Notes:

Probability and Statistics

Key Learning Points

1. Describe the importance of using probability distributions.
2. Explain how to evaluate probability from a distribution.
3. Utilize probability in improvement projects.

What is Probability?

The Normal Probability Plot is another way besides the histogram to plot data and look for normality. Normal data, when plotted with the data value on the X-axis and specially spaced percentiles of the normal distribution on the Y-axis, will fall on a straight line.

- Formula: A probability distribution function is a mathematical formula that relates the values of the characteristics with their probability of occurrence in the population.
- Distribution: The collection of these probabilities is called a probability distribution.
- Discrete or Continuous: A probability distribution may be discrete or continuous depending on the possible values of the data.

Probability Distributions

Continuous Probability: Continuous probability distributions model situations where the outcome of interest can take on values in a continuous range.

Discrete Probability: Discrete probability distributions are used to model situations where the outcome of interest can take on only discrete values (such as 0 or 1 for failure or success, or 0,1,2,3…as a number of occurrences of some event of interest).

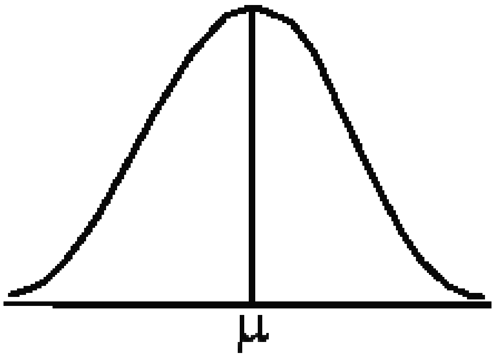
Continuous Distributions

The Normal Distribution

The normal distribution is applicable when there is a concentration of observations about the average and it is equally likely that observations will occur above and below the average. Variation in observations is usually the result of many small causes.

μ = Mean
σ = Standard Deviation

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

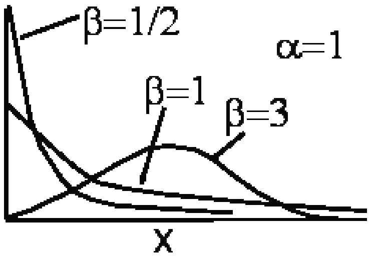


The Weibull Distribution

The normal distribution is applicable when describing a wide variety of patterns of variation, including departures from the normal and exponential.

$\gamma=\alpha\beta(X-\gamma)^{\beta-1}e^{-\alpha(x-\gamma)^{\beta}}$

α = Scale Parameter
β = Shape Parameter
γ = Location Parameter

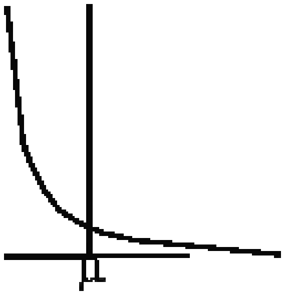


Notes:

The Exponential Distribution

The normal distribution is applicable when there is a concentration of observations about the average and it is equally likely that observations will occur above and below the average. Variation in observations is usually the result of many small causes.

$$y = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$



Discrete Distributions

The Binomial Distribution

The binomial distribution is applicable in defining the probability of r occurrences in n trials of an event which has probability of occurrence of p on each trial.

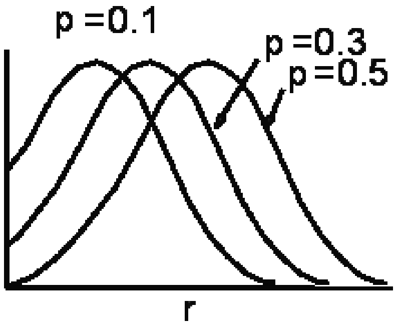
$$y = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$n =$ Number of trials

$r =$ Number of occurences

$p =$ Probability of occurence

$q = 1 - p$



The binomial distribution is discrete, but it is shown as a continuous curve for ease of comparison.

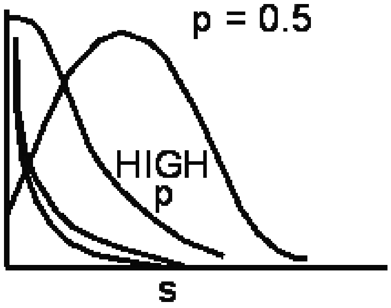
Notes:

The Negative Binomial Distribution

The negative binomial distribution is applicable in defining the probability that r occurrences will require a total of r+s trials of an event which has a probability of occurrence of p on each trial. Note that the total number of trials n is r+s.

$$y = \frac{(r+s-1)!}{(r-1)!(s!)} p^r q^s$$

r = Number of Occurrences
s = Difference Between the Number of Trials and the Number of Occurences
p = Probability of Occurences
q = 1-p

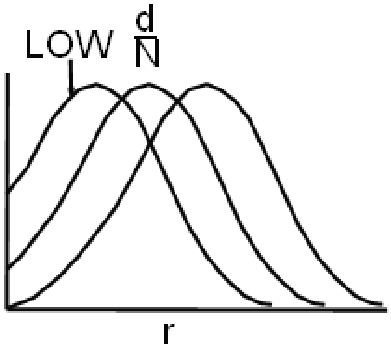


The negative binomial distribution is discrete, but it is shown as a continuous curve for ease of comparison.

The Hypergeometric Distribution

The hypergeometric distribution is applicable in defining the probability of r occurrences in n trials of an event when there are a total of d occurrences in a population of N.

$$Y = \frac{\left(\frac{d}{r}\right)\left(\frac{N-d}{n-r}\right)}{\left(\frac{N}{n}\right)}$$



The hypergeometric distribution is discrete, but it is shown as a continuous curve for ease of comparison.

Notes:

The Poisson Distribution

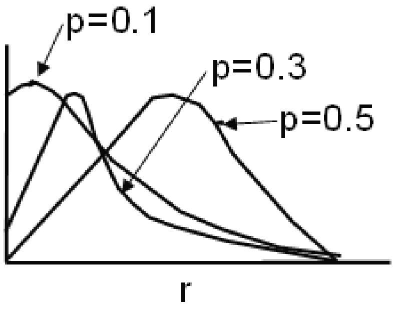
The poisson distribution is the same as the binomial distribution, but is particularly applicable when there are many opportunities for occurrence of an event, but a low probability (less than 0.10) on each trial.

$$y = \frac{(np)^r e^{-np}}{r!}$$

n = Number of trials

r = Number of occurrences

p = Probability of occurrences



The poisson distribution is discrete, but it is shown as a continuous curve for ease of comparison.

Formulas

N! is called "N factorial"
$N!=N(N-1)(N-2)\ldots1$
$0!=1$
The expression $\binom{N}{n}$ is the combination of N items taken n at a time.

Its value is: $\frac{N!}{[n!(N-n)!]}$

Identify the Distribution

Construct a histogram of each of the data sets and compare them with a normal curve.

- Minitab: Graph > Histogram
  - Select "With Fit"
  - Select Items to Plot

The plot shows the actual distribution with a histogram and a normal distribution line graph with the same mean and standard deviation.

Notes:

The Central Limit Theorem

Notes:

In probability theory, the Central Limit Theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (or bell curve) even if the original variables themselves are not normally distributed.

CLT Example: Normal Data

This exercise generates random data, so each run of this exercise will differ.

Step 1: Create Simulated Data

Use the following commands to create 9 columns of numbers from a normal distribution in Minitab.

- Minitab: Calc > Random Data > Normal
  - Generate 250 Rows
  - Store in C1-C9
  - Mean = 70
  - Standard Deviation = 9

Step 2: Calculate Row Statistics

Use the following commands to create “Row Statistics” in column 10.

- Minitab: Calc > Row Statistics
  - Select Columns C1-C9
  - Select Mean
  - Store Result in C10

Step 3: Calculate Statistics on All Columns

Use the following commands to see descriptive statistics.

- Minitab: Stat > Display Descriptive Statistics
  - Select All 10 Columns

Step 4: Display Data Graphically

Use the following commands to produce a Dot Plot.

- Minitab: Graph > Dotplot >Multiple Ys
  - Select Columns C1 and C10

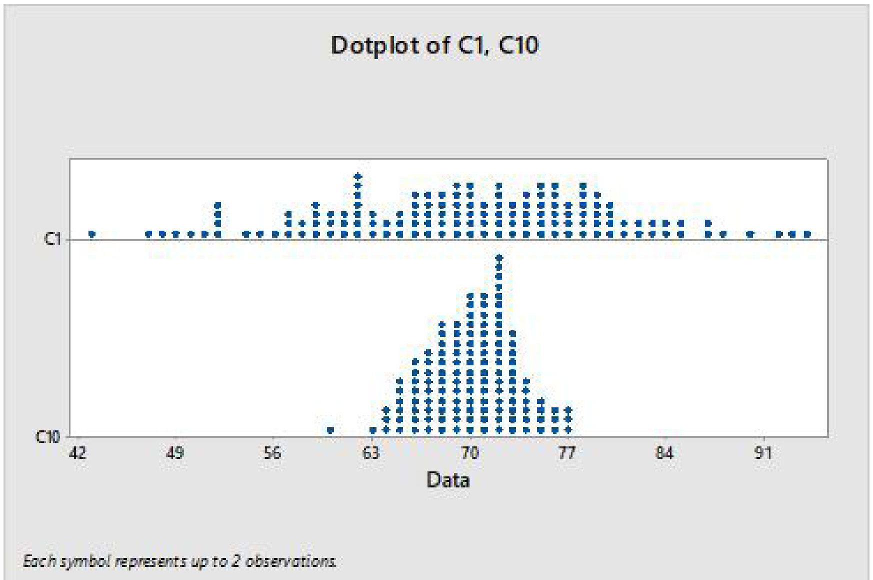
Step 5: Interpret Data

Notice the drastic reduction in variance!

Notes:

The top dotplot is the distribution of individual observations. Variation is quantified as σx, standard deviation of the data.

The bottom dotplot is the distribution of simple means. Variation is quantified as σxbar, standard deviation of the data.



CLT Example: Non-Normal Data

This exercise generates random data, so each run of this exercise will differ.

Step 1: Create Simulated Data

Use the following commands to create 9 columns of numbers from a normal distribution in Minitab.

- Minitab: Calc > Random Data > Chi-Square
  - Generate 250 Rows
  - Store in C1-C9
  - Degrees of Freedom = 2

Step 2: Calculate Row Statistics

Use the following commands to create “Row Statistics” in column 10.

- Minitab: Calc > Row Statistics
  - Select Columns C1-C9
  - Select Mean
  - Store Result in C10

Step 3: Calculate Statistics on All Columns

Use the following commands to see descriptive statistics.

- Minitab: Stat > Display Descriptive Statistics
    - Select All 10 Columns

Step 4: Display Data Graphically

Use the following commands to produce a Dot Plot.

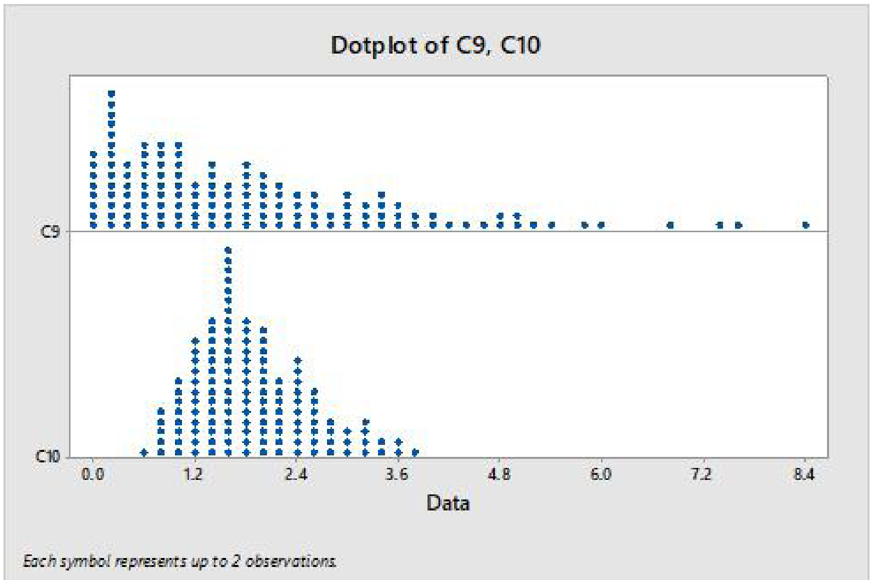
- Minitab: Graph > Dotplot >Multiple Ys
    - Select Columns C1 and C10

Step 5: Interpret Data

Notice the distribution is much more bell-shaped.

The top distribution of individual observations is heavily skewed to the right.

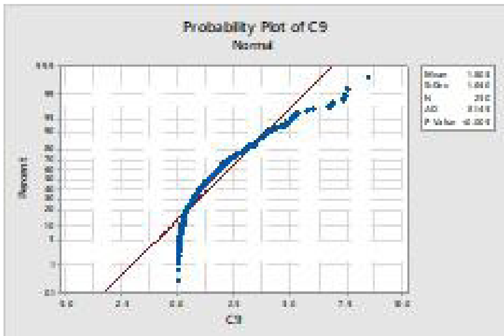
The bottom distribution of sample means is more Normally distributed.



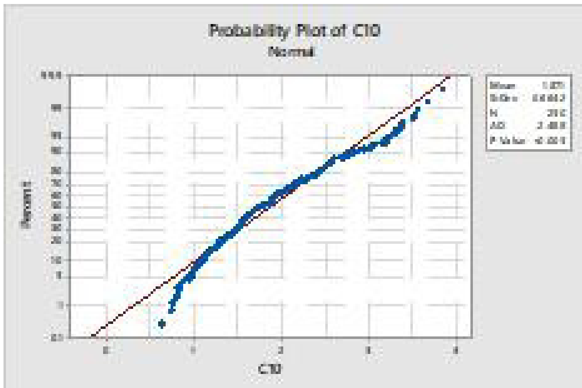
Are the Distributions Normal?

To test the normality of the distribution for columns C9 and C10.

- Minitab: Stat > Basic Statistics > Normality Test
    - Run the test for Column C9, and C10



Notes:


Probability Plot of C10
Normal
Percent
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
0
1
2
3
4
C10
Mean
StDev
N
AD
P-Value

Notes: