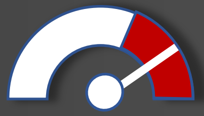
Distribution Refresher

Become Future Fit

Objective

- Ability to identify different types of statistical distributions
- Apply the concept of distributions to practical situations

Level of Difficulty



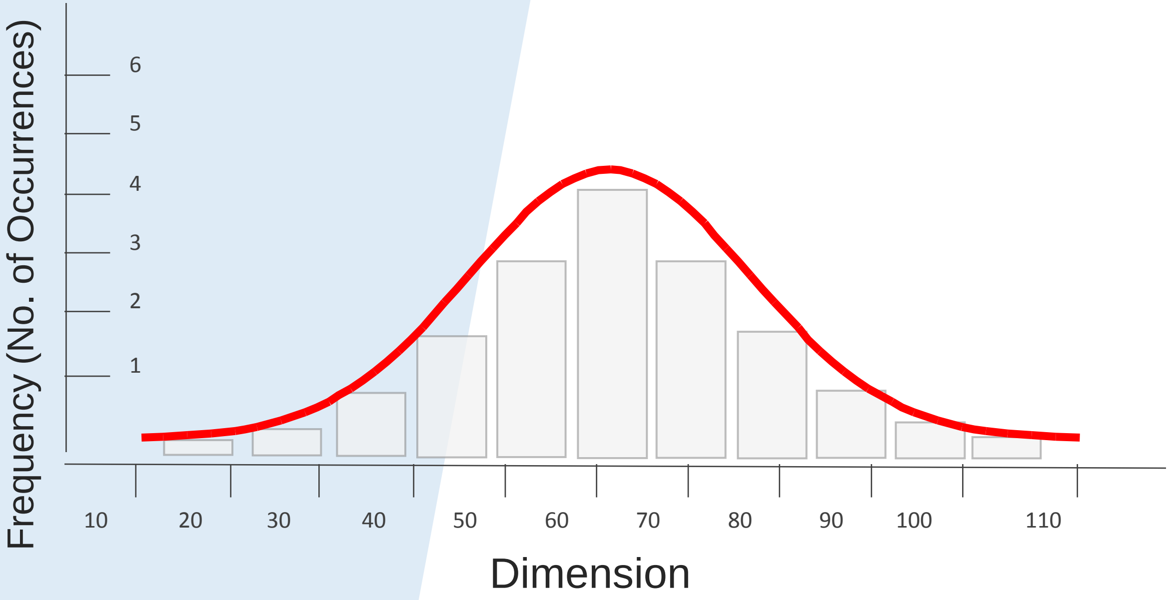
High

Terms to Refresh

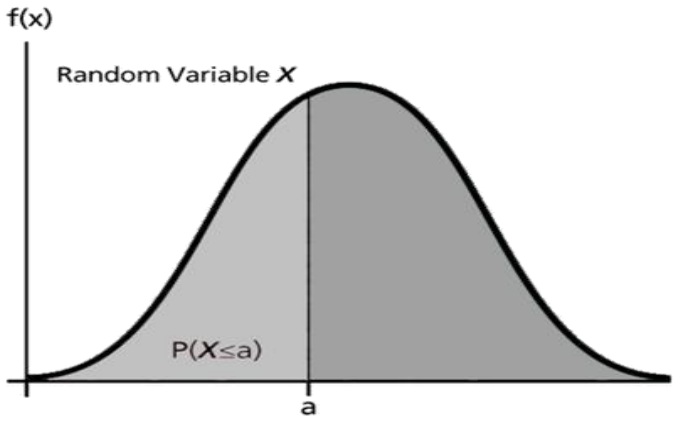
1. Random Variable
2. $P(X)$ represents the probability of X.
3. $P(X = x)$ refers to the probability that the random variable X is equal to a particular value, denoted by x.
4. A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Histogram

Canopus Business Management Group

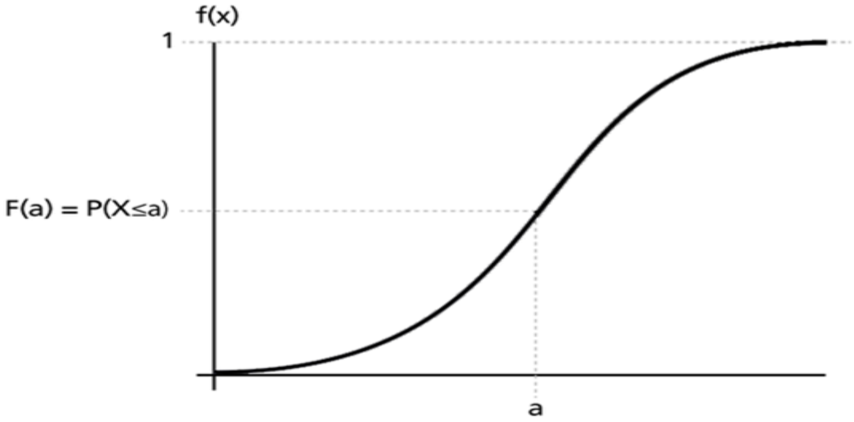


Probability Distribution Function

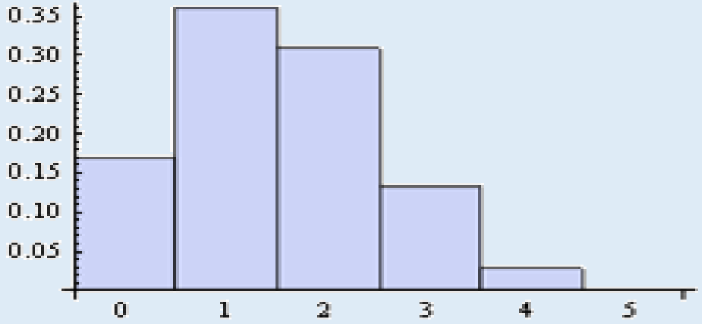


PDF represents the relative frequency of the random variable as a function of time.

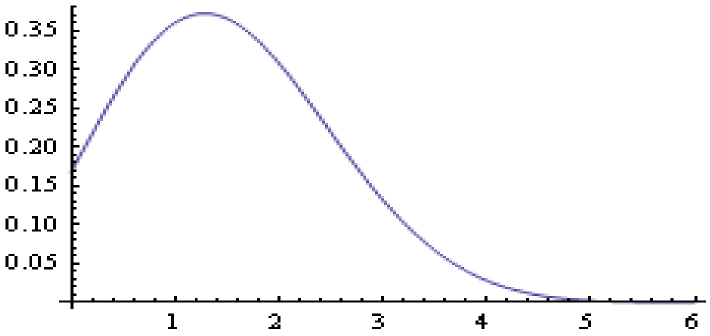
Cumulative Distribution Function



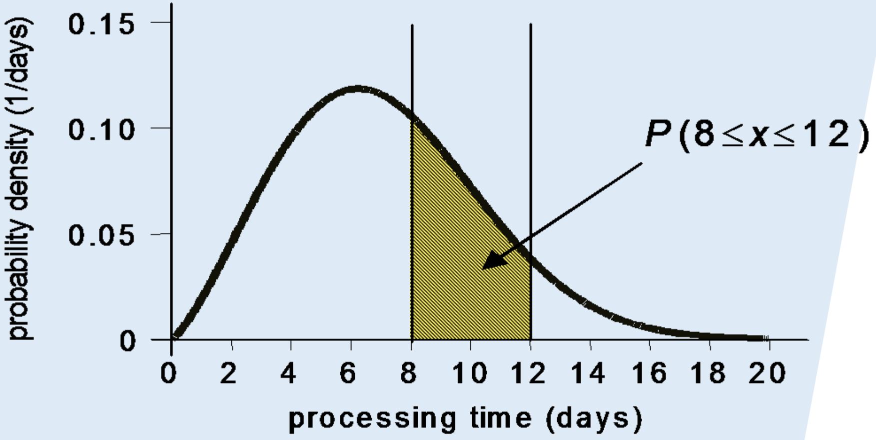
CDF represents the cumulative values of the PDF



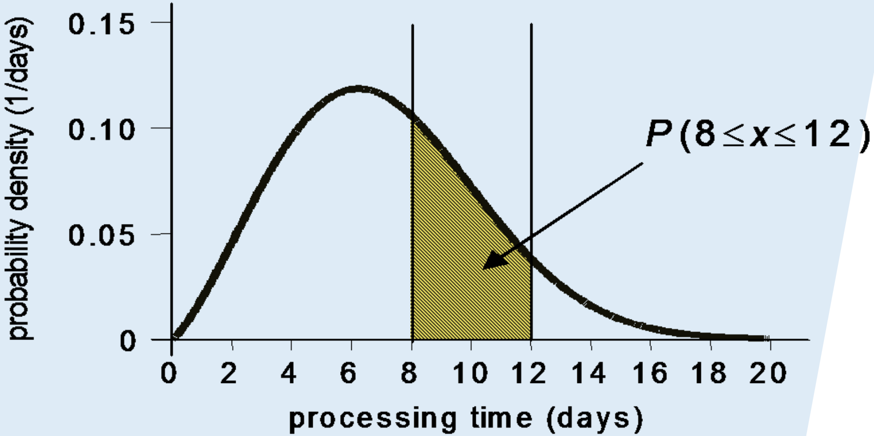
Probability Mass Function



Probability Density Function



The probability that the sample will arrive between 8th & 12th of this month.



Probability that the sample will arrive before12th of this month.

Probability that the sample will arrive after 15th of this month.

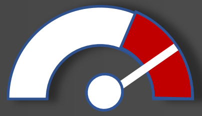
Standard Normal Variable

Become Future Fit

You will learn

Understand what is Standard Normal Variable

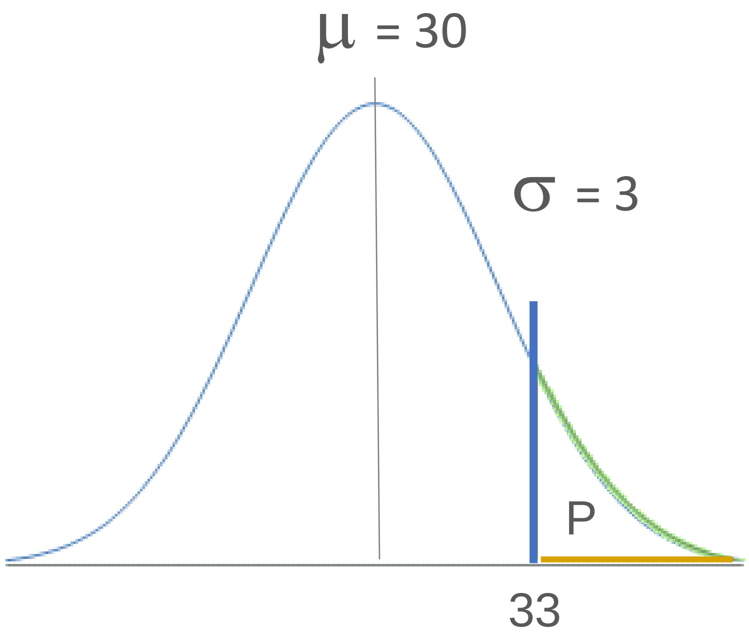
Level of Difficulty



High

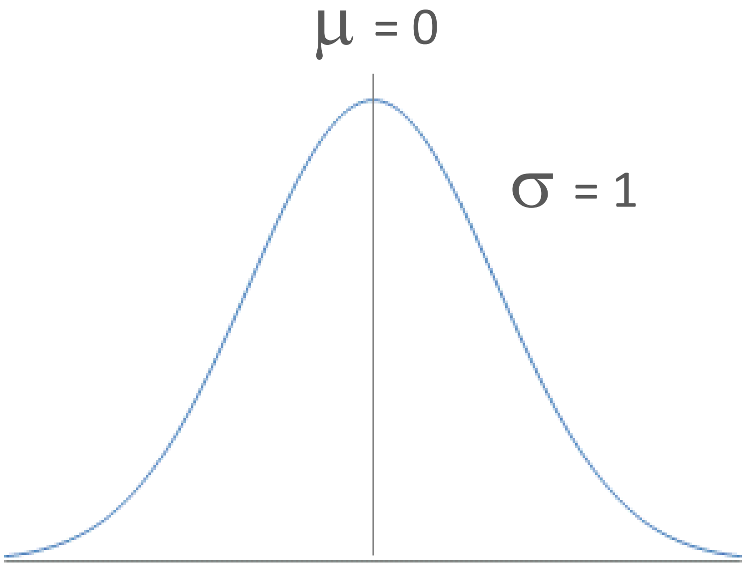
Our process data is Normal with Mean (30) & SD (3).

We would like to find out the probability of having a value greater than 33.

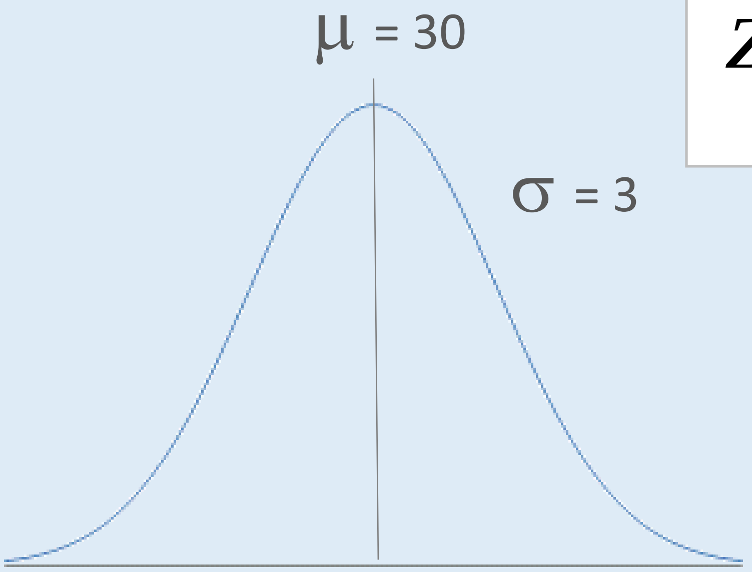


Standard Normal Distribution

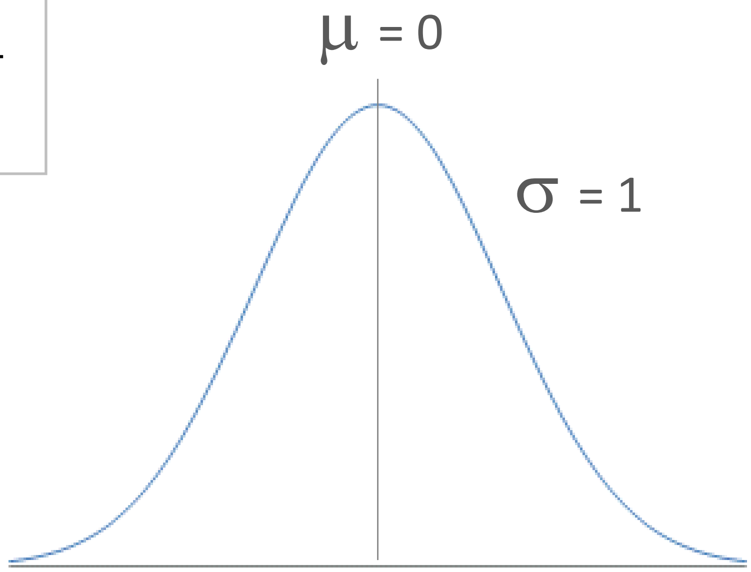
Normal distribution defined by Gauss



Normalization with Standard Normal Variable

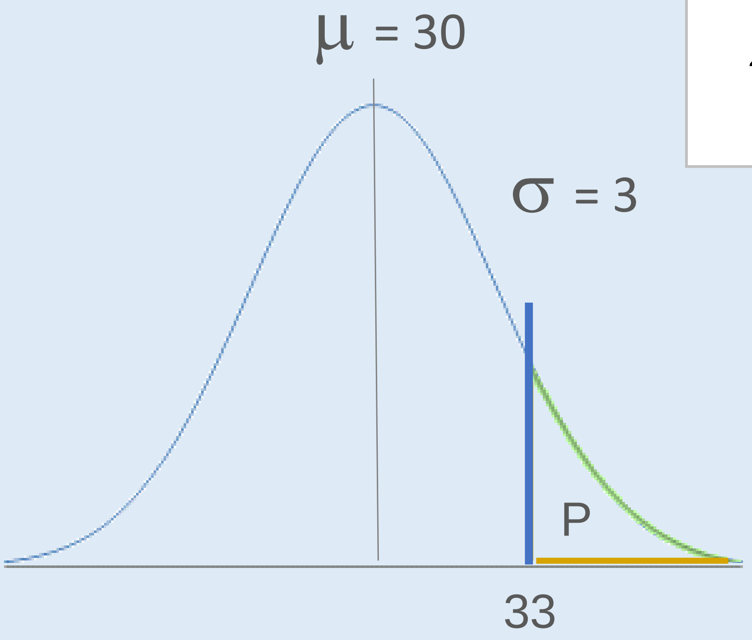


$$Z = \frac{x - \mu}{\sigma}$$

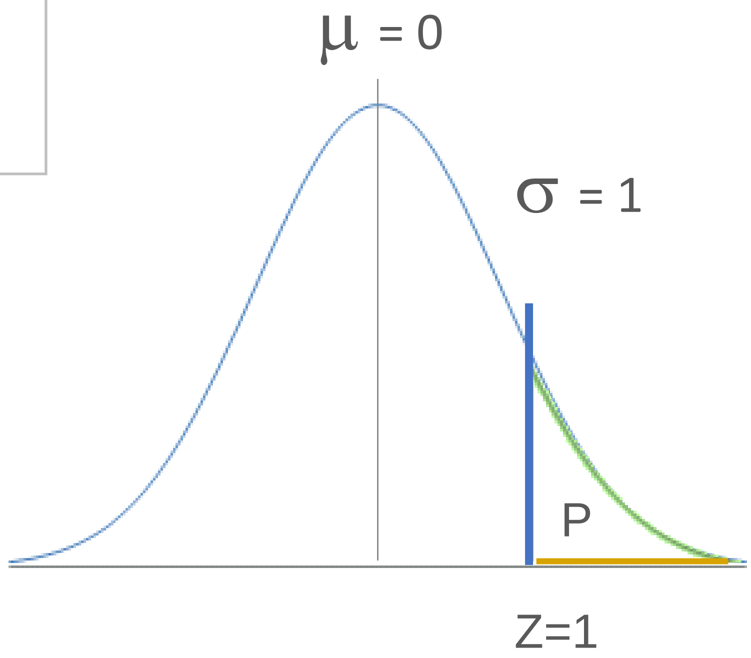


Standard normal variable is a normally distributed random **variable** with Mean = 0 and SD = 1.

Normalization with Standard Normal Variable



$$Z = \frac{33 - 30}{3}$$



Activity in minitab for finding the probability.

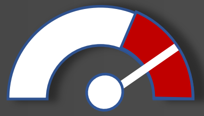
Application of Probabilities

Become Future Fit

You will learn

Application of the concept of probability to business scenarios

Level of Difficulty



High

Finding Probability of Defectives

A critical parameter was produced in a particular equipment. Data collected from the samples had Mean = 20.2 ; SD=0.01. Assume Normal Data, then:

1. If 20.222 is **USL**, what % of units produced are likely to be defective

Cumulative Distribution Function

Normal with mean = 20.2 and standard deviation = 0.01

x	P(X ≤ x)
20.222	0.986097

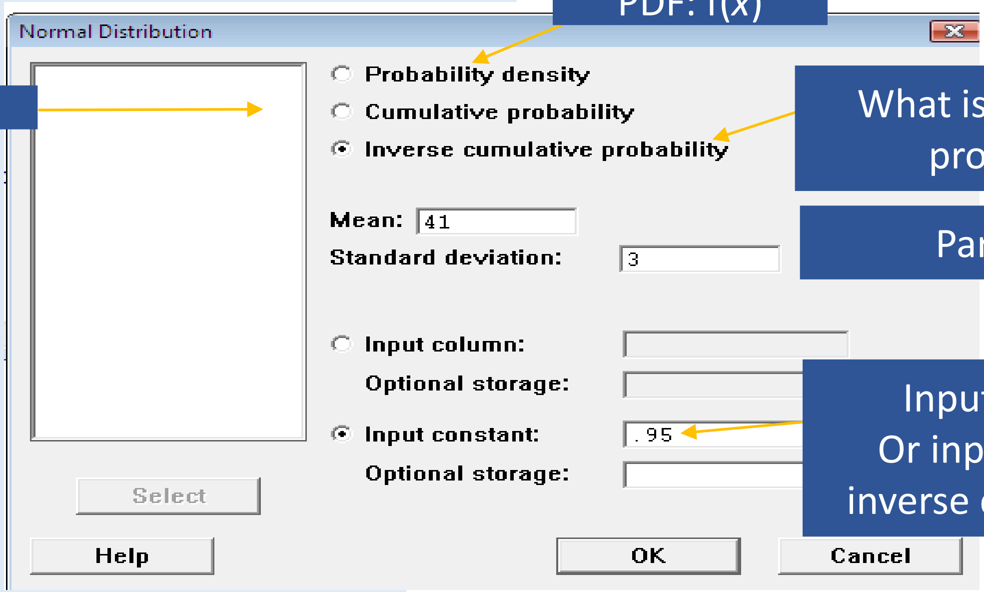
ANS:(1-0.986097) = 1.3%

Canopus Business Management Group

Calc>Probability Distributions> **<Distribution Type : Normal>**

PDF: f(x)

CDF: F(x)



What is x for a given probability?

Parameters

Input x for PDF and CDF.
Or input probability for the inverse cumulative probability.

Estimation for Trial Batch

A critical parameter was produced in a particular equipment. Data collected from the samples had Mean = 20.2 ; SD=0.01. Assume Normal Data, then:

2. For a trial batch, you want only units with dimensions between 20.18 & 20.22. What % of units are you likely to get?

Canopus Business Management Group

Cumulative Distribution Function

Normal with mean = 20.2 and standard deviation = 0.01

x	P(X ≤ x)
20.18	0.0227501

Cumulative Distribution Function

Normal with mean = 20.2 and standard deviation = 0.01

x	P(X ≤ x)
20.22	0.977250

ANS:(97.725-2.27501) = 95.45 %

Planning for Future

A critical parameter was produced in a particular equipment. Data collected from the samples had Mean = 20.2 ; SD=0.01. Assume Normal Data, then:

3. If wish to have 99.9% acceptable rate from this process, what should be **USL**?

Canopus Business Management Group

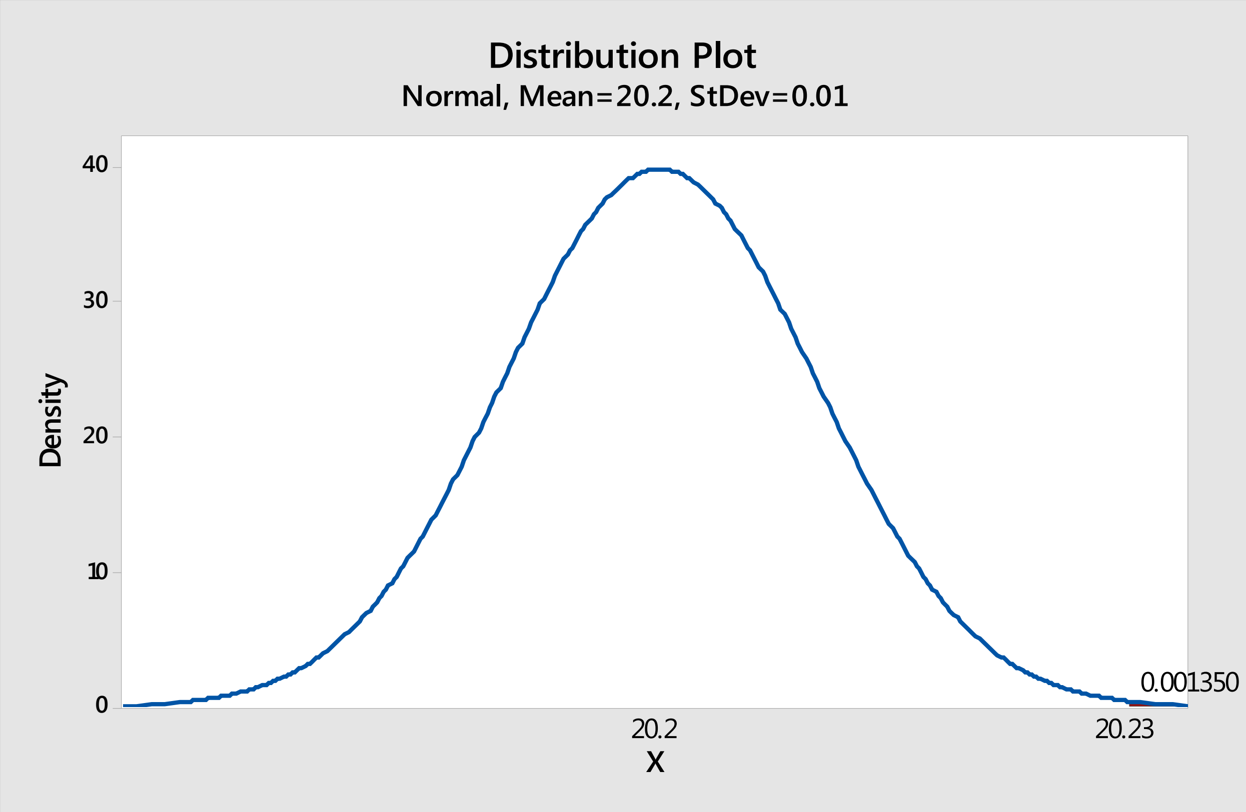
Inverse Cumulative Distribution Function

Normal with mean = 20.2 and standard deviation = 0.01

$P(X \le x)$	x
0.999	20.2309

ANS: 20.2309

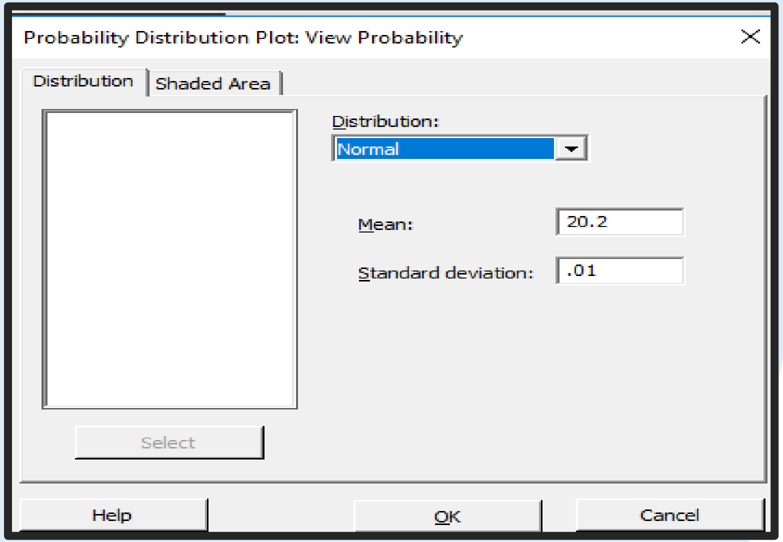
Solution

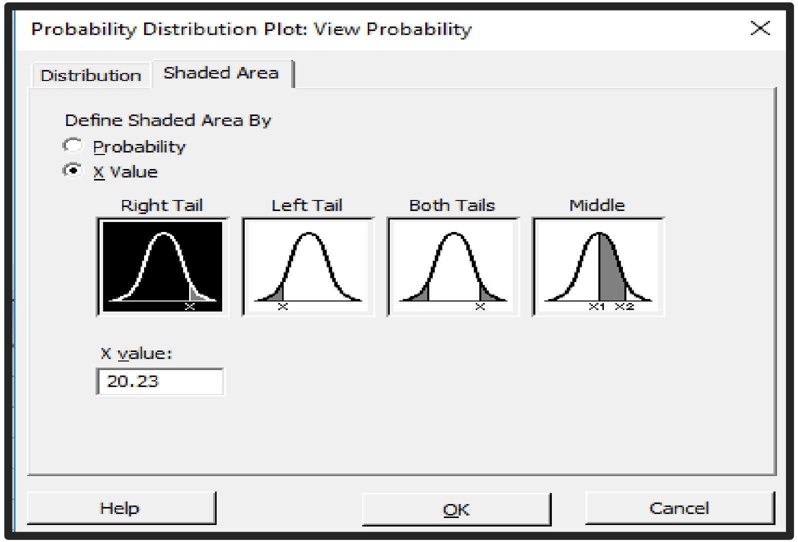


Distribution Plot
Normal, Mean=20.2, StDev=0.01
Density
40
30
20
10
0
0.001350
20.2
20.23
X

Canopus Business Management Group

Graph > Probability Distribution Plot > View Probability





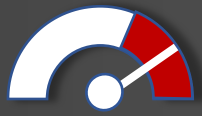
Types of Distribution

Become Future Fit

You will learn

Learn about different types of distributions

Level of Difficulty



High

Continuous Distributions

- Normal
- Exponential
- Weibull
- Chi Squared
- T
- F
- ❖ Lognormal
- ❖ Gamma
- ❖ Beta

To be covered in later lesson

Not covered in BB class

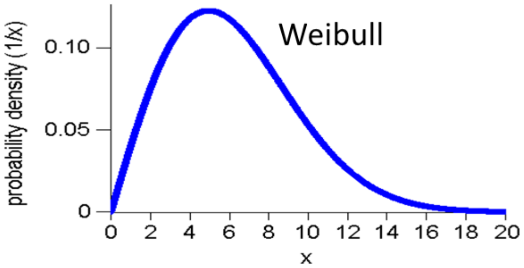
Discrete Distributions

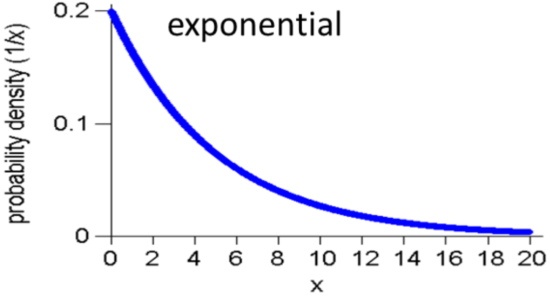
- Binomial
- Poisson
- ❖ Geometric
- ❖ Negative Binomial
- ❖ Hyper geometric

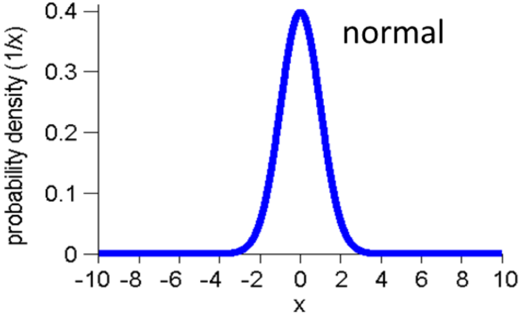
To be covered in later lesson

Not covered in BB class

Gallery of Continuous Distribution

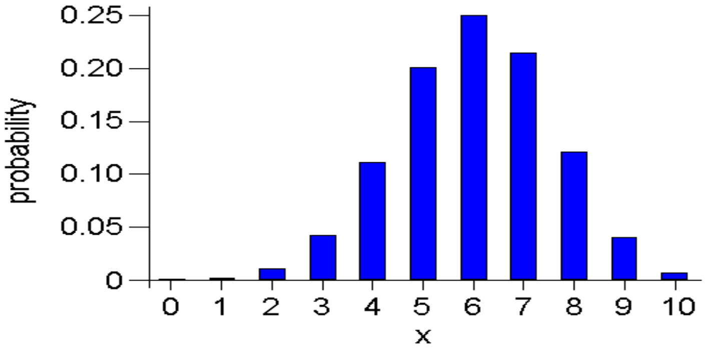




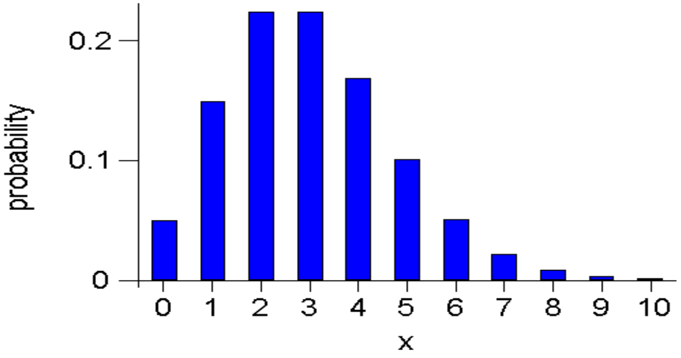


Gallery of Discrete Distribution

Binomial



Poisson



Distribution Parameters

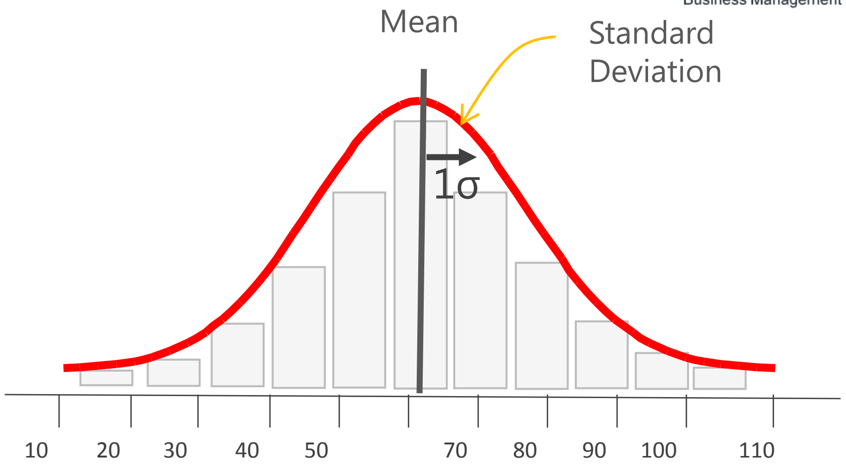
To characterize the distributions -

Location Parameter **(γ)**

Scale Parameter **(η)**

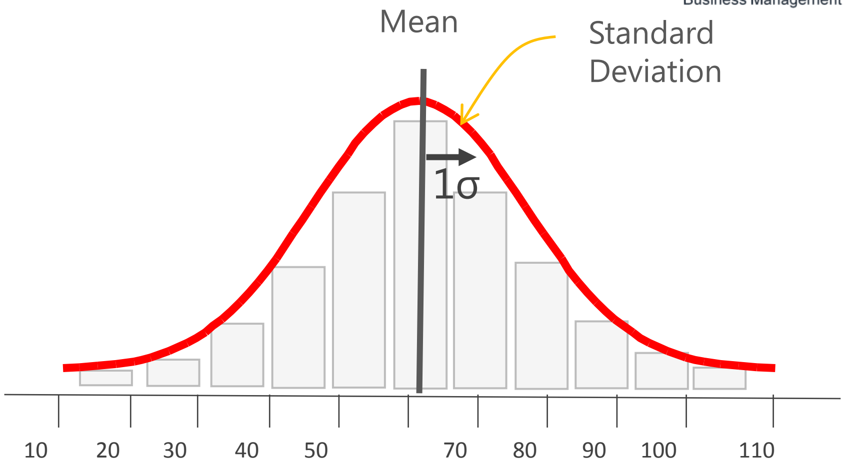
Shape Parameter **(β)**

Location Parameter



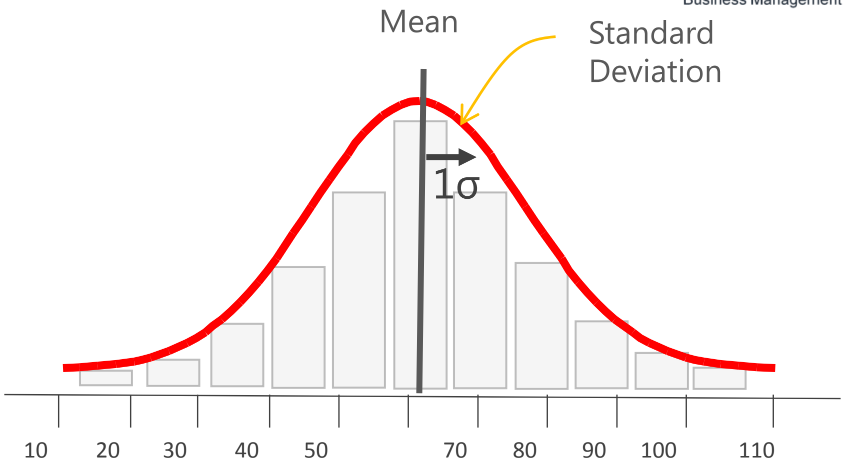
- Represented by γ
- The lower or midpoint (as prescribed by the distribution) of the range of the random variable. E.g., for a normal distribution, the mean.

Scale Parameter



- Represented by η
- Determines the scale of measurement for x (magnitude of the x-axis scale). E.g., for a normal distribution, the standard deviation.

Shape Parameter



- Represented by β
- Defines the PDF shape within a family of shapes. E.g., for a t distribution, the degrees of freedom.

Application of Distributions

- To **design the process** based on distribution properties (Capacity, Planning, Prediction, Scenario evaluation, etc)
- **In planning** by compute probabilities and process capabilities accurately
- To calculate **confidence intervals** for parameters (target setting) and to calculate critical regions for **hypothesis tests**.

Process Lead Time Data

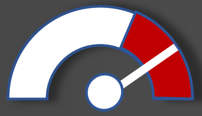
Become Future Fit

You will learn

How to deal with process lead time data?

Answer important questions for business

Level of Difficulty



High

Process Lead Time

You have process lead time (TAT) data. You want to find probabilities to make some important decisions.

- What proportion of items take greater than 2880 mins (48 hrs)?

Process Lead Time

You have process lead time (TAT) data.

You want to find probabilities to make some important decisions.

What proportion of items take greater than 2880 mins (48 hrs)?

Check for Normality

Process Lead Time

You have process lead time (TAT) data.

You want to find probabilities to make some important decisions.

What proportion of items take greater than 2880 mins (48 hrs)?

Non-Normal Data

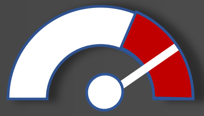
Weibull Distribution

Become Future Fit

You will learn

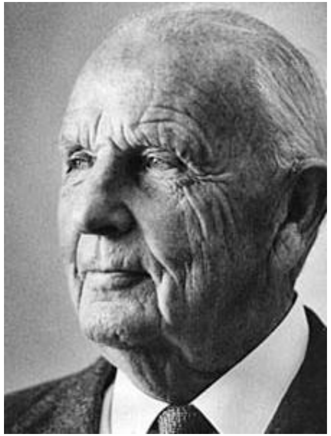
Learn about Weibull Distribution & its characteristics

Level of Difficulty



High

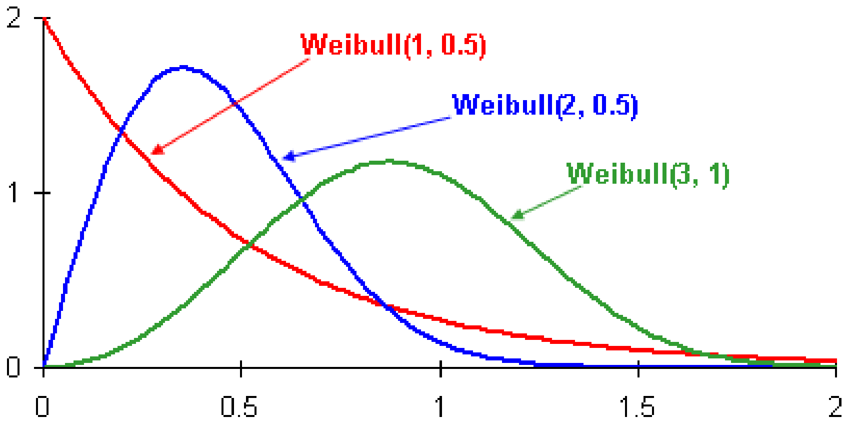
Weibull Distribution



Waloddi Weibull

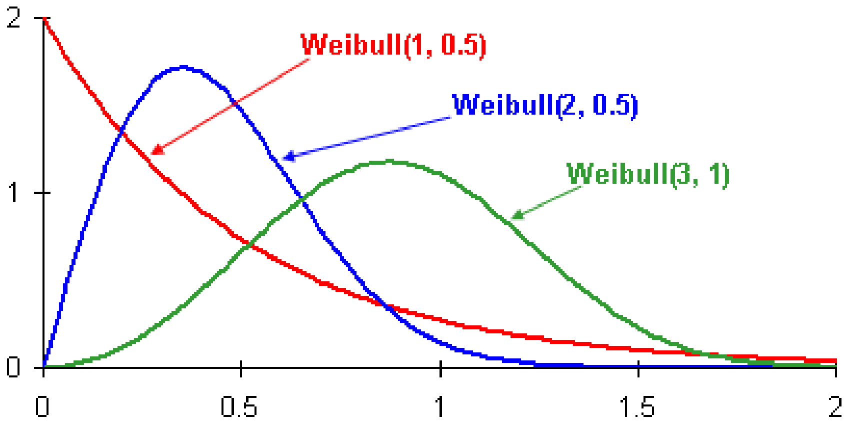
- Many continuous data sets fit to Weibull distribution.
- Used Reliability Data like failure times, process times, etc

- Weibull is a family of distributions
- It can take can shape
- Depending on the parameter values, it approximate an exponential, a normal or a skewed distribution



Weibull Parameters

- **3 Parameter Weibull Distribution**
  - β is the shape parameter
  - η is the scale parameter
  - γ is the location parameter
- **2 Parameter Weibull Distribution**
  - β is the shape parameter
  - η is the scale parameter
- Shape & Scale are expressed as a function of Standard Deviation



Weibull Parameters

Lets see if we can generate Weibull distribution for different Parameters

Weibull Distribution

1. Any Non-normal Continuous distribution can be treated as Weibull .
2. If you don't know which distribution this data belongs to, treat it as Weibull Distribution

Can Process Time Data belong to Weibull family?

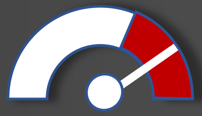
Identifying Distributions

Become Future Fit

You will learn

How to identify what distribution a given data belongs to?

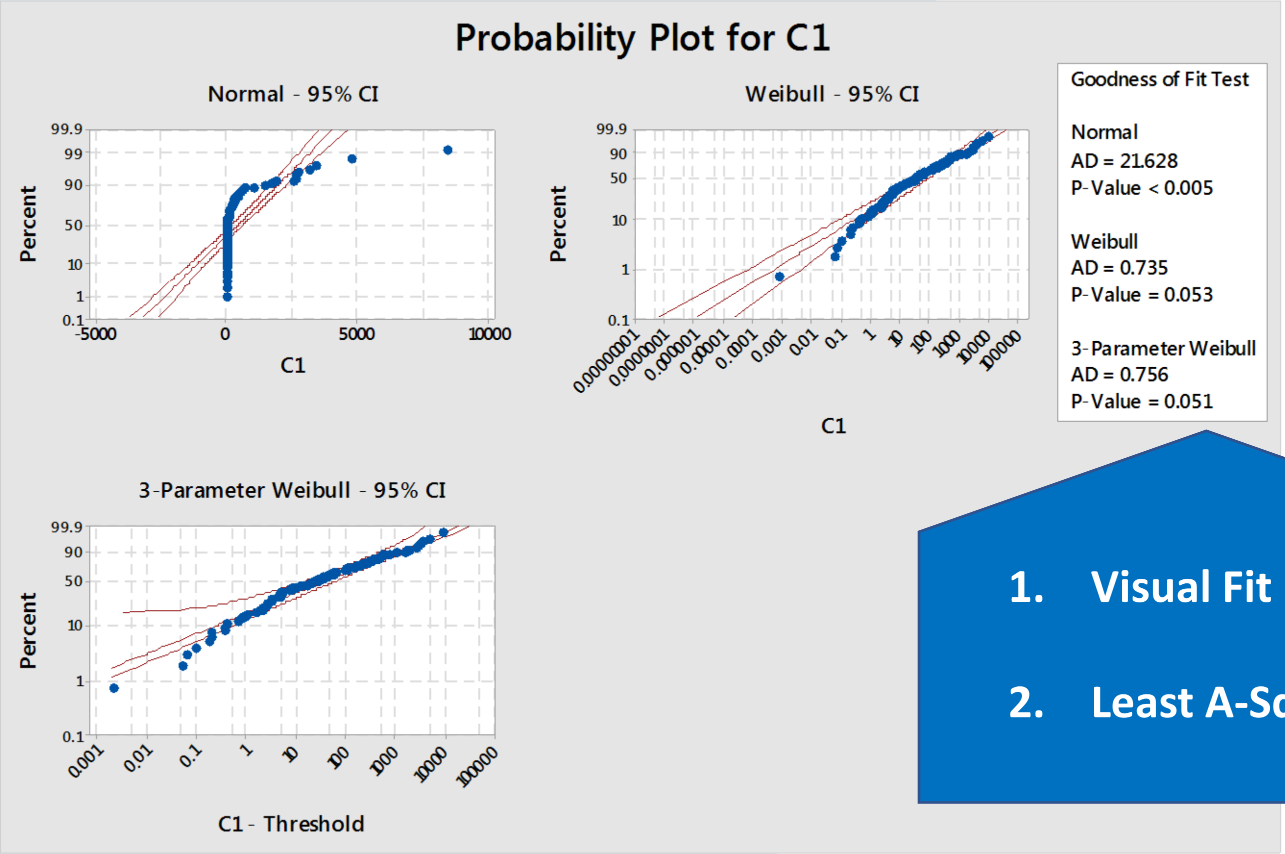
Level of Difficulty



High

Process Lead Time Data Case

Let's find out if the Process Time Data belongs to Weibull….



1. **Visual Fit**
2. **Least A-Sq value**

Process Lead Time Data Case

Let's find out if the Process Time Data belongs to Weibull….

Process Lead Time

You have process lead time (TAT) data.

You want to find probabilities to make some important decisions.

- What proportion of items take greater than 2880 mins (48 hrs)?

Process Lead Time

You have process lead time (TAT) data.

You want to find probabilities to make some important decisions.

- What proportion of items take greater than 2880 mins (48 hrs)?

We yet don't the Weibull Parameters…

Finding Weibull Parameters

- Use the data and perform Weibull Process Capability
- Weibull Shape & Scale Parameters will be available there

Finding Weibull Parameters

- Use the data and perform Weibull Process Capability
- Weibull Shape & Scale Parameters will be available there

Shape =0.377
Scale = 105.2

Process Lead Time

You have process lead time (TAT) data.

You want to find probabilities to make some important decisions.

- What proportion of items take greater than 2880 mins (48 hrs)?

P(X>=2880) =0.03073 (3.07%)

Weibull Applications

Become Future Fit

Canopus
Business Management Group

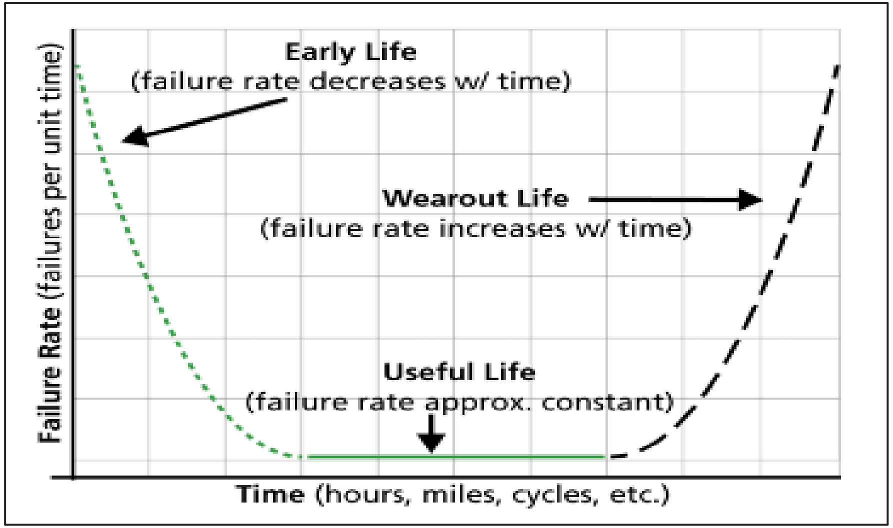
You will learn

Learn about the application of Weibull Distribution

Level of Difficulty

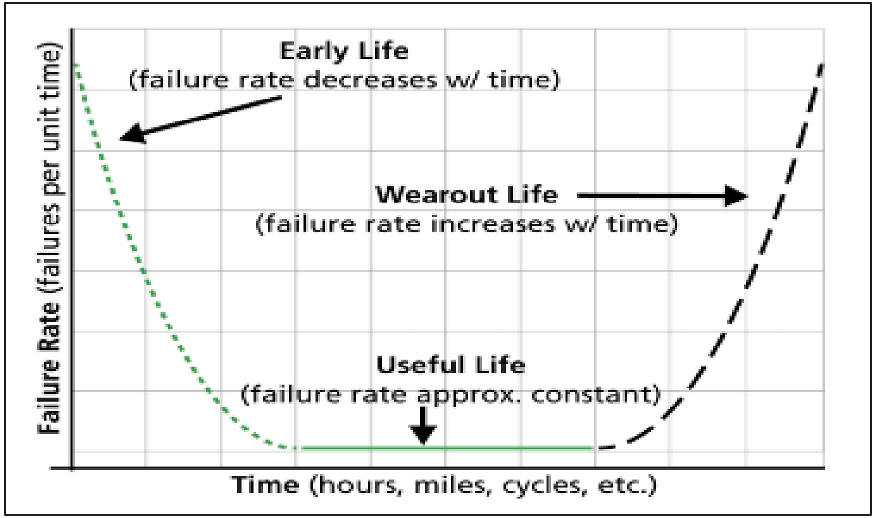
Medium

Bath-tub Curve

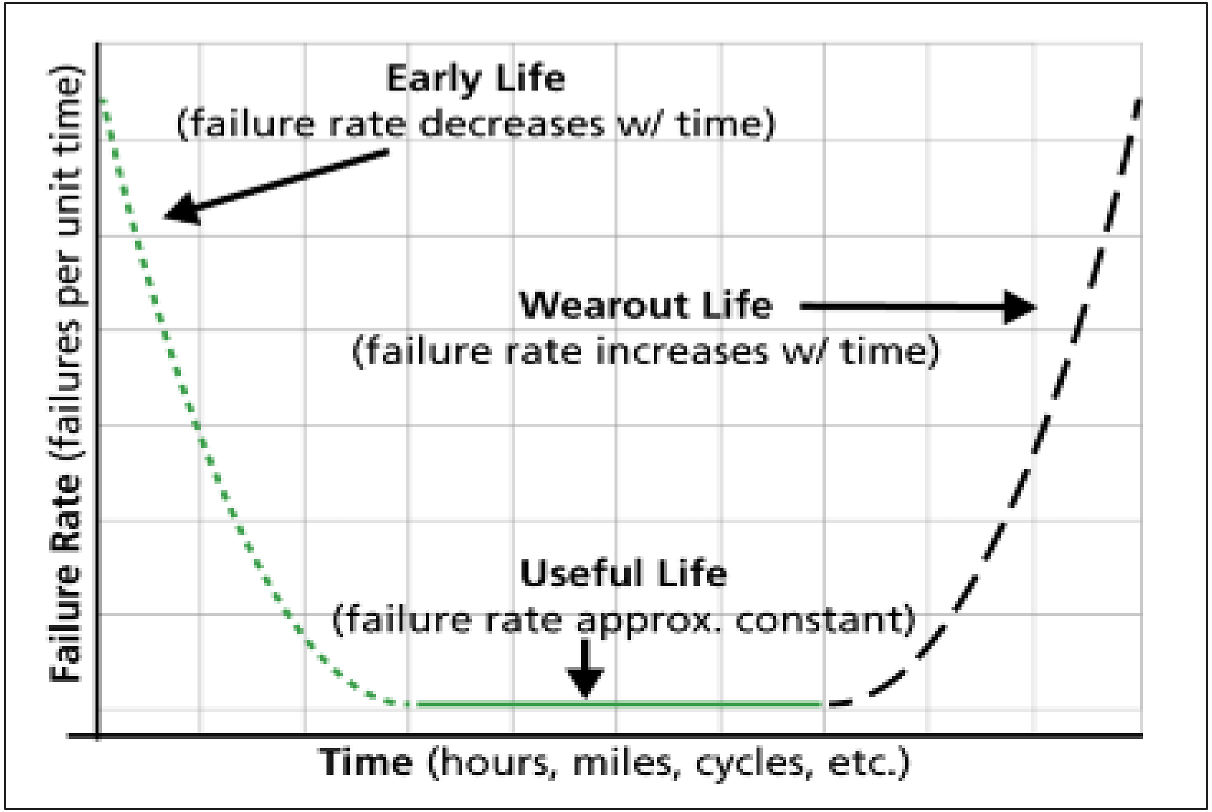


Bath-tub Curve

- Hazard Function Curve
- Shows Rate of Failure over Time



Canopus Business Management Group



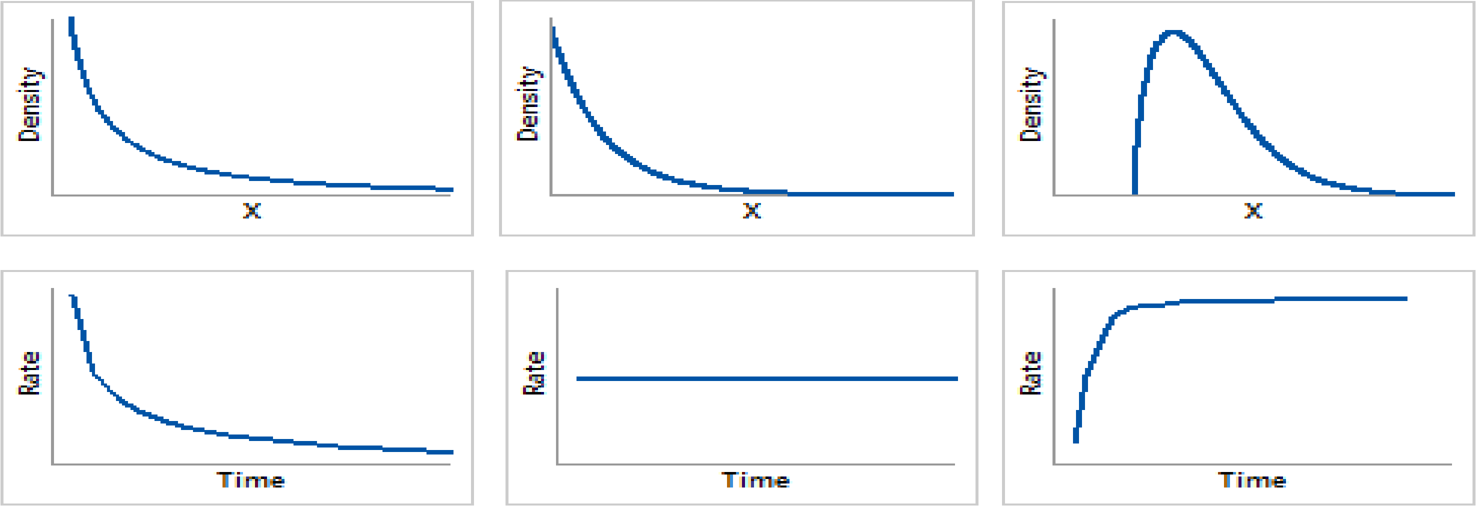
Weibull Variants

Business Management Group

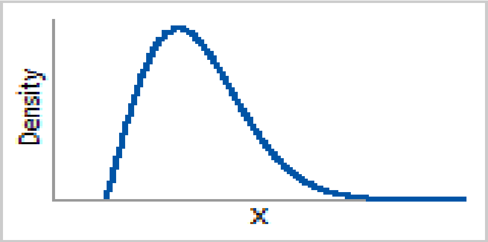
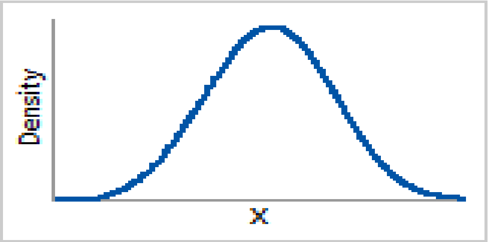
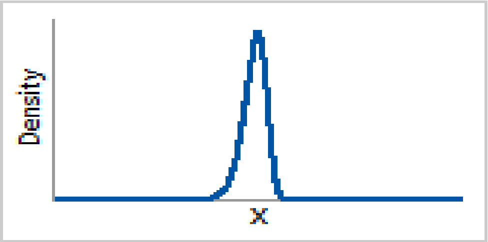
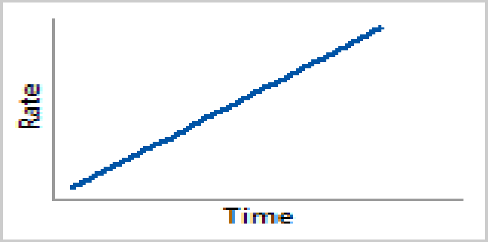
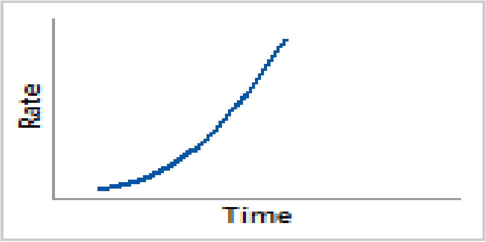
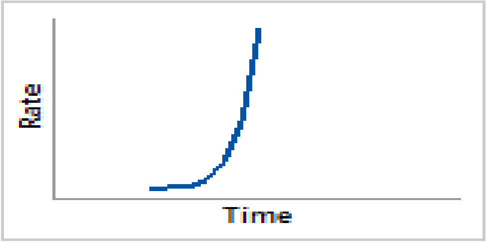
Early Life	Useful Life	Early wear-out failure
$0 < ß < 1$	$ß = 1$	$ß = 1.5$

Probability Plot

Hazard Function



Weibull Variants

	Wear-out ß = 2	Final period of product life 3 ≤ ß ≤4	The end ß > 10
Probability Plot			
Hazard Function			

Weibull Application across Industries

- Manufacturing
- Insurance
- Weather Forecasting
- Communications Engineering

Weibull Applications

- Failure Studies
- Reliability Studies
- Warranty Studies
- Maintenance Planning
- Equipment Life Studies

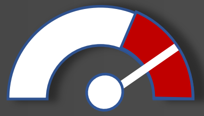
Exponential Distribution

Become Future Fit

You will learn

Learn about Exponential Distribution and its applications

Level of Difficulty

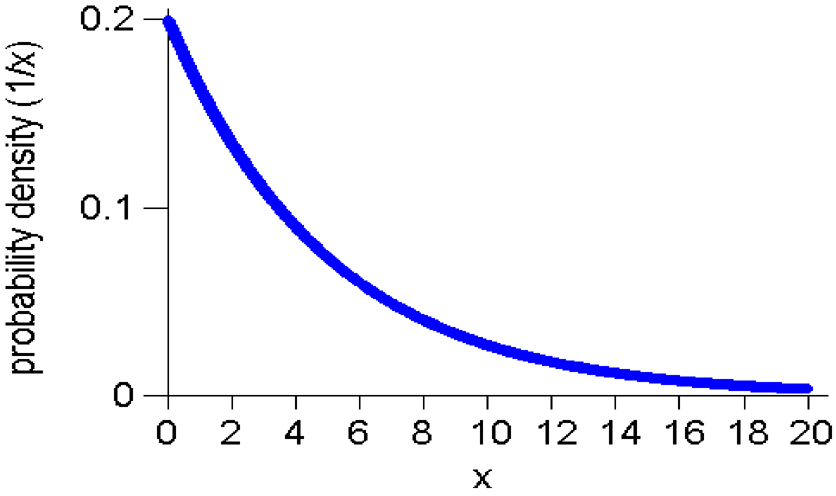


High

Search Time in Library

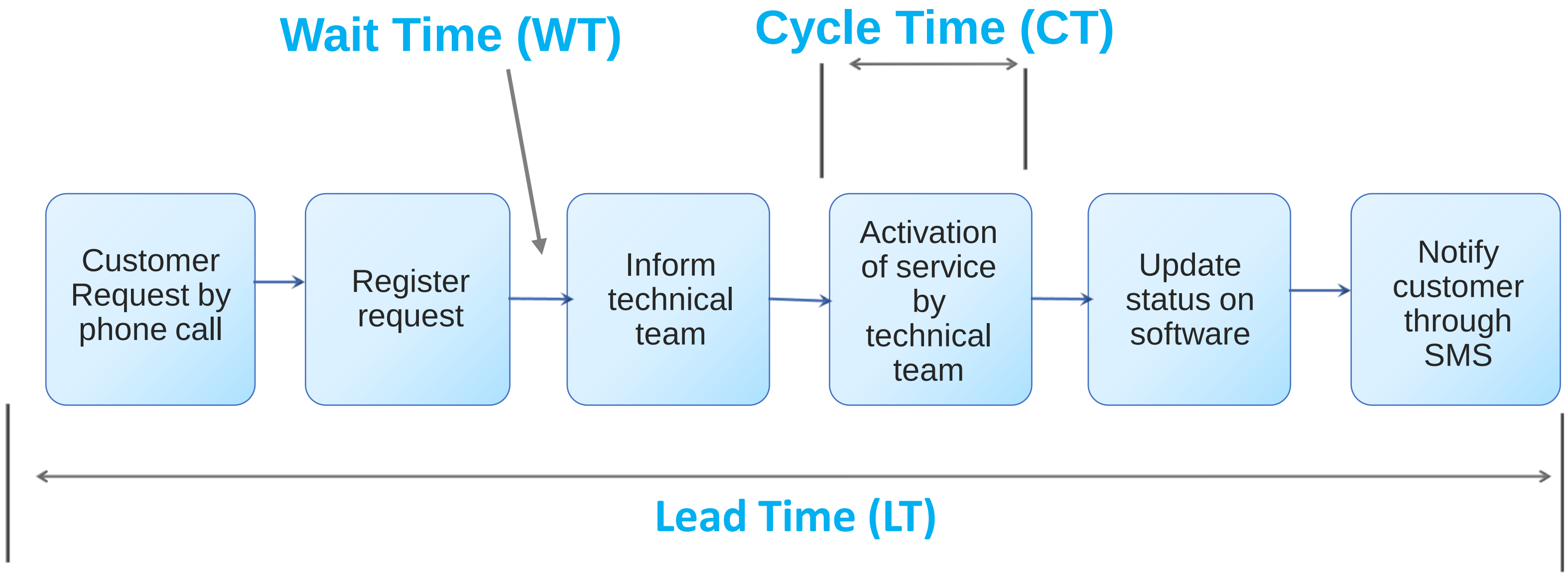
A start-up wants to identify time taken in Library to identify books to develop a innovative solution.

- Part of Weibull Family
- Maximum at x = 0, decays steadily as x increases.
- Approaches zero as $x \to \infty$
- x must be non-negative
- Memory less distribution



Mean = Std Dev

Non-normal Distribution



Applications

- Lead Times
- Process Times

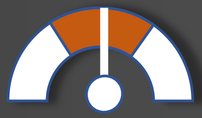
Binomial Approximation

Become Future Fit

You will learn

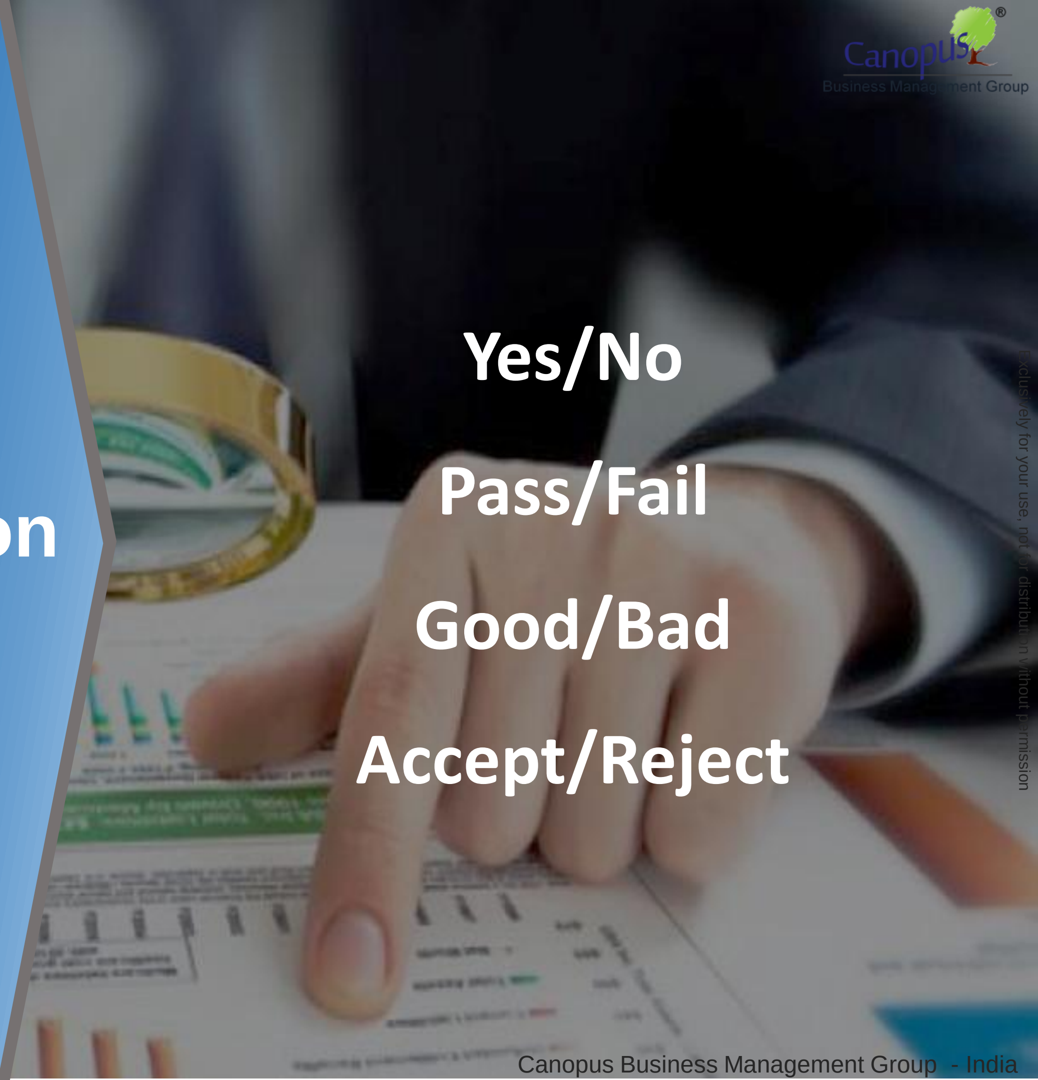
Learn about Binomial Distribution

Level of Difficulty



Medium

Binomial Distribution



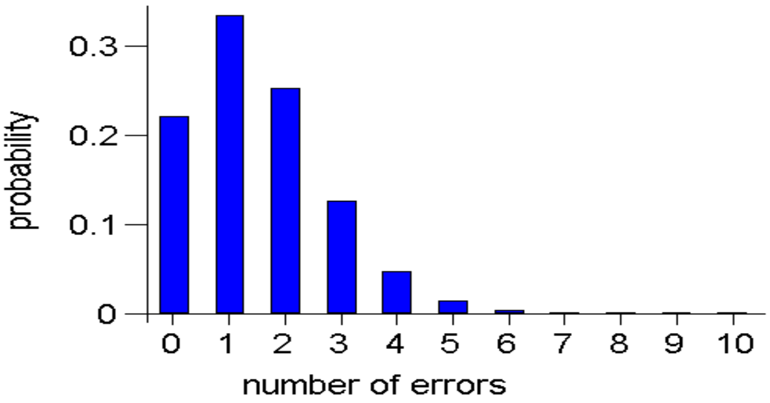
Yes/No

Pass/Fail

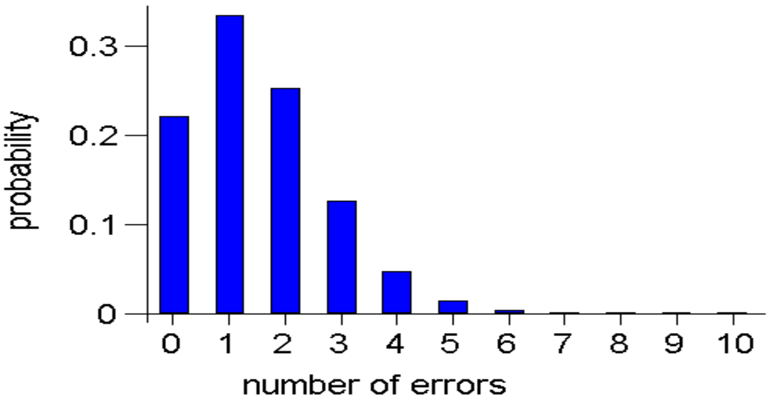
Good/Bad

Accept/Reject

Binomial Distribution



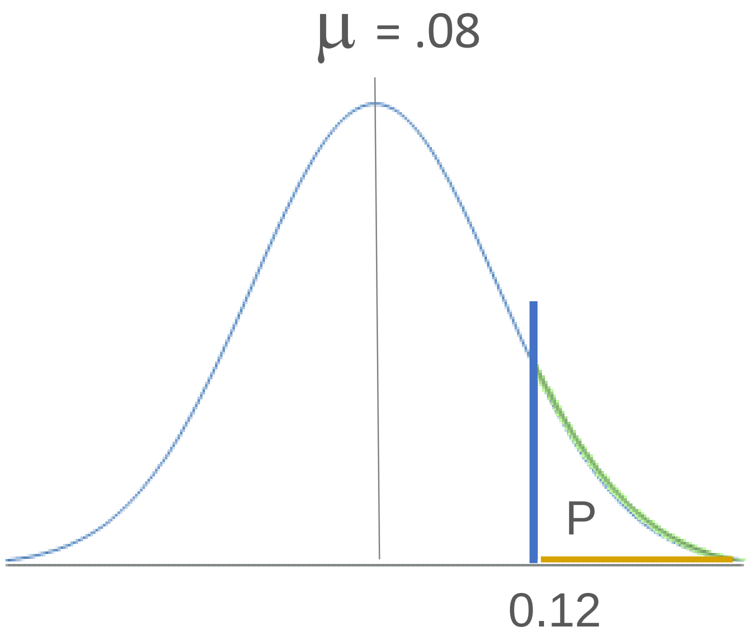
- Number of trials are defined
- Just 2 outcomes for each trial
- Trials are independent
- Probability of an outcome does not change from trial to trial



QC Audit on Process

Binominal Distribution - Example

- Process has 8% historic defect rate and the target is to be less than 12%.
- What is the chance of having greater than 12 defects if 100 components are checked the QC personnel now?



- Process has 8% historic defect rate and the target is to be less than 12%.
- What is the chance of having greater than 12 defects if 100 components are checked the QC personnel now?

Importance :

If 12 or more defects are found, then its likely that QC person will escalate the issue stating defect rate has increased compared to historic rate of 8%

Canopus Business Management Group

- Process has 8% historic defect rate and the target is to be less than 12%.
- What is the chance of having greater than 12 defects if 100 components are checked the QC personnel now?

Cumulative Distribution Function

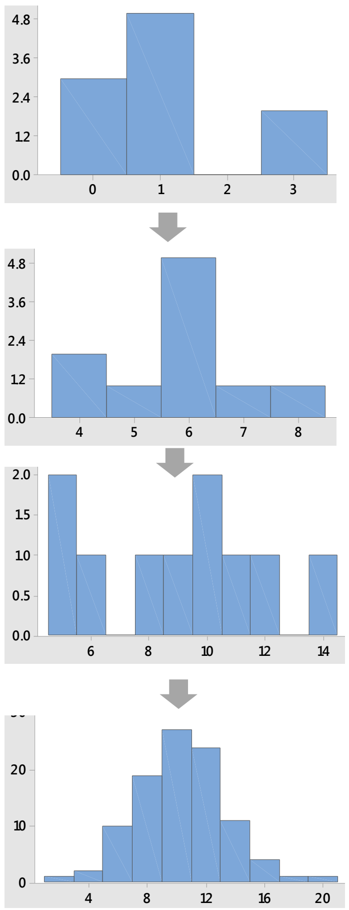
Binomial with n = 100 and p = 0.08

x	P(X ≤ x)
12	0.944120

P (X> 12) = 1 - 0.944120
= 0.05588

Binomial Approximations

- As n increases, Binomial Distribution tends to Normal
- With sufficient sample size, Discrete data can be analyzed with principles of Normal Distribution



Binomial Approximations

Canopus Business Management Group

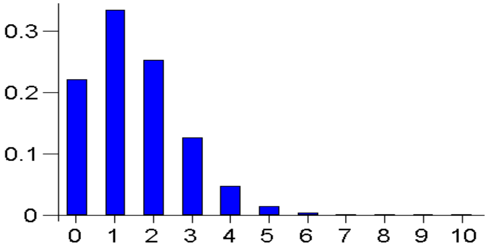
If n> 5/min(p,q)

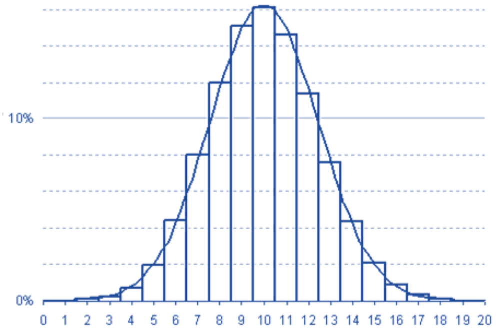
Binomial tends to Normal

Mean (μ) = n*p ; SD (σ) = sqrt[n*p*q]

If n> 20 & p < 0.05

Binomial tends to Poisson





Binomial Distribution Application

Become Future Fit

Canopus Business Management Group

You will learn

Application of Binomial Distribution

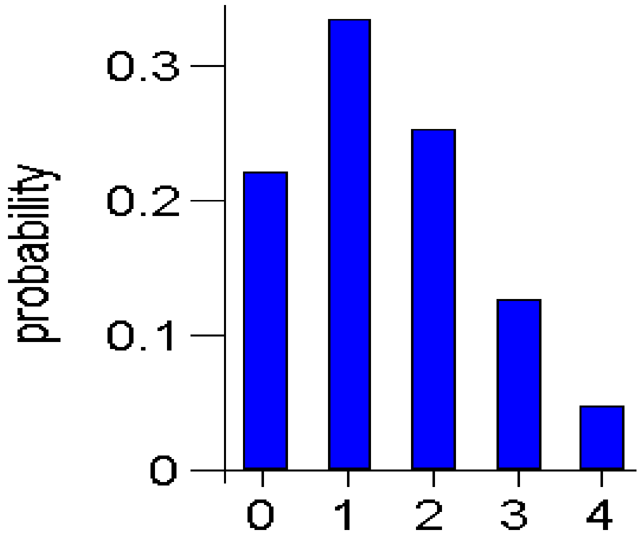
Level of Difficulty

High

IT Infra Planning Problem

Infra team procured servers. The manufacturer states that 95% of the servers last at least 10000 hours.

- What are the chances that all four of the servers will last at least 10000 hours?
- What is the probability that three will last that long & so on?



Infra team procured servers. The manufacturer states that 95% of the servers last at least 10000 hours.

- What are the chances that all four of the servers will last at least 10000 hours?
- What is the probability that three will last that long & so on?

Probability Density Function

Binomial with n = 4 and p = 0.95

x	P(X = x)
1	0.000475
2	0.013538
3	0.171475
4	0.814506

Poisson Distribution

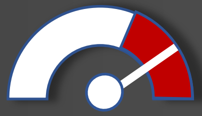
Become Future Fit

Canopus Business Management Group

You will learn

Learn about Poisson Distribution

Level of Difficulty



High

Poisson Distribution

Poisson distribution is used to represent process outcomes measured in rates.

Weibull

Time between Failure

Poisson

Failure Rate

Canopus
Business Management Group

Poisson Data

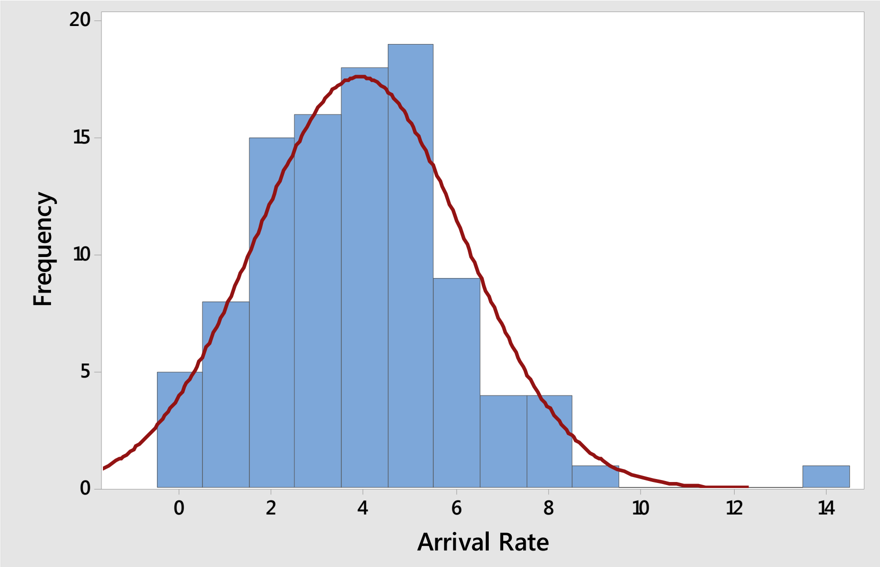
Defects/Opportunities

Arrivals per minute

Units per hour

Poisson Distribution

Data of arrival rates of customers into the stores



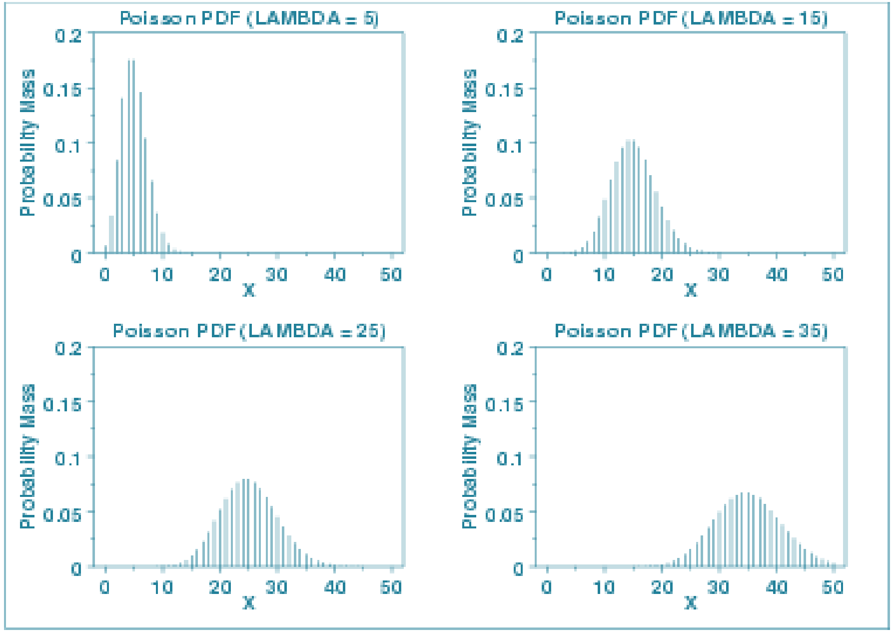
Mean = 3.9

SD = 2.26

St dev = Sqrt (Mean)

Poisson Distribution - Assumptions

- Equal Opportunity of Occurrences in the Area of Measurement
- Occurrences are independent



Applications

- Capacity problems
- Planning
- Floor Management
- Quality Management

Applications

- Sales – Enquiry Arrival Rate
- Marketing – Click Rates
- Front Office – Customer Arrivals & Queue Management
- Customer Service – Call Arrival Rates, Complaint Arrival Rates
- Supply Chain – Replenishment Rates
- Quality – Defect Rates
- IT – Outage or Issue Rates

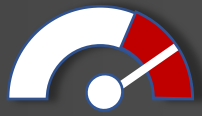
Poisson Distribution Application

Become Future Fit

You will learn

Application of Poisson Distribution

Level of Difficulty

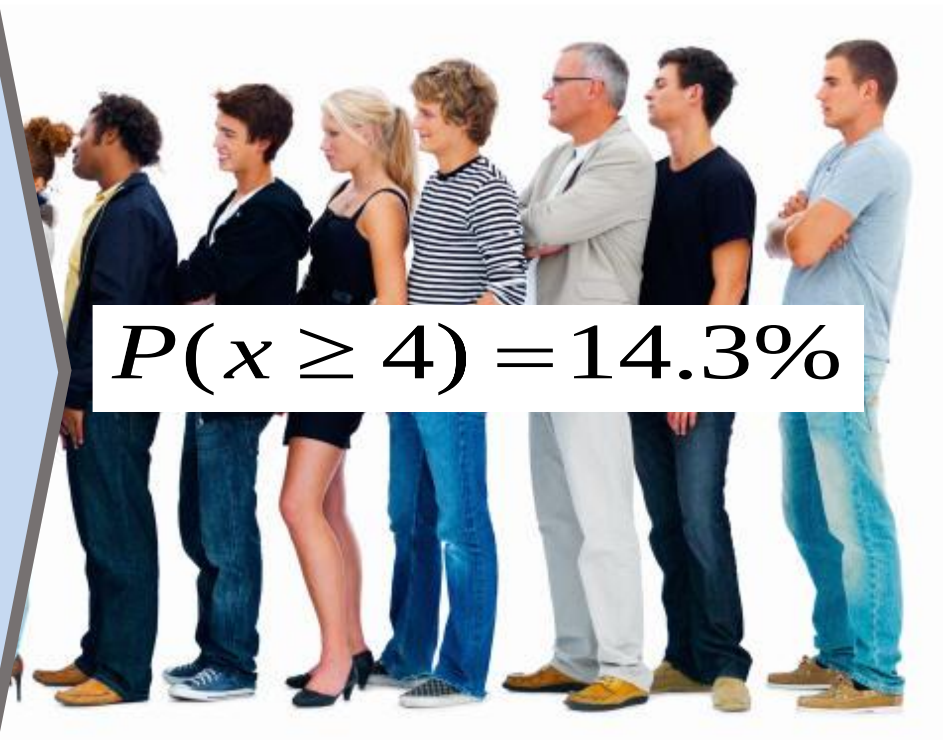


High

Front Desk Capacity Issue Problem

At a Front office desk, the customers arrive on average 2 per 10 mins during the morning. The store manager wants to determine staffing level for morning hour. In order to do that, she wishes to know what is the probability that there will be 4 or more customers will arrive in a 10 mins window?

At a Front office desk, the customers arrive on average 2 per 10 mins during the morning. The store manager wants to determine staffing level for morning hour. In order to do that, she wishes to know what is the probability that there will be 4 or more customers will arrive in a 10 mins window?

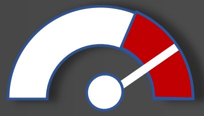


$$P(x \geq 4) = 14.3\%$$

You will learn

Learn about the impact of Sample Size on Distributions

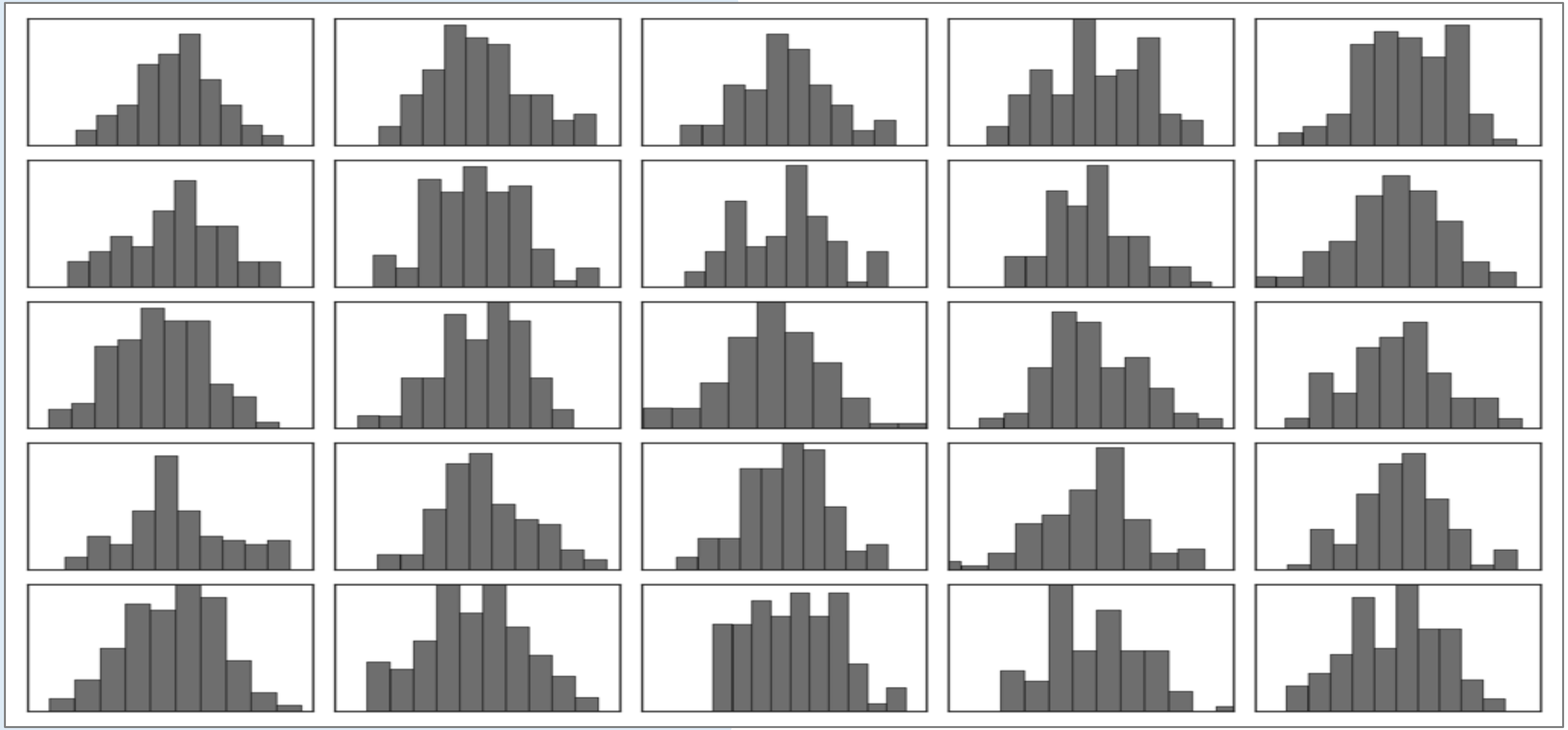
Level of Difficulty



High

Variability with Sample Size

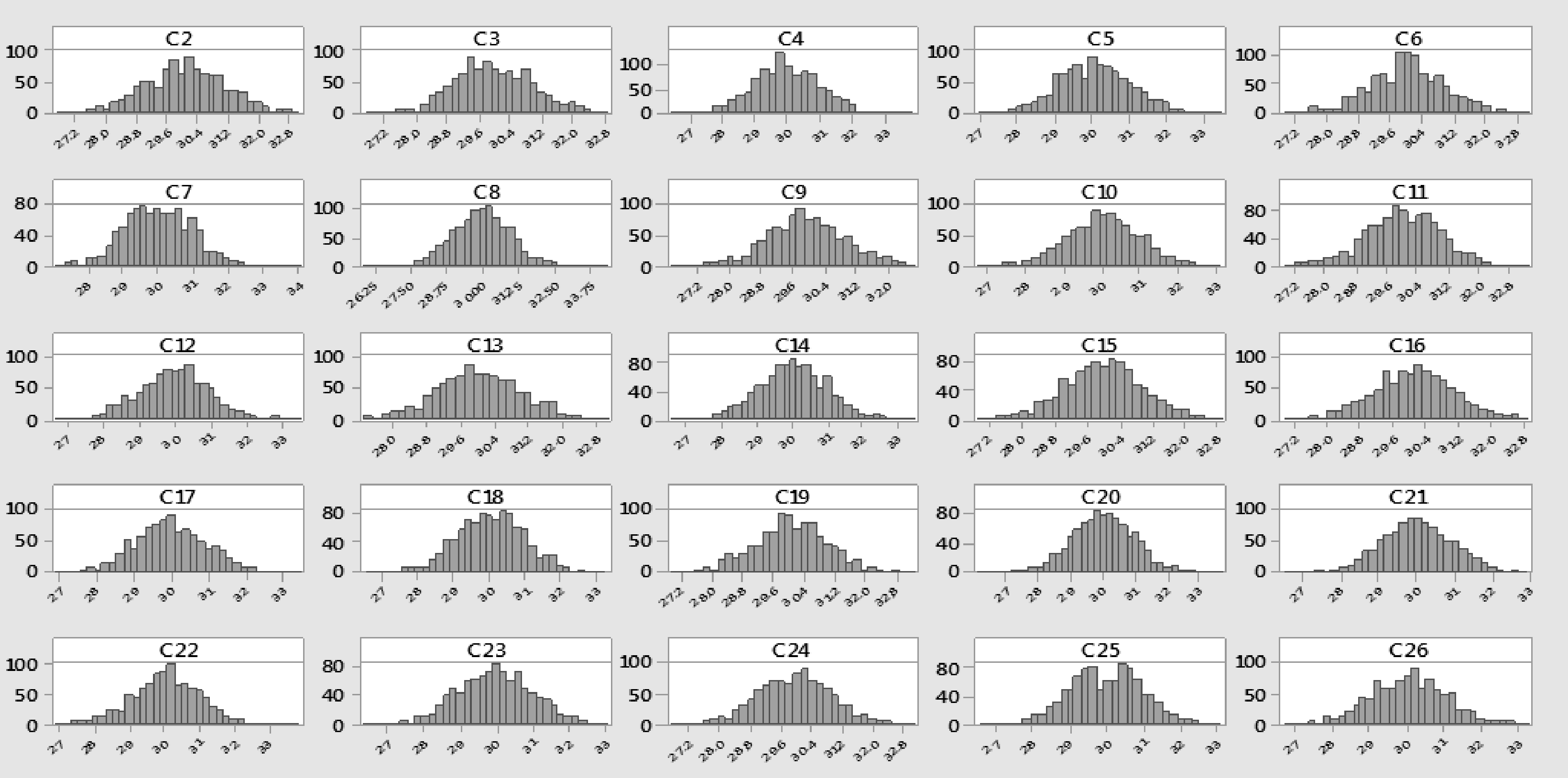
N= 100



Variability with Sample Size

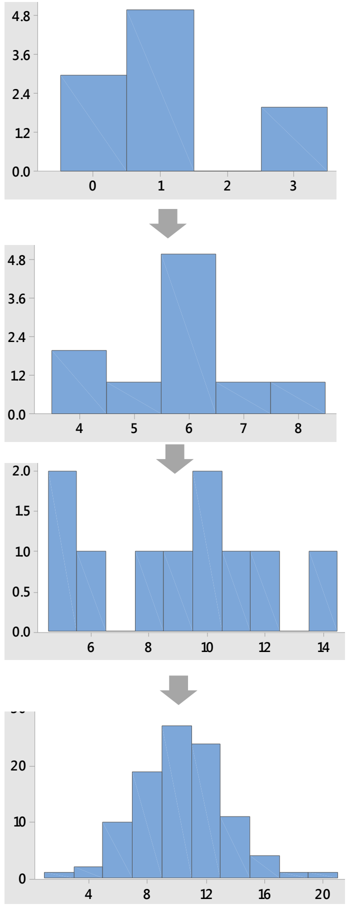
Canopus Business Management Group

N= 1000



Binomial Approximations

- As n increases, Binomial Distribution tends to Normal
- With sufficient sample size, Discrete data can be analyzed with principles of Normal Distribution



Binomial Approximations

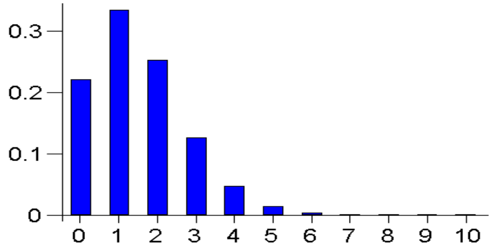
If n> 5/min(p,q)

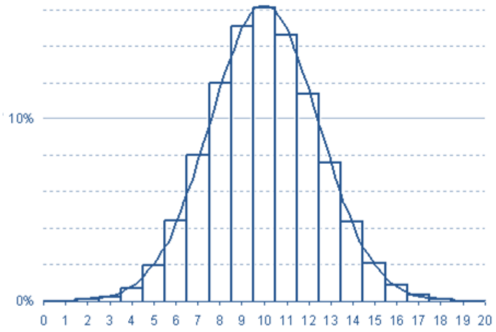
Binomial tends to Normal

Mean (μ) = n*p ; SD (σ) = sqrt[n*p*q]

If n> 20 & p < 0.05

Binomial tends to Poisson





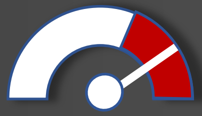
Central Limit Theorem

Become Future Fit

You will learn

What is Central Limit Theorem and its relevance to Business

Level of Difficulty



High

Letter of Credit Processing Data

The time to process a letter of credit by a banking officer is collected.

- Each day 51 samples of data are collected
- In total, 50 days data is collected

Let's check out the data now…..

Inference

Immaterial of the original distribution, the average of samples tends to be Normal.

Letter of Credit Processing Data

Now lets assume, we had lesser samples per day for letter of credit process:

- Each data 3 samples of data are collected
- In total, 50 days data is collected

Let’s check out the data now…..

Inference

Two things should be noted about the effect of increasing N :

- Distributions becomes more and more normal
- Spread of the distributions decreases

Central Limit Theorem

The mean of the SAMPLE MEANS:

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

The standard deviation of the SAMPLE MEAN:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

The standard deviation of the SAMPLE MEAN:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

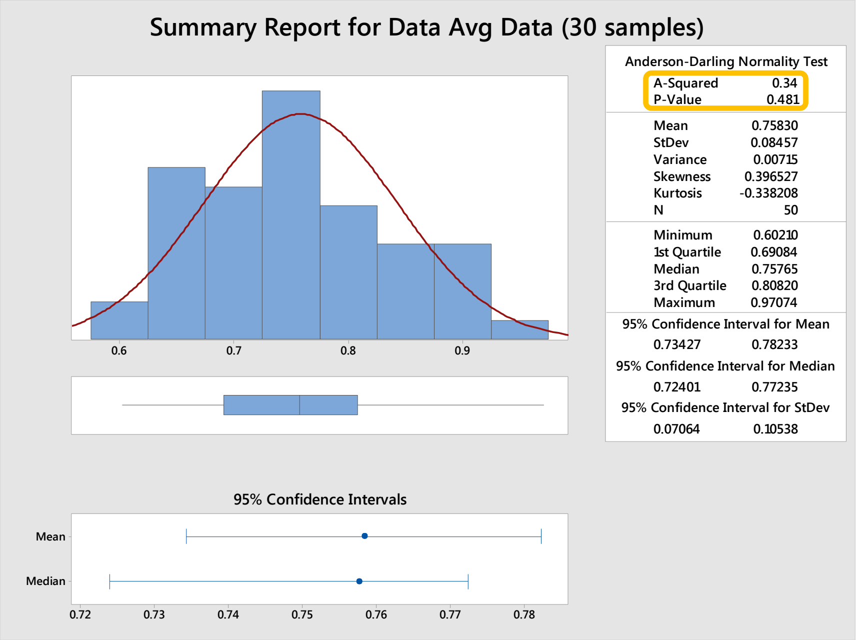
Standard Error of Mean (SE)

Letter of Credit Processing Data

Now lets assume, we had **30 samples** per day for letter of credit process.

Let’s check out the data now…..

Letter of Credit Processing Data



Inference

At around 30 Samples, the magic starts to happen!

Application

Original Distribution is Unknown or Non-Normal

- We can collect N samples, and apply the principles of Normal Distribution on sample data
- We can use a Sample Average & Standard Deviation to estimate the Population Average & Standard Deviation