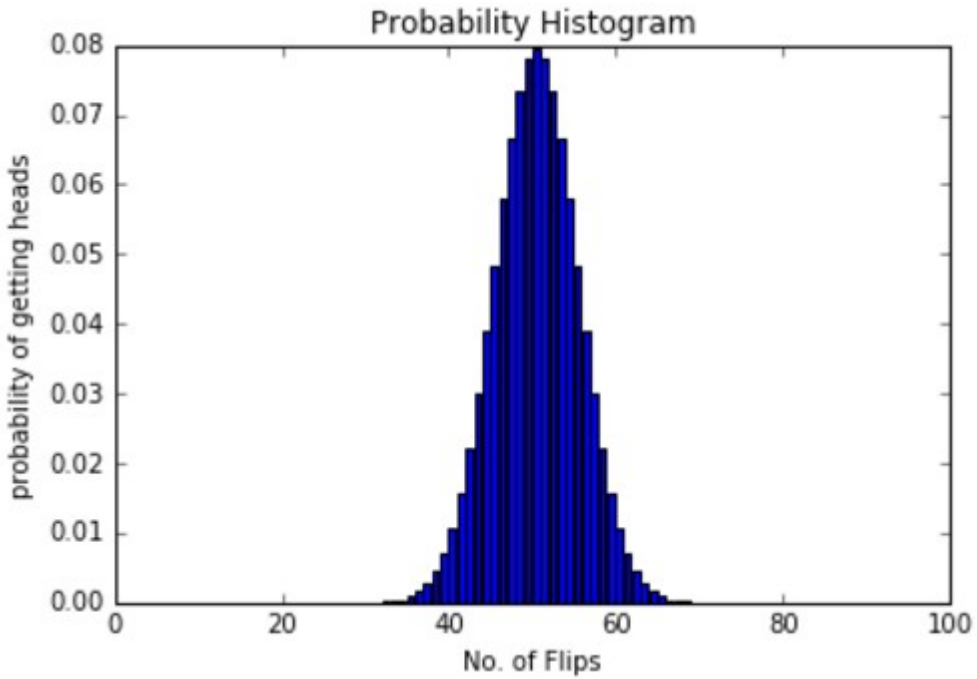
Exactly 60 times.

# Problem 3 - Solution

n = 100
x = 60

Exactly 60 heads:

$$P(X = 60) = {}^{100}C_{60} * (0.5)^{60} * (0.5)^{40}$$

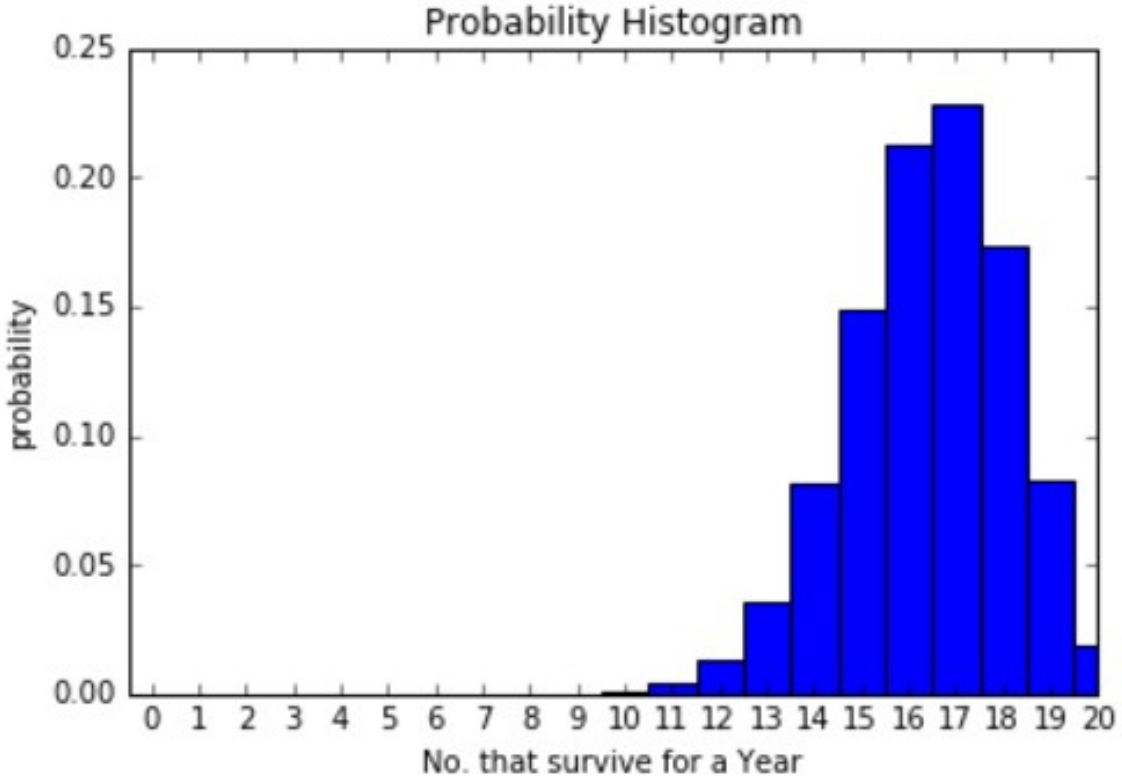


# Problem 4

According to Statistics Canada life tables, the probability a randomly selected 90 year-old Canadian male survives for at least another year is approximately 0.82.

If twenty 90 year-old Canadian males are randomly selected, what is the probability atleast 18 survive for at least another year?

# Problem 4 – Solution




$$P(X >= 18) = P(X = 18) + P(X = 19) + P(X = 20)$$
$$= 0.173 + 0.083 + 0.019$$
$$= 0.275$$

# Problem 5

The ratio of boys to girls at birth in Singapore is quite high at 1.09:1.

What proportion of Singapore families with exactly 6 children will have at least 3 boys? (Ignore the probability of multiple births.)

# Problem 5 - Solution

The probability of getting a boy is $\frac{1.09}{1.09 + 1.00} = 0.5215$

Let $X$ = number of boys in the family.

Here,

$$n = 6,$$
$$p = 0.5215,$$
$$q = 1 - 0.52153 = 0.4785$$

When $x = 3$:

$$P(X) = C_x^n p^x q^{n-x} = C_3^6 (0.5215)^3 (0.4785)^3 = 0.31077$$

When $x = 4$:

$$P(X) = C_4^6 (0.5215)^4 (0.4785)^2 = 0.25402$$

When $x = 5$:

$$P(X) = C_5^6 (0.5215)^5 (0.4785)^1 = 0.11074$$

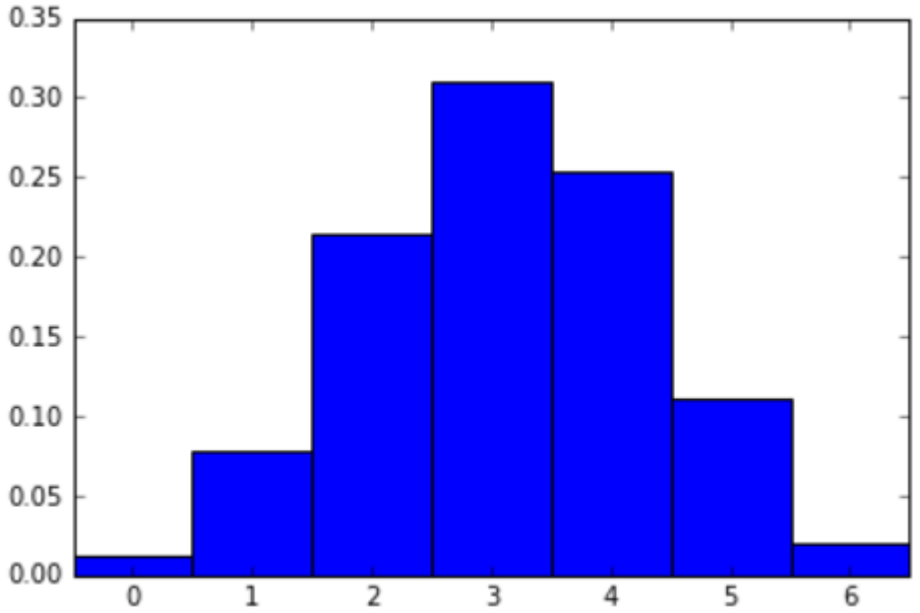
When $x = 6$:

$$P(X) = C_6^6 (0.5215)^6 (0.4785)^0 = 2.0115 \times 10^{-2}$$

So the probability of getting at least 3 boys is:

$$\text{Probability} = P(X \geq 3)$$
$$= 0.31077 + 0.25402 + 0.11074 + 2.0115 \times 10^{-2}$$
$$= 0.69565$$

| X | p(x) |
|---|---|
| 0 | 0.012003051035714148 |
| 1 | 0.07849017072256961 |
| 2 | 0.21385905972737748 |
| 3 | 0.3107697656505115 |
| 4 | 0.2540226219227927 |
| 5 | 0.11074006046623731 |
| 6 | 0.020115270474797196 |
| total | 0.9999999999999999 |



# What if p is not given? (Success probability of a Bernoulli Trial is unknown)

For example: Let's say we want to know the percentage of people in the population that are left-handed.

It would be impossible to measure every single person in the world, so we take a sample of 500 people and create a proportion

# Estimating p – Sample proportion $\hat{p}$

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

$$\text{Sample proportion} = \frac{\text{Count the number of success}}{\text{number of trials}}$$

$$\hat{p} = \frac{X}{n} \text{ where } X = \text{number of successes in sample}$$

- Sample proportion is **just an estimate of p and is not equal to p.**
- If we take **another sample**, the **value of sample propotion** might come out **different**ly. Hence, there might be some **uncertainity in the estimated value**

# Computing bias of $\hat{p}$

**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.

$$\mu_{\hat{p}} - p$$

# Computing uncertainity of $\hat{p}$

**Uncertainity** – is the standard deviation of sample proportion.

$$\sigma_{\hat{p}} = 1/n * \sigma_x$$

$$\sigma_{\hat{p}}^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

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

As p is unknown when computing uncertainity in sample proportion , we approximate p with $\hat{p}$

# Sampling Distribution of $\hat{p}$

**Use csv : height–weight.csv**

**Problem statement : to find out the proportion of people who are overweight**

sampling(SampleSize, numberOfSamples)

For each item in a sample,

1) Count the no of people who have weight > 65 Kg
2) Find the sample proportion = count / SampleSize

Append the value of sample proportion to a list called **p_hats**

Print mean and SD of p_hats

Print expected SD of p_hats using

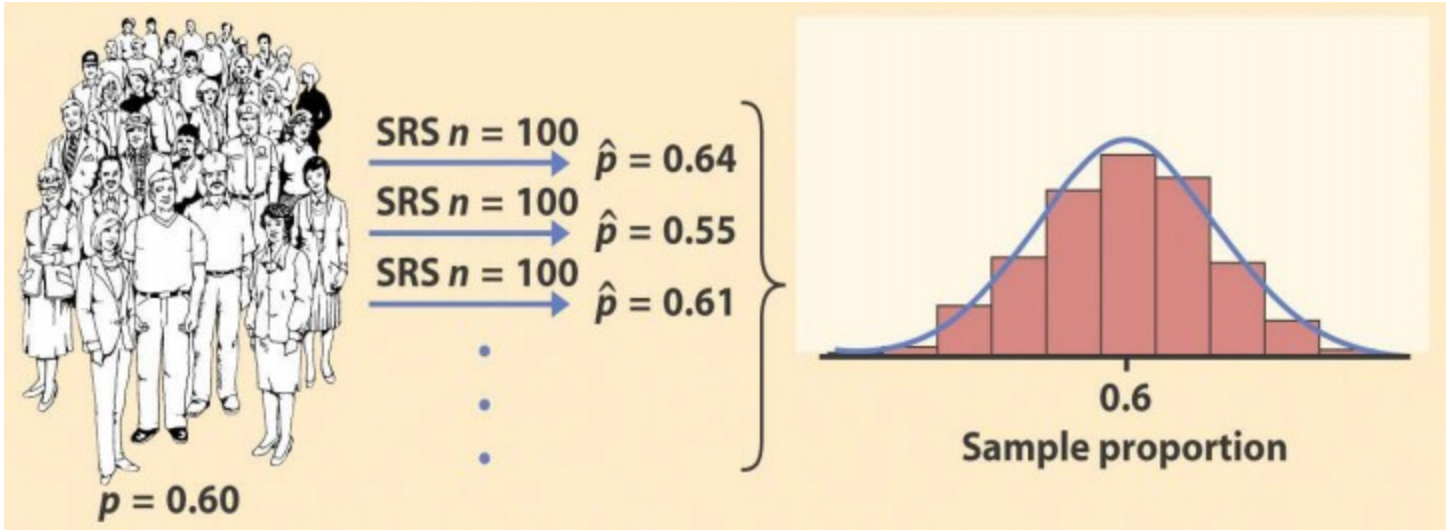
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Call sampling with following cases :

Case 1: noOfSamples : 100 , SampleSize : 50, 100, 1000

Case 2: noOfSamples : 1000 , SampleSize : 50, 100, 1000

# Uncertainity in the Sample proportion

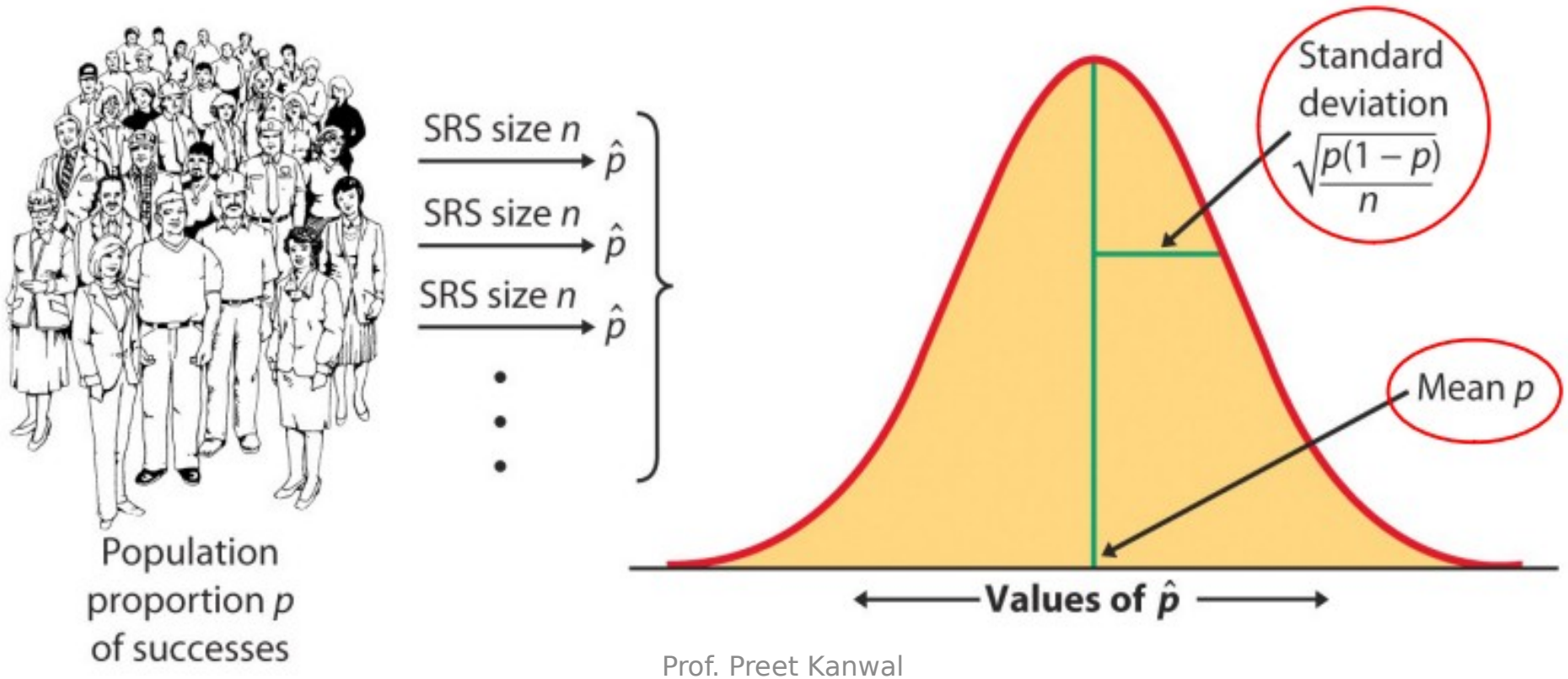



- Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic.
- This is called sampling variability.
- If we take a lot of random samples of the same size from a given population, the variation from sample to sample—the 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.

# 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.



# Note:

- The larger the sample size (n) or the closer p is to 0.50, the closer the distribution of the sample proportion is to a normal distribution.

- We refer to the standard deviation of a sampling distribution as the standard error.

# Problem 6

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%?

# Problem 6 – Solution

Solution

1) Sample proportion = 92/100

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

3) Given,

Uncertainity = 0.01

P =0.92

n = ?

n = p ( 1- p) / square(uncertainity)
= 0.92 * 0.08 / square(0.01)
= 736

# Problem 7

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.
(a) 0.0200
(b) 0.0469
(c) 0.0634
(d) 0.0125

2) Calculate uncertainity in the estimate.

# Problem 7 - Solution

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.

(a) 0.0200

**(b) 0.0469**

(c) 0.0634

(d) 0.0125

# Problem 7 - Solution

1) $\hat{p} = X / n = 15/320 = 0.0469$ - option (b)

2) SE or SD or uncertainity = 0.0118 as n = 320, $\hat{p} = 0.0469$

# Problem 8

In a random sample of 1000 students, $\hat{p} = 0.80$ (or 80%) were in favor of longer hours at the school library. The standard error of $\hat{p}$ (the sample proportion) is

A. 0.013

B. 0.160

C. 0.640

D. 0.800

# Problem 9

Consider a random sample of 100 females and 100 males. Suppose 15 of the females are left-handed and 12 of the males are left-handed. What is the estimated difference between population proportions of females and males who are left-handed (females − males)? Select the choice with the correct notation and numerical value.

A. $p_1 - p_2 = 3$

B. $p_1 - p_2 = 0.03$

C. $\hat{p}_1 - \hat{p}_2 = 3$

D. $\hat{p}_1 - \hat{p}_2 = 0.03$

# Problem 10

Which of the following is NOT true about the standard error of a statistic?

A. The standard error measures, roughly, the average difference between the statistic and the population parameter.

B. The standard error is the estimated standard deviation of the sampling distribution for the statistic.

C. The standard error can never be a negative number.

D. The standard error increases as the sample size(s) increases.

# Problem 11

Which of the following binomial distributions has the largest standard deviation?

1) N = 10, p = 0.5

2) N = 10, p = 0.75

3) N = 20, p = 0.5

4) N = 20, p = 0.75

# Problem 11 - Solution

Which of the following binomial distributions has the largest standard deviation?

1) N = 10, p = 0.5

2) N = 10, p = 0.75

**3) N = 20, p = 0.5**

4) N = 20, p = 0.75

The smaller the difference between p and 0.5 and the larger the N, the larger the standard deviation.

# Application – Binomial Distribution

In studies of public health, the binomial distribution is used to know the number of times a particular event will occur in a sequence of observations. I**nterested 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.

# Geometric Distribution

Prof. Preet Kanwal

# Geometric Distribution : X ~ Geom(p)

A series of independent Bernoulli trials are conducted until a success occurs, and a random variable X is defined as either:
The number of trials in the series, or
The number of failures in the series.
In either case, the geometric distribution is defined as the probability distribution of X.

X – represents the number of trials up to and including the first success.

For the first success to occur on the $i^{th}$ trail:
1.The first x – 1 trials must be failures.
2.The $x^{th}$ trial must be success.

X is a discrete random variable, which has geometric distribution with parameter p,

$$X \sim Geom(p)$$

## If X ~ Geom(p), PMF of X :

$$p(x) = P(X = x) = \begin{cases} (1-p)^{x-1} * p & \text{for } x = 1, 2, 3..... \\ 0 & \text{otherwise} \end{cases}$$

## If X ~ Geom(p), CDF of X :

$$F(x) = P(X <= x) = 1 - (1-p)^{x} \qquad \text{for } x = 1, 2, 3.....$$

## If X ~ Geom(p), Mean and Variance of X :

$$\text{Mean} = 1/p$$

$$\text{Variance} = (1-p)/p^2$$

# Binomial vs. Geometric Distribution

| Binomial | Geometric |
| --- | --- |
| Has a **FIXED number of trials** before the experiment begins and X counts the number of successes obtained in that fixed number. | Has a **fixed number of successes** (ONE...the FIRST) and counts the number of trials needed to obtain that first success. |
| **Variable of interest :** no of successes in a given no of trials. | **Variable of interest** : no of trials required to get first success |
| X – represent no of successes | X – represents no of trials |
| **Example:**<br>1)Getting 3 heads when a coin is tossed 10 times.<br>2)Getting a 5 twice in 3 rolls of a die. | **Example:**<br>1) flip a coin UNTIL you get a head<br>2) roll a die UNTIL you get a 3<br>1)attempt a three-point shot in basketball UNTIL you make a basket |

# Problem 1

A doctor is seeking an anti-depressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is p=0.6. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on?

What is the expected number of drugs that will be tried to find one that is effective?

# Problem 1 - Solution

The probability that any given drug is effective (success) is p=0.6.

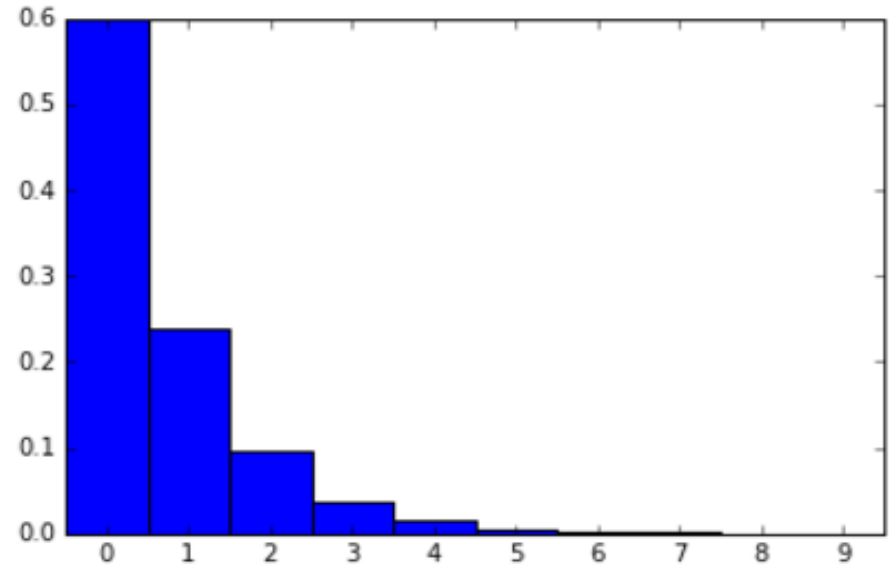
The probability that a drug will not be effective (fail) is q = 1 – p = 1 – 0.6 = 0.4.

Here are probabilities of some possible outcomes.

(i) The first drug works.
(ii) The first drug fails, but the second drug works.

Expected number of drugs that will be tried to find one that is effective = Mean = 1/ p = 1/0.6 = 1.667

| X | p(x) |
|---|---|
| 1 | 0.6 |
| 2 | 0.24 |
| 3 | 0.09600000000000002 |
| 4 | 0.03840000000000001 |
| 5 | 0.01536000000000002 |
| 6 | 0.006144000000000001 |
| 7 | 0.0024576000000000008 |
| 8 | 0.0009830400000000003 |
| 9 | 0.0003932160000000002 |
| 10 | 0.00015728640000000008 |



# Problem 2

A representative from the National Football League's Marketing Division randomly selects people on a random street in Kansas City, until he finds a person who attended the last home football game.

Let p, the probability that he succeeds in finding such a person, equal 0.20.

And, let X denote the number of people he selects until he finds his first success.

1)What is the probability that the **4$^{th}$ person selected** by the marketing representative is the one who attended the last home football game?

2)What is the probability that the marketing representative **must select more than 6** people before he finds one who attended the last home football game?

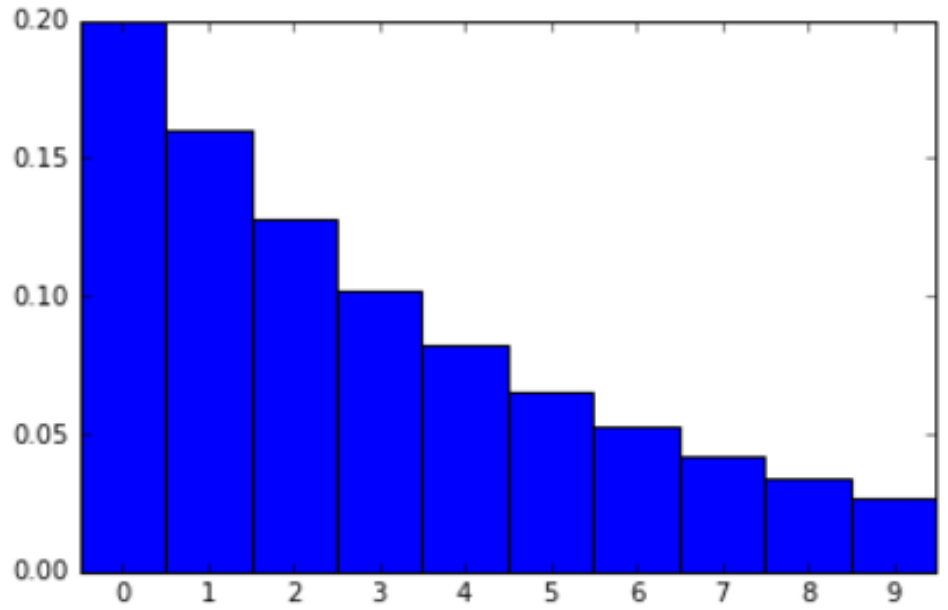
# Problem 2 – Solution part a

To find the desired probability, we need to find P(X = 4), which can be determined readily using the p.m.f. of a geometric random variable with
p = 0.20
1−p = 0.80
and x = 4:

$P(X=4)=0.80^3 * 0.20=0.1024$

There is about a 10% chance that the marketing representative would have to select 4 people before he would find one who attended the last home football game.

| X | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.16000000000000003 |
| 3 | 0.12800000000000003 |
| 4 | 0.10240000000000003 |
| 5 | 0.08192000000000002 |
| 6 | 0.06553600000000002 |
| 7 | 0.052428800000000025 |
| 8 | 0.041943040000000015 |
| 9 | 0.033554432000000016 |
| 10 | 0.026843545600000015 |



# Problem 2 – Solution part b

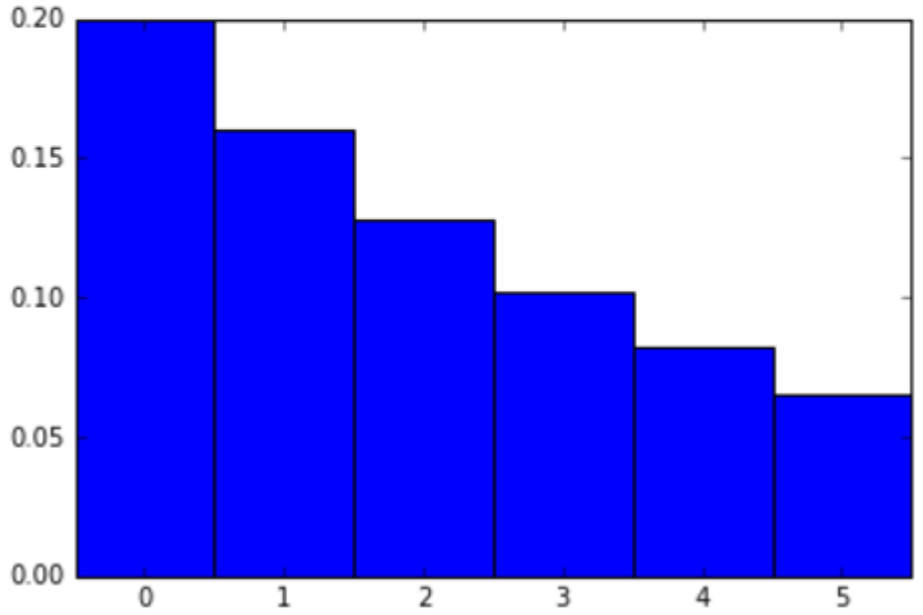
We need to find $P(X > 6) = 1 - P(X \leq 6)$,

which can be determined readily using the c.d.f. of a geometric random variable with
$1 - p = 0.80$, and $x = 6$:

$$
\begin{aligned}
P(X > 6) &= 1 - P(X \leq 6) \\
&= 1 - (1 - [1 - 0.20]^6) \\
&= 1 - (1 - 0.80^6) \\
&= 1 - 0.7378 \\
&= 0.262
\end{aligned}
$$

There is about a 26% chance that the marketing representative would have to select more than 6 people before he would find one who attended the last home football game.

| X | p(x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.16000000000000003 |
| 3 | 0.12800000000000003 |
| 4 | 0.10240000000000003 |
| 5 | 0.08192000000000002 |
| 6 | 0.06553600000000002 |
| total | 0.7378560000000002 |



# Problem 3

In a large population of adults, 30% have received CPR training.

If adults from this population are randomly selected, what is the probability that the $6^{th}$ person  sampled is the first that has received CPR training.

# Problem 3 – Solution

In a large population of adults, 30% have received CPR training.

$\Rightarrow \quad p = 0.3$

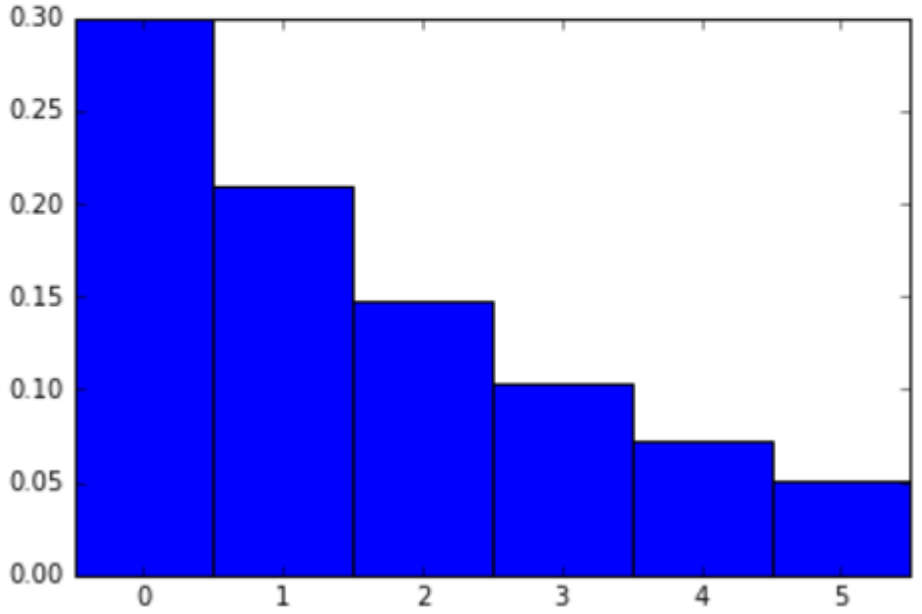
$\Rightarrow 1- p = 0.7$

If adults from this population are randomly selected, what is the probability that the $6^{th}$ person sampled is the first that has received CPR training.

$p(6) = P(X = 6) = 0.7^{6-1} * 0.3$

$= 0.0504$

| X | p(x) |
|---|---|
| 1 | 0.3 |
| 2 | 0.21 |
| 3 | 0.14699999999999996 |
| 4 | 0.10289999999999998 |
| 5 | 0.07202999999999998 |
| 6 | 0.05042099999999998 |



# Application – Geometric Distribution

Since we are seeking the first success at whatever trial it occurs, geometric simulations are called **"waiting time" simulations.**

It is useful for modelling situations in which it is necessary to know **how many attempts are likely necessary for success,** and thus has applications to population modelling, econometrics, ROI of research, and so on.

Example: A patient is waiting for a **suitable matching kidney donor for a transplant**. If the probability that a randomly selected donor is a suitable match is p=0.1, what is the expected number of donors who will be tested before a matching donor is found?

# Negative Binomial Distribution

Prof. Preet Kanwal

# Properties of Negative Binomial Distribution

- The experiment consists of **X repeated trials.**
- Each trial can result in just **two possible outcomes**. We call one of these
- outcomes a success and the other, a failure.
- The probability of success, denoted by **p, is the same on every trial.**
- The **trials are independent**; that is, the outcome on one trial does not affect the outcome on other trials.
- The experiment continues **until r successes are observed**, where r is specified in advance.

**Negative binomial probability -** the probability that rth success occurs on the xth trial, when the probability of success on an individual trial is p.

# Example

What is the probability of getting $3^{rd}$ head on $7^{th}$ toss of a coin??

X represent the no of trials requires to get $3^{rd}$ success.

$X \sim NB(3, 0.5)$

This means first two heads would have occurred in first six flips.

${}^{6}C_{2} * p^{2} * (1 - p)^{4}$

$P(X = 7) = {}^{6}C_{2} * p^{2} * (1 - p)^{4} * p$

$= {}^{6}C_{2} * p^{3} * (1 - p)^{4}$

$= {}^{(x-1)}C_{(r-1)} * p^{r} * (1 - p)^{x-r}$

## If X ~ NB(r, p), PMF of X :

$$p(x) = P(X = x) = \begin{cases} {}^{(x-1)}C_{(r-1)} * p^r * (1-p)^{x-r} & \text{for } x = r, r+1, r+2.... \\ 0 & \text{otherwise} \end{cases}$$

Smallest possible value of X is r .

As it requires atleast r trails to produce r successes.

=> **NB(1, p) = Geom(p)**

# Negative Binomial Random Variable = Sum of Geometric Random Variables

## If X ~ NB(r, p)

$$X = Y_1 + Y_2 + .... + Y_r$$

Where, $Y_i \sim Geom(p)$

$Y_i$ are independent random variables each with Geom(p) distribution.

$Y_1$ denotes no of trials upto and including the first success.

$Y_2$ denotes no of trials starting with first trail after the first success, upto and including the second success.

So on....

# If X ~ Geom(p), Mean and Variance of X :

Mean = r/ p

Variance = r [(1 – p )/ $p^2$)]

# Problem 1

Bob is a high school basketball player. He is a 70% free throw shooter. That means his probability of making a free throw is 0.70. During the season,

1) What is the probability that Bob makes his third free throw on his fifth shot?

2) What is the probability that Bob makes his first free throw on his fifth shot?

# Problem 1 – Part a Solution

Probability that Bob makes his third free throw on his fifth shot?

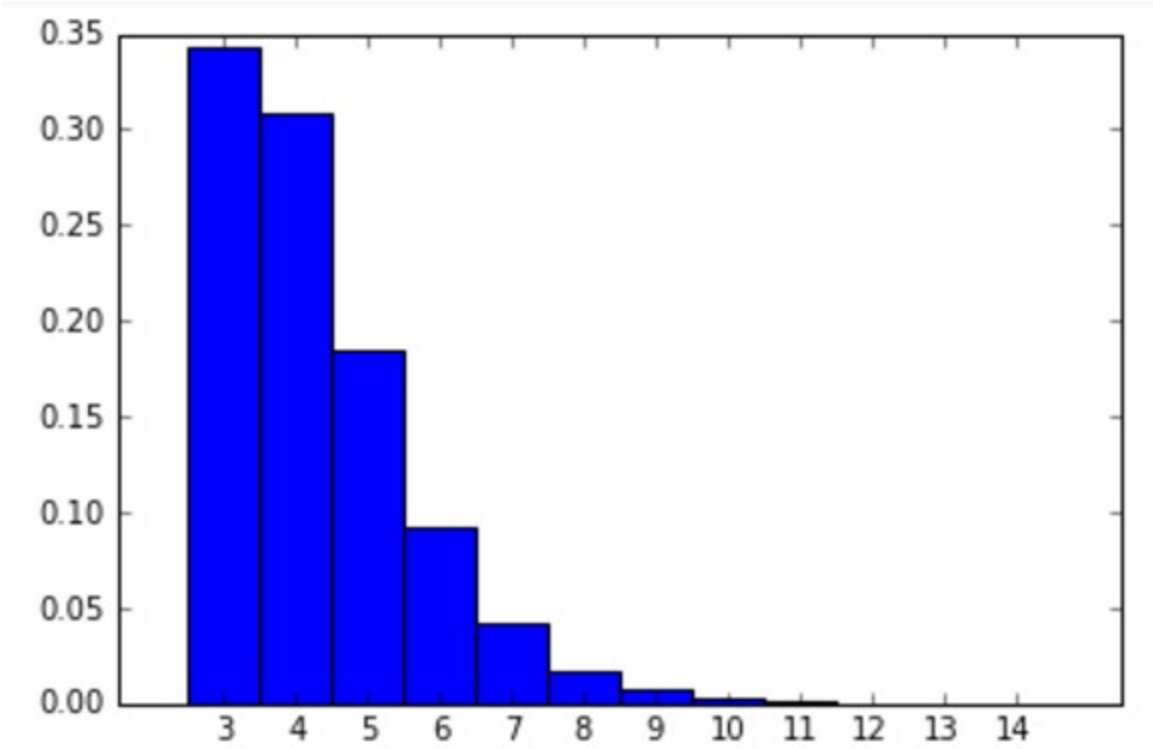
$X \sim NB(3, 0.70)$

$p = 0.70$

$x = 5$

$r = 3$

$3^{rd}$ free throw occurred on fifth shot.




This means first two free throw would have occurred in first four shots.

$^4C_2 * p^2 * (1 - p)^2$

$P(X = 5) = {}^4C_2 * p^2 * (1 - p)^2 * p = 0.18522$

# Problem 1 – Part b Solution

Probability that Bob makes his first free throw on his fifth shot?

**X ~ geom(0.70)**

P = 0.70 , x = 5

$P(X = 5) = (1 - p)^4 * p$

$= (0.3)^4 * 0.7$

$= 0.00567$

**X ~ NB(1, 0.70)**

p = 0.70 ,x = 5, r = 1

$P(X = 5) = {}^4C_0 * p^1 * (1 - p)^4 = 1 * 0.7 * (0.3)^4 = 0.00567$

# Problem 2

A traffic light at certain intersection is
Green – 50% of the time
Yellow – 10 % of the time
Red – 40 % of the time

A car approaches this intersection once each day.

Let X represent the number of days that pass upto and including the first time the car encounters a red light.

Let Y represent the number of days upto and including the third day on which a red light is encountered.

Assume that each day represents an independent trail.

Find P(X <=3).

Find P(Y =7 ).

Find mean and Variance of X and Y.

# Problem 2 – Solution

a) $X \sim Geom(0.4)$

$P(X <= 3)$

b) $Y \sim NB(3, 0.4)$

$P(Y = 7)$

# Application – Negative Binomial Distribution

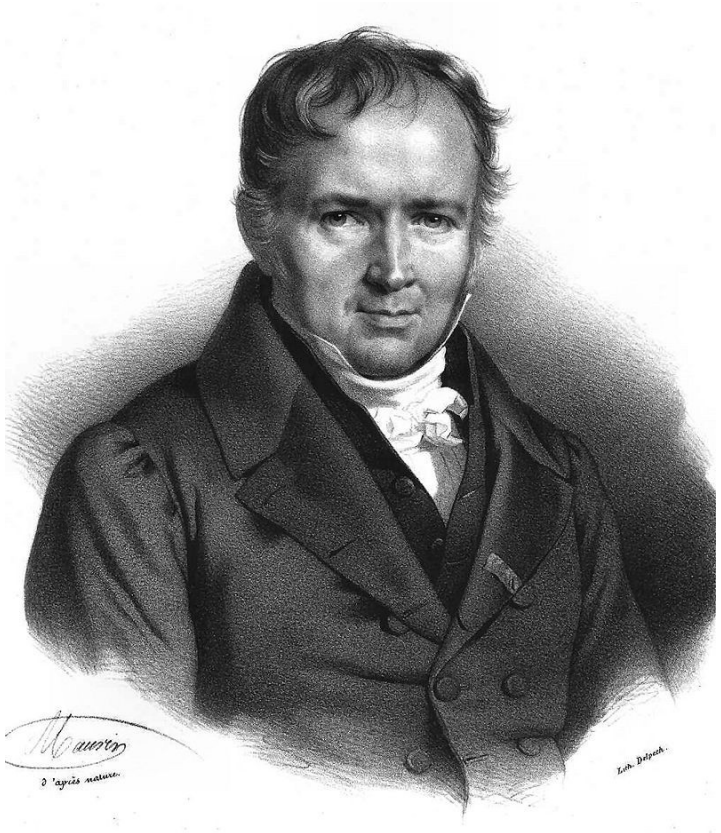
In industrial production processes items not meeting production specifications occasionally occur, but the detection of one such item may not in itself be a cause for alarm.

Instead we may wait for a certain number of events to occur. The random variable involved is called a negative binomial random variable.

# Poisson Distribution

Prof. Preet Kanwal

# Poisson Distribution



**Siméon Denis Poisson (1781–1840)**

**First derived Poisson distribution in 1837**

Poisson distribution is used to describe number of occurrences of a **rare** event that occur randomly during a specified interval.
The interval may be time, distance, area, or volume.

It describes the frequency of “successes” in a test where a “success” is a rare event.

Events with low frequency in a large population follow a Poisson distribution

# Poisson Distribution - Examples

1) The number of deaths by horse kicking in the Prussian army (First application).
2) The number of cyclones in a season.
3) Arrival of Telephone calls, Customers, Traffic, Web requests.
4) Estimating the number of mutations of DNA after exposure to radiation.
5) Rare diseases (like Leukemia(cancer of the blood cells), but not AIDS because it is infectious and so not independent).
6) The number of calls coming per minute into a hotels reservation Center.
7) The number of particles emitted by a radioactive source in a given time.
8) The number of births per hour during a given day.
9) The number of patients arriving in an emergency room between 11 and 12 pm.
10) The number of car accidents in a day.

In such situations we are often interested in whether the events occur randomly in time or space.

# The Law of Small Numbers by von Bortkiewicz (1898)

Showed how the Poisson distribution could be used to explain statistical regularities in the occurrence of rare events.

As examples von Bortkiewicz considered:
1) The number of suicides of children in different years
2) The number of suicides of women in different states and years
3) The number of accidental deaths in different years and
4) **The number of deaths from horse kicks in the Prussian Army in different years.**

The last example provided the most extensive data and has become a classical example of the Poisson distribution.

He extracted from official records the number of deaths from horse kicks in 14 army corps over the 20 year period 1875-1894, obtaining 280 observations in all.

He argued that the chance that a **particular soldier should be killed by a horse in a year was extremely small ,** but the no of men in corps was very large, so that the no of deaths in a cavalry corps in a year should follow the Poisson distribution.

# Student

**William Sealy Gosset (pen name Student)**
**(1876 – 1937)**
**English statistician**
**Famous for Student's t-distribution**

The first biological application of the Poisson distribution was given by 'Student' (1907) in his paper on the **error of counting yeast cells in a haemocytometer**(instrument for counting the no of cells in a cell suspension.), although he was unaware of the work of Poisson and von Bortkiewicz and derived the distribution afresh.

W.S. Gosset who was employed by Messrs Guinness in Dublin, Ireland.

Guinness is one of the most successful beer brands worldwide.

# Student's t- distibution

A normal distribution describes a full population, t-distributions describe samples drawn from a full population;

accordingly, the t-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution.

The t-distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the  statistical significance of
-> the difference between two sample means,
->  the construction of confidence intervals for the difference between two population means, and
-> in linear regression analysis.

# Coding Assignment

As lambda increases, the graphs begin to resemble a normally distributed curve:

Mean = 0.5, 2, 4, 10

# Properties of Poisson distribution

Events occur independently

The probability that an event occurs in a given interval of time is constant (does not change with time)

Events occur randomly and independently.

Then, X, the number of events in a fixed unit of time, has a Poisson distributin.

$$P(X = x) = f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, x = 0,1,2,...$$

Where $\lambda$ *is* mean of the distribution or the average rate at which events are occuring in a given interval of time or space.

# X ~ Poisson ( $\lambda$ )

Has a single parameter (mean of the distribution)
Theoretical range of the random variable is zero to infinity

Mean and Variance

$$\mu_X = \lambda \qquad \sigma_X^2 = \lambda$$

In Poisson distribution, mean = variance = λ. This is the **acid test** to be applied to any data which might appear to confirm to Poisson distribution

# Problem 1

If electricity power failures occur according to a Poisson distribution with an average of 3 failures every twenty weeks, calculate the probability that there will not be more than one failure during a particular week.

# Problem 1 - Solution

Average no of failures per 20 weeks = $\lambda = 3$

X denote the number of failures per week

X ~ Poisson($\lambda t$)

*X ~ Poisson(3/ 20) => X ~ Poisson(0.15)*

"Not more than one failure" means we need to include the probabilities for "0 failures" plus "1 failure".

$$P(x_0) + P(x_1) = \frac{e^{-0.15}0.15^0}{0!} + \frac{e^{-0.15}0.15^1}{1!} = 0.98981$$

# Problem 2

Vehicles pass through a junction on a busy road at an average rate of 300 per hour.

1)Find the probability that none passes in a given minute.

2)What is the expected number passing in two minutes?

3)Find the probability that this expected number actually pass through in a given two-minute period.

# Problem 2 - Solution

Average no of vehicles passing per hour = 300
Let X denotw Average no of vehicles passing per minute
X ~ Poisson( $\lambda t$) => X ~ Poisson( 300 / 60) => X ~ Poisson(5)

1) P(none pases in a given minute) = P(X = 0) = $P(x_0) = \frac{e^{-5}5^0}{0!} = 6.7379 \times 10^{-3}$

2) Expected no passing in two minutes = E(X) = 5 * 2 = 10.

3) Average no of vehicles that pass in 2 min period = $\lambda$ = 10

P( X =10 actually pass) = $P(x_{10}) = \frac{e^{-10}10^{10}}{10!} = 0.12511$

# Problem 3

A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell

1)Some policies

2)2 or more policies but less than 5 policies.

3)Assuming that there are 5 working days per week, what is the probability that in a given day he will sell one policy?

# Problem 3 - Solution

$\lambda = 3$

1) Some policies means P(X > 0)

P(X > 0) = 1 – P(X = 0)

$$P(x_0) = \frac{e^{-3}3^0}{0!} = 4.9787 \times 10^{-2}$$

2) 2 or more policies but less than 5 policies.

P( 2 <= X < 5) = P(X =2) + P(X =3) + P(X = 4)

$$= P(x_2) + P(x_3) + P(x_4)$$

$$= \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!}$$

$$= 0.61611$$

3) Assuming that there are 5 working days per week, what is the probability that in a given day he will sell one policy?

Average number of policies sold per day: $\frac{3}{5} = 0.6$

on a given day, $P(X) = \frac{e^{-0.6}(0.6)^1}{1!} = 0.32929$

# Problem 4

Albinism is a rare genetic disorder that affects one in 20,000 Europeans.

People with albinism produce little or none of the pigment melanin.

In a random sample of 1000 Europeans, what is the probability that exactly 2 have albinism?

# Problem 4 - Solution

**Binomial Distribution:**

$p = 1/20{,}000 = 5 \times 10^{-5}$

$X \sim Bin(1000, 5 \times 10^{-5})$

$P(X = 2) = {}^{1000}C_2 \ (5 \times 10^{-5})^2 \ (1 - 5 \times 10^{-5})^{1000 - 2}$

$= 0.001187965053$

# Problem 4 - Solution

**Poisson Distribution:**

$\lambda = n * p = 1000 * (5 \times 10^{-5}) = 0.05$

$X \sim \text{Poisson}(0.05)$

$$P(X = 2) = \frac{e^{-0.05}\ 0.05^2}{2!} = 0.001189036781$$

# Relationship between Binomial & Poisson Distribution

The binomoial distribution tends toward the Poisson distribution as

$$n \to \infty \qquad p \to 0$$

and

np stays constant

The Poisson distribution with $\lambda = np$ = closely approximates the binomial distribution if n is large and p is small.

# Problem 5

Plutonium- 239 is an isotope of plutonium that is used in nuclear weapons and reactors.

One nanogram of Plutonium-239 will have an average of 2.3 radioactive decays per second and the number of decays will follow a Poisson distribution.

What is the probability that in a 2 second period there are exactly 3 radioactive decays?

# Problem 5 - Solution

Let X represent no of decays in 2-second period.

Average no of decays in a 2 second period = $\lambda = 2 * 2.3 = 4.6$

$X \sim Poisson(4.6)$

$$P(X = 3) = \frac{e^{-4.6}\ 4.6^3}{3!} = 0.163$$

# What if $\lambda$ is unknown? (mean number of events that occur in one unit of time or space)

# Estimating $\lambda$

- Let $\lambda$ denote the mean number of events that occur in one unit of time or space.

Let X denote the number of events that are observed to occur in t units of time or space.

Then,

$$X \sim \text{Poisson}(\lambda t)$$

Where $\lambda$ is estimated with $\hat{\lambda} = X / t$

# Computing bias of $\lambda\hat{}$

**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.

$$\mu_{\hat{\lambda}} - \lambda$$

# Computing uncertainity of λ^

**Uncertainity** – is the standard deviation of sample proportion.

$$\sigma_{\lambda^{\wedge}} = \sigma_x / t$$

$$= \text{sqrt}\,(\lambda\, t)/t$$

$$= \text{sqrt}(\lambda/t)$$

As λ is unknown when computing uncertainity , we approximate it with λ^

# Problem 6

A microbiologist wants to estimate the concentration of a certain type of bacterium in a wastewater sample.

She puts a 0.5 mL sample of the waste-water on a microscope slide and counts 39 bacteria.

1) Estimate the concentrration of bacteria per mL, in this waste-water.

2) Find the uncertainity in the waste.

# Problem 6 - Solution

Let X represent the number of bacteria observed in 0.5 mL.

Let $\lambda$ represent the true concentration in bacteria per mL.

Then $X \sim \text{Poisson}(0.5\lambda)$.

The observed value of X is 39.

1) The estimated concentration is $\lambda = 39/0.5 = 78$.

2) The uncertainty is = sqrt (78/0.5) = 12.49

$\lambda = 78 \pm 12.49$

# Problem 7 - Description

A physicist wants to estimate the rate of emissions of alpha particles from a certain source.

He makes two counts.

First he measures the background rate by counting the number of particles in 100 seconds in absence of the source. He counts 36 background emissions.

Then, with the source present, he counts 324 emissions in 100 seconds. This represents the sum of source emissions plus background emissions.

# Problem 7 - Questions

1) a) Estimate the background rate in emissions per second.
   b) Find the uncertainity in the estimate.

2) a) Estimate the sum of source plus background rate in emissions per second.
   b) Find the uncertainity in the estimate.

3) Estimate the rate of source emissions in particles per second.

# Problem 7 - Solution - Part 1

Let λB represent the background rate,
let λS represent the source rate, and
let λ T = λ B + λ S represent the total rate, all in particles per second.

1) Let X be the number of background events counted in 100 seconds.
Then X ~ Poisson(100λ B ).
The observed value of X is 36.

a) The estimated background rate is λ^B = X/100 = 36/100 = 0.36.

b) The uncertainty is = sqrt (λ B /100) .
Replacing λ B with λ^B yields
Uncertainity = sqrt(0.36/100) = 0.06.
λ B = 0.36 ± 0.06

# Problem 7 - Solution - Part 2

Let λB represent the background rate,
let λS represent the source rate, and
let λ T = λ B + λ S represent the total rate, all in particles per second.

2)Let Y be the number of events, both background and source, counted in 100 seconds.
Then Y ~ Poisson(100λ T ).
The observed value of Y is 324.

a) The estimated total rate is λ^T = Y /100 = 324/100 = 3.24.

b) The uncertainty is = sqrt (λ T /100) .
Replacing λT with λ^T yields
Uncertainity = sqrt( 3.24/100) = 0.18.
λT = 3.24 ± 0.18

# Problem 7 – Solution - Part 3

Let λB represent the background rate,
let λS represent the source rate, and
let λ T = λ B + λ S represent the total rate, all in particles per second.

1) a) Estimate the rate of source emissions in particles per second.

λ S = λ T − λ B ,
so λ^S = λ^T − λ^B = 3.24 − 0.36 = 2.88.

# Hypergeometric Distribution

Prof. Preet Kanwal

# Example

An Urn contains 6 red balls and 14 yellow balls.

5 balls are randomly drawn without replacement.

What is the probability exactly 4 red balls are drawn?

# Can we use Binomial Distribution here?

Since the sampling is done without replacement,

**The trials are not independent!!**

(Probability of success on any individual trial depends upon what happened in the previous trial)

Hence the number of successes does not follow a binomail distribution.

# Example - Solution

P( exactly 4 red balls are drawn) =

$$= \frac{\text{No. of samples that result in 4 red and 1 yellow balls}}{\text{No. of possible samples of size 5}}$$

$$= \frac{{}^{6}C_{4} * {}^{14}C_{1}}{{}^{20}C_{5}} = 0.01354$$

# Properties of Hypergeometric distribution

Assume a finite population – N items.

R successes
N – R failures.

n items are sampled from this population without replacement.

X – denote the number of successes in the sample. Then,

$$X \sim H(N, R, n)$$

# PMF of X ~ H(N, R, n)

$$P(X = x) = \frac{{}^{R}C_{x} * {}^{N-R}C_{n-x}}{{}^{N}C_{n}}$$

# Problem 1

Of 50 buildings in an industrial park, 12 have electrical code violations.

If 10 buildings are selected at random for inspection, what is the probability that exactly 3 of the 10 have code violations?

# Problem 1 – Solution

$X \sim H(50, 12, 10)$

$$P(X = 3) = ({}^{12}C_3 * {}^{38}C_7) / {}^{50}C_{10}$$

$$= 0.2703$$

# Problem 2

20 air-conditioning units have been brought in for service.

12 of them have broken compressors, 8 have broken fans.

7 units are chosen at random to be worked on.

What is the probability that 3 of them have broken fans?

# Problem 2 – Solution

$X \sim H(20, 8, 7)$

$$P(X = 3) = ({}^{8}C_{3} * {}^{12}C_{4}) / {}^{20}C_{7}$$

$$= 0.3576$$

# Problem 3

There are 30 restaurants in a certain town.

Assume that 4 of them have health code violations.

A health inspector chooses 10 restaurants at random to visit.

1) What is the probability 2 of the restaurants with health code violations will be visited?
2) What is the probability none of the restaurants that are visited have health code violations?

# Problem 3– Solution

$X \sim H(30, 4, 10)$

1) $P(X = 2) = ({}^{4}C_{2} * {}^{26}C_{8}) / {}^{30}C_{10}$

$= 0.31199$

2) $P(X = 0) = ({}^{4}C_{0} * {}^{26}C_{10}) / {}^{30}C_{10}$

$= 0.1768$

# Problem 4

Suppose a large high school has 1100 female students and 900 male students.

A random sample of 10 students is drawn.

What is the probability exactly 7 of the selected students are female?

# Problem 4 – Soln using Hypergeometric Distribution

$$P(X = 7 \text{ Females}) = ({}^{1100}C_7 * {}^{900}C_3 ) / {}^{2000}C_{10}$$

$$= 0.166478$$

# Problem 4– Solution using Binomial Distribution

$$p = 1100 / 2000 = 0.55$$

$$P(X = 7 \text{ Females}) = {}^{10}C_7 (0.55)^7 (1 – 0.55)^3$$

$$= 0.166478$$

# Relationship between Binomial & Hypergeometric Distribution

Population Size = N
No of Successes in N = R
No of Failures in N = N – R

Sample Size = n
Success probability, p = R/N

Sample Size(n) is small compared to Population Size(N) [that is, no more than 5%], The difference between sampling with replacement and without replacement is too small(negligible). Then,

Bin(n, R/N) is a good approximation to H(N ,R ,n)

# If X ~ H(N, R, n), Mean and Variance of X :

Mean of Bin(n, p) = np

Mean of Bin(n, R/N) = **nR/N = Mean of H(N, R, n)**

---

Variance of Bin(n, p) = np [1 -p]

Variance of Bin(n, R/N) = n (R/N) [ 1 – (R/N)]

Variance of H(N, R, n) = **n (R/N) [ 1 – (R/N)] [(N – n) /(N – 1)]**

When n is small relative to N the quantity [(N – n) /(N – 1)] is close to 1.

# Problem: Multivariate Hypergeometric Distribution

Suppose a business employs 12 Democrats, 24 republicans, 8 independents.

If a random sample of 6 employees is drawn without replacement,

What is the probability there are:

3 Democrats, 2 Republicans and 1 independent in the sample.

# Solution : Multivariate Hypergeometric Distribution

Total Employees = 12 Democrats+ 24 republicans + 8 independents = 44

Sample Size = 6

P( 3 Democrats, 2 Republicans and 1 independent in the sample) =

$$= \frac{{}^{12}C_3 * {}^{24}C_2 * {}^{8}C_1}{{}^{44}C_6}$$

# Multinomial Distribution

Prof. Preet Kanwal

# Properties of Multinomial distribution

Generalization of Bernoulli trial.

Multinomial trial is a process that can result in any of r outcomes, where r>=2.

For each outcome i,

$X_i$ denote no of trials that result in that outcome.

$$(X_1, X_2, \ldots X_r) \sim MN(n, p_1, p_2 \ldots . p_r)$$

The entire collection $(X_1, X_2, \ldots X_r)$ is said to have multinomial distribution instead of single $X_i$

and,

For each i,

$$X_i \sim Bin(n, p_i)$$

# Note:

The number of ways of dividing a group of n objects into r groups $x_1$, $x_2$, ... $x_r$ objects, where $x_1 + x_2 + .. + x_r = n$ is

$$\frac{n!}{x_1 ! \, x_2 ! .. \, x_r !}$$

# PMF of $(X_1, X_2, \ldots X_r) \sim MN(n, p_1, p_2 \ldots p_r)$

Suppose $p_1, \ldots, p_r$ are nonnegative numbers such that $p_1 + \cdots + p_r = 1$. Random variables $X_1, \ldots, X_r$ have a *multinomial distribution with parameters* $n, p_1, \ldots, p_r$ if

$$P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

for nonnegative integers $x_1, \ldots, x_r$ such that $x_1 + \cdots + x_r = n$.

We write $(X_1, \ldots, X_r) \sim \text{Multi}(n, p_1, \ldots, p_r)$.

# Problem 1

An Urn contains 8 red balls, 3 yellow balls, and 9 white balls.

6 balls are randomly selected **with replacement.**

What is the probability 2 are red, 1 is yellow, and 3 are white?

# Problem 1– Solution

$$(X_1, X_2, X_3) \sim MN(6, 8/20, 3/20, 9/20)$$

$$P(X_1 = 2,\ X_2 = 1,\ X_3 = 3) = \frac{6!}{2!\ 1!\ 3!} * (8/20)^2 * (3/20)^1 * (9/20)^3$$

$$= 0.13122$$

# Problem 2

Of customers ordering a certain type of personal computer,
20% order an upgraded graphics card,
30% order extra memory,
15% order both upgraded graphics card and extra memory,
35% order neither.

15 orders are selected at random. Let $X_1$, $X_2$, $X_3$ ,$X_4$ denote respective numbers of orders in the four given categories.

1) Find $P(X_1 = 3, X_2 = 4 X_3 = 2, X_4 = 6)$.

2) Find $P(X_1 = 3)$

# Problem 2– Solution

(a) $(X_1, X_2, X_3, X_4) \sim MN(15, 0.20, 0.30, 0.15, 0.35)$

$$P(X_1 = 3, X_2 = 4, X_3 = 2, X_4 = 6) = \frac{15!}{3!4!2!6!} (0.20)^3 (0.30)^4 (0.15)^2 (0.35)^6$$

$$= 0.0169$$

(b) $X_1 \sim Bin(15, 0.2)$.

$$P(X_1 = 3) = \frac{15!}{3!(15-3)!} (0.2)^3 (1 - 0.2)^{15-3}$$

$$= 0.2501$$

# Thank you!