STATISTICS

Semester IV

STAT SEC402/ Statistical Computing (Minitab)

Topics:-

- **Box plot ,Stem-leaf ,Frequency polygon**
- **Descriptive statistics , Correlation and Regression.**
- **Random number generation,Fitting of polynomials**
- **Confidence interval**

Ms. Moon

Head,Department of Statistics

E-mail : moon.stat@patnawomenscollege.in

Regression:-

Regression analysis is a mathematical measure of the average relationship between two or more variables.

Use Regression to fit least squares models when you have continuous and/or categorical predictors. We can:

- fit interaction and polynomial terms
- perform stepwise regression
- store regression statistics
- examine residual diagnostics
- perform the pure error lack-of-fit test when our data contain replicates
- transform highly skewed data

The default model contains the variables that we enter in **Continuous predictors** and **Categorical predictors**. If we want to add interaction and polynomial terms, use the tools in the **Model** subdialog box.

Minitab stores the last model that we fit for each response variable. We can use the stored models to quickly generate predictions, contour plots, surface plots, overlaid contour plots, factorial plots, and optimized responses.

Responses: Select the continuous variable(s) that we want to explain or predict with the predictors (X). The response is also called the Y variable. If there is more than one response variable, Minitab fits separate models for each response.

Continuous predictors: Select the continuous variables that explain changes in the response. The predictor is also called the X variable.

Categorical predictors: Select the categorical classifications or group assignments, such as a type of raw material, that explain changes in the response. The predictor is also called the X variable.

Use 1: To determine how a single response variable is related to a variety of predictor variables

An agricultural researcher knows that a number of predictor variables (temperature, rainfall, type of fertilizer, and so on) can affect crop yield. If she understands how these predictors combine to affect crop yield, she can maintain high productivity regardless of weather conditions.

Fit Regression Model is a versatile tool for investigating relationships between a response variable and both categorical and continuous predictor variables.

Use 2: To perform a regression analysis of nonlinear relationships

A doctor studies the relationship between the number of bacteria in a throat culture and two predictors: body temperature and antibiotic dosage. Though certain that a strong relationship exists, she is unable to obtain a good fit with linear regression.

While it's common to model linear relationships between a response and its predictors, sometimes the true relationship is not linear, but curved. Use Fit Regression Model to model these relationships with quadratic, cubic, or other polynomial terms.

Polynomial regression models

we can fit the following linear, quadratic, or cubic regression models:

Model type	**Order**	**Statistical model**
Linear	first	$Y = \beta_0 + \beta_1 X + e$
quadratic	second	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$
Cubic	third	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$

Correlation:-

Stat > Basic Statistics > Correlation

A correlation coefficient measures the extent to which two variables tend to change together. Minitab offers two different correlation analyses:

- **Pearson product moment correlation** - The Pearson correlation evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other.

For example, we might use a Pearson correlation to evaluate whether increases in temperature at our production facility are associated with decreasing thickness of our chocolate coating.

- **Spearman rank-order correlation** (also called Spearman's rho) - The Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman correlation is often used to evaluate relationships involving ordinal variables. For example, we might use a Spearman correlation to evaluate whether the order in which employees complete a test exercise is related to the number of months they have been employed.

There are a few points to keep in mind when performing or interpreting a correlational analysis:

- It is always a good idea to examine the relationship between variables with a scatterplot. Correlation coefficients only measure linear (Pearson) or monotonic (Spearman) relationships. Other relationships are possible.

- It is never appropriate to conclude that changes in one variable cause changes in another based on a correlation alone. Only properly controlled experiments allow you to determine if a relationship is causal.

- The Pearson correlation coefficient is very sensitive to extreme values. A single value that is very different from the others in a data set can change the value of the coefficient a great deal.

Variables: Choose the columns containing the variables we want to correlate. When we list two columns, Minitab calculates the correlation coefficient for the pair. When we list more than two columns, Minitab calculates the correlation for every possible pair, and displays the lower triangle of the correlation matrix (in blocks if there is insufficient room to fit across a page).

Method

Pearson correlation: Calculate the linear correlation coefficient for each pair of variables.

Spearman rho: Calculate the rank-order correlation coefficient for each pair of variables.

Display p-values: Check to display p-values for the hypothesis test .For a coefficient, ρ, the hypothesis are: H_0: $\rho = 0$ versus H_1: $\rho \neq 0$.

Correlation Coefficients

The correlation coefficient can range in value from –1 to +1, and tells you two things about the linear (Pearson) or monotonic (Spearman) relationship between two variables:

- Strength – The larger the absolute value of the coefficient, the stronger the relationship between the variables. A value of 0 indicates the absence of a relationship.

For the Pearson correlation, an absolute value of 1 indicates a perfect linear relationship. For the Spearman correlation, an absolute value of 1 means that the rank-ordered data are perfectly aligned. For example, a Spearman correlation of -1 means that the highest value for Variable A is associated with the lowest value for Variable B, the second highest value for Variable A is associated with the second lowest value for Variable B and so on.

Whether an intermediate value is interpreted as a weak, moderate, or strong correlation depends on our objectives and requirements.

- Direction – The sign of the coefficient indicates the direction of the relationship. If both variables tend to increase or decrease together, the coefficient is positive. If one variable tends to increase as the other decreases, the coefficient is negative.

There are a few points to keep in mind when performing or interpreting a correlational analysis:

- It is always a good idea to examine the relationship between variables with a scatterplot. Correlation coefficients only measure linear (Pearson) or monotonic (Spearman) relationships. Other relationships are possible.

- It is never appropriate to conclude that changes in one variable cause changes in another based on a correlation alone. Only properly controlled experiments allow we to determine if a relationship is causal.

- The Pearson correlation coefficient is very sensitive to extreme values. A single value that is very different from the others in a data set can change the value of the coefficient a great deal.

Example

	Hydrogen	Porosity
Porosity	**0.625**	
	0.017	
Strength	**-0.790**	**-0.527**
	0.001	0.053

Cell Contents: Pearson correlation

P-Value

Interpretation

The results indicate the following:

- The Pearson correlation coefficient for the relationship between hydrogen content and porosity is **0.625**. This means that when the hydrogen content increases, the porosity of the aluminum casting also tends to increase.

- In contrast, the correlations between hydrogen and strength (**-0.790**) and between porosity and strength (**-0.527**) are both negative. This means that as hydrogen content or porosity increase, strength tends to decrease.

The correlations do not prove that increased hydrogen causes increased porosity or that increased porosity allows the entry of more hydrogen into the aluminum. For example, both phenomena may be caused by a third variable such as temperature.

Display descriptive statistics:-

Summarize our data with display descriptive statistics,such as a mean and standard deviation and display the result in session window .

Stat > Basic Statistics > Display Descriptive Statistics

Produces descriptive statistics for each column, or for each level of a By variable

To calculate descriptive statistics individually and store them as constants, see Column Statistics. To store many different statistics, use Store Descriptive Statistics.

Use Display Descriptive Statistics to produce statistics for each column or for subsets within a column. The data columns must be numeric.

Displays a histogram, a histogram with a normal curve, an individual value plot, and a boxplot.

Histogram of data: Choose to display a histogram for each variable.

Histogram of data, with normal curve: Choose to display a histogram with a normal curve for each variable.

Individual value plot: Choose to display an individual value plot for each variable.

Boxplot of data: Choose to display a boxplot for each variable.

We can display our data in a histogram, a histogram with normal curve, a dotplot, or a boxplot, or display a graphical summary. The displayed statistics are listed in Descriptive Statistics Available for Display or Storage.

The graphical summary includes a table of descriptive statistics, a histogram with normal curve, a boxplot, a confidence interval for the population mean, and a confidence interval for the population median. Minitab can display a maximum of 100 graphs at a time. Therefore, the graphical summary will not work when there are more than 100 columns, 100 distinct levels or groups in a By column, or the combination of columns and By levels is more than 100.

There is no restriction on the number of columns or levels when producing output in the Session window.

Store descriptive statistics:-

Summarize our data with display descriptive statistics,such as a mean and standard deviation and display the result in worksheet.

Stat > Basic Statistics > Store Descriptive Statistics

Stores descriptive statistics for each column, or for each level of one or more By variables.

To calculate descriptive statistics individually and store them as constants, see Column Statistics. To display many different statistics in the Session window, see Display Descriptive Statistics. For a list of statistics available for display or storage, see Descriptive Statistics Available for Display or Storage.

Variables: Enter the column(s) containing the data we want to describe.

By variables (optional): Enter the column(s) containing the By variable

The data columns must be numeric.

Minitab automatically names the storage columns with the name of the stored statistic and a sequential integer starting at 1. For example, suppose we enter two columns in **Variables** and choose to store the default mean and sample size. Minitab will name the storage columns Mean1 and N1 for the first variable and Mean2 and N2 for the second variable. If we use two By variables, Minitab will store the distinct levels (subscripts) of the By variables in columns named ByVar1 and ByVar2, with the appended integer cycling as with the stored statistics.

Boxplot:-

Boxplot is a method for graphically depicting groups of numerical data through their quartiles.Boxplots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles.

Use to a boxplot to:

- Quickly compare distributions
- View the central tendency of the data
- Highlight the variability of the data
- Determine whether a sample distribution is symmetric or skewed.
- Check for outliers .

In minitab,the response variable data on the y-axis

- Groups based on categorical grouping variables along the x-axis

- A rectangular box for each group representing the middle 50% (interquartile range) of the data.

- The median value indicated by the horizontal line inside the box.

- Lines (called "whiskers") extending from the box representing the upper and lower 25% of the distribution (excluding outliers).

- Outliers indicated by asterisks beyond the whiskers

Use the boxplot to examine and compare the central tendency and variability of one or more distributions and to identify any outliers. Minitab can produce multiple boxplots on one chart using a categorical grouping variable, allowing we to compare several groups of data. Display options include symbols for the mean and boxes for the median confidence intervals.

Stem and leaf plot:- It is device for presenting quantitative data in a graphical format similar to a histogram.

In Minitab,Use a stem-and-leaf plot to:

- Display the actual data values in a binned format.

- Highlight the central tendency of the data.

- Emphasize the variability of the data.

- Determine whether a sample distribution is symmetrical or skewed .

Minitab displays a stem-and-leaf plot in the Session window. The layout of the plot is similar to a histogram rotated 90 degrees to the right; however, instead of bars, digits from the actual data values indicate the frequency of each bin (row).

To construct a steam and leaf display,the observations must first be sorted in ascending order.This can be done most easily by constructing a draft of the stem and leaf display with the leaves unsorted ,then sorting the leaves to produce the final stem & leaf display.

For eg:- 44,46,47,49,63,64,66,68,68,72,72,75,76,81,84,88,106.

In this ,leaf contains the last digits of the number & the stem contains all the other digits.

Stem	leaf
4	4 6 7 9
5	0
6	3 4 6 8 8

7	2 2 5 6
8	1 4 8
9	0
10	6

❖ How do we generate a random data?

Calc > Random Data > Sample From Columns

Randomly samples the same rows from one or more columns. We can sample with replacement (select the same row more than once), or without replacement (select each row only once).

Number of rows to sample: Specify the number of rows to randomly select.

From columns: Enter the column(s) we want to sample from. If sample from several columns at once, they must all have the same length.

Store samples in: Specify the column(s) where we want to store the sampled values. The number of storage columns must be the same as the number of columns sampled from.

Sample with replacement: Check to sample with replacement. Leave unchecked to sample without replacement (sample size must be less than or equal to the length of the columns).

Try to mention (explain it) ,sample with replacement and sample without replacement.

❖ Display Descriptive Statistics : Topics

Central tendency

Mean

Median

Dispersion

Standard error of the mean (SE Mean)

Minimum and maximum

First and third quartiles (Q1 and Q3)

Graphs

Histogram of data

Histogram of data with normal curve

Individual value plot

Boxplot of data

- The median (also called the 2nd quartile or 50th percentile) is the midpoint of the data set: half the observations are above it, half are below it. It is determined by ranking the data and finding observation number [N + 1] / 2. If there are an even number of observations, the median is extrapolated as the value midway between that of observation numbers N / 2 and [N / 2] + 1.

  The median is less sensitive to extreme values than the mean . Therefore, the median is often used instead of the mean when data contain outliers, or are skewed.

- The standard deviation (StDev) is a measure of how far the observations in a sample deviate from the mean . It is analogous to an average distance (independent of direction) from the mean. The standard deviation is the most commonly reported measure of dispersion. It also serves as an estimate of the dispersion in the broader population from which a sample is taken. Like the mean, the standard deviation is very sensitive to extreme values.

  If the data are normally distributed, then the standard deviation and mean can be used to determine what proportion of the observations fall within any given range of values. For example, 95% of the values in a normal distribution fall within ± 1.96 standard deviations of the mean.

Example:-

Variable	Mean	SE Mean	StDev	Minimum
Precipitation	3.636	0.717	**2.378**	1.000

Variable	Q1	Median	Q3	Maximum
Precipitation	2.000	3.000	4.000	10.000

Interpretation:-The standard deviation for the precipitation data is **2.378**. This tells we that on average, the values in the data set tend to differ from the mean by ± 2.378.

The large value of 10 days with precipitation for April increases the standard deviation quite a bit. Without this value, the standard deviation would be 1.155 instead of **2.378**. Conversely, if April had 30 days of rain, the standard deviation would be 8.210.

- ➢ The standard error of the mean (SE Mean) is not often used as a descriptive statistic, but it is important in hypothesis testing. It is an estimate of the dispersion that we would observe in the distribution of sample means , if we continued to take samples of the same size from the population.

  The standard error of the mean is the standard deviation divided by$\sqrt{N}$.

Dispersion:-

One of the simplest ways to assess dispersion in our data is to compare the minimum and maximum . The minimum is the smallest value in a data set, and the maximum is the largest value.

Minimum and maximum are used to calculate the range , which is a statistic that is often used to describe dispersion in data sets. The range is simply the maximum – minimum. Notice that the range is very sensitive to extreme values.

- Exactly 25% of your data is less than the first quartile (Q1, also called the 25th percentile). It equals the data value at position (N + 1) / 4. If this position number is not an integer, Minitab extrapolates between the two observations on either side of that position.
- Exactly 75% of your data is less than the third quartile (Q3, also called the 75th percentile). It equals the data value at position 3(N + 1) / 4. If this position number is not an integer, Minitab extrapolates between the two observations on either side of that position.
- Q1 and Q3 are often used to calculate the interquartile range (IQR) ,which is another statistic used to describe dispersion . The IQR is the range of the middle 50% of the values and is calculated by the formula Q3 – Q1. The IQR is relatively insensitive to extreme values.

- A confidence interval is an interval used to estimate a population parameter from sample data. The upper and lower bounds of the confidence intervals for µ (mu), σ (standard deviation), and the median are displayed in the graphical summary. In addition, the confidence intervals for µ and the median are displayed graphically.

❖ Confidence Intervals for Mean, Standard deviation, and Median

Confidence intervals are composed of two basic parts:

- **Point estimate** – a single value computed from the sample data. This value is considered to be an estimate of the parameter of interest, however it is unlikely that the point estimate is equal to the parameter. Therefore, to account for the possibility of estimation error, the error margin is included in the confidence interval to provide a range of possible parameter values.

- **Error margin** – determines the width of the confidence interval through the use of probability. To construct the confidence interval, we simply add and subtract the error margin from the point estimate.

If a 95% confidence interval is selected, the method used to construct the interval has a probability of 0.95 of producing an interval containing the parameter of interest. In other words, we can be 95% confident that the true value of the parameter is within the interval. Thus, if one hundred 95% confidence intervals were constructed, we would expect around 95 of the intervals to contain the parameter.