BASICS OF STATISTICS

OVERVIEW

Statistics is concerned with the study and interpretation of data samples to make informed business decisions. A typical application of statistical methods is in quality control of a production run, often referred to as statistical quality control. Statistical methods are a mainstay in quality assurance and control enabling engineers to manufacture a consistent product, detect problems and predict product life.

Statistics also finds applications in marketing, for example, to study the effect of advertising, for conducting sample surveys to gather customer feedback, perform field trials in test markets to assess product feasibility and marketability. Other applications include Government, Health and Medicine, natural resource industry, etc. Irrespective of the nature of the application, Statistics involves evaluation of the data source and sampling prior to designing the statistical tests for obtaining quantitative and qualitative insights to aid decision making.

EXAMPLE - 1: Acceptance Sampling

A tool bit producer supplies bulk packages of 1000 tool bits in a box. A consumer conducts a random sample test before accepting the package. What could be a possible acceptance sampling plan?

Solution: The consumer could sample 20 tool bits at random from the package and decide to accept the lot only if the sample contains no more than one defective item. This means that on average, there would be no more than 5% of defective pieces. Such acceptance sampling plans must be agreed *a priori* by the producer and consumer.

DATA ANALYSIS

Numerical Data

These are considered statistical data values of a variable quantified in terms of measurements. They can be either discrete or continuous. Discrete numerical data values are integers such as the number of students in a class, the number of workers in a company, etc. Continuous numerical data values are real numbers such as money exchange rates, the tensile strength of steel etc.

Histograms

Histograms are extremely useful in statistical data analysis for displaying the relative distribution of a numerical data set. They are essentially a plot of the frequency distribution of the data over pre-defined data ranges/ bins.

EXAMPLE - 2: Histogram

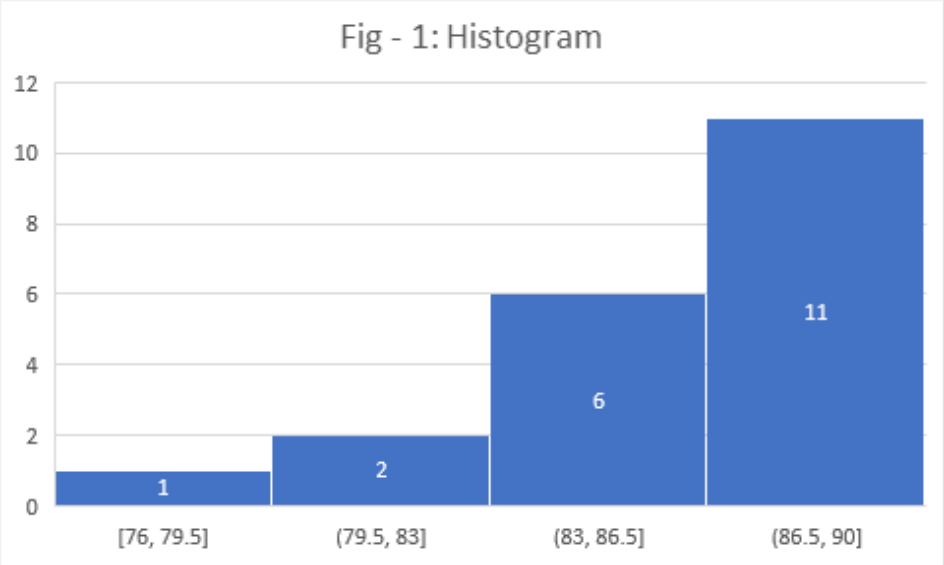
Plot the histogram for the following data:

Carbon content [%] of coal

89	90	89	84	80	88	90	89	88	90
85	87	86	82	85	76	89	87	86	86

Solution

We simply let the automatic creation of bins based on the minimum and maximum values. The resulting histogram is:



As can be seen from the histogram over 50% of data values are over 86.5% Carbon content.

Centre and Spread of data – Median and Quartiles

The median value is the centre value with 50% of the data samples on either side. For even number of data samples, the median is defined as the average of the two centre values.

Range of data values = max value – min value. Splitting into quartiles (25% data values), the data spread is split by min value, 1st quartile value, median value, 3rd quartile value and max value. 50% of data values around the median is between the 1st and 3rd quartile values. This is defined as the interquartile range (IQR), i.e., IQR = 3rd quartile value – 1st quartile value.

Outliers are considered as data that is significantly different from the rest. A particular definition of outliers is data that lies beyond 1.5 IQR from the edge of the quartiles on either side.

Box Plots

Box Plots help graphically depict the centre and spread of data for multiple data sets in one plot for easy visualization, including the quartiles.

EXAMPLE - 3: Box Plot

Consider the following two data sets comprising of measurement of the tensile strength of sheet steel in kg/mm^2

Tensile strength of Steel (Data Set 1)

89	84	87	81	89	86	91
90	78	89	87	83	89	

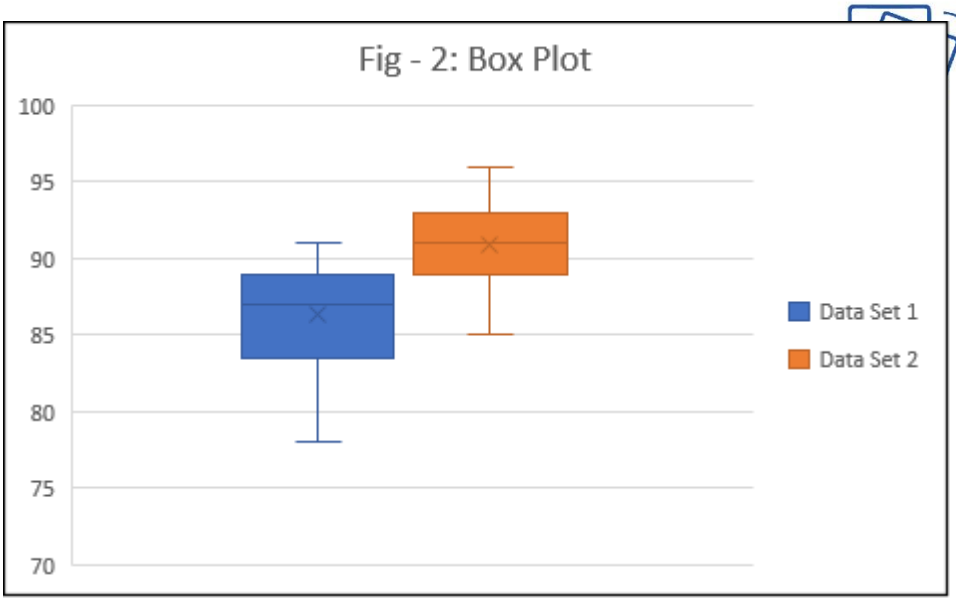
Tensile strength of Steel (Data Set 2)

91	89	93	91	87	94	92
85	91	90	96	93	89	

Determine the min, max, and quartile values and graph box plots for the two data sets. Does the min and max values lie within outlier definition of 1.5 IQR?

Solution

Tensile Strength of Steel (Data Set 1, sorted)	78	81	83	84	86	87	87
	89	89	89	89	90	91	
Tensile Strength of Steel (Data Set 2, sorted)	85	87	89	89	90	91	91
	91	92	93	93	94	96	
	Min.	Q_1	Median	Q_3	Max.		
Tensile Strength of Steel (Data Set 1, sorted)	78	83.5	87	89	91		
Tensile Strength of Steel (Data Set 2, sorted)	85	89	91	93	96		



The box plot as depicted shows the full range of data spread including the quartiles. Formal box plots, however, show only the quartile portion of the data spread. The box is then extended with single lines on either side representing data spread to outliers.
For Data Set 1, 1.5 * IQR = 1.5 x (89 – 83.5) = 8.25 Outlier limits are 83.5 – 8.25 = 75.25 & 89 + 8.25 = 97.25.
The min value 78 and max value 91 lie within this range.

For Data Set 2. IQR = 1.5 x (93 - 89) = 6 Outlier limits are 89 - 6 = 83 & 93 + 6 = 99.
The min value 85 and max value 96 lie within this range.

Mean, Standard Deviation & Variance

While median and quartiles help describe data spread, mean and standard deviation help quantify the data spread in terms of an average value (mean) and variability of the data around the mean value (standard deviation)

The mean is defined as:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}\,(x_1 + x_2 + \cdots + x_n)$$

……. eq (1)

The standard deviation s or s^2, the variance, is given by:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\,[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

……. eq (2)

EXAMPLE - 4: Mean and Standard Deviation

Calculate the mean and standard deviation for Carbon content [%] of coal in Example - 2

Solution

From eq (1), mean = $\frac{1}{20}(89 + 90 + \cdots + 86) = \frac{1726}{20} = 86.3$

From eq (2), variance $s^2 = \frac{1}{19}[(89-86.3)^2 + (90-86.3)^2 + \cdots + (86-86.3)^2 = \frac{250.2}{19} = 12.51$

Standard Deviation = $\sqrt{250.2/19} = 3.536$

NORMAL DISTRIBUTION

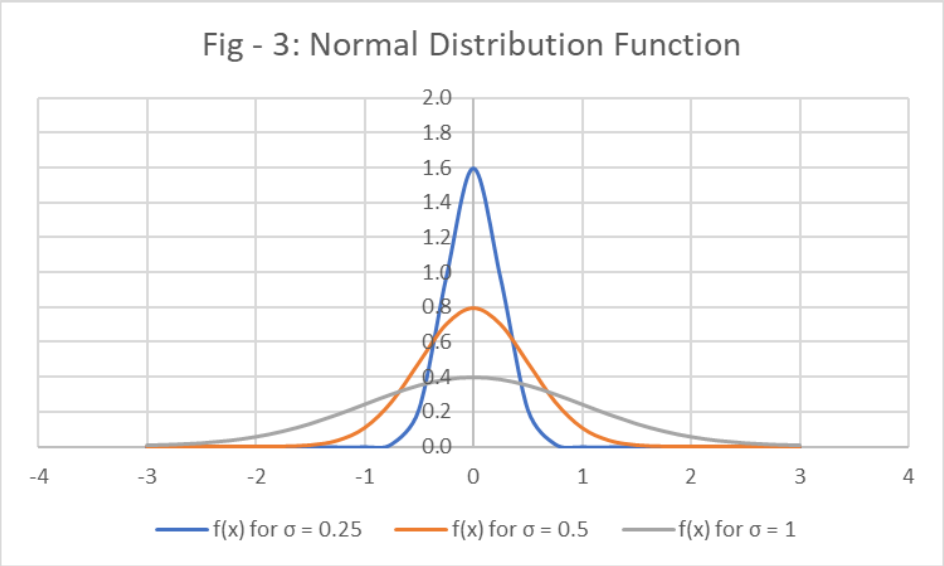
As was seen from the example 4, a data variable measurement tends to be distributed around a mean value. This distribution often tends to be symmetric and follows a bell-shaped pattern on either side of the mean. The bell curve is often well approximated by a normal distribution or gauss distribution with the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

……. eq (3)

where exp is the exponential function with base e= 2.718 … The quantity $\frac{x-\mu}{\sigma}$ is referred to as the z-score of the distribution function. The quantity $\frac{1}{\sigma\sqrt{2\pi}}$ is a factor that makes the total area under the probability density function equal to 1.

Fig-3 shows the Normal Distribution Function for μ = 0 and σ = 0.25, 0.5 & 1.



x	f(x) for σ = 0.25	f(x) for σ = 0.5	f(x) for σ = 1
-3	8.58E-32	1.22E-08	0.00443
-2.75	8.47E-27	2.15E-07	0.00909
-2.5	3.08E-22	2.9731E-06	0.01753
-2.25	4.11E-18	3.19641E-05	0.03174
-2	2.02E-14	0.000267632	5.40E-02
-1.75	3.65E-11	1.75E-03	0.08627
-1.5	2.43E-08	0.008862757	0.1295
-1.25	5.9462E-06	0.035052886	0.18263
-1	0.000535264	0.107970489	0.24195
-0.75	0.017725515	0.259007739	0.30111
-0.5	0.215940978	0.48389016	0.35203
-0.25	0.967780321	0.704056029	0.38663
0	1.5956	0.7978	0.3989
0.25	0.967780321	0.704056029	0.38663
0.5	0.215940978	0.48389016	0.35203
0.75	0.017725515	0.259007739	0.30111
1	0.000535264	0.107970489	0.24195
1.25	5.9462E-06	0.035052886	0.18263
1.5	2.43E-08	0.008862757	0.1295
1.75	3.65E-11	1.75E-03	0.08627
2	2.02E-14	0.000267632	5.40E-02
2.25	4.11E-18	3.19641E-05	0.03174
2.5	3.08E-22	2.9731E-06	0.01753
2.75	8.47E-27	2.15E-07	0.00909
3	8.58E-32	1.22E-08	0.00443

The probability distribution function F(x) for the normal distribution can then be written as:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left[-\frac{1}{2}\left(\frac{x' - \mu}{\sigma}\right)^2\right] dx'$$

……. eq (4)

This is the probability the random variable will take a value between -∞ and x for a normal probability distribution function.

The standardized normal probability distribution function is obtained by setting mean μ = 0 and standard deviation σ = 1 in eq (4) and denoted by Φ(z), i.e.,

$$\Phi(z) = \int_{-\infty}^{x} e^{-u^2/2}\, du$$

……. eq (5)

By using substitution of variables z = (x-μ)/σ, it also follows that

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

……. eq (6)

In terms of actual probabilities, the following three equations follow from eq (5) & eq (6)

$$P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

……. eq (7)

$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

……. eq (8)

$$P(X \geq b) = 1 - P(X \leq b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

……. eq (9)

spanning the entire range of values of random variable X from -∞ to ∞. All probability calculations involving the normal probability distribution function for a random variable X is done using eq (7), (8) & (9) with the help of tabulated values of the integral Φ(z) in eq (5) from published tables as given in Table A7 & A8 in Appendix.

EXAMPLE -5: Three-sigma limits

Prove the following for the normal distribution using the tables given in Appendix

$$(a)\ P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\%$$
$$(b)\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.5\%$$
$$(c)\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.7\%$$

Illustrate the solution for case (b) graphically

Solution

Case (a) – This is the probability that X will lie between $\mu \pm \sigma$

Setting b = μ + σ and a = μ - σ in eq (8) and using Table A7, we get

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \Phi(1) - \Phi(-1) = 2 \times 0.8413 - 1 = 0.6826 \approx 68\%$$

Case (b) – This is the probability that X will lie between $\mu \pm 2\sigma$

Setting b = μ + 2σ and a = μ - 2σ in eq (8) and using Table A7, we get

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \Phi(2) - \Phi(-2) = 2 \times 0.9772 - 1 = 0.9544 \approx 95\%$$

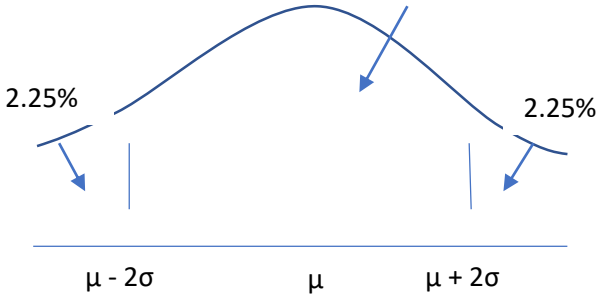
Case (c) – This is the probability that X will lie between $\mu \pm 3\sigma$

Setting b = μ + 3σ and a = μ - 3σ in eq (8) and using Table A7, we get

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 2 \times 0.9987 - 1 = 0.9974 \approx 99\%$$

Case (b) Graphic Illustration

95.5%



EXAMPLE -6: 95 to 99.9% Probabilities

Prove the following for the normal distribution using the tables given in Appendix

$$(a)\ P(\mu - 1.96\sigma \le X \le \mu + 1.96\sigma) = 95\%$$
$$(b)\ P(\mu - 2.58\sigma \le X \le \mu + 2.58\sigma) = 99\%$$
$$(c)\ P(\mu - 3.29\sigma \le X \le \mu + 3.29\sigma) = 99.9\%$$

Solution

Case (a) – This is the probability that X will lie between $\mu \pm 1.96\sigma$

Setting b = μ + 1.96σ and a = μ - 1.96σ in eq (8) and using Table A7, we get

$$P(\mu - 1.96\sigma \le X \le \mu + 1.96\sigma) = \Phi(1.96) - \Phi(-1.96) = 2 \times 0.9750 - 1 = 0.95 = 95\%$$

Case (b) – This is the probability that X will lie between $\mu \pm 2.58\sigma$

Setting b = μ + 2.58σ and a = μ - 2.58σ in eq (8) and using Table A7, we get

$$P(\mu - 2.58\sigma \le X \le \mu + 2.58\sigma) = \Phi(2.58) - \Phi(-2.58) = 2 \times 0.9951 - 1 = 0.9902 = 99\%$$

Case (c) – This is the probability that X will lie between $\mu \pm 3.29\sigma$

Setting b = μ + 3.29σ and a = μ - 3.29σ in eq (8) and using extended Table A7, we get

$$P(\mu - 3.29\sigma \le X \le \mu + 3.29\sigma) = \Phi(3.29) - \Phi(-3.29) = 2 \times 0.9995 - 1 = 0.999 = 99.9\%$$

EXAMPLE -6: Defects

In a production of iron rods let the diameter X be normally distributed with mean 2 in. and standard deviation 0.008 in.

(a) What percentage of defective pieces can we expect if we set the tolerance limits at 2 ± 0.02 in.?
(b) How should we set the tolerance limits to allow for 4% defectives?

Solution

Using Eq.8 in conjunction with Table 7 as before, it is found:

(a) $P(1.98 \le X \le 2.02) = \Phi\left(\frac{2.02-2}{0.008}\right) - \Phi\left(\frac{1.98-2}{0.008}\right) = \Phi(2.5) - \Phi(-2.5) = 2 \times 0.9938 - 1 = 0.9876 \approx 98\%,$

So, the tolerance covers 98% of the acceptable iron rods. So, there are still 2% defective pieces to be expected.

(b) $0.96 = P(2 - c \le X \le 2 + c) = \Phi\left(\frac{2+c-2}{0.008}\right) - \Phi\left(\frac{2-c-2}{0.008}\right) = \Phi\left(\frac{c}{0.008}\right) - \Phi\left(-\frac{c}{0.008}\right) = 2 \times \Phi\left(\frac{c}{0.008}\right) - 1; Therefore, \Phi\left(\frac{c}{0.008}\right) = 0.98,$

$\frac{c}{0.008} = 2.06, c \approx 0.0164$. So, we should set a tolerance level of 0.0164 inches.

SAMPLING AND CONFIDENCE INTERVALS

Study of sample data characteristics and their distribution is a key step towards understanding observable reality. Statistical inferences are basically an extrapolation of the results of the sample data analysis towards understanding a larger set of data values comprising a population.

Statistical inferences are based on "randomness" of the sampling process and fair sample sizes. In other words, larger the sample size, greater is the accuracy of typical inferences such as mean of the population. The key metrics of statistical inferences from a given sample is the ability to estimate confidence intervals relating to the population. This sample metrics forms the basis for hypothesis testing of statements about the characteristics of a population.

Confidence Interval for μ of the Normal Distribution of a random variable with known σ^2

Let a sample data set comprise of n variable values $X_1, \ldots, X_n$. The sample mean is obtained as:

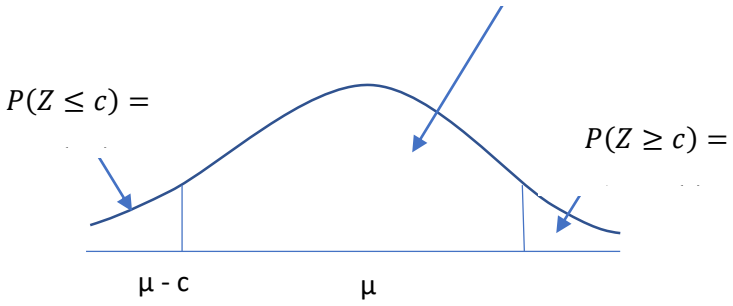
$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

The following random variable $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has a normal distribution with mean 0 and variance 1.

Refer to figure and table below for the normal distribution:

$$P(-c \leq Z \leq c) = \Phi(c) - \Phi(-c)$$

$$2\,\Phi(c) - 1 = \gamma$$



γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

It then follows:

$$P(\bar{X} - k \leq \mu \leq \bar{X} + k) = \gamma; k = \sigma c/\sqrt{n}$$

……. eq (10)

In other words, the confidence interval is the sample mean $\bar{X}$ lies within +/- k from the population mean μ with probability γ (also called significance level 90%, 95% etc.,).

EXAMPLE -7: Confidence Interval (known σ)

Determine a 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of 50 values with mean $\bar{x} = 5$. What is the confidence interval if the sample size is increased to 100?

Solution

From the table and eq (10), for a sample size of 50, $k = 3 \times \frac{1.960}{\sqrt{50}} = 0.8316; confidence\ interval\ is\ 5 - 0.8316 \leq \mu \leq 5 + 0.8316, i.e., 4.168 \leq \mu \leq 5.8316$

If the sample size is increased to 100, it is found , $k = 3 \times \frac{1.960}{\sqrt{100}} = 0.588; confidence\ interval\ is\ 5 - 0.588 \leq \mu \leq 5 + 0.588, i.e., 4.412 \leq \mu \leq 5.588$

The logical statistical inference is that with an increase in sample size, the sample mean is much closer to the population mean resulting in a narrower confidence interval.

Confidence Interval for μ of the Normal Distribution of a random variable with unknown σ^2

The previous case of sampling involved a normal distribution with known σ^2. Frequently, in practice, σ^2 is unknown. In such cases, the Student's t-Distribution function is used as illustrated below.

Let a sample data set comprise of n variable values $X_1, \ldots, X_n$. The sample mean and standard deviation is obtained as:

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

The following random variable $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ has a student's t-distribution with n-1 degrees of freedom.

EXAMPLE -7: Confidence Interval (unknown σ)

Using the following 10 sample values of Copper content (%) of brass (66, 66, 65,64, 66, 67, 64, 65, 63, 64), estimate the 99% confidence interval for the mean

Solution

Sample mean = $\frac{1}{10}(66 + 66 + \cdots + 64) = \frac{650}{10} = 65$

Sample variance $S^2 = \frac{1}{10-1}[(66-65)^2 + (66-65)^2 + \cdots + (64-65)^2 = \frac{14}{9} = 1.556$

Sample Standard Deviation S = $\sqrt{1.556} = 1.247$

Chosen confidence level γ = 99%

Consider the equation $F(c) = \frac{1}{2}(1 + \gamma) = \frac{1}{2}(1 + 0.99) = 0.995$

The value of c is obtained from Table-9 in Appendix for n-1 = 9 degrees of freedom for F(c) = 0.995 as c = 3.25

The value of k is obtained as $k = \frac{\sigma c}{\sqrt{n}} = 1.247 \times \frac{2.576}{\sqrt{10}} = 1.02$; confidence interval is $65\text{-}1.02 \leq \mu \leq 65 + 1.02, i.e., 63.98 \leq \mu \leq 66.02$

Confidence Interval for μ of the Normal Distribution of a random variable with unknown σ^2

The previous case of sampling involved a normal distribution with known σ^2. Frequently, in practice, σ^2 is unknown. In such cases, the Student's t-Distribution function is used as illustrated below.

Let a sample data set comprise of n variable values $X_1, \ldots, X_n$. The sample mean and standard deviation is obtained as:

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

The following random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a student's t-distribution with n-1 degrees of freedom.

CORRELATION AND LINEAR REGRESSION ANALYSIS

Correlation Analysis

Correlation analysis from a statistical perspective is concerned with the relation between X and Y, where the pair (X, Y) is a two-dimensional random variable. A sample consists of *n* ordered pairs of values $(x_1, y_1), \ldots (x_n, y_n)$

The interrelation between the x and y values in the sample is measured by the sample correlation coefficient given by

$$r = \frac{s_{xy}}{s_x s_y}$$

……. eq (11)

Where the sample covariance s_{xy} is defined as

$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)\right]$$

……. eq (12)

The sample standard deviation of random variable X of the ordered pair is defined as

$$s_x{}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j{}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]$$

……. eq (13)

The sample standard deviation of random variable Y of the ordered pair is defined as

$$s_y{}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j{}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right]$$

Correlation Coefficient

The following is key to the interpretation of the correlation coefficient.

(a) $-1 \ll r \leq 1$
(b) $if\ r = \pm 1$, it indicates perfect linear (positive or negative, respectively) correlation between x and y
(c) $if\ r = 0$, it indicates x and y are uncorrelated
(d) $Intermediate\ values\ 0 \leq r \leq 1\ or\ -1 \leq r \leq 0$ indicates, weak to strong positive and negative correlations

This statistical interpretation of the relationship between ordered pair of two-dimensional random variables X and Y is useful in many applications.

EXAMPLE -8: Perfect and no-correlation

Calculate and interpret the correlation coefficient for the following data value sets of an ordered pair of random variables X and Y.

(a) {(2, 1), (4, 2), (6, 3), (10, 5), (12, 6), (14, 7), (16, 8), (18, 9), (24, 12), (26, 13)}
(b) {(2, 8), (4, 2), (6, 7), (10, 4), (12, 6), (14, 12), (16, 2), (18, 11), (24, 2), (26, 7)}

Solution

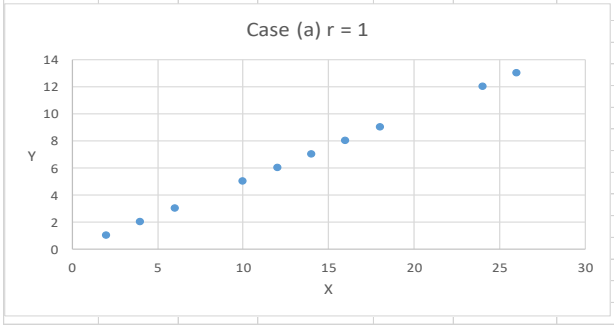
Case (a): Using eq (11), (12), (13) & (14) it is found

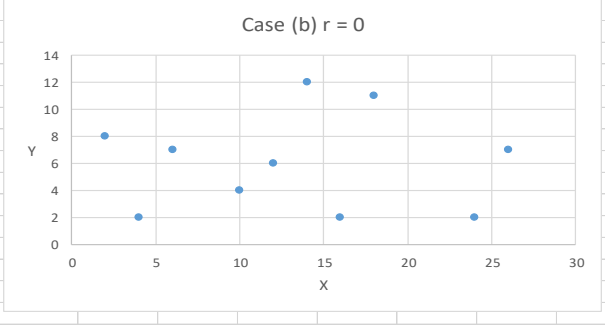
$$r = \frac{s_{xy}}{s_x s_y} = \frac{+\,292.8}{\sqrt{(585.6 * 146.4)}} = 1, perfect\ correlation$$

Case (b): Using eq (11), (12), (13) & (14) it is found

$$r = \frac{s_{xy}}{s_x s_y} = \frac{+\,0.8}{\sqrt{(585.6 * 118.9)}} = 0.003 \approx 0, \qquad no\ correlation$$

S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy	S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy
1	2	1	-11.2	-5.6	125.44	31.36	62.72	1	2	8	-11.2	1.9	125.44	3.61	-21.28
2	4	2	-9.2	-4.6	84.64	21.16	42.32	2	4	2	-9.2	-4.1	84.64	16.81	37.72
3	6	3	-7.2	-3.6	51.84	12.96	25.92	3	6	7	-7.2	0.9	51.84	0.81	-6.48
4	10	5	-3.2	-1.6	10.24	2.56	5.12	4	10	4	-3.2	-2.1	10.24	4.41	6.72
5	12	6	-1.2	-0.6	1.44	0.36	0.72	5	12	6	-1.2	-0.1	1.44	0.01	0.12
6	14	7	0.8	0.4	0.64	0.16	0.32	6	14	12	0.8	5.9	0.64	34.81	4.72
7	16	8	2.8	1.4	7.84	1.96	3.92	7	16	2	2.8	-4.1	7.84	16.81	-11.48
8	18	9	4.8	2.4	23.04	5.76	11.52	8	18	11	4.8	4.9	23.04	24.01	23.52
9	24	12	10.8	5.4	116.64	29.16	58.32	9	24	2	10.8	-4.1	116.64	16.81	-44.28
10	26	13	12.8	6.4	163.84	40.96	81.92	10	26	7	12.8	0.9	163.84	0.81	11.52
Sum	132	66	0	0	585.6	146.4	292.8	Sum	132	61	0	3.55E-15	585.6	118.9	0.8
Mean	13.2	6.6			8.06639118	4.03319559	32.53333	Mean	13.2	6.1			8.066391	3.634709	0.088889
Corr_Coeff							1	Corr_Coeff							0.003032





EXAMPLE -9: Strong and weak positive correlation

Calculate and interpret the correlation coefficient for the following data value sets of an ordered pair of random variables X and Y.

(a) {(2, 1), (4, 2), (6, 4), (10, 4), (12, 7), (14, 8), (16, 7), (18, 9), (24, 12), (26, 13)}
(b) {(2, 6), (4, 1), (6, 4), (10, 3), (12, 9), (14, 8), (16, 2), (18, 11), (24, 11), (26, 8)}

Solution

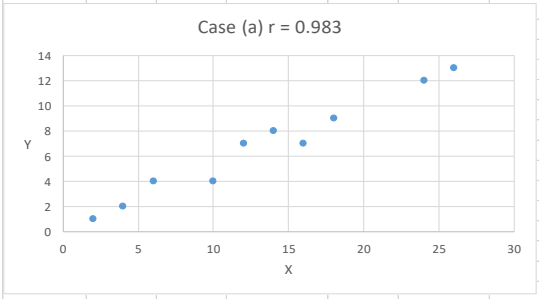
Case (a): Using eq (11), (12), (13) & (14) it is found

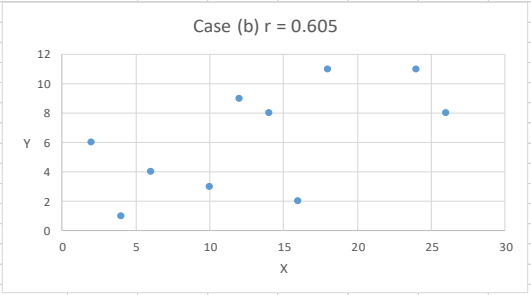
$$r = \frac{s_{xy}}{s_x s_y} = \frac{+285.6}{\sqrt{(585.6 * 144.1)}} = 0.983, strong\ positive\ correlation$$

Case (b): Using eq (11), (12), (13) & (14) it is found

$$r = \frac{s_{xy}}{s_x s_y} = \frac{+160.4}{\sqrt{(585.6 * 120.1)}} = 0.605, weak\ positive\ correlation$$

S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy	S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy
1	2	1	-11.2	-5.7	125.44	32.49	63.84	1	2	6	-11.2	-0.3	125.44	0.09	3.36
2	4	2	-9.2	-4.7	84.64	22.09	43.24	2	4	1	-9.2	-5.3	84.64	28.09	48.76
3	6	4	-7.2	-2.7	51.84	7.29	19.44	3	6	4	-7.2	-2.3	51.84	5.29	16.56
4	10	4	-3.2	-2.7	10.24	7.29	8.64	4	10	3	-3.2	-3.3	10.24	10.89	10.56
5	12	7	-1.2	0.3	1.44	0.09	-0.36	5	12	9	-1.2	2.7	1.44	7.29	-3.24
6	14	8	0.8	1.3	0.64	1.69	1.04	6	14	8	0.8	1.7	0.64	2.89	1.36
7	16	7	2.8	0.3	7.84	0.09	0.84	7	16	2	2.8	-4.3	7.84	18.49	-12.04
8	18	9	4.8	2.3	23.04	5.29	11.04	8	18	11	4.8	4.7	23.04	22.09	22.56
9	24	12	10.8	5.3	116.64	28.09	57.24	9	24	11	10.8	4.7	116.64	22.09	50.76
10	26	13	12.8	6.3	163.84	39.69	80.64	10	26	8	12.8	1.7	163.84	2.89	21.76
Sum	132	67	0	0	585.6	144.1	285.6	Sum	132	63	0	0	585.6	120.1	160.4
Mean	13.2	6.7			8.06639118	4.00138865	31.73333	Mean	13.2	6.3			8.066391	3.653005	17.82222
Corr_Coeff							0.983163	Corr_Coeff							0.604829





EXAMPLE -10: Strong and weak negative correlation

Calculate and interpret the correlation coefficient for the following data value sets of an ordered pair of random variables X and Y.

(a) {(2, 9), (4, 7), (6, 8), (10, 4), (12, 6), (14, 6), (16, 3), (18, 4), (24, 2), (26, 1)}
(b) {(2, 8), (4, 2), (6, 7), (10, 4), (12, 6), (14, 8), (16, 3), (18, 7), (24, 2), (26, 5)}

Solution

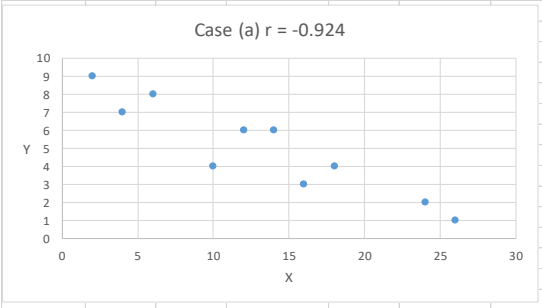
Case (a): Using eq (11), (12), (13) & (14) it is found

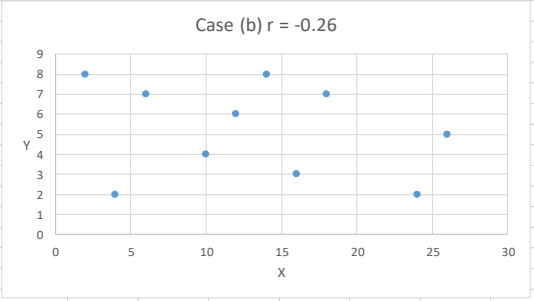
$$r = \frac{s_{xy}}{s_x s_y} = \frac{-176}{\sqrt{(585.6 * 62)}} = -0.924, strong\ negative\ correlation$$

Case (b): Using eq (11), (12), (13) & (14) it is found

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-44.4}{\sqrt{(585.6 * 49.6)}} = -0.26, weak\ negative\ correlation$$

S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy	S_No	X	Y	X-MeanX	Y-MeanY	Square for Sx_srqd	Square for Sy_srqd	Product for Sxy
1	2	9	-11.2	4	125.44	16	-44.8	1	2	8	-11.2	2.8	125.44	7.84	-31.36
2	4	7	-9.2	2	84.64	4	-18.4	2	4	2	-9.2	-3.2	84.64	10.24	29.44
3	6	8	-7.2	3	51.84	9	-21.6	3	6	7	-7.2	1.8	51.84	3.24	-12.96
4	10	4	-3.2	-1	10.24	1	3.2	4	10	4	-3.2	-1.2	10.24	1.44	3.84
5	12	6	-1.2	1	1.44	1	-1.2	5	12	6	-1.2	0.8	1.44	0.64	-0.96
6	14	6	0.8	1	0.64	1	0.8	6	14	8	0.8	2.8	0.64	7.84	2.24
7	16	3	2.8	-2	7.84	4	-5.6	7	16	3	2.8	-2.2	7.84	4.84	-6.16
8	18	4	4.8	-1	23.04	1	-4.8	8	18	7	4.8	1.8	23.04	3.24	8.64
9	24	2	10.8	-3	116.64	9	-32.4	9	24	2	10.8	-3.2	116.64	10.24	-34.56
10	26	1	12.8	-4	163.84	16	-51.2	10	26	5	12.8	-0.2	163.84	0.04	-2.56
Sum	132	50	0	0	585.6	62	-176	Sum	132	52	0	-1.8E-15	585.6	49.6	-44.4
Mean	13.2	5			8.06639118	2.62466929	-19.5556	Mean	13.2	5.2			8.066391	2.347576	-4.93333
Corr_Coeff							-0.92367	Corr_Coeff							-0.26052





Linear Regression Analysis

In linear regression analysis, a linear relationship is established between X and Y using the least squares error principle. As before, the pair (X, Y) is a two-dimensional random variable and the sample consists of *n* ordered pairs of values (x_1, y_1), … (x_n, y_n). Either of X or Y could be an independent variable, with the other depending on its value, or both could be independent of each other. The straight regression line is given by:

$$y = k_0 + k_1 x$$

……. eq (15)

The constants k_0 and k_1 are determined by minimizing the sum of the squares of errors in y from the straight line to be fitted. This holds true for X being an independent variable, hence implying that all the errors are in Y.

It can be shown that:

$$k_1 = \frac{s_{xy}}{{s_x}^2};\ k_0 = \mu_y - k_1\mu_x;$$

……. eq (16)

The strength of the linear relationship is established as before by:

$$r = \frac{s_{xy}}{s_x s_y} \approx \pm 1$$

EXAMPLE -11: Ohm's Law

It is believed that the applied Voltage *V* across a resistor generates a current *I* given by Ohm's law, *V* = *IR*. Check this result using the following measurements of voltage and current across the circuit and estimate the value of R (Ω, Ohms)

Voltage y[V]	40	40	80	80	110	110
Current x[A]	5.1	4.8	10.0	10.3	13.0	12.7

Solution

It is first shown the relationship is linear subject to random errors in the measurement of voltage and current.

Case (a): Using eq (11), (12), (13) & (14) it is found

$$r = \frac{s_{xy}}{s_x s_y} = \frac{+\,561.333}{\sqrt{(64.628 \times 4933.333)}} = 0.9941 \approx 1$$

From eq (16), $k_1 = \frac{s_{xy}}{{s_x}^2} = R = +\frac{561.333}{64.628^2} = 0.1343\ (\Omega, Ohms)$