Review of useful probability & statistics concepts

Simulation Modeling and Analysis

Jayendran V

IEOR @ IIT Bombay

Random variable

- Random variable is a function that assigns a real number to each point in sample space

- Example: Rolling a pair of dice
  - Sample Space =
  - Let random variable X be sum of two dice

    Outcome Value assigned to X

We denote..
- Random variables by capital letters, X, Y, ..
- Values that random variables take by lowercase letters *x*, *y*, …

Cumulative Distribution function (CDF)

- Cumulative distribution function $F(x)$ of a random variable X is defined as probability of event $\{X \leq x\}$

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

- Properties of cdf
  1. $0 \leq F(x) \leq 1$ for all x
  2. $F(x)$ is non-decreasing
  3. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$

Discrete random variable

- X is a discrete random variable if it can take on at most countable number of values

- Probability mass function (pmf)

$$p(x_i) = P\{X = x_i\}, \forall\, i = 1, 2, \ldots, n$$

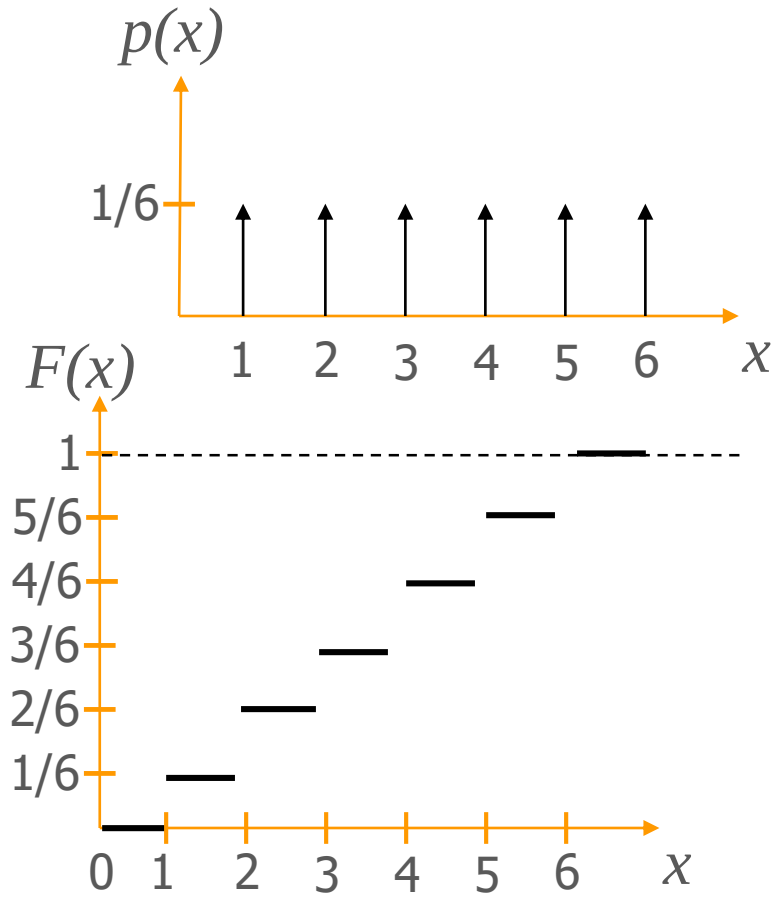
r.v. X takes on finite values x_1, x_2, .., x_n

- Cumulative distribution function

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

Example 2d

Number appearing on rolling a die



Continuous random variable

- X is continuous if there exists a nonnegative function $f(x)$ such that for any set of real numbers C: $P\{X \in C\} = \int_C f(x)dx$
- Probability density function (pdf), $f(x)$
- Interpretation of density function
  - Probability associated with each value x is zero
  - Density: probability that X will be very near x
- Cumulative Distribution function, CDF

$$F(X) = P\{X \leq x\} = \int_{-\infty}^{x} f(x)dx\,, \qquad -\infty \leq x \leq \infty$$

Mean or Expectation

- Expectation of *X* or Expected value of *X* or Mean of *X* *or* *E*[*X*] is weighted average of the possible values that X can take on
  - weights = probability
- When *X* is discrete random variable, with pmf p(x);

$$E[x] = \sum_i x_i p(x_i)$$

  - Example 2e: Suppose demand is a random variable that takes on the value 1, 2, 3, 4 with respective probabilities 1/8, 1/2, 1/8, ¼. Then, *E*[*X*] =
- When *X* is continuous random variable with pdf f(x);

$$E[x] = \int_{-\infty}^{\infty} x f(x) dx$$

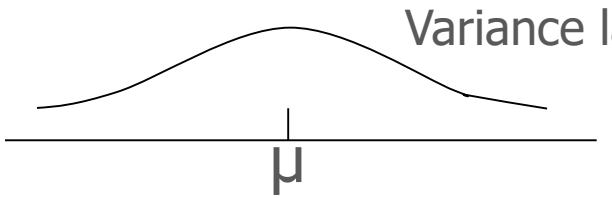
  - Example 2f: Suppose pdf of *X* is given by f(x) = $4x^3$ if 0<x<1, & 0 otherwise

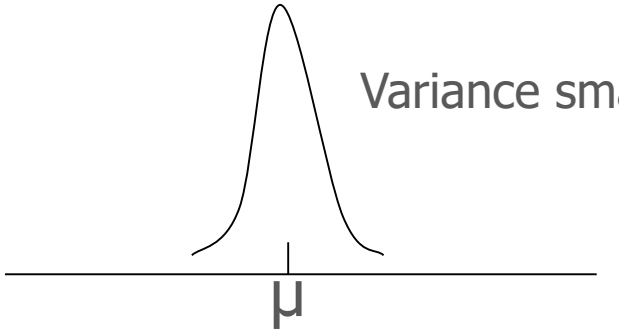
Variance, Var(X)

- Variance is the average value of the square of the difference between X and $E[X]$

$$Var(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$$

- Variance is the measure of dispersion





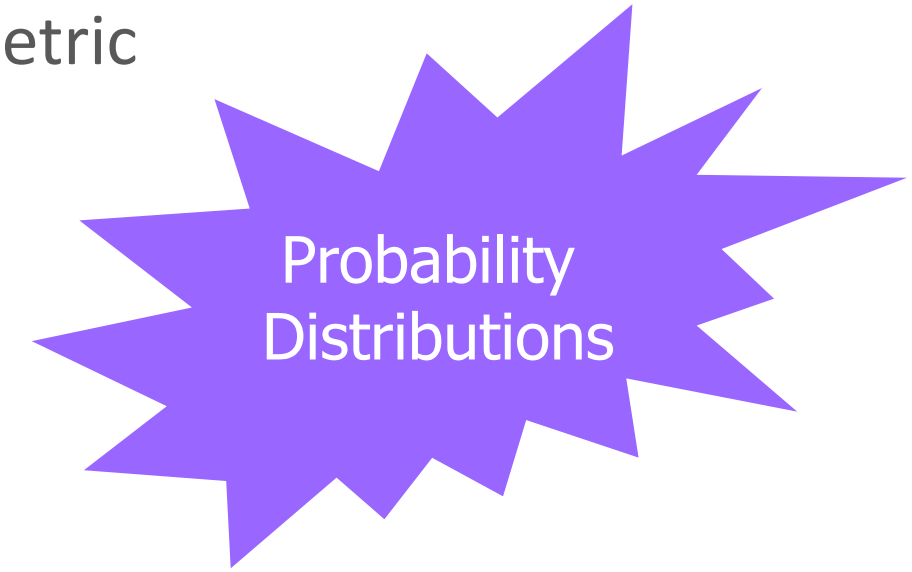
- Standard deviation
- Coefficient of variation, $\rho = \sigma/\mu$
  - Normalized measure of dispersion

Helpful to learn the properties of Mean and Variance

Special Random Variables

DISCRETE RANDOM VARIABLES

- Discrete
- Bernoulli
- Binomial
- Poisson
- Geometric
- …



CONTINUOUS RANDOM VARIABLES

Uniform

Exponential

Normal

Triangular

Beta

Gamma

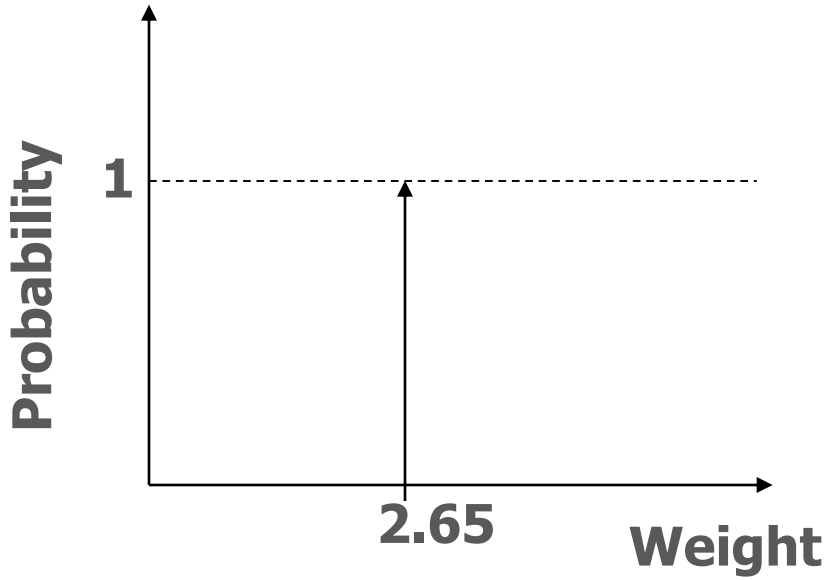
Erlang

LogNormal

…

Deterministic

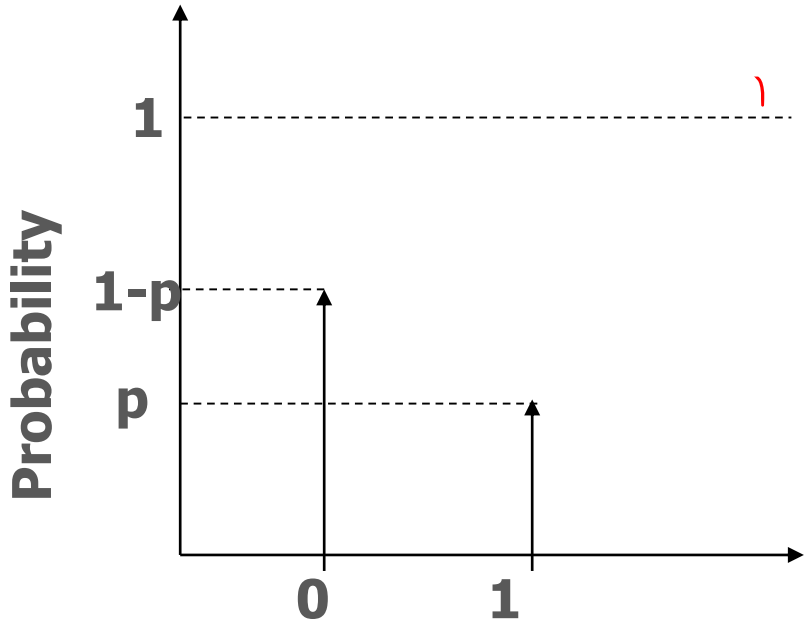
- No distribution at all!
- If deterministic value is used then it is assumed NO uncertainty exists
- Example: Weight = 2.65 Kgs



Bernoulli Random Variable

- X ~ Bernoulli(p) → parameter.
  - Discrete random variable *X* takes on value 1 with success probability p and value 0 with failure probability q=1-p

- PMF $p(1) = p$, $p(0) = 1-p = q$



- Mean = p
- Variance = $p(1-p)$

Binomial random variable

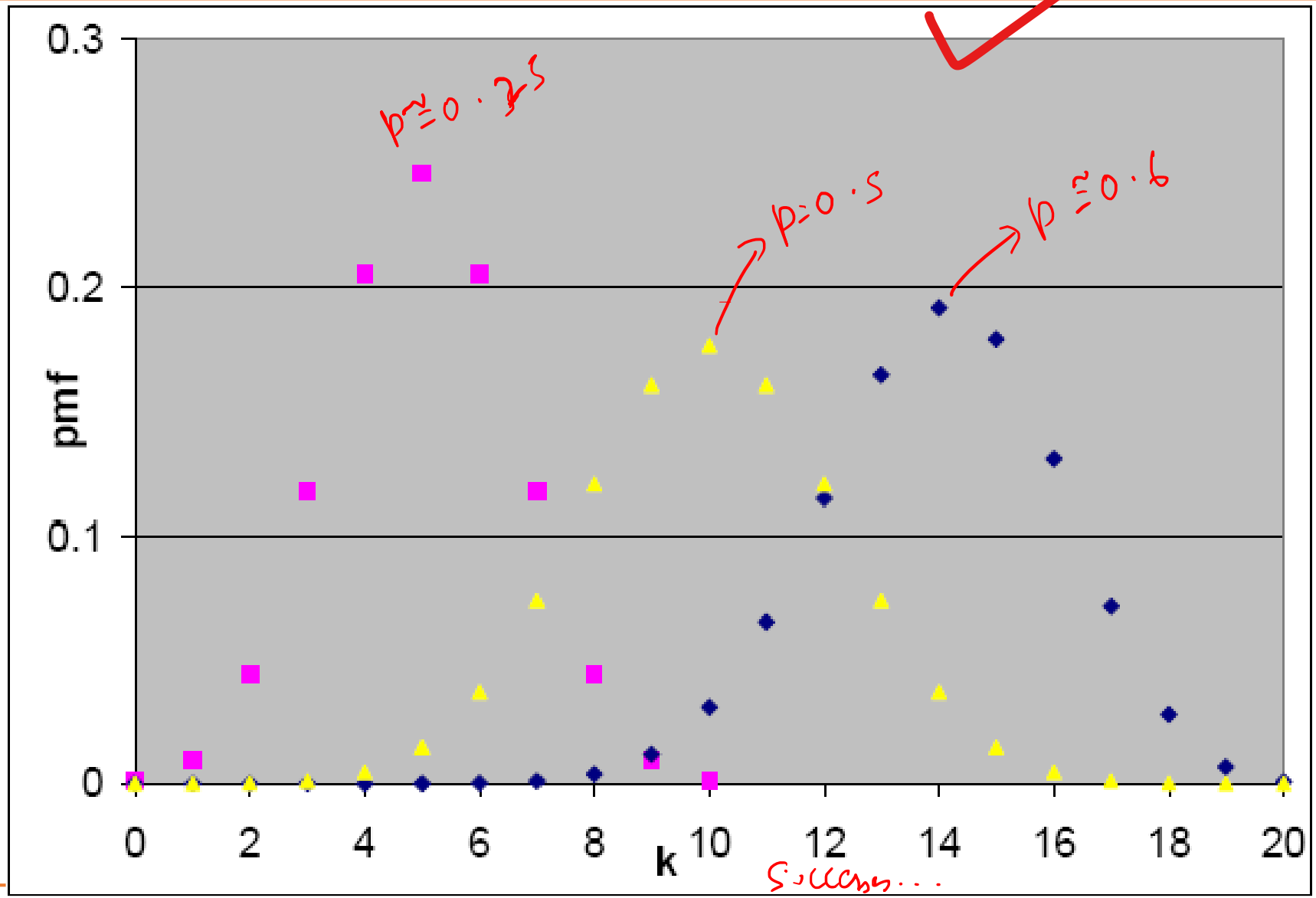
- X ~ Binomial(n, p)
- X is the number of successes in *n* trials, where each trial can result in success with probability *p* and failure with probability *q* = 1-p

- Probability mass function

prob. of k successes in n trials $\Big\}\ p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$

- Mean = np
- Variance = $np(1-p)$

Binomial distribution pmf

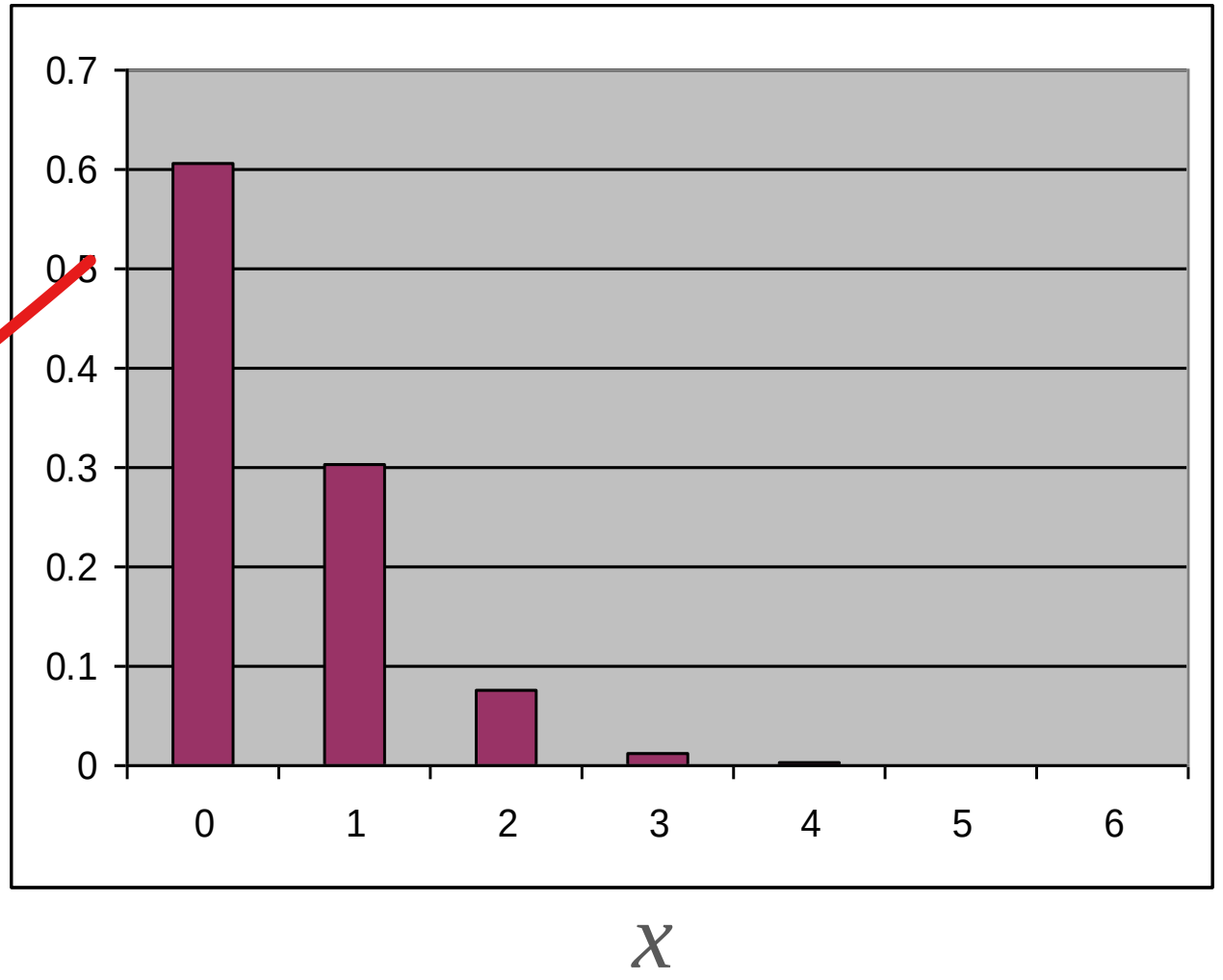


Poisson random variable

- X ~ Poisson(λ)
- Discrete random variable takes on one of the values 0, 1, 2, … is said to be Poisson random variable with parameter λ (rate)
  - Number of events on an interval of time
  - Number of items demanded from inventory

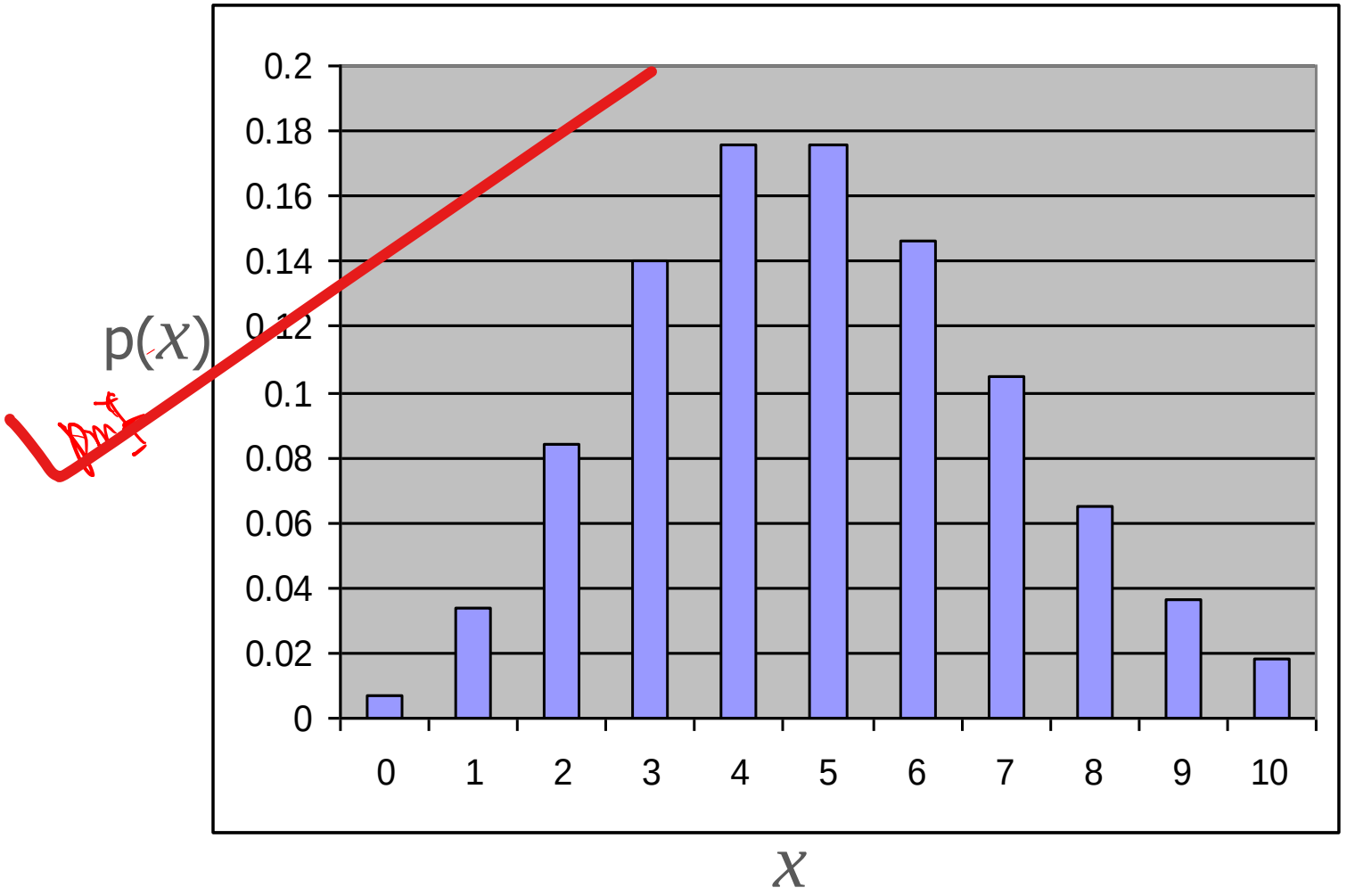
- PMF: $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$
- Mean = λ
- Variance = λ

Poisson random variable (2)



With
$\lambda=0.5$

Poisson random variable (3)



With
$\lambda=5$

Exponential random variable

- Continuous random variable $x \geq 0$
- PDF $f(x) = \lambda e^{-\lambda x}, x \geq 0$
- CDF $F(x) = 1 - e^{-\lambda x}, x \geq 0$
- Mean & Variance $\rightarrow 1/\lambda$, $\rightarrow 1/\lambda^2$
- Exponential random variables are the only continuous random variable with the "memoryless" property

Normal Distribution

Continuous domain. $-\infty < x < \infty$

- Many natural phenomena can be approximated by Normal distribution
  - Two parameter distribution
  - Bell shaped curve symmetric about the mean
- Probability density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $-\infty < x < \infty$

No closed form CDF

- Mean = μ
- Variance = σ^2

Standard Normal Distribution

- N(0, 1) distribution is called standard normal distribution
- Distribution function
- $X \sim N(\mu, \sigma^2)$ can be mapped to $Z \sim N(0,1)$ as follows:

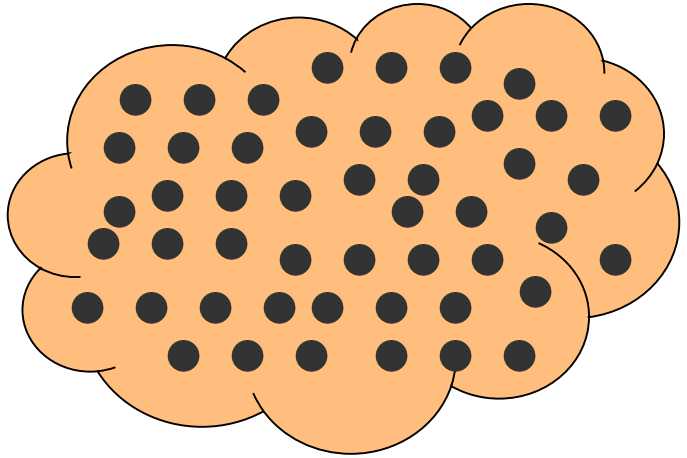
$$Z = (X - \mu) / \sigma$$

- Now can evaluate all probabilities concerning X in terms of Φ_Z (table)

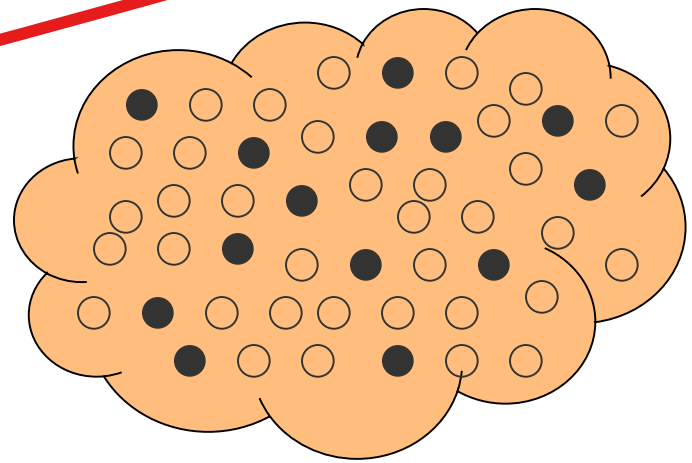
Statistics Overview

Population vs. Sample

- **Population** is the universal set of all things under study
  - Eg: all students @ IITB
  - Eg: all MTech students @ IITB



- **Sample** is the portion of the population selected for analysis
  - Eg: Randomly chosen from population



Parameter vs. Statistic

- Parameter describes a characteristic of full population
  - Example: 65% of all IITBombay students are males
- Statistic describes a characteristic of the sample
  - Example: 83% of students sampled are males

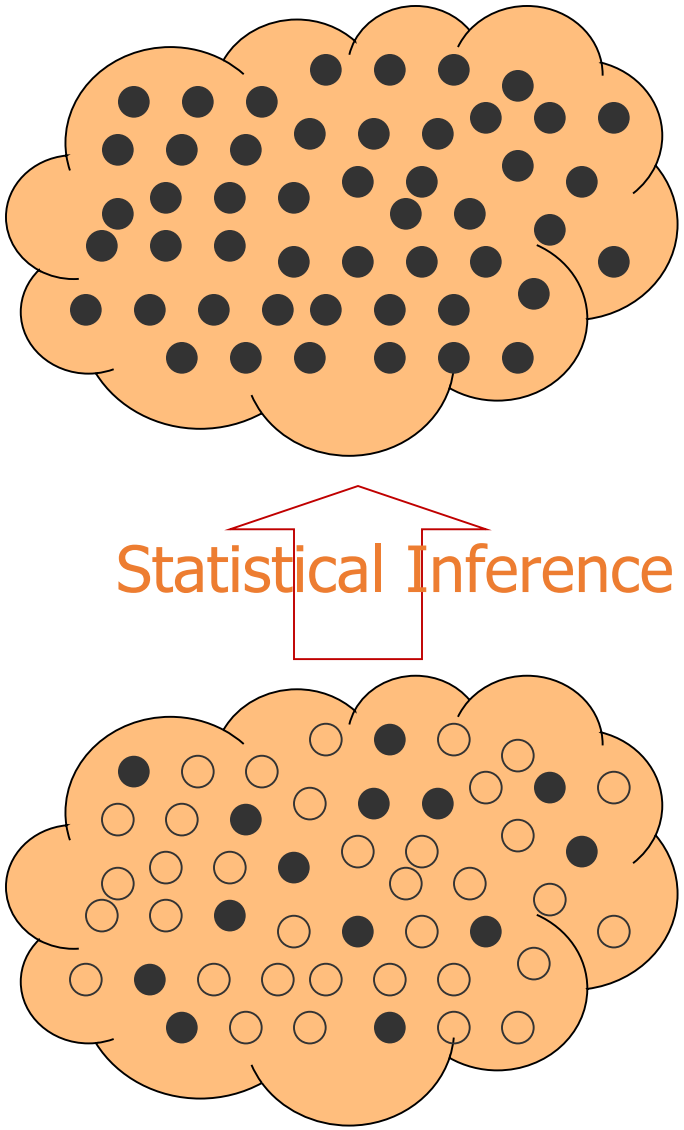
Generally

- Greek letters represent population parameter
- Roman letters represents sample statistic

Descriptive & Inferential Statistics

- Descriptive statistics are methods used to summarize & interpret collected data

- Inferential statistics are methods used to estimate unknown population characteristics based on sample results



Independent Identically Distributed (i.i.d)

- Independence
  - Two events are independent if occurrence of one does not influence the occurrence of the other
  - Two random variables X_1 and X_2 are independent if and only if for any numbers a_1 and a_2, the events $\{X_1 \leq a_1\}$ and $\{X_2 \leq a_2\}$ are independent
- Identically distributed
  - Random variables are said to be identically distributed if all have the same probability distribution
- *i.i.d random variables simplifies statistical inference*

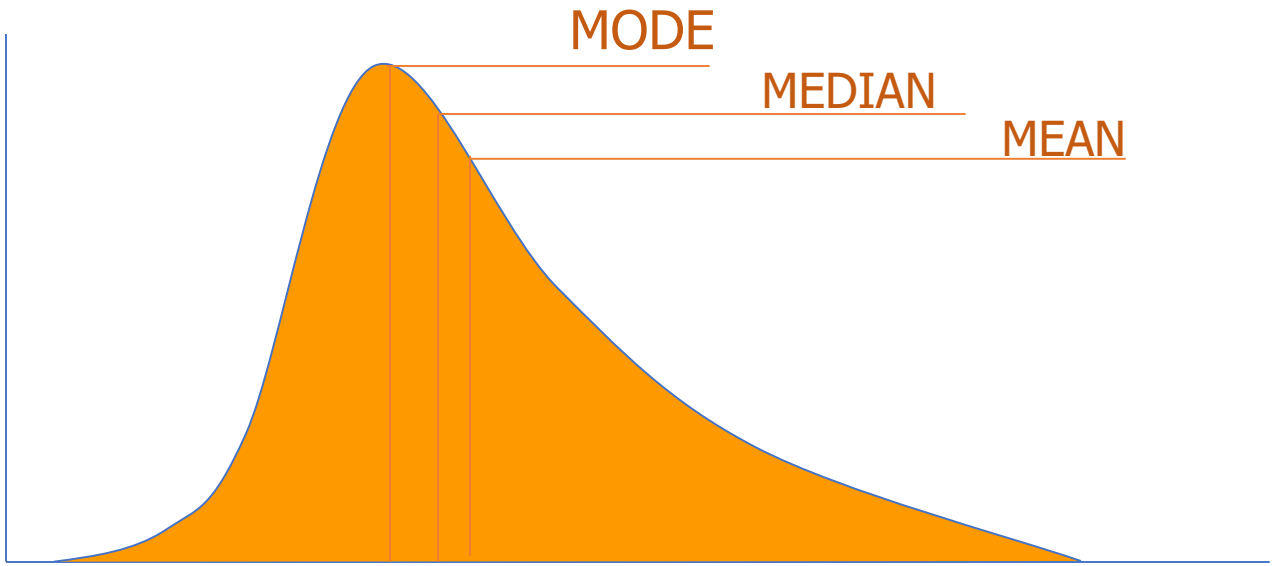
Dataset

- Suppose we have collected a set of data

- We want to describe the data collected
  - Measure of Central Tendency
  - Dispersion Statistics
  - Histogram

Measures of Central Tendency

- Central tendency is the measure of "middle" of a data set
- Mean: Arithmetic average of data set
- Median: Middle value of ordered data set
- Mode: Most frequent value(s) in data set



Sample Mean

- Sample Mean is the arithmetic average of the data set
- Suppose $X_1,\ldots, X_n$ are iid random variables drawn from a population with (unknown) mean μ and (unknown) variance σ^2
- We write the above statement as:

$$X_1, X_2, \ldots, X_n \sim \text{iid r. v.}, \text{with mean } \mu \text{ and variance } \sigma^2$$

Calculate Sample Mean as $\quad \bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

Sample Mean

- Sample mean is used to estimate the population mean (μ)
  - Sample mean is centered about population mean but the spread (variance) reduces as sample size increases

$$E[\bar{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}(n\mu) = \mu$$

and

$$\text{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

Sample Median

- Median is the middle observation of ordered data set
- Example 2g
  1. Median of 9, 3, 6, 4, 2 is *4*

     Order the set: 2, 3, 4, 6, 9; choose middle value
  2. Median of 2, 3, 4, 6, 7, 9 is *(4+6)/2=5*

- Median is resistant to outliers (unlike mean)
- Example 2h
  - Median of 2, 3, 4, 6, 7, 840 is *5*

Sample Mode

- Mode is value of data set that occurs most frequently
  - Example 2i: Mode of 2, 4, 4, 6, 7, 8 is *4*
- Mode need not be unique
  - Example 2j: Mode of 2, 2, 4, 6, 7, 7, 8, 11 is *2,7*
- Data can have no MODE
  - Example: 3, 5, 6, 8, 9

Dispersion Statistics

- Mean, Median & Mode insufficient to characterize data sets.
  - Example 2k:
    - Data Set 1: 98, 99, 100, 101, 102
    - Data Set 2: 20, 50, 100, 130, 200
  - Mean & Median are the same (no mode) --> 100
  - But, doesn't the data sets look different?
- Need to quantify the spread of the data set
  - ✓ Range
  - ✓ Variance
  - ✓ Standard Deviation

Sample Range

- Range is the difference between the largest and smallest observations made

- Example
  - Data Set 1: 98, 99, 100, 101, 102 -> Range: 4
  - Data Set 2: 20, 50, 100, 130, 200 -> Range: 180
  - Both dataset has the same mean

Sample Variance & Standard Deviation

- Suppose $X_1,\ldots, X_n$ are iid random variables drawn from a population with (unknown) mean μ and (unknown) variance σ^2
- Sample Variance is defined as: $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
- Sample variance is used to estimate the population variance (σ^2)
  - $E[S^2] = \sigma^2$
  - Note: (n-1) is used to compute S^2 to get an unbiased estimator of σ^2
- Standard deviation is square root of variance
  - Same units as mean

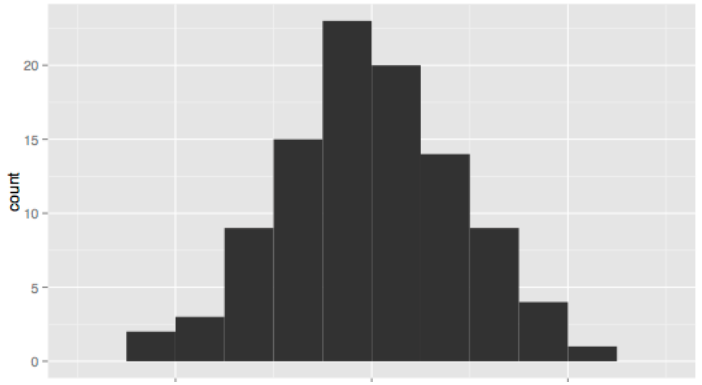
Histogram

- Histogram is an approximate representation of distribution of the data

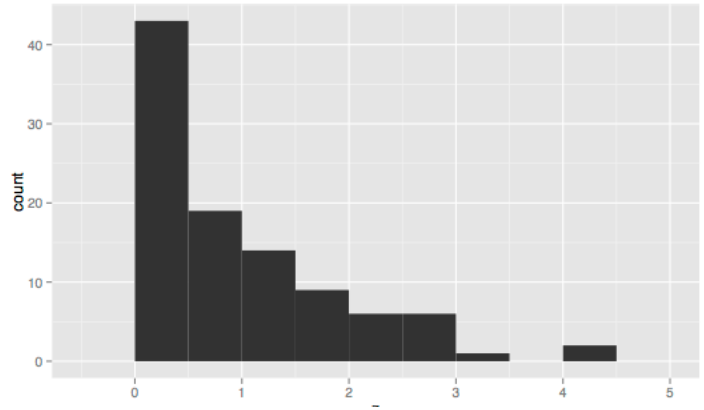
- To build histogram
- Create bins or buckets or interval for entire range of values
  - Bins to be non-overlapping, but consecutive
- Count how many values fall into each bin or interval
- Plot x-axis (bins) vs. y-axis (count or frequency)

https://commons.wikimedia.org/w/index.php?curid=3619268

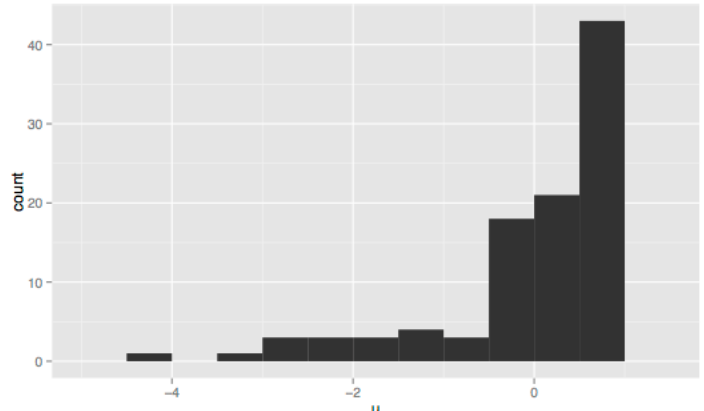
Histograms comes in various shapes



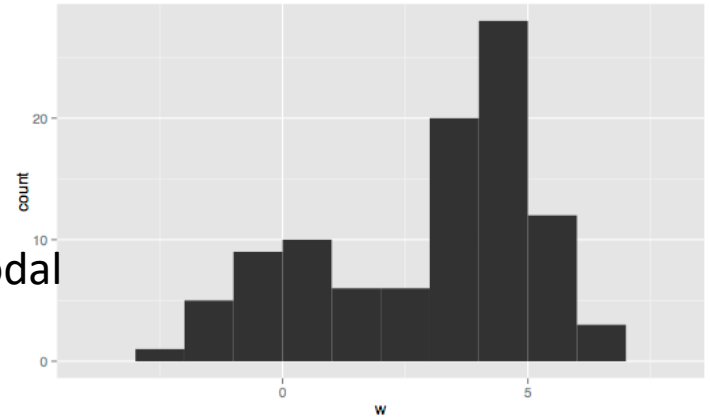
- Symmetric
- (unimodal)



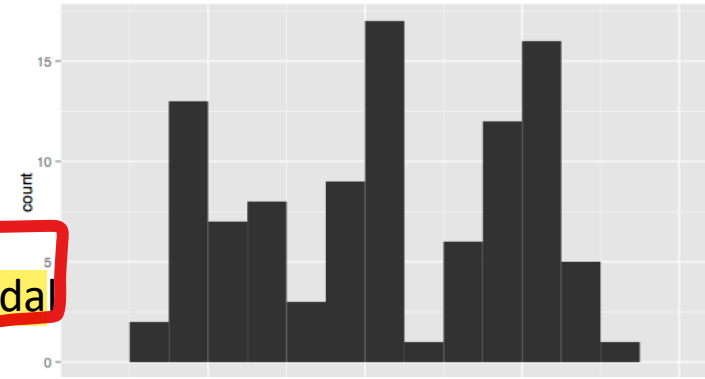
- Skewed right



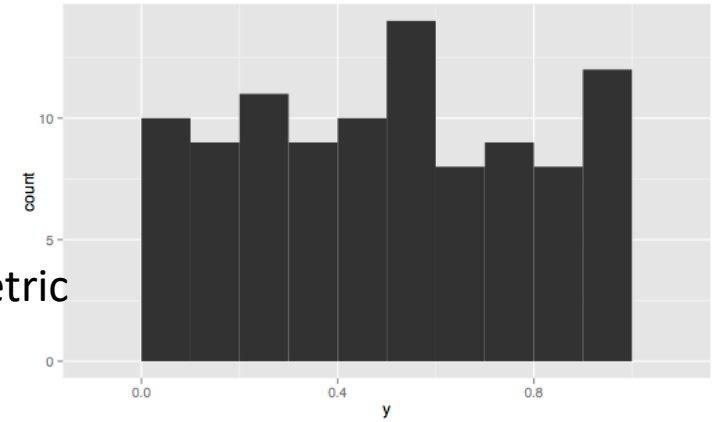
- Skewed left



Bimodal



Multi-modal



Symmetric

Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be a sequence of IID random variables each with mean μ and variance σ^2. Then for large *n*, the distribution of $(X_1 + X_2 + \ldots + X_n)$ is approximately Normal with mean $n\mu$ and variance $n\sigma^2$
  - That is, the sum of the random variables:

$$(X_1 + X_2 \ldots + X_n) \sim \text{Normal}(n\mu, n\sigma^2)$$

  - Or can be interpreted as:

$$\left(\frac{X_1 + X_2 \ldots + X_n}{n}\right) = \bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$$

  - Or,

$$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \sim \text{Normal}(0, 1) \sim Standard\ Normal$$

How large a sample is required?

- How large a sample is required for Normal approximation to be valid?
  - When underlying population is Normal…
    - Any sample size is ok
  - When underlying population is not Normal…
    - At least a sample size of 30

Sampling from Normal Distribution

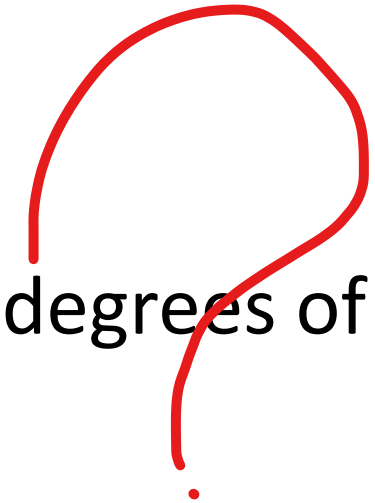
- Suppose $X_1, X_2, \ldots, X_n$ be IID sample from Normal population having mean μ and variance σ^2

  then, $\bar{X}$ and S^2 are independent random variables,

  distribution of sample mean, $\bar{X}$ is $\boldsymbol{N}(\mu, \sigma^2/n)$,

  and

  distribution of $(n-1)S^2/\sigma^2$ is **chi-square** with $n-1$ degrees of freedom



Student t-distribution

- Suppose $X_1, X_2, \ldots, X_n$ be IID sample from Normal population with mean μ (σ^2 unknown)

  then, $\left(\frac{\bar{X} - \mu}{S/\sqrt{n}}\right) \sim t_{n-1}$

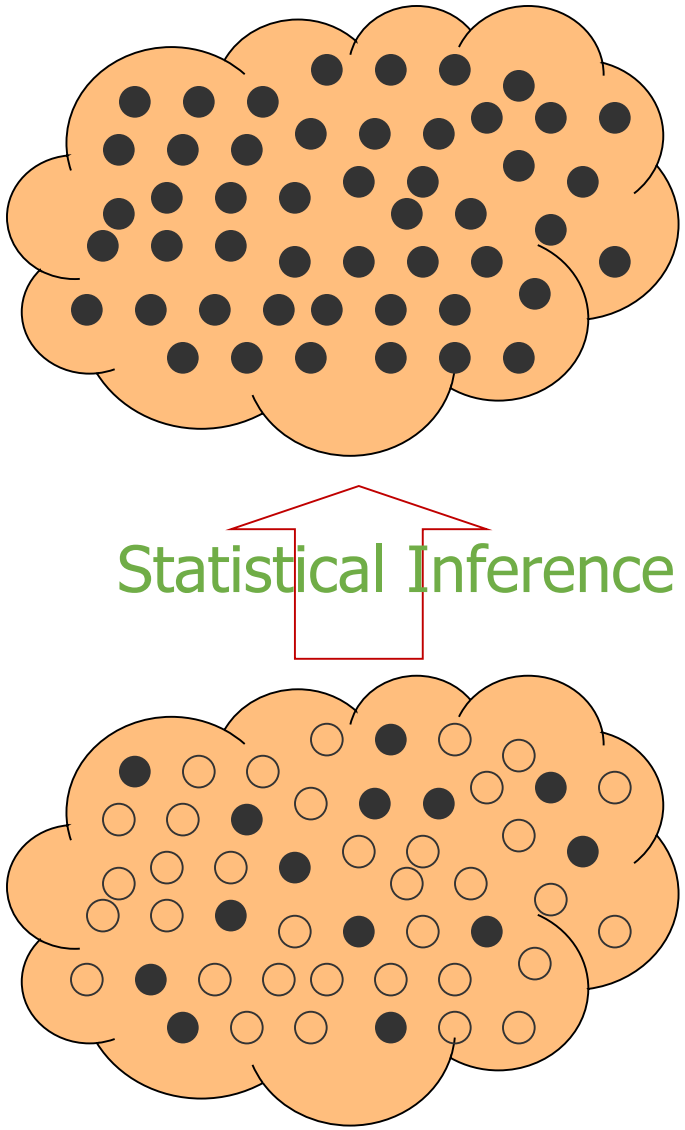
where, t_{n-1} : t-distribution with n-1 degrees of freedom

Student-t distribution, developed by William Gosset while at Guinness!
Gosset was Head Experimental Brewer, experimenting on barley for best yielding varieties.
Developed various statistical methods, design of experiments, significance testing, etc

Statistical Inference

- How to use observed data (sample) to make inferences about the unknown population parameters?

- Let $X_1, X_2, \ldots, X_n$ be random sample from population distribution F_θ where θ is vector of unknown parameters



Estimators

- Any statistic (based on sample) used to estimate the value of unknown parameter θ (of population) is called an *estimator of* θ

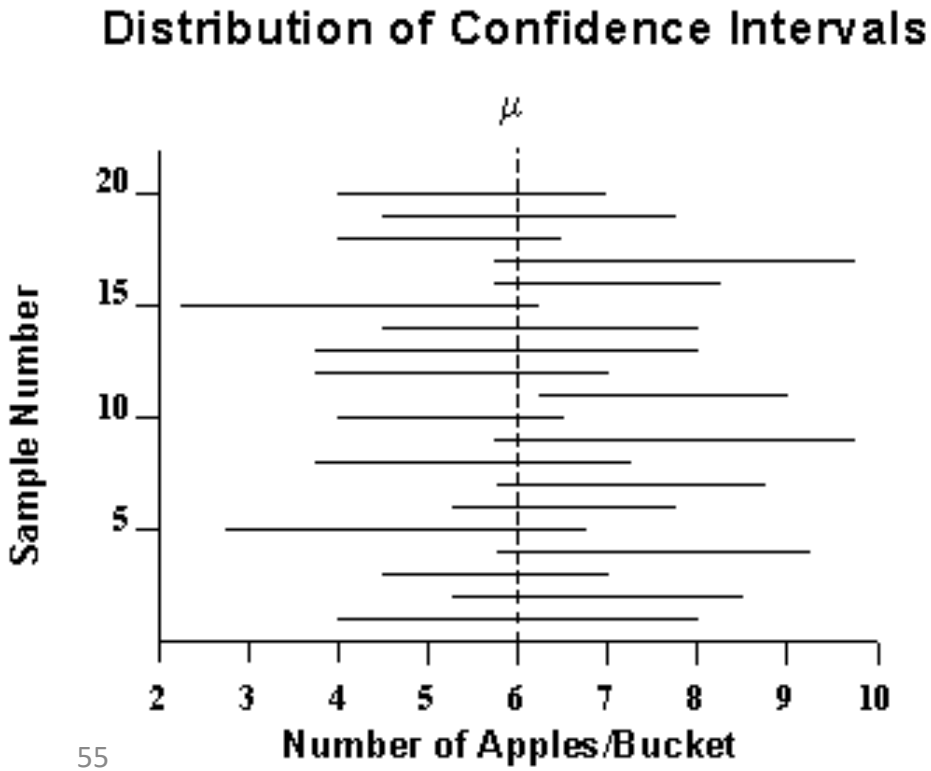
- Maximum Likelihood Estimation (MLE)
  - Popular method of parameter estimation
  - *"If I assume that the underlying population distribution is __so-and-so__ then my best guess for the parameters will be __such-and-such__"*

- Point Estimates: $\bar{X}, S^2$
  - Gives **no** information on how accurate the estimate is from true value of unknown parameter
- Interval Estimates: Confidence & Prediction intervals

Confidence Interval Estimates

- Interval within which we have certain level of confidence that true mean (μ) falls.
  - CI bounds the error between sample mean and true population mean
  - 100(1 – α)% CI

If we construct a number of 100(1-α)% CIs, each based on n observations, the probability that the CI will contain true mean μ is (1-α).



Confidence Interval Estimates

- $X_1, \ldots, X_n$ be samples from Normal population
  1. 100(1 – α)% CI when σ^2 known

$$\mu \quad \epsilon \quad \bar{X} \pm z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}$$

  2. 100(1 – α)% CI when σ^2 unknown

$$\mu \quad \epsilon \quad \bar{X} \pm t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$$

Statistical Hypothesis Testing

- Hypotheses: Are claims about population characteristics
  - E.g.: It can be about the nature of distribution, or parameters of population distribution
- Hypothesis testing
  - “Can the difference between observed result and expected result be attributed to chance?”
  - Testing by contradiction
  - Remember: We take a decision about the population characteristics based on samples

Statistical Hypothesis Testing (2)

- Null hypothesis, H_0
  - Observation or Samples are purely by chance
- Alternate Hypothesis, H_A
  - Typically, Not-H_0
  - Observation or Samples are result of some real effect

- Define a test statistic
  - If test statistic lies in some predefined region *C* then 'reject' null hypothesis, else 'cannot reject' H_0.
  - Many types of test statistics

Statistical Hypothesis Testing (3)

- Types of errors
  - Type I error: Rejecting H_0 when it is correct
  - Type II error: Accepting H_0 when it is false

- Classical way:
  - Specify a value α (level of significance of test)
  - Now, we require that the probability of type I error < α

Statistical Hypothesis Testing (4)

- *p*-value
  - It is the probability of getting a result at least as extreme as the observed result.
  - Lower the p-value, more ‘significant’ the result
  - Statistical significant : it is unlikely to have occurred by chance
  - H_0 is accepted if $\alpha <$ p-value, else rejected

Summary

- Many concepts in Probability and Statistics that are useful for modelling and analysing simulation models discussed