Lean Six Sigma Green Belt Certification Course

simplilearn

Statistical Distributions

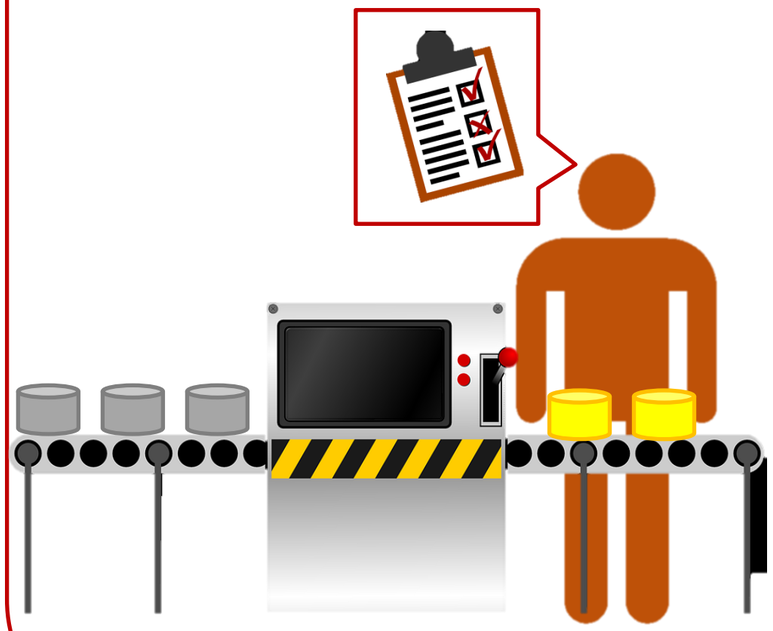
Learning Objectives

By the end of this lesson, you will be able to:

- Identify the types of statistical distributions
- Explain the Central Limit Theorem

Scenarios

What is the likelihood that there are no more than 10 defects per day?



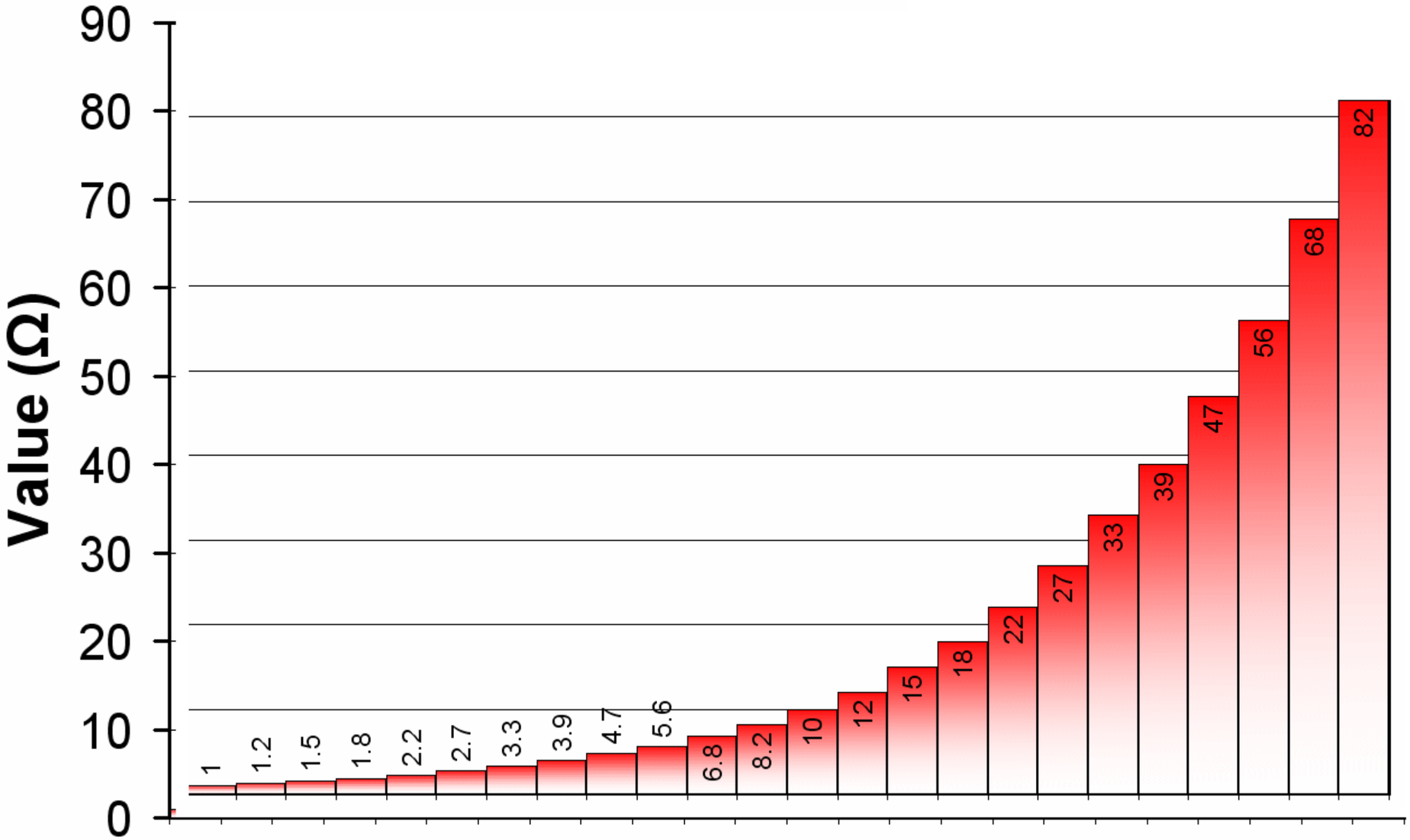
What is the probability that he achieves a 30-minute turn-around-time?

What is the chance that a customer will browse through the merchandise in the store before making a purchase?

DIGITAL OPERATIONS

Statistical Distributions

Classes of Distribution

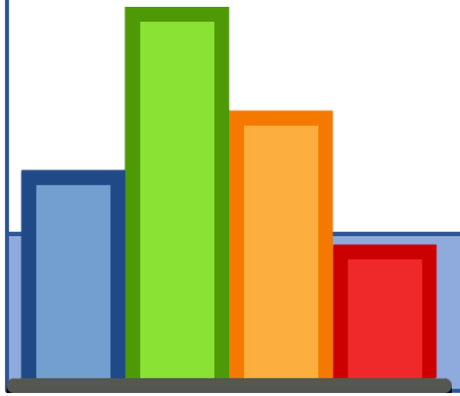


Classes of Distribution

- Based on assumed model of distribution
- Used to find the chances of a certain outcome/event to occur

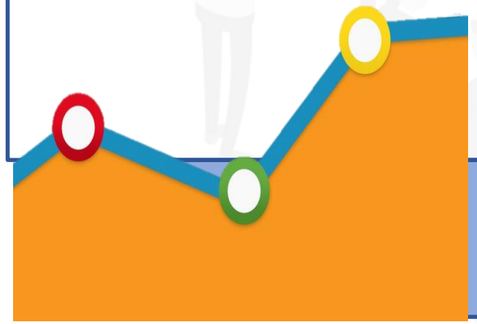
Probability

- Use the measured data to determine a model to describe the data used



Statistics

- Describes the population parameters based on the sample data using a particular distribution model

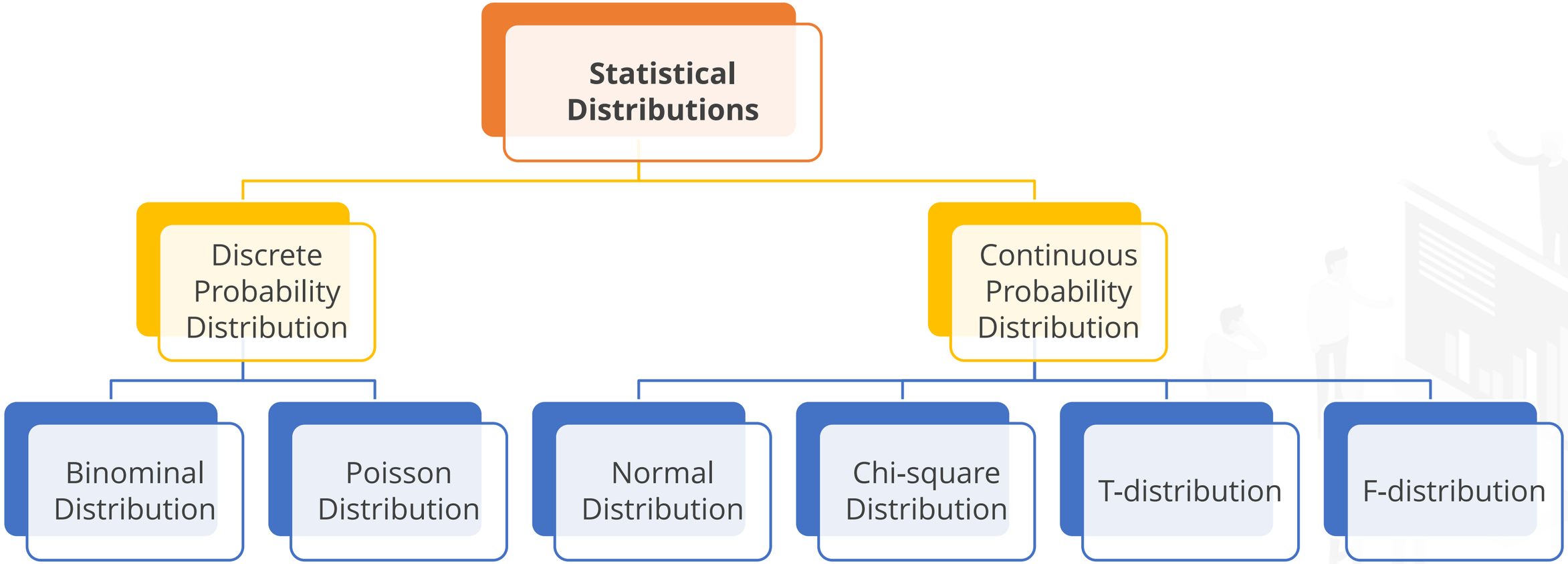


Inferential Statistics

Key Terms

Distribution of a variable	A distribution of a variable is a description of the relative number of times each possible outcome will occur in a number of trials.
Probability function	A function or equation that describes the probability a certain value will occur is called a probability function.
Probability mass function	Probability Mass Function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.
Probability density function	Probability Density Function is a function that gives the probability that a continuous random variable is exactly equal to some value.
Cumulative distribution function	A function or equation that describes the cumulative probability of a given value or any value smaller than it will occur is called the cumulative distribution function.

Types Of Statistical Distribution



Discrete Probability Distribution

Discrete probability distribution is characterized by the probability mass function.

These distributions help in predicting the sample behavior that has been observed in a population.

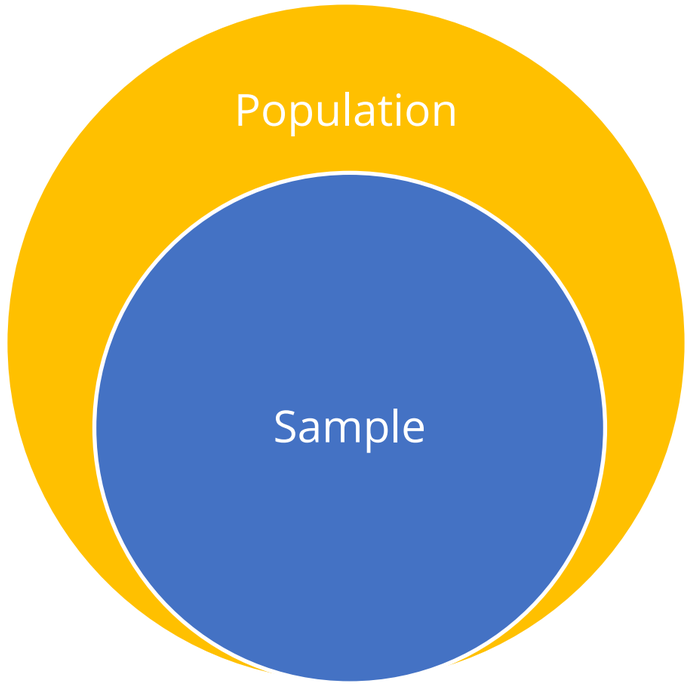
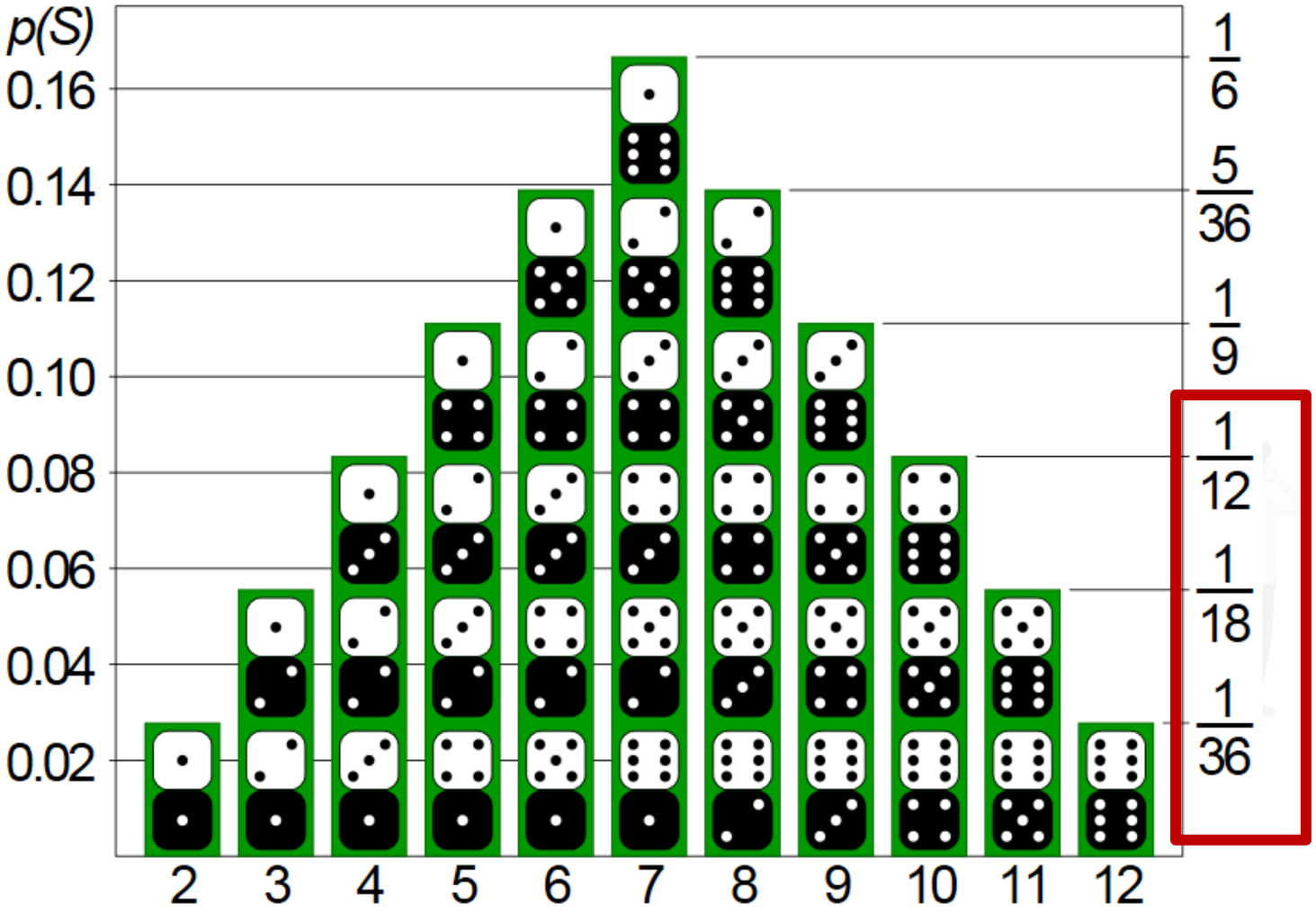


Illustration of the Probability Mass Function (PMF)

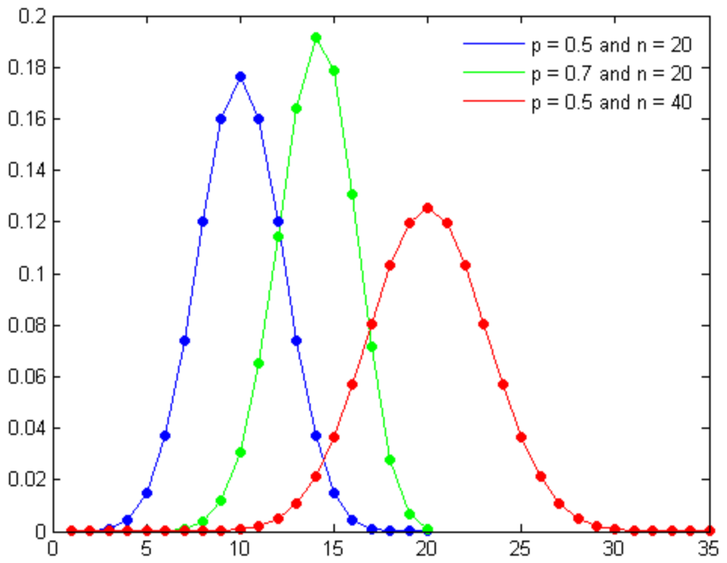


Binomial Distribution

Binomial distribution is a probability distribution for the discrete data.

Based on the Bernoulli process to predict sample behavior

Used to deal with defective items



Used best in situations where the sample size is less than 30 and less than 10% of population

Expressed as a percentage of non-defective items

Binomial Distribution: Formula

$$P(R) = {}^{n}C_{r} * p^{r} * (1-p)^{n-r}$$

Where,

P(R) = probability of exactly (r) successes out of a sample size of (n)

p = probability of success

r = number of successes desired

n = sample size

Key Calculations of Binomial Distribution

Mean

$\mu = np$

n = Sample size
p= Probability of success

Standard deviation

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

Sample factorial calculation

$5! = 5 * 4 * 3 * 2 * 1 = 120$

$4! = 4 * 3 * 2 * 1 = 24$

Binomial Distribution: Example

Using binomial distribution formula, find the probability of getting 6 heads in 10 coin tosses.

Outcomes are statistically independent.
Therefore,

Heads or tails?

p = probability of success = 0.5 (this remains fixed over time)
n = sample size = 10
r = number of successes desired = 6

$$P(R) = \binom{10}{6} * 0.5^6 * (1 - 0.5)^{10-6} = 0.205078 = \textbf{20.5\%}$$

=BINOM.DIST(successes, trails, probability for success, cumulative distribution?)
=BINOM.DIST(6, 10, 0.5, FALSE) = **20.5%**

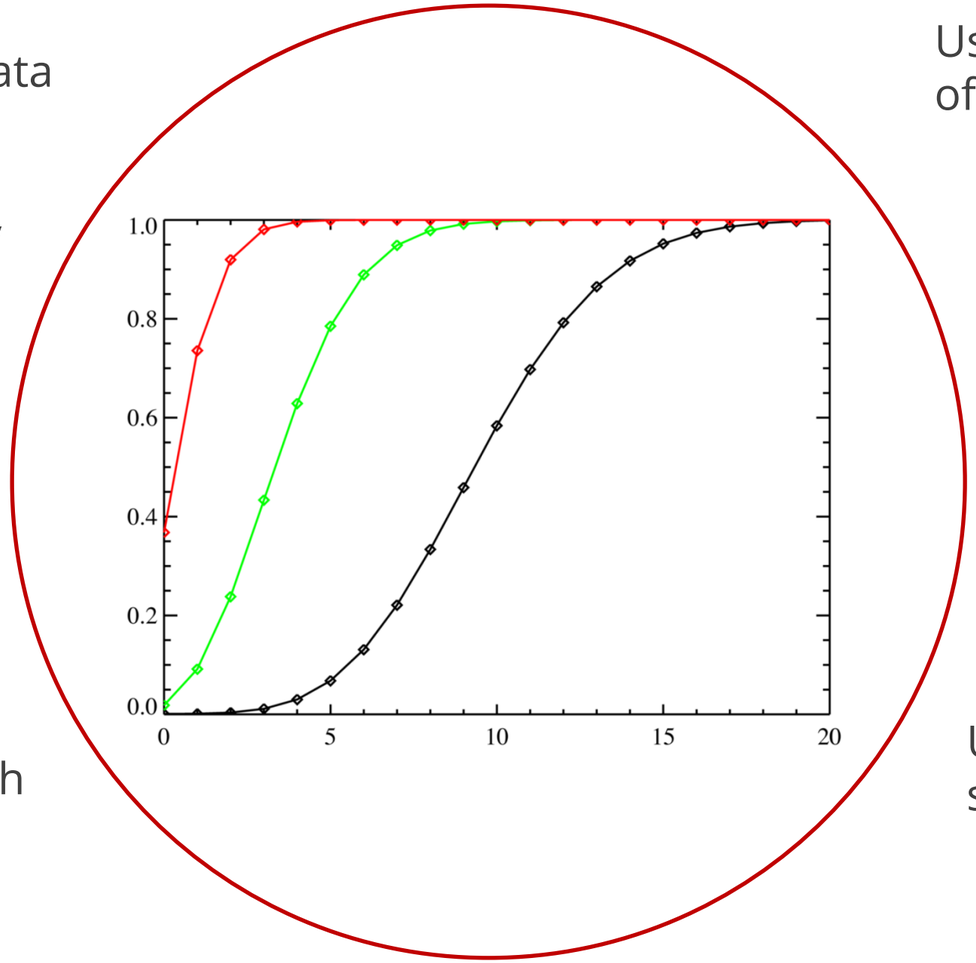
Poisson Distribution

Poisson distribution is an application of the population knowledge to predict the sample behavior.

Used for discrete data

Used to predict the number of defects

Used to describe the probability distribution of an event with respect to time or space



Used where the number of trials is large and tends towards infinity

Describes discrete data resulting from a process

Deals with integers which can take any value

Used where the probability of success in each trail is very small

Poisson Distribution: Formula

$$P(x) = \frac{\lambda^{X} * e^{-\lambda}}{x!}$$

Where,

P(x) = Probability of exactly (x) occurrences in a Poisson distribution (n)

λ = Mean number of occurrences during interval

x = Number of occurrences desired

e = Base of the natural logarithm (equals 2.71828)

Mean of a Poisson Distribution (μ) = λ

Standard Deviation of a Poisson Distribution (σ) = $\sqrt{\lambda}$

Poisson Distribution: Example

The past records of a traffic intersection shows that the mean number of accidents every month is three at this junction. Assume that the number of accidents follows a Poisson distribution and calculate the probability of number of accidents happening in a month.

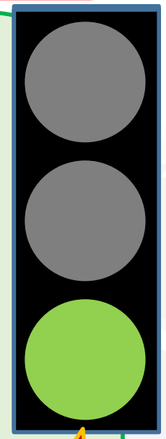
Given: λ=15 per month

Now, probability of zero accidents per week $P(0) = \frac{3^0 * e^{-3}}{0!} = 0.0498 = 5.0\%$

Probability of exactly one accident per week $P(1) = \frac{3^1 * e^{-3}}{1!} = 0.1494 = 15.0\%$

Probability of more than two accidents per week = 1 – [P(0)+P(1)+P(2)] = 1 – [0.0498+0.1494+0.2240]

= 0.5768 = **57.7%**



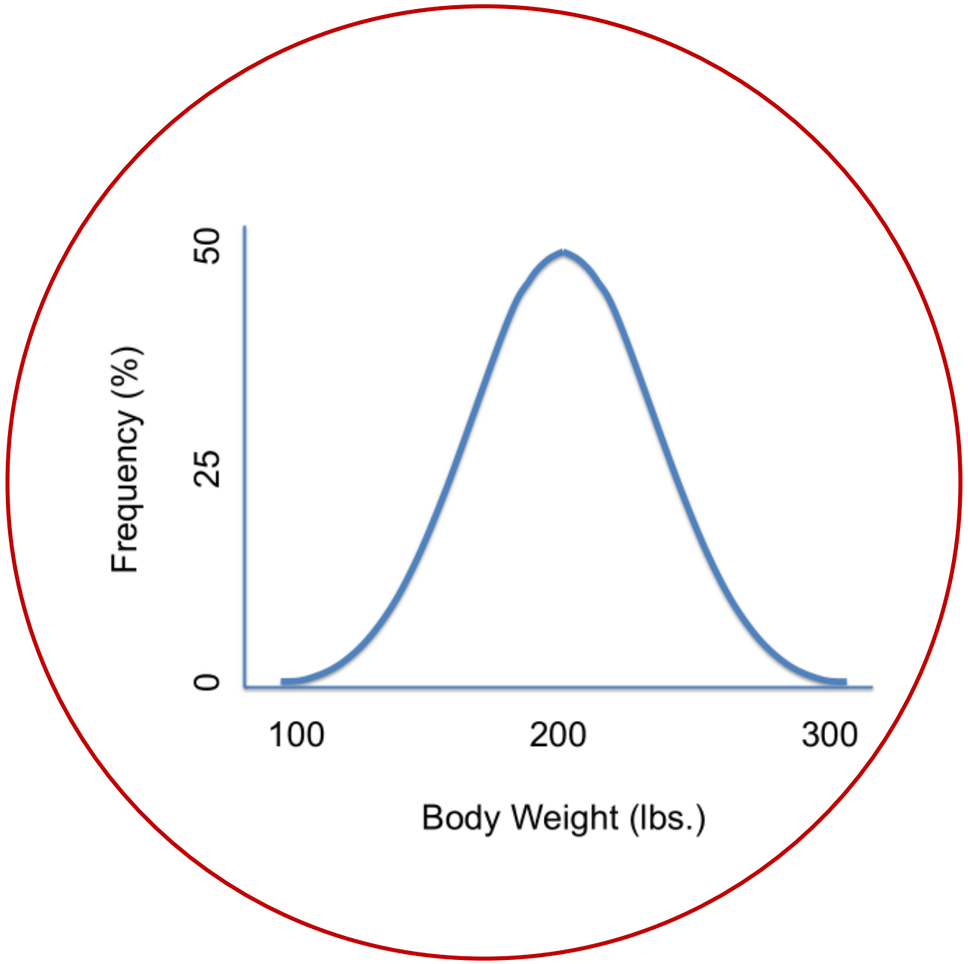
=POISSON.DIST(successes desired, average of past successes, cumulative distribution?)

=POISSON.DIST(0, 3, FALSE) = 5.0% OR POISSON.DIST(2,3,TRUE) = **42.3%**

Continuous Probability Distribution

Continuous probability distribution is characterized by the probability density function.

Variable is continuous if the range of possible values fall along an infinite continuum



Distributions help to predict the sample behavior observed in a population

Normal Distribution

The Normal or Gaussian distribution is a continuous probability distribution.

Illustrated as N (μ, σ)

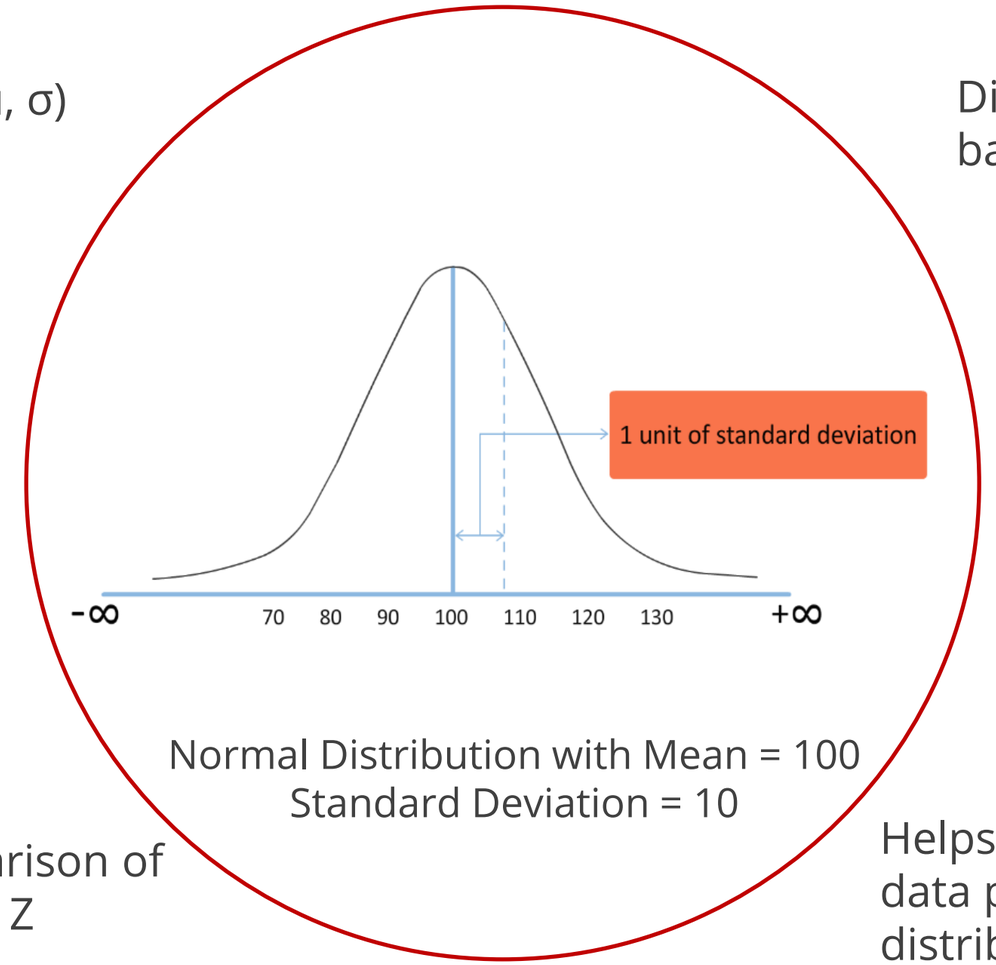
Distribution is bell or upside down bath tub shaped and symmetrical

Higher frequency of values around the mean and lesser occurrences away from it

The total area under the normal curve p(x which is found in the distribution) = 1.

Used as a first approximation to describe real-valued random variables

Distribution is continuous and symmetrical



Normal Distribution with Mean = 100
Standard Deviation = 10

To standardize comparison of dispersion a standard Z variable is used

Helps in finding probabilities for data points anywhere within the distribution

Normal Distribution: Formula

$$Z = \frac{(Y - \mu)}{\sigma}$$

Where,

Z = number of standard deviations between Y and the μ

Y = Value of the data point in concern

μ = Mean of the population

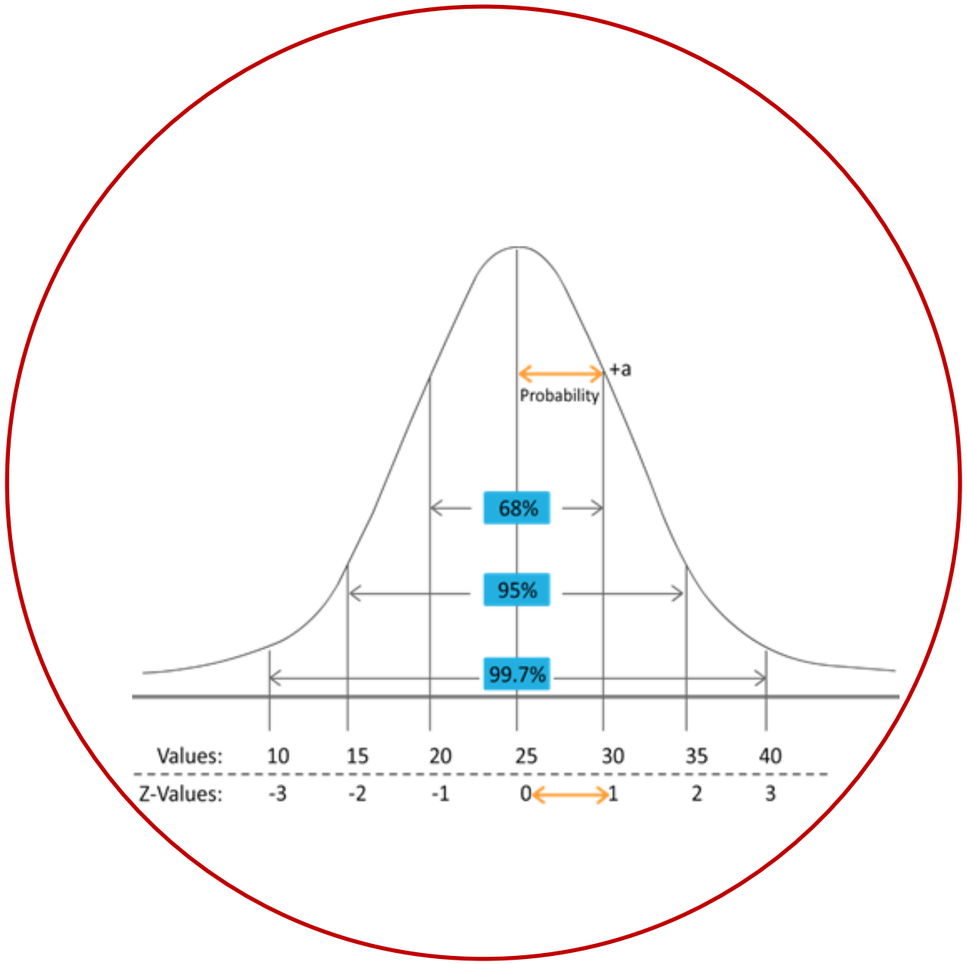
σ = Standard deviation of the population

Normal Distribution: Z-Table

The probability of areas under the curve is 1.

This probability is the area under the curve to the left of point a+ to 0.

Normal distribution
Uses actual data:
Average = 25
Standard deviation = 5



Standard normal distribution
Standardized data to:
Average = 0
Standard deviation = 1

Normal Distribution: Left Tail Z-Table

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5348	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	08869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Normal Distribution: Example 1

Find the value of P of (Z less than 0).

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.534	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.58								0.6141
0.3	0.6179	0.621								0.6517
0.4	0.6554	0.659								0.6879
0.5	0.6915	0.695								0.7224
0.6	0.7257	0.729								0.7549
0.7	0.7580	0.761								0.7852
0.8	0.7881	0.791								0.8133
0.9	0.8159	0.818								0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	08869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

There is no need of the table to find the answer once you know that the variable Z takes a value of less than (or equal to) zero.

The area under the curve is 1.

The curve is symmetrical about Z = 0. Hence, there is above 0.5 (or 50%) chance of Z = 0 and below 0.5 (or 50%) chance of Z = 0.

Normal Distribution: Example 2

Find the value of P of (Z greater than 1.12).

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500									
0.1	0.5398									
0.2	0.5793									
0.3	0.6179									
0.4	0.6554									
0.5	0.6915									
0.6	0.7257									
0.7	0.7580									
0.8	0.7881									
0.9	0.8159	0.8186	0.8212		0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	08869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

The opposite or complement of an event A occurring is the event A not occurring.

P(not A) = 1 – P(A)

P(Z greater than 1.12) = 1 – P(Z less than 1.12)

Using the table:

P(Z < 1.12) = 0.8686

Hence, P(Z > 1.12) = 1 – 0.8686 = **0.1314**

Normal Distribution: Example 3

Find the value of P of (Z lies between 0 and 1.12).

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500									
0.1	0.5398									
0.2	0.5793									
0.3	0.6179									
0.4	0.6554									
0.5	0.6915									
0.6	0.7257									
0.7	0.7580									
0.8	0.7881									
0.9	0.8159	0.8186	0.8212		0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	08869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

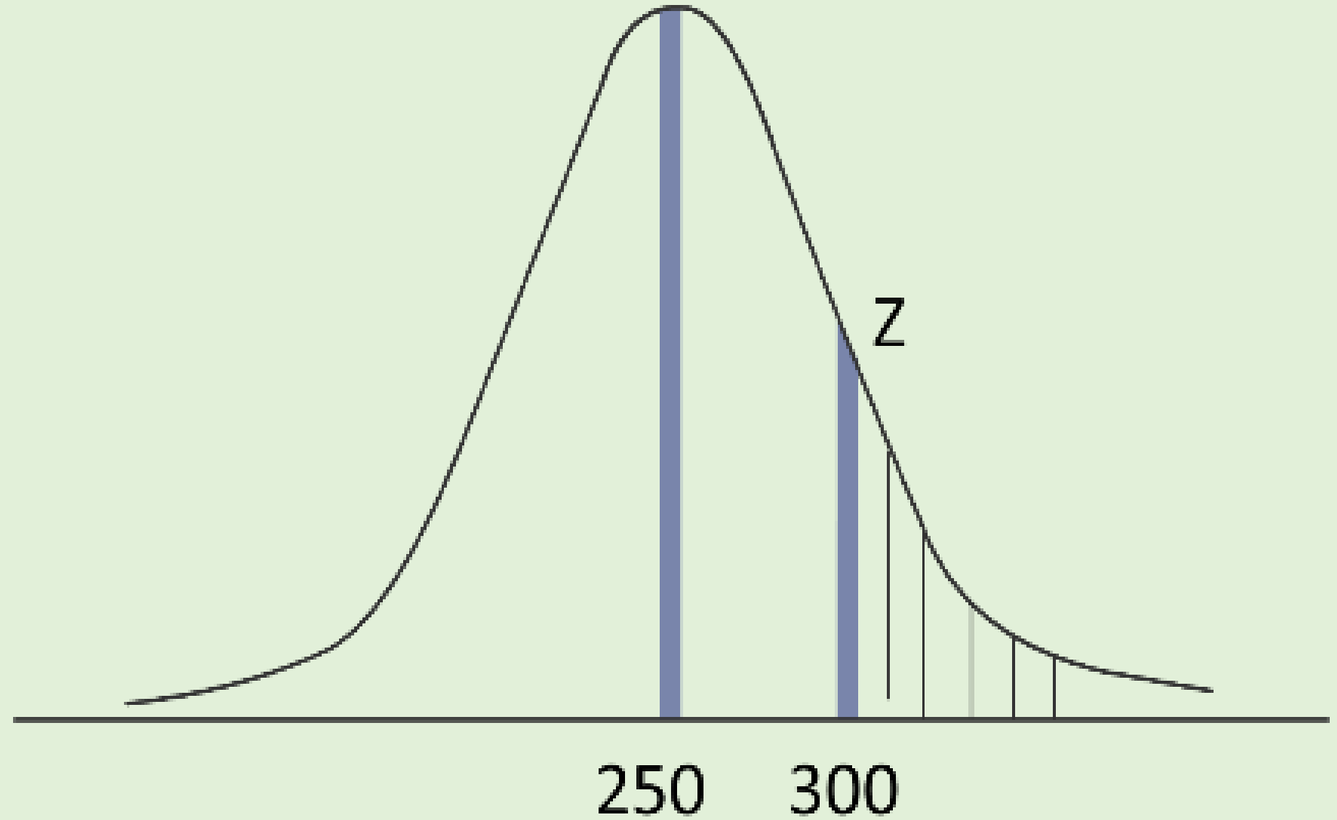
Z falls within an interval.

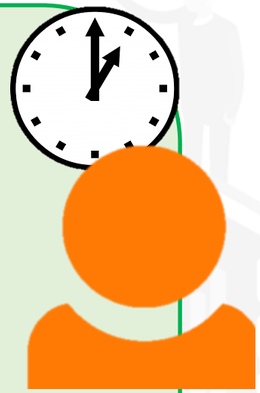
Using the table:

P(Z lies between 1.12 and 0) $= P(Z < 1.12) - P(Z > 0) = 0.8686 - 0.5 =$ **0.3686**

Normal Distribution: Example 4

Suppose the time taken to resolve customer complaint follows a normal distribution with the mean value of 250 hours and standard deviation value of 23 hours. What is the probability that a problem resolution will take more than 300 hours?





Normal Distribution: Example 4

Suppose the time taken to resolve customer complaint follows a normal distribution with the mean value of 250 hours and standard deviation value of 23 hours. What is the probability that a problem resolution will take more than 300 hours?

Given: Y = 300; μ = 250; σ = 23

METHOD 1:

=NORM.DIST(value of interest, average, standard deviation, cumulative distribution?)

=NORM.DIST(300, 250,23, TRUE) = 0.985 = 98.5%

1 – 98.5% = **1.5%**

Normal Distribution: Example 4

Suppose the time taken to resolve customer complaint follows a normal distribution with the mean value of 250 hours and standard deviation value of 23 hours. What is the probability that a problem resolution will take more than 300 hours?

Given: Y = 300; μ = 250; σ = 23

METHOD 2:

Using the Standard Z formula: $Z = \frac{(Y-\mu)}{\sigma} = \frac{(300-250)}{23} = 2.17$

- The Z value of 2.17 covers an area of 0.98499 under itself.
- The probability that a problem can be resolved in less than 300 hours is 98.5%
- The chances of a problem resolution taking more than 300 hours is **1.5% (1 - 0.985)**

Normal Distribution: Example 4

Suppose the time taken to resolve customer complaint follows a normal distribution with the mean value of 250 hours and standard deviation value of 23 hours. What is the probability that a problem resolution will take more than 300 hours?

Given: Y = 300; μ = 250; σ = 23

METHOD 3:

=NORM.S.DIST(Z score, cumulative distribution?)

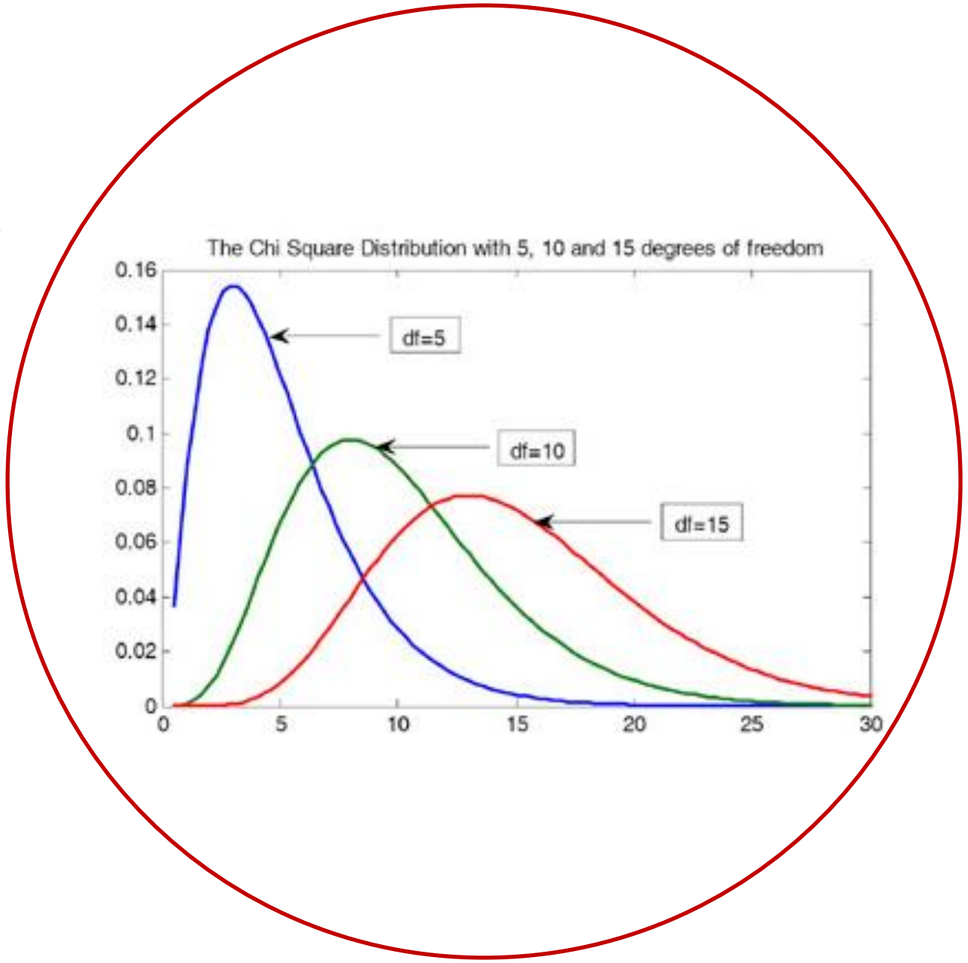
=NORM.S.DIST(2.17, TRUE) = 98.5%

1 - 98.5 = **1.5%**

Chi-Square Distribution

Chi-square distribution (chi-squared or χ^2 distribution) with k-1 degrees of freedom is the distribution of the sum of the squares of k independent standard normal random variables.

Used in inferential statistics



Used in hypothesis test

Degree of freedom is k-1, where k is the sample size

Chi-Square Distribution: Formula

$$\chi^2_{\text{calculated}} = \Sigma = \frac{(f_O - f_e)^2}{f_e}$$

Where,

$\chi^2_{\text{calculated}}$ (Σ) = chi-square index

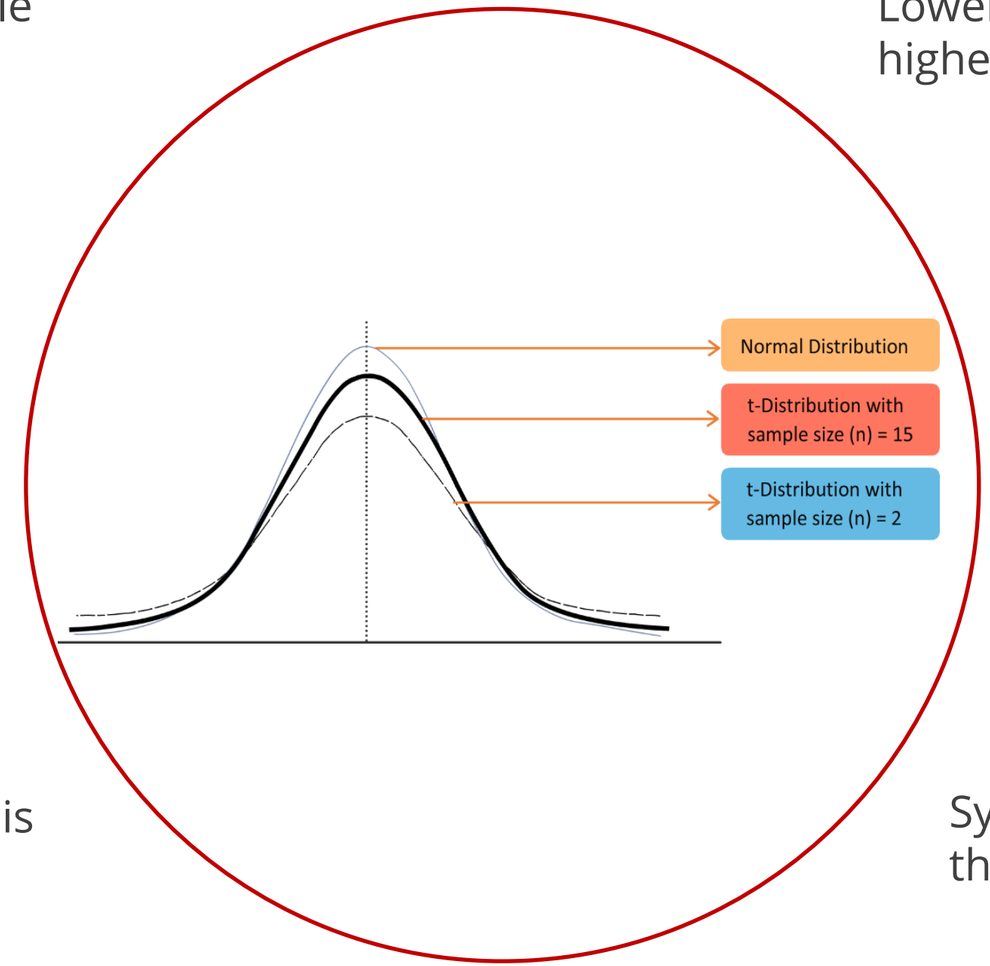
f_O = observed frequency

f_e = expected frequency

T-Distribution

Used when the sample size is less than 30

Lower at the mean and higher at the ends



Used when the population standard deviation is not known

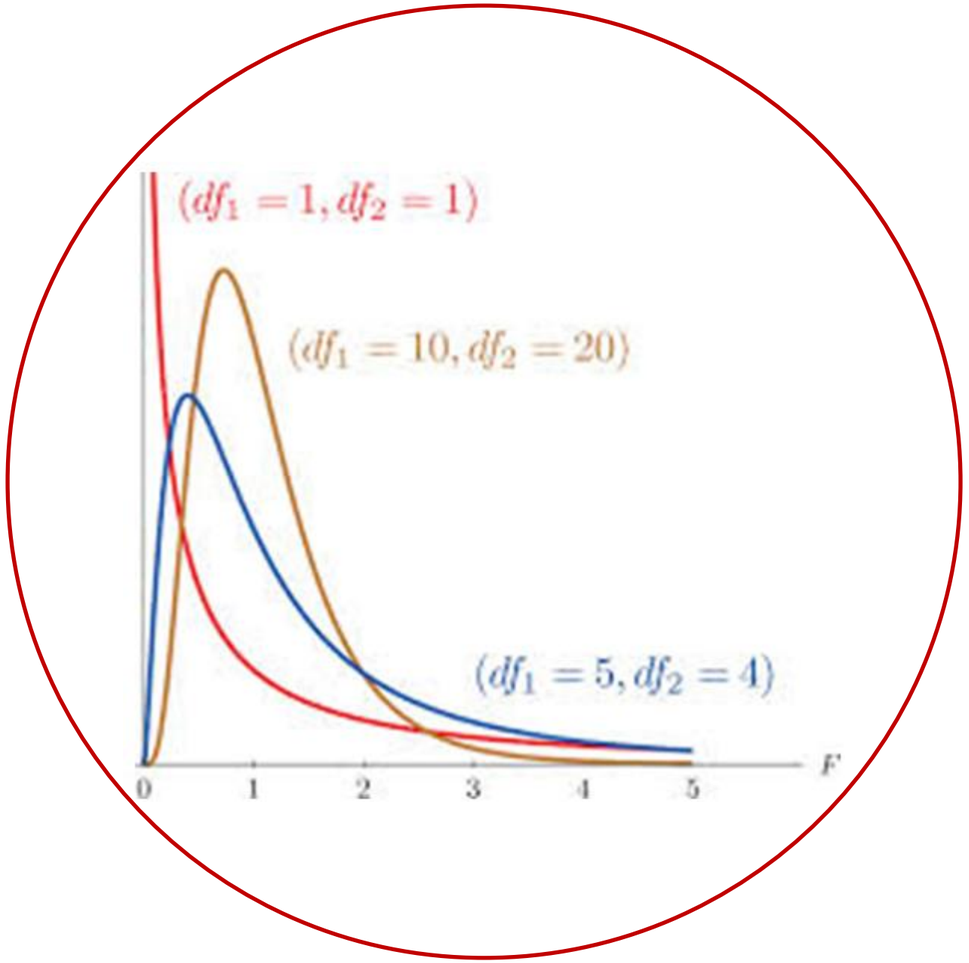
Used for hypothesis testing

Used when the population is approximately normal

Symmetrical in shape but flatter than the normal distribution

F-Distribution

The F-distribution is a ratio of two Chi-square distributions. A specific F-distribution is denoted by the degrees of freedom for the numerator Chi-square and the degrees of freedom for the denominator Chi-square.



- Calculates and observes if the standard deviations or variance for two processes are significantly different

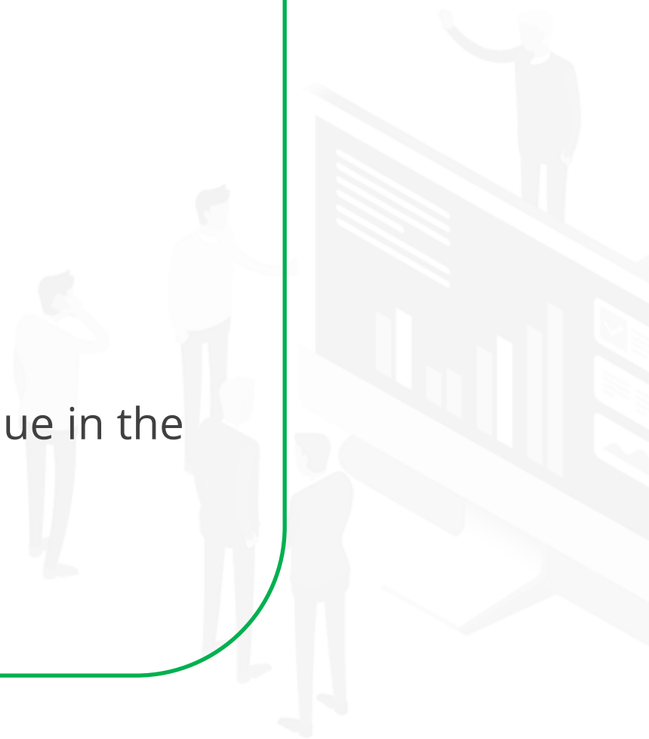
F-Distribution: Formula

$$F_{calculated} = \frac{S_1^2}{S_2^2}$$

Where,

S_1 and S_2 = standard deviations of the two samples

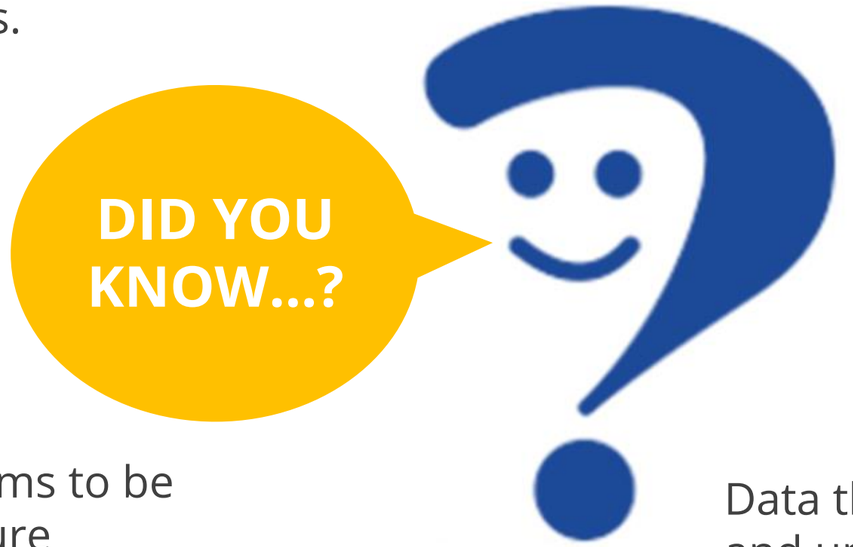
- If $F_{calculated}$ is 1, there is no difference in the variance
- The larger variance should be placed in numerator and the smaller value in the denominator
- $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$)



Fun Facts

There are over 100 distribution models.

Run data through a normality plot to see if it is normally distributed first.



The normal distribution seems to be everywhere from temperature fluctuations, student test scores, and time taken to complete a task. It is the average result of other factors.

Data that are influenced by many small and unrelated random effects are approximately normally distributed.

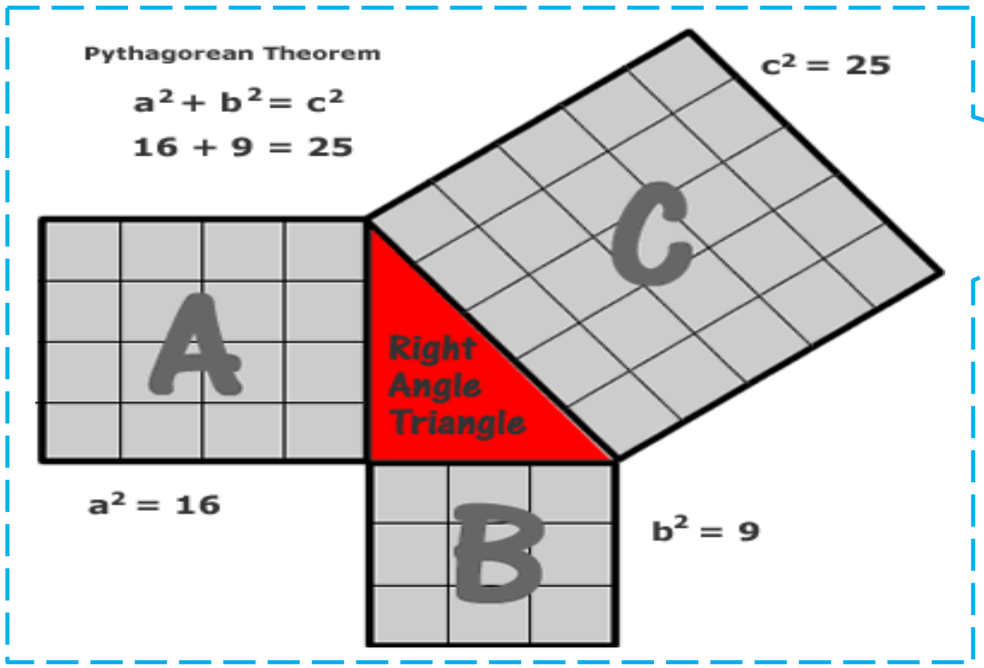
Central Limit Theorem

Meaning of Theorem

A theorem is a general proposition not self-evident but proved by a chain of reasoning.

It is a truth established by means of accepted truths.

Example

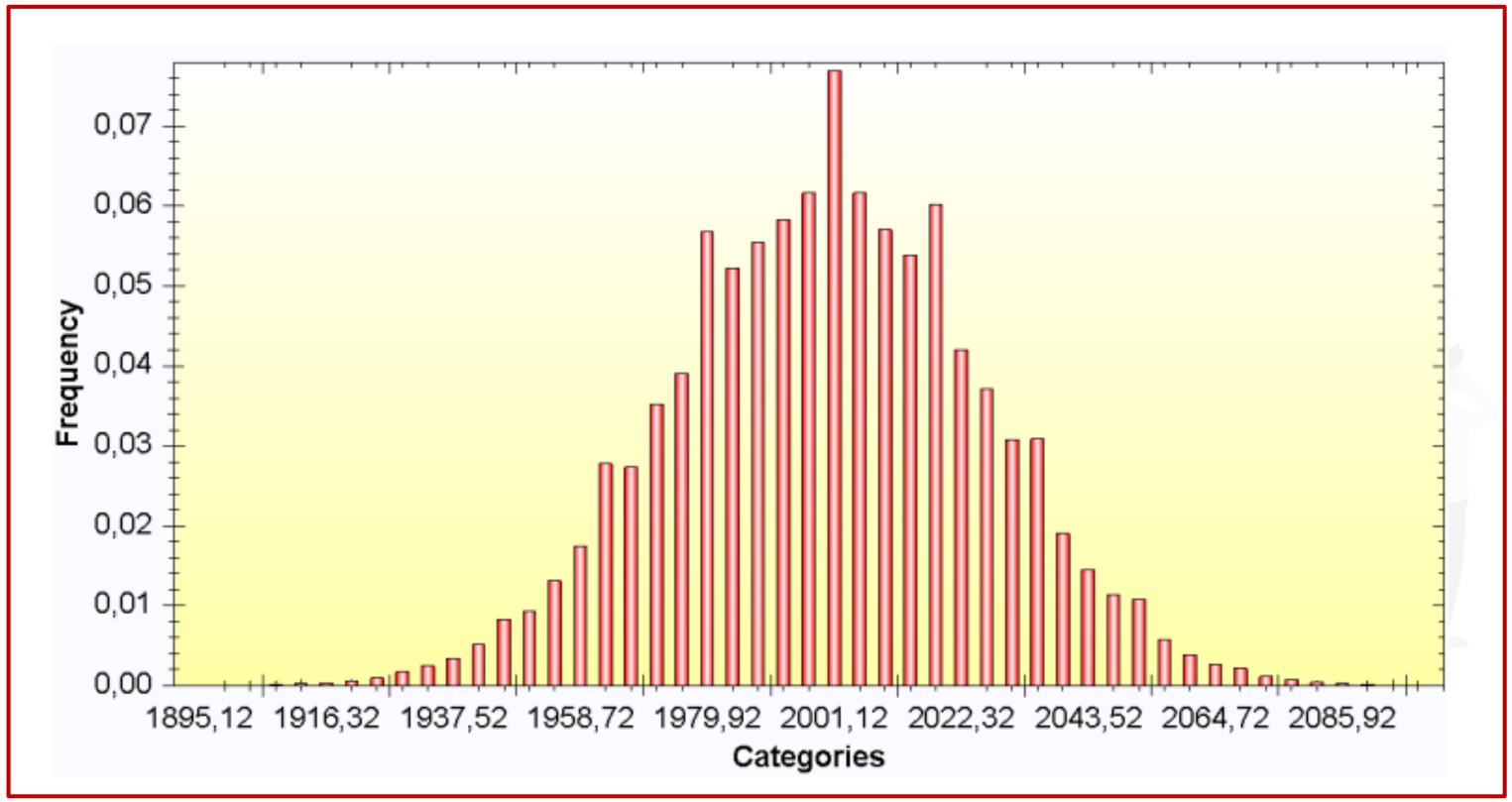


Pythagoras' Theorem

In a right-angled triangle, the square of the long side is equal to the sum of the squares of the other two sides.

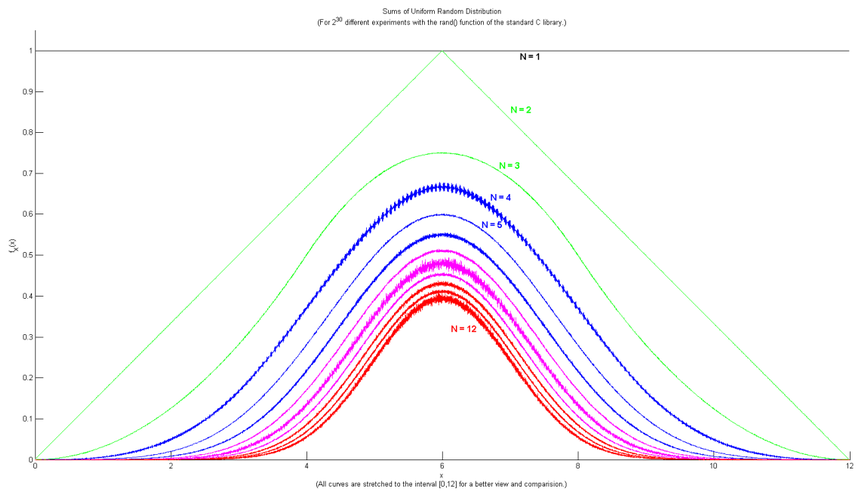
Central Limit Theorem (CLT)

CLT takes any data, with enough samples, and applies normal distribution principles.



Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) states that the means of random samples drawn from any distribution with mean μ and variance σ^2 will have an approximately normal distribution with a mean equal to μ and a variance equal to σ^2 / n, as n increases greater than 30.

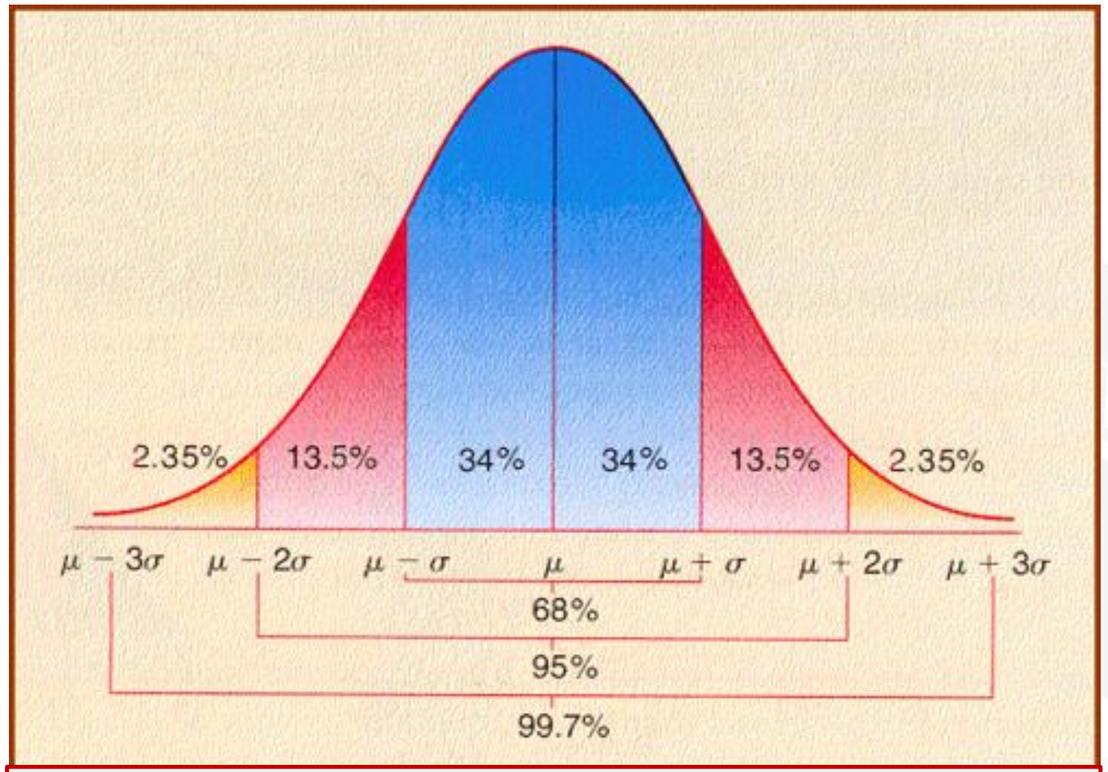


Importance of CLT

CLT implies that the distribution of the sample means will approach a normal distribution regardless of what the population distribution looks like.

CLT makes probability statements about the possible range of values the sample mean may take.

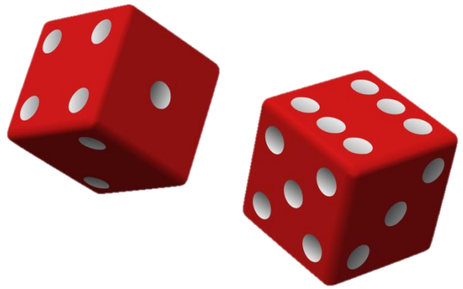
The normal distribution has a useful property called the Empirical Rule.



The Empirical Rule

As a Six Sigma practitioner remember the CLT forms the basis of inferential statistics.

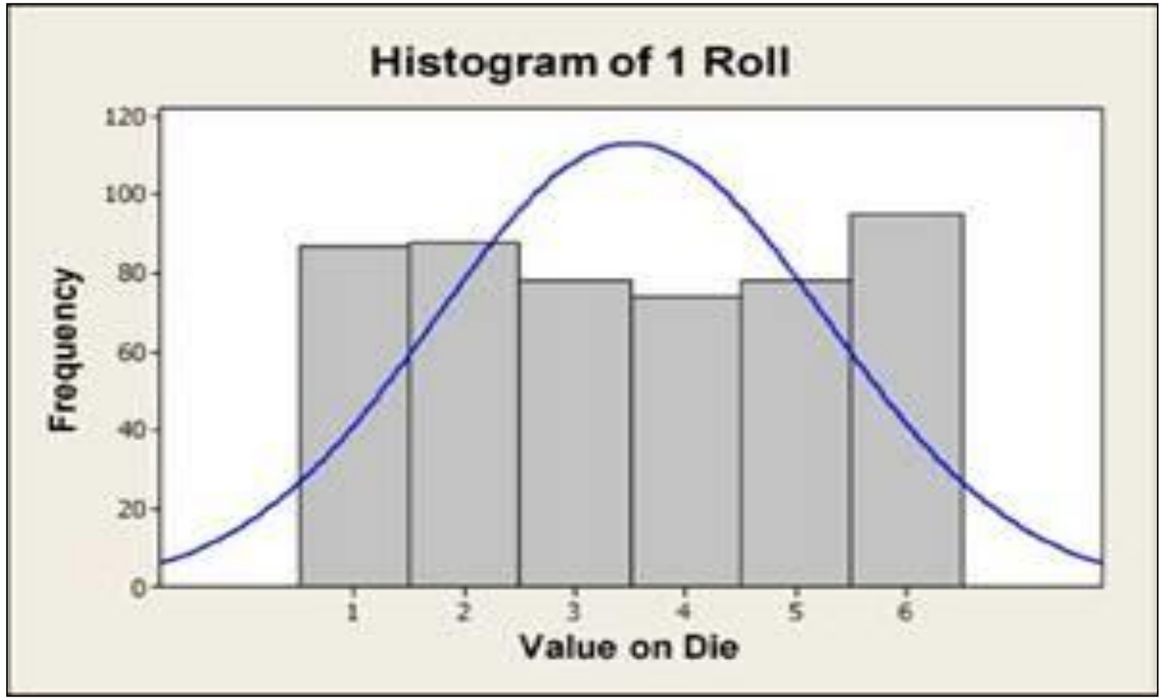
How CLT Works: Illustration



The probability of the dice landing on any one side is equal to the probability of it landing on any of the other five sides.

How CLT Works: Experiment 1

500 times



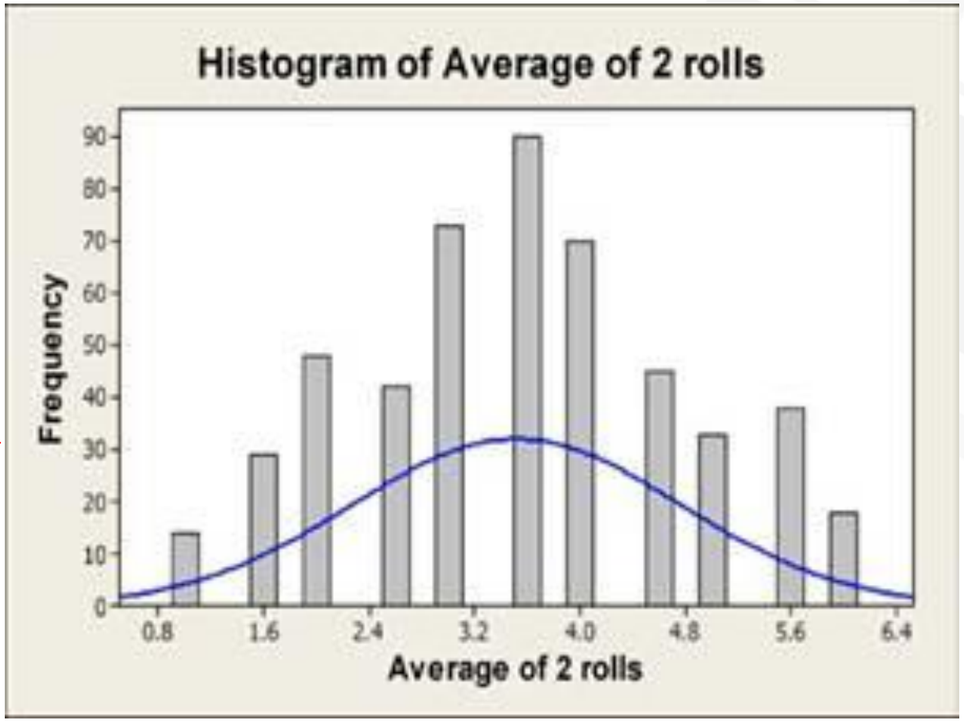
Source: https://www.minitab.com/uploadedFiles/Content/Academic/CentralLimitTheorem.pdf

How CLT Works: Experiment 2

2 rolls

↓	C1	C2	C3
	1st roll	2nd roll	Average of 2 rolls
1	4	4	4.0
2	4	3	3.5
3	2	6	4.0
4	5	5	5.0
5	6	3	4.5
6	4	5	4.5
7	1	4	2.5

500 trials

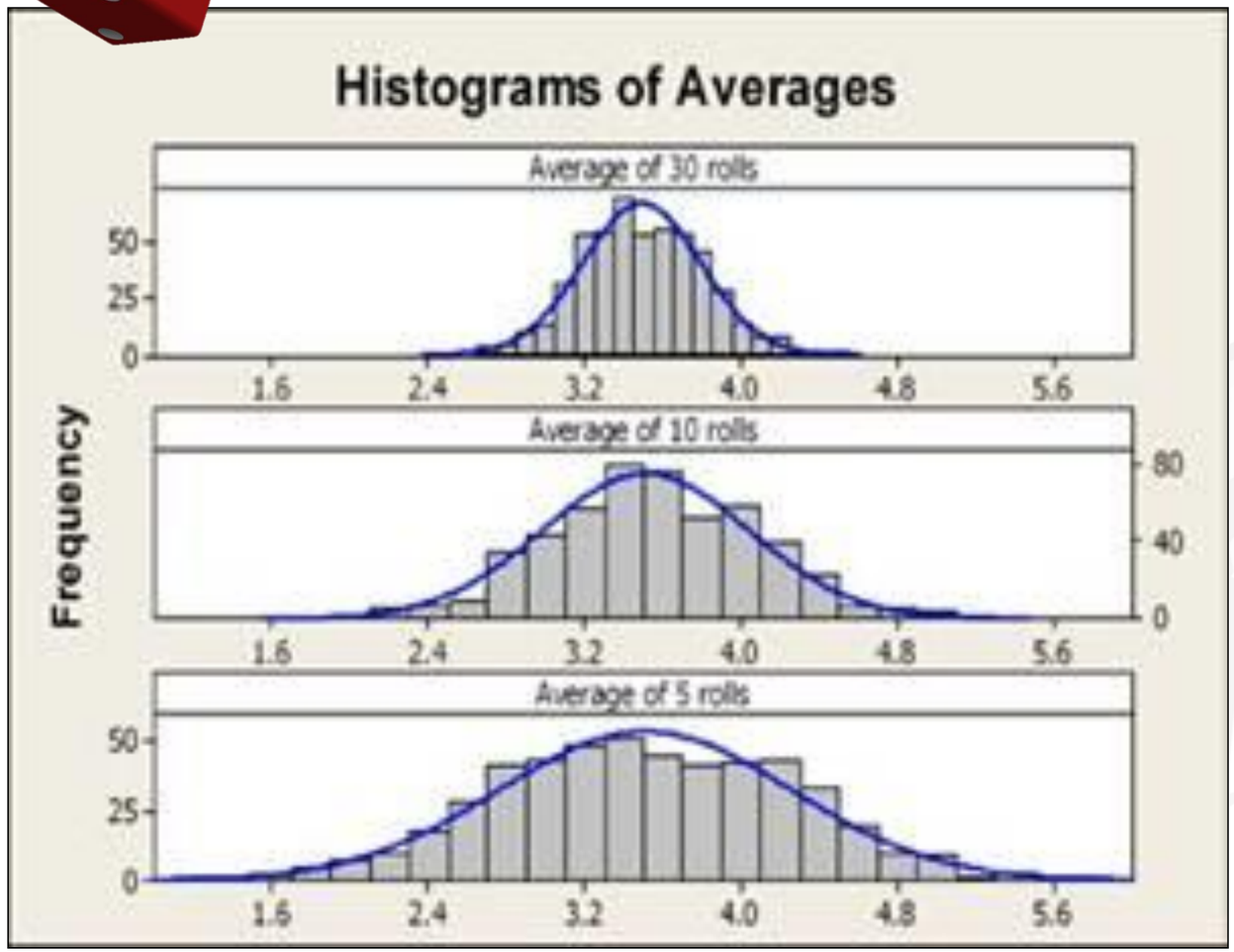


How CLT Works: Experiment 3

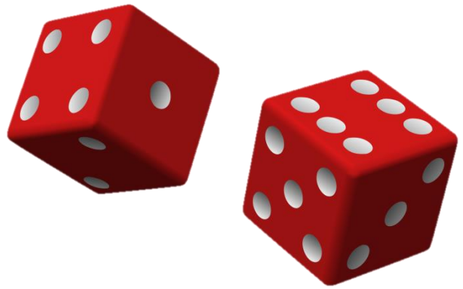
Average of **30** rolls per trial

Average of **10** rolls per trial

Average of **5** rolls per trial



How CLT Works: Experiment 3



The central limit theorem states that for a large enough n, X-bar can be approximated by a normal distribution with mean μ and standard deviation σ/√n.

How CLT Works: Conclusion

The population mean for a six-sided die is:
(1+2+3+4+5+6)/6 = 3.5

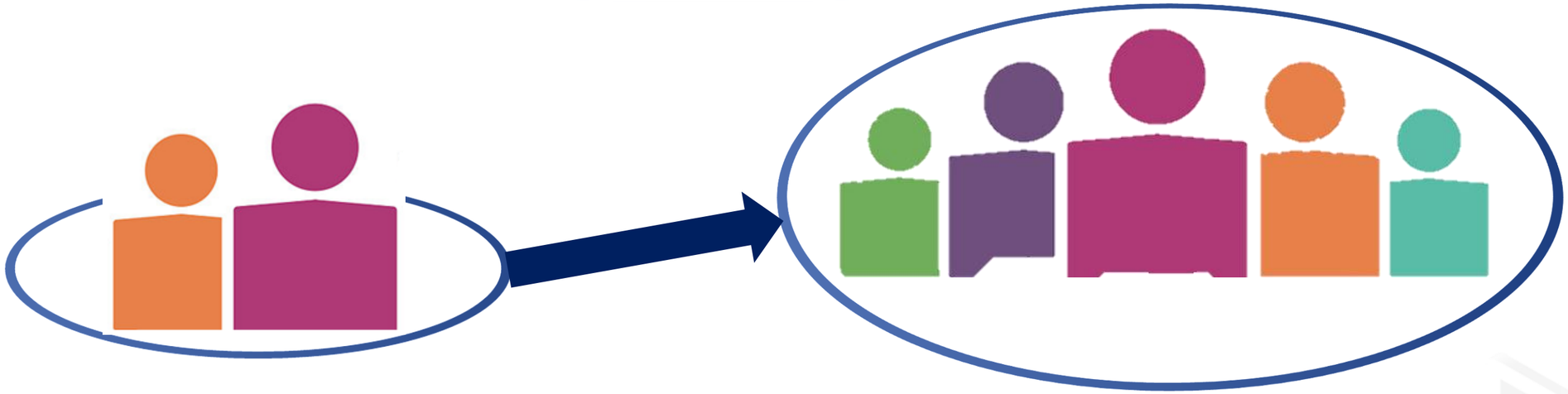
The population standard deviation = 1.708

Standard deviation 1.708/√30 = **0.31** ✓

Average of the 30 averages = 3.49

Standard deviation = **0.30** ✓

Good to Know



The normal model for the sample mean is good when the sample has at least 30 independent observations.

As the sample size increases, the mean has more of a normal distribution.

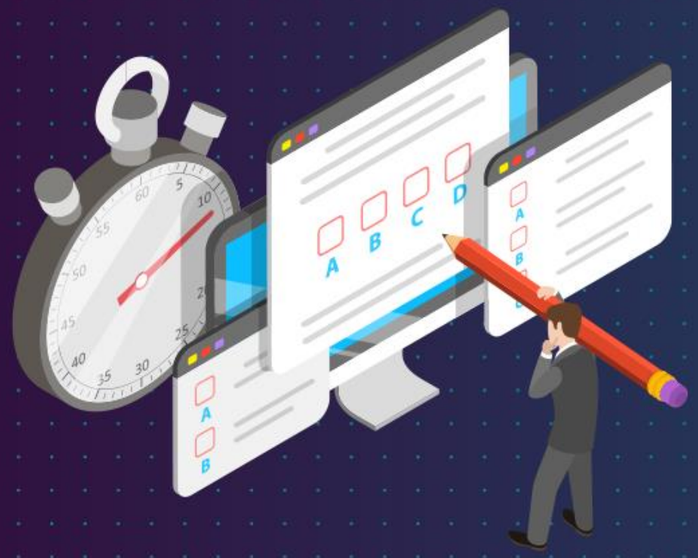
Caution

Observations that are independent may not occur consecutively.

If more outliers are present, it is likely that more than 30 observations will be needed to use the normal distribution.

Key Takeaways

- Probability, statistics, and inferential statistics describe the parameters for the classes of distributions.
- A theorem is a general proposition not self-evident but proved by a chain of reasoning.
- CLT takes any data, with enough samples, and applies normal distribution principles.



Knowledge Check

Knowledge Check

1

What is the similarity between the Binomial and Poisson Distributions?

A. Focus on defective items

B. A part of the discrete distribution family

C. The calculation of the average or expected value

D. The number of trials tend towards infinity

Knowledge Check

1

What is the similarity between the Binomial and Poisson Distributions?

A. Focus on defective items

B. A part of the discrete distribution family

C. The calculation of the average or expected value

D. The number of trials tend towards infinity

The correct answer is **B**

A Binomial is a discrete distribution that focuses on defective items, as a small number of trials and the calculation for the expected value is n*p; whereas the Poisson is a discrete distribution that focuses on defects, the number of trials tends towards infinity, and the expected value is λ.

Knowledge Check

2

What is the probability of P(Z<2.4)?

A. 99.18%

B. 0.81%

C. 95%

D. 5%

Knowledge Check

2

What is the probability of P(Z<2.4)?

A. 99.18%

B. 0.81%

C. 95%

D. 5%

The correct answer is **A**

Looking up the Z value of 2.4 in a left-tailed Z-table gives the probability of 99.18%.

Knowledge Check

3

Which distribution is based on the Bernoulli process to predict sample behavior?

A. Poisson

B. Binomial

C. F-distribution

D. Normal

Knowledge Check

3

Which distribution is based on the Bernoulli process to predict sample behavior?

A. Poisson

B. Binomial

C. F-distribution

D. Normal

The correct answer is **B**

The binomial distribution is based on the scenario where the output has only two options and probability remains consistent over time. This scenario is called the Bernoulli process.

Knowledge Check 4	If the output value is 45, with process average of 40 and standard deviation of 2, what is the Z score value?

A. 5

B. 2

C. 2.5

D. 3

Knowledge Check 4	If the output value is 45, with process average of 40 and standard deviation of 2, what is the Z score value?

A. 5

B. 2

C. 2.5

D. 3

The correct answer is **C**

$$Z = \frac{(Y-\mu)}{\sigma} = \frac{(45-40)}{2} = 2.5$$

Knowledge Check

5

Given a normal distribution, what is the probability of having a Z score value smaller than 2.5?

A. 0.6%

B. 99.4%

C. 90.2%

D. 2.5%

Knowledge Check

5

Given a normal distribution, what is the probability of having a Z score value smaller than 2.5?

A. 0.6%

B. 99.4%

C. 90.2%

D. 2.5%

The correct answer is **B**

: Using a left tailed Z-table or the Excel function "= NORM.S.DIST() will provide a result of 99.4%.

Knowledge Check 6	The central limit theorem states that the means of random samples drawn from any distribution with mean μ and variance σ2 will have an approximately _____ distribution with a mean equal to μ and a variance equal to σ2 / n, as n increases greater than ___.

A. Normal, 10

B. T, 30

C. Normal, 25

D. Normal, 30

Knowledge Check

6

The central limit theorem states that the means of random samples drawn from any distribution with mean μ and variance σ2 will have an approximately _____ distribution with a mean equal to μ and a variance equal to σ2 / n, as n increases greater than ___.

A. Normal, 10

B. T, 30

C. Normal, 25

D. Normal, 30

The correct answer is **D**

The central limit theorem states that the means of random samples drawn from any distribution with mean μ and variance σ2 will have an approximately normal distribution with a mean equal to μ and a variance equal to σ2 / n, as n increases greater than 30.