CLASS LECTURE NOTES

- What is Applied Statistics?
- Applications from various fields.
- What is statistics?
- What is probability?
- Relationship between probability and Statistics.
- **Review of Probability and Statistics.**
- Introduction to MINITAB.

WHAT IS APPLIED STATISTICS?

- Collection of (statistical) techniques used in practice.
- Range from very simple ones such as **graphical display, summary statistics, and time-series plots**, to sophisticated ones such as **design of experiments, regression analysis, principal component analysis, and statistical process control.**
- Successful application of statistical methods depends on the close interplay between theory and practice.
- There should be interplay (communication and understanding) between engineers and statisticians.
- Engineers should have adequate statistics background to (a) know what questions to ask; (b) mix engineering concepts with statistics to optimize productivity; (c) get help and understand the implementation.
- The object of statistical methods is to make the scientific process as efficient as possible. Thus, the process will involve several iterations, each of which will consist of an “hypothesis”, data collection, and “inference”. The iterations stop when satisfactory results are obtained.

WHY WE NEED STATISTICS?

Quality is something we all look for in any product or service we get.

- What is Quality?

It is a measure of the extent to which customer expectations and satisfaction are met.

- It is not static and changes with time.
- Continuous quality improvement program is a **MUST** to stay competitive in these days.
- Final quality and cost of a product are pretty much dependent on the (engineering) designs and the manufacture of the products.
- Variability is present in machines, materials, methods, people, environment, and measurements.
- Manufacturing a product or providing a service involves at least one of the above 6 items (may be some other items in addition to these)
- Need to understand the variability.
- Statistically designed experiments are used to find the optimum settings that improve the quality.
- In every activity, we see people use (or abuse?) statistics to express satisfaction (or dissatisfaction) towards a product.
- There is no such a thing as good statistics or bad statistics.
- It is the people who report the statistics manipulate the numbers to their advantage.
- Statistics properly used will be more productive.

EXPLORE, ESTIMATE and CONFIRM

Statistical experiments are carried out to

EXPLORE: gather data to study more about the process or the product.
ESTIMATE: use the data to estimate various effects.
CONFIRM: gather additional data to verify the hypotheses.

EXAMPLE 1 (EEC)

Bonding Example: An engineer working for a chemical company has the following *diary* of activities with regard to a “new bonding method” that is under consideration by the company.

Hypothesis 1: A new bonding method to bond two films is expected to yield a higher bonding strength compared to the current method.

In order to verify the hypothesis, I started by gathering a list of key variables in consultation with the group involved in the project. The key variables identified are: bonding glue, temperature, density and thickness of the films, and pressure setting. I ran the following experiment first.

Experiment 1: Two films were bonded together by choosing bonding glue type A, temperature level

to be 300^{o}C, the thickness of the two films to be 4 mils, and a pressure setting to be 200 psi.
Data 1: The bonding strength measured was lower than the current method.

Question 1: Why is data 1 not supportive of the hypothesis 1?

After thinking over this experiment I arrived at the following conclusion.

Induction 1: The temperature setting may be low causing the glue to perform at below optimum level.

Experiment 2: Three sets of two films were bonded together by choosing bonding glue type A, the thickness of the two films to be 4 mils, and a pressure setting to be 200 psi. The temperature settings for these three sets were taken to be 400^{o}C, 450^{o}C and 500^{o}C, respectively.

Data 2: The bonding strengths for the three specimens were as follows:
At 400^{o}C the strength was still lower than the current one;
At 450^{o}C the strength was higher than the current one;
At 500^{o}C the strength was lower than the current one;

Induction 2: The temperature setting at 450^{o}C seems to give a better bonding strength when all other variables are set at the above mentioned levels.

The above investigation in various steps illustrates the basic ideas in a statistical experiment conducted in a scientific way. The remaining series of steps, with possible modifications including varying the settings of the variables simultaneously, form the basis of an experimental design. This will be seen in great detail later.

The basic ideas of this experiment can be summarized as follows.

- **Constraint:** the films should not peel off under “normal” usage.
- **Key variables**: bonding glue, temperature, density and thickness of the films, and pressure setting.
- **Goal**: the effectiveness of such bonding method.
- **Procedure**: All possible configurations in actual production setup should be considered in the study.

EXPLORE: Bond specimens of films at several settings and measure the bonding strength.

ESTIMATION: Suppose our study shows that the bonding strength is affected by glue, temperature and setting, then we would like to estimate the strength.

CONFIRMATION: Once we find the optimal settings, we run additional experiments to verify that the settings are in fact “best”.

Recommendation: If the study is done scientifically, then we may have one of the following:
(a) Continue with the production.
(b) Not to use the method.
(c) Suggest appropriate modification in the process.
However, if it is not scientifically done, the conclusion may be totally false.

APPLICATIONS

- Statistical methods have applications in many areas: industrial, medical, behavioral, sociological and economic.
- General principles and strategies to be adopted in these areas will all be the same. However, certain problems can call for some special techniques.

Some detailed engineering applications are given below. You may want to add more to these as we go along.

ENGINEERING APPLICATIONS EXAMPLES

The following are examples taken mainly from past students= class projects. Additional ones are taken from published sources, such as journals, magazines, and technical reports.

1. The cradle mount system, consisting of both rubber and metal assemblies, is used in automobiles to reduce the vibration and noise from the engine and from the ride. This will improve the ride characteristics of the automobile. In one study, the objective was to find an optimum configuration for the three variables: neck type, rate ring and material, that will give the best vertical dynamic rate. In this study these three variables were used at 2 levels each. The statistical tools used here are: Design of experiments and regression analysis. The student was able to collect the data and came up with a recommendation for the best design. In another study involving the cradle mount system, the student=s objective was to minimize the damping of a rubber component. Two factors: temperature and the time were considered at three levels. Using 3^2 factorial design, the student recommended an optimum setting for the temperature and the time.

2. This project involves the peening process, where in the operation involves moving aluminum material around a stainless steel ball. The ball is used to plug entrance holes where drilling operations are required. The objective of the study is to establish a planned replacement for the peen tools and to increase tool life and thus to increase machine uptime. Before the project was undertaken, there was no established tool life for the process and only visual inspection of the part was used to determine when a tool needs to be changed. He chose a set of four balls and used initial data to come up with some interesting observations. He indicates that this will be a starting point for further analysis involving other locations as well as using some additional factors in the study.

3. Paint process is a costly and a very involved one. Air spray pain guns and electrostatic rotary atomizers (referred to as bells) are two of the many ways to paint an automobile. Air spray guns use high air pressure to atomize and direct the flow of paint, and thus leads to a lot of wastage of the paint. In this project, a student was involved in finding optimal initial process parameters to achieve a target paint film thickness using “bells”. The process parameters chosen for the study were: distance from bell to body, linear speed of the robot and bell, rotational speed of the turbine, paint flow rate, and shaping air pressure. Using design of experiments and regression analysis, the student identified optimal factor settings that resulted in film builds that were extremely close to the target value.

4. As is known Sulfur dioxide (SO_2) is one of the contributors to the “greenhouse effects”. EPA is very strict about the amount of SO_2 released from the manufactruing processes. Many assembly plants and other component facilities produce the much needed steam through gas, oil or coal fired boilers. As is known, when burning coal, on average 98 % of the sulfur present in coal will be emitted as SO_2. The amount of SO_2 emitted is dependent only on the percentage of the sulfur in the coal, and not on the boiler size, the firing configuration, or the operation of the boiler. In this project, the student was interested in the amount of sulfur and the energy content of the coal supplied to her company. There were two coal suppliers to her company, one of which is believed to supply a better quality. She was interested in testing this claim as well as to determine whether one supplier has superior emission characteristics over the other, and if so, which parameter can be attributed to this phenomenon. Using statistical analysis (including 2-sample tests) the student came up with a number of interesting conclusions.

5. This project deals with evaluation of a prototype mechanism for fuel economy improvement. That is, a particular device, which is claimed to improve fuel economy by converting engine pumping loss into electrical output to supplant the engine driven alternator, and thereby reducing fuel consumption, was under study. The objective of this project was to determine whether the vehicle test data that were previously collected supports the need for the device. Using statistical tools (tests of hypotheses, regression analysis) the student concluded that the device does have an impact on the engine performance for combustion and emissions; and also the device could increase MPG.

6. This project deals with tube size variation study. Tube sizing is a process in which the diameter is changed for a given piece of tubing. This feature is critical in exhaust system applications. Tubes that are too large may not fit into mating tubes and thus prevent assembly. On the other hand, tubes that are too small will fit, but may not seal exhaust gases. Thus, it is very important to study the variation in the tube sizing process. After a brain storming session, the student identified 5 key variables (all at two levels) that might explain the variability and also help to find the best configuration setting. Using design of experiments, the student summarized his findings and recommendations for future work.

7. This deals with the problem of part thickness. A flat metallic part (which comes in varying thickness) is pressed to a specified thickness. Since parts were coming with various thickness initially, it is important to adjust the pressing mechanism to get the final product to be of the same thickness. The student used a linear regression model to predict the relationship between change in part thickness and the pressure applied.

8. Spark plug insulators are extremely abrasive and hence results in tool wear out prematurely. In this project, the student was interested in determining whether one or two different coatings will help extend the life of the nests that hold the parts. Using one machine and testing 3types of nests, it was determined that these three types do not significantly extend the life and it was recommended to test using different materials and coatings.

9. This project deals with window regulator pivot joint=s effect on the sash angle. The angle of the glass sash in the full up position is an important parameter and indicates how well the glass is sealing in the header. This is important as any leak will lead to structural problem. The student first developed a regression model of the current pivot design and then developed a similar one for an alternate pivot design. He concluded that the alternate design could control the variance better than the current one, but requires additional testing in view of an increase in the alternate design=s diameter.

10. In order to compare the system throughput before and after the initiation of process improvement efforts within the body shop of a leading automobile company, this student used standard statistical methodology to conclude that the process improvement team continue the efforts and also monitor the throughput to establish certain factors, such as cycle time and machine downtime.

11. Over the last few years the use of heated exhaust oxygen sensors has grown due to the emissions related benefits. With the increasing emphasis on quality improvements and process/product variation reduction, it has become very essential to look at and analyze process control data to ensure that key product characteristics are being met consistently, and the process performance is meeting or exceeding requirements. In this project, the student uses the data on ceramic heater resistance, which was compiled for a three-month time period, to make a number of useful recommendations and conclusions.

12. Testing electronic components to assure the resistive and capacitive values are correct for the components requires statistical analysis. In this project, the student was interested to see whether or not the test values obtained are consistent from tester to tester as well as from fixture to fixture within the testers. Using Analysis of variance techniques, some interesting conclusions and recommendations are drawn here.

13. In this project, the student=s interest was to verify whether the product HVM (heating and ventilating module) for cars, is of high quality and free from defects. He used basic statistical methods as well as his own program to analyze the data and make useful recommendations.

14. Color is one of the main ingredients in marketing strategy. We see these in food items, medicine, containers, packing materials, and so on. Consumers are attracted to color products more than plain ones. The product may not be good, but the presentation of the product may have more influence in people acquiring the product. One of the major problems in coloring products is environmental in nature. For example, in plastic industries heavy-metal pigment approach for coloring has been very popular until 1980's, when environmentally safer coloration began to emerge. One such method is using mixed-metal oxides (MMO's), which is quite efficient and more expensive than the traditional method. However, due to federal and state regulations, MMO's are being used. Statistics (DOE) is used to determine the optimal composition of MMO's.

15. Composite materials are used as pressure release systems in many devices. In order to improve the reliability assessment of these materials, statistics (regression analysis) is used.

16. Any company that produces a product invests time and money to study the product's reliability. That is, the company is interested to (a) know (predict) the lifetime of the product, (b) find how long a warranty, if any, should be given to the product, (c) calculate the probability that the product will function without any problem during the warranty period, (d) identify the causes of different types of failures of the product, (e) study alternative method, if any needed, that might improve the product, and so on. Problems of this nature fall under the general theme of "reliability data analysis" or "life time data analysis". Analyzing the reliability data is one of the major functioning components of the engineering and management department of the R&D of any company. Much of the analysis requires knowledge from probability and statistics.

17. Steel containers are used in transporting many goods including oil and nuclear power plant waste. Hence, it is very important to develop acceptance resistance criterion for the containers to brittle failures. The main variables of interest are the crack size and the stress intensity. Since these are random variables, the study of these requires probability theory and statistics.

18. Vinyls are used as siding in the houses to protect the brick walls (or dry walls). Since colored PVC can reach temperatures high enough to cause early failure of the vinyl siding. Hence, it is imperative to know the maximum temperature that a piece of pigmented PVC can reach when exposed to sunlight. Statistics (regression analysis) is used to predict the heat buildup of pigmented PVC as a function of the amount of sunlight absorbed by the PVC panels.

19. PCR resin is used in the manufacture of trash bags/liners. From environmental point of view it is very important the bags/liners have very high recyclable contents. This requires the study of the physical properties of the bags/liners that have different grade of polyethylene films and different resins. Again, DOE is used to determine the maximum performance characteristics.

20. Thin films are used in many diverse fields such as food packaging industries, housing industries, paper industries, and medicine. The film obtained from the blown film is known to be influenced by many variables. For example, in the polymer the variables are MW, MWD, density and branching; in the equipment, the die size, the die gap, the air ring, IBC, and IRIS are some of the variables; the output, melting temperature, and frost line height will be the variables in the process. The DOE is used to determine the effects of fabrication variables: the blown-up ratio (bubble radius/die radius), the specific output rate, the melting temperature, and the frost line height.

21. Cycle time (time between the completion of the last product and the completion of the next product) is an important quantity in the manufacturing processes. Depending on the product and the layout of the manufacturing area (some products such as a car and a semiconductor chip may require many different processing steps before rolling out as a final product)the cycle time, in reality, is a difficult one to quantify. The manufacturing company would like to estimate this variable as accurately as possible, to figure out the cost, isolate any problem, identify source, if any, of variability and so on. Probability and statistics are the essential tools for such study.

22. Package delivery companies are interested in scheduling the delivery trucks and the drivers to pick up and deliver packages. The companies have several hubs (central processing places) where the packages are sorted. One of the important factors in efficient delivery of packages is to estimate the number of drivers needed in each of the hubs for each of the working days as well as the season of the year. Regressions analysis and time-series analysis are used in developing statistical models for predicting the number of drivers as well as the number of packages. Also from the past data, ANOVA techniques are used to allocate the resources among the various divisions in each hub.

23. Devices such as the pillars supporting the platform of an offshore drilling unit, the pillars of a bridge, electronic items mounted on outboard the ship, and others, are exposed to several hazards in a marine environment. These hazards in the form of chemical, electrical, thermal, or mechanical stresses lead to degradation and eventual failure of the device. Maintenance and replacement of such devices are time-consuming and expensive. Hence, it is imperative to estimate the reliability of such devices as functions of time.

24. In order to determine whether there is any statistical difference between metal and ceramic substrates on the emission performance of automotive exhaust systems, a design of experiment was conducted. A fractional factorial design (2^{5-1} design) was used to conclude that metal substrate may only be marginally better than ceramic substrate.

25. The assembly of a blower motor into a fan coil blower requires that a band be assembled around the motor, which captures and pins the assembly legs against the motor. The band is fastened using a bolt and nut through the ears of the band. It is noted that if the bold and nut assembly is over torqued, the band, which is fabricated from sheet metal, fractures in the bend forming the ears. This leads to eventual band failure resulting in more damage to the blower, its housing and the evaporator coil. Hence, a design of experiment was conducted to identify the optimum settings for nut material, bolt material, lubricant/no lubricant, and, nut and bolt size.

26. Rubber weather strips are used on automobiles to seal out water, dust, and vehicle and road noise. The seal around a vehicle door are some of the most important ones as the weather strips are easily noticeable by the occupants' of the vehicle. During the introduction of a new model car, an unacceptable noise level was noticed and a team was assembled to resolve this problem. The team identified four key variables (automotive coating type, thickness, degree of cure, and weather strip (rubber) type. Using a 2^4 design, optimum settings were identified and the production process used these settings to reduce the noise level.

27. An engineer in a small production process within the manufacturing plant was interested in comparing two 10.5-hour shifts with respect to the production output. These two shifts meet the same daily production requirements. In order to determine whether one shift cost less than the second one to operate (using electrical power consumption as the differentiating factor), the engineer collected data on a number of key variables. Using ANOVA, interesting conclusions were drawn.

28. An ABS manufacturing company was asked by one of its customers to reduce the base brake leakage in their ABS braking system solenoid values. To accomplish this, the engineer set up a DOE with many key factors and came up with very interesting findings.

29. A high percentage of "pole height" rejects was a concern of an engineer working in a manufacturing plant of sensors. A team consisting of 9 members identified key factors using fishbone diagram and conducted a series of experiments iteratively. The problem (at least at the time of submitting the report) was not solved. However, the conclusion of this report underscores the importance of applied statistics and so is quoted directly from the student's report as follows. "It was a very good thing for me to be working in this project, because I was able to apply what I saw in the class and really understand the results and the steps we had to follow to determine the solution. It is very easy in the production real world that when you have a problem you start to shoot everything you think can be affecting, and sometimes you spend more time on this because you do not have the right tools of you just do not understand them. The problem has not been solved yet, but the variability was brought down to two major reasons and thanks to all the statistical tools that each member of the team knew how to use."

30. The primary function and design of the core crimp is to optimize the electrical performance of he cable/terminal interface for the service life of the terminal. The core crimp design utilizes the set of relationships between the terminal, the cable, the core crimp subtools, and the crimp die to provide this high quality connection needed for sufficient power and signal distribution in automotive applications. For signal distribution, or low energy systems operating under 5 volts, the core crimp connection must sustain optimal electrical performance in the face of extreme environmental conditions. This investigation uses regression analysis to establish a linear relationship between environmental conditions and core crimp geometry and their effects on core crimp resistance.

31. In this project dealing with the study of an automobile glove box door alignment, the student uses a 2^4 design to identify the optimal settings that improves the door parallelism to measure within a given tolerance limit.

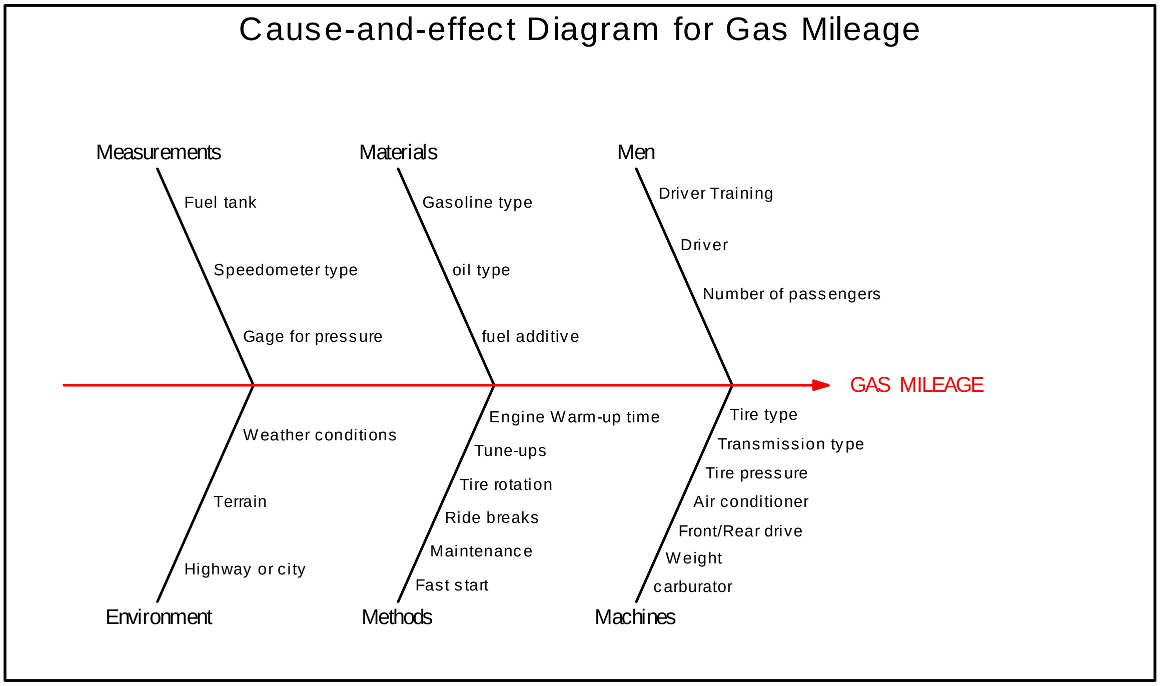
32. A multiple linear regression analysis was used to determine the effects of the duration of laser heat and the tension force on the taper length of an optical fiber

EXAMPLE 2 (Application)

- Nashua Corporation (in NH) manufactures carbonless paper.
- 1100 ft/min; used 3.6 lbs of chemicals/3000 sq.ft.
- 3.6 lbs of chemicals were high. Idea to buy a costly coating head.
- Operator was adjusting constantly.
- Statistics was used to determine that adjustments were made based delayed data and so it didn't pertain to current conditions.
- New operating instructions led to fewer adjustments and reduced the average to 2.6lbs
- Resulted in savings of $800,000/year in addition to not buying a new machine.

BRAINSTORMING

- This is a starting point for any analysis, more so in a statistical study.
- Gather information about the problem by assembling a group of people involved.
- Simple statement of the problem; get all ideas; group these into several classes.
- Draw a cause-and-effect diagram. The following is an example.


Cause-and-effect Diagram for Gas Mileage

EXAMPLE 3 (Illustrative)

- To illustrate and to motivate the need for probability and statistics, we will take a concrete example from practice. Automotive news is one of the magazines that publish all kinds of data about automobiles. Some of these include information about
- MPG's of cars (different types) and trucks;
- Dimensions of wheelbase, overall length, overall width, overall height, curb height, curb weight, front-seat head room, and rear tread.
- Engine data such as type, number of valves, bore and stroke, displacement, compression ratio and net HP.
- Capacities such as fuel tank, cooling system, and crankcase.
- Miscellaneous data about the transmission, axle ratio, drive system and number of passengers.
- A sample data sheet is attached at the end.
- Suppose that MPG of a particular mid-size car (say, Olds Ciera) is to be "estimated". Note the word estimated!
- We cannot find the exact value of the MPG of this type of car (WHY?). Obviously we cannot test all the cars rolling out of that particular model. We need to select a "representative" sample of cars from this model and estimate the "true" average MPG. Note that true average will be known only when **all** the cars in that model are tested. As we all know, this is not practical. So we can only estimate the true value. We can estimate without testing a single car (by simply using the past performance?) But this will not take into account any improvement made on the model. So, what is the best way to do?

Well, we take a sample of n cars. Test them and find the average MPG. Which n cars to be selected for testing? Is it important to know which cars are selected? How are the cars tested? Does the same driver drive these? Are all these cars driven under one type of road condition? These are some of the important questions and should be addressed properly if the estimate is to make any sense! These questions can only be properly addressed using probability and statistics. Statistics also provides a wide variety of techniques, which can be applied to build quality into data, performance and decisions. Statistics also helps us to detect, estimate, control, compare and measure effectiveness of key elements.

Before we proceed further, let us see the definition of data.

WHAT IS DATA?

- Data is collection of information pertaining to a specific problem under study.
- For example, referring to MPG example above, the data would be the miles per gallon of the cars that were tested.
- Suppose we are interested in the braking distance (at 35mph) of that particular model car, then the data would comprise the braking distances of the tested cars.
- As another example if one wishes to study the income level of people in a city (to see whether it is profitable to start a new business), the data for this would be the income of all people living in the city.
- A quality control engineer is interested to see whether the machines under his supervision are producing items within the specifications. Naturally he has to collect data to verify this.
- A pharmaceutical company is planning to introduce a new drug into the market and is interested in seeing the reception for the drug. The company performs a pilot study through contacting a number of physicians and gathers information (data) to see the impact of the drug.
- The data can be quantitative or qualitative.
- Variables, such as the MPG of a new model car, number of defective in a lot sampled, the weight of a cereal box, etc, is quantitative.
- Quantitative variable can be discrete or continuous.
- Variables, which cannot be quantified such as the color of the eyes, location, etc., are classified as qualitative variables.
- A qualitative variable which can be ordered (according to some scale)is referred to as ordinal.
- An unordered qualitative variable (such as the color of the hair) is referred to as nominal.
- In dealing with data one has to be aware of major types of problems such as data errors, outliers and missing observations.
- A data error is an observation that is incorrectly recorded.
- Recording error, typing error, transcription (copying) error, repetition error and deliberate (falsification) error.
- An outlier is an observation that falls away from the rest of the data.
- Missing observations arise for a number of reasons.
- For example, in response to a questionnaire people may forget to answer some questions. In agricultural experiments the crops may suddenly die in some plots leading to no yield, which cannot be taken as 0 yield.
- Some analysis becomes more involved due to missing observations. We shall see more of these in the discussion of Analysis of variance techniques.
- There are two kinds of data: raw and grouped.

- Raw data: not compiled in any way.
- Grouped data: classified into several groups or classes according to some criteria.

UNI- AND MULTI-VARIATE DATA

If the study involved is only on one variable, such as the MPG of a new model car as a function of the size of the car then we are dealing with univariate data. However, if the study deals with more than one variable at a time, then we are dealing with multivariate data. For example, if we are interested in the study of MPG as a function of the engine size, HP, passenger capacity, fuel capacity, etc, then the study deals with multivariate data.

MULTIVARIATE ANALYSIS

Multivariate analysis deals with study involving simultaneous measurements on many variables. Multivariate statistical techniques differ from univariate in the sense that the attention is drawn away from the analysis of mean and variance of a single variable. Instead, the attention is focused on:
(i) Data reduction;
(ii) Sorting and grouping;
(iii) Study of dependence among variables.
There are several multivariate techniques available for investigating the above three areas. These include: (a) multiple regression; (b) discriminant analysis; (c) multivariate ANOVA; (d) correlation analysis; (e) logit analysis; (e) principal component analysis; (f) factor analysis; (g) cluster analysis; (h) metric multidimensional scaling.
In this course, we will concentrate only on the principal component analysis. First, you need to review your matrix algebra, specifically the concepts of linear combination, eigenvalues, eigenvectors, and positive definite matrices.

PRINCIPAL COMPONENT ANALYSIS

The main idea of principal component analysis deals with explaining the variance-covariance structure of the variables under study through a small number of linear combinations of the original variables. These linear combinations are constructed in such a way that much of the total variation in the original variables is explained by these few linear combination ones. Thus, we are concerned with data reduction (in the form of identifying a small number of linear combinations of the original variables) and interpretation. It should be noted that principal components serve as intermediate steps in multiple regression (to be seen later), factor analysis and cluster analysis.

HOW TO USE STATISTICS (efficiently)?

- What is the main objective of the study?
  Then, we ask:

(a) What information is available on this problem?

(b) Do we have data on this problem? If so how the data was selected?

(c) Has any study been done on this problem before?

INVESTIGATION STAGES

Proper statistical study of a problem involves:

Stage 1: Proper understanding of the problem and the goals of the study.

Stage 2: Determine the type of data to be used for the study.

Stage 3: We need to assess the structure and the quality of the data.

Stage 4: Perform an initial examination of the data.

Stage 5: Carry out a number of formal statistical procedures to analyze the data.

Stage 6: Compare with any previous findings.

Stage 7: Summarize the findings through report writings and presentation of important graphs to highlight the crucial findings.

The process of analyzing the data involves the study of populations, samples, variation and other concepts.

Population: is a collection of all units defined by some characteristic, which is the subject under study.

- In the study of the MPG of a new model car, the population consists of the MPG's of all cars of that model.
- To study the income level of a particular city the population consists of the incomes of all working people in that city.

Sample: is a subset (part) of the population.

Since it is infeasible (and impossible in many cases) to study the entire population, one has to rely on samples to make the study.

- Samples have to be as representative as possible in order to make valid conclusions about the populations under study.
- Contain more or less the same type of information that the population has.
- For example if workers from three shifts are involved in assembling cars of a particular model,

then the sample should contain units from all three. -Simple random sampling is the most commonly used one, although there are several types of sampling possible.

- These include cluster sampling, stratified sampling, convenience sampling and judgement sampling.
- Once the data has been gathered, what do we do next?

EXPLORATORY DATA ANALYSIS

Before any formal statistical inference through estimation or test of hypotheses is conducted, exploratory data analysis should be employed. This is a procedure by which the data is carefully looked for patterns, if any, and to isolate them. It often provides the first step in identifying appropriate model for further analysis including prediction. The main difference between exploratory data analysis and the conventional data analysis is while the former, which is more flexible (in terms of any assumptions on the nature of the populations from which the data are gathered) emphasizes on searching for evidence and clues for the patterns, the latter concentrates on evaluating the evidence and the hypotheses on the nature of the parameters of the population(s) under study.

DESCRIPTIVE STATISTICS

- Deals with characterization and summary of key observations from the data.
- Quantitative measures: mean, median, mode, standard deviation, percentiles, etc.
- Graphs: histogram, Box plot, scatter plot, Pareto diagram, stem-and-leaf plot, etc.
- Here one has to be careful in interpreting the numbers. Usually more than one descriptive measure will be used to assess the problem on hand.
- Before we se various descriptive quantitative measures, let us first give a brief introduction to MINITAB and then discuss the graphical display of the data.

DIFFERENT TYPES OF PLOTS

1. Point plot: The horizontal axis (x-axis) covering the range of the data values and vertically plot the points, stacking any repeated values.

2. Time series plot: x-axis corresponds to the number of the observation or the time of the observation or the day and so on and the y-axis will correspond to the value of the observation.

3. Scatter plot: Construct x-axis and y-axis that cover the ranges of two variables. Plot (x_i, y_i) points for each observation in the data set.

4. Histogram: This is a bar graph, where the data is grouped into many classes. The x-axis corresponds to the classes and the y-axis gives the frequency of the observations.

5. Stem-and-leaf plot: Data is plotted in such a way the output will look like histogram and also features a frequency distribution. The idea is to use the digits of the data to illustrate its range, shape and density. Each observation is split into leading digits and trailing digits. All the leading digits are sorted and listed to the left of a vertical line. The trailing digits are written to the right of the vertical line.

6. Pareto Diagram: Named after the Italian economist. This is a bar diagram for qualitative factors. This is very useful to identify and separate the commonly occurring factors from the less important ones. Visually it conveys the information very easily.

7. Box plot: is due to J. Tukey and provides a great deal of information. A rectangle whose lower and upper limits are the first and third quartiles, respectively, is drawn. The median is given by a horizontal line segment inside the rectangle box. The average value is marked by a symbol such as "x" or "+". All points that are more extreme are identified.

8. Quantile plot: This plot is very useful when we want to identify/ verify an hypothesized population distribution from which the data set could have been chosen. A quantile, Q(r), is a number that divides a sample (or population) into two groups so that the specified fraction r of the data values is less than or equal to the value of the quantile.

9. Probability plot: This involves plotting the cumulative probability and the observed value of the variable against a suitable probability scale which will result in linearization of the data. The basic steps involved here are: (a) Sorting the data into ascending order; (b) Computing the plotting points; (c) Selecting appropriate probability paper; (d) Plot the points; (e) Fitting a "best" line to data.

In the following we will use 93CAR data to illustrate the use of MINITAB.

MINITAB

- First make sure that you have installed the software correctly [Refer to the user’s guide for full details].
- READ Session one (pp3.1-3.21) in the User’s guide and Chapter 1 (pp1.1-1.28) in the Reference Manual.
- A brief introduction of how to get into MINITAB and manipulate the data will be given in the class.
- Remember that unless and until you try on your own, it is difficult to follow the class demonstration.

STEM & LEAF DISPLAY

Here the data is plotted in such a way the output will look like histogram and also features a frequency distribution. The basic idea is to use the digits of the data to illustrate its range, shape and density. Each observation is split into leading digits and trailing digits. All the leading digits are sorted and listed to the left of a vertical line. The trailing digits are written to the right of the vertical line in the appropriate places. The main steps are: (a) Choose a suitable unit to divide the data into leading and trailing digits; (b) List the sets of all possible leading digits in a column in ascending order; (c) For each observation list the trailing digits on the same line.

```
Stem-and-leaf of MPG_C      N  = 93
Leaf Unit = 1.0

    2    1 55
   13    1 66677777777
   35    1 8888888888889999999999
  (14)   2 00000000111111
   44    2 222222233333333
   29    2 44444555555
   18    2 66
   16    2 88999999
    8    3 011
    5    3 23
    3    3
    3    3
    3    3 9
    2    4
    2    4 2
    1    4
    1    4 6
```

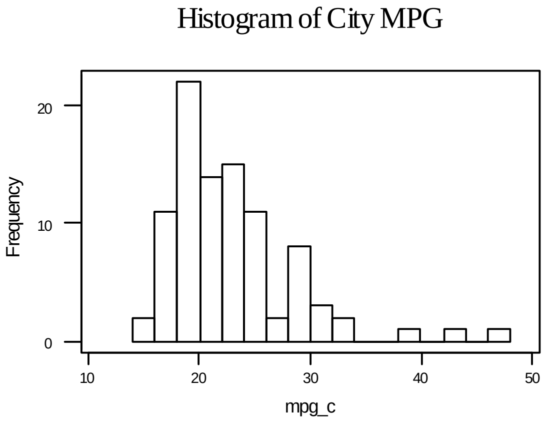
DOT PLOT

Here construct the horizontal axis (x-axis) covering the range of the data values and vertically plot the points, stacking any repeated values.

```
Size              .
Compact         : :  : .
             .: : :: : :.
-------+---------+---------+-------MPG_C
    18.0      24.0      30.0
```

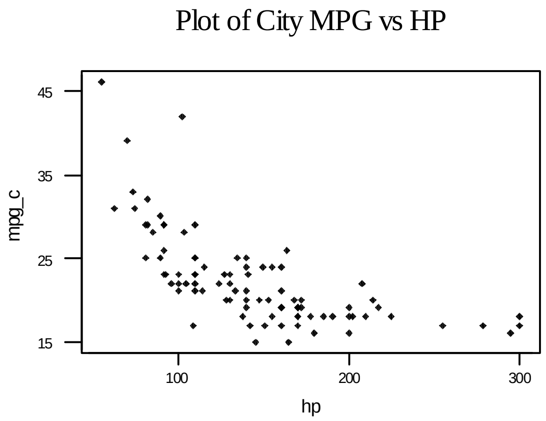
HISTOGRAM

This is a bar graph, where the data is grouped into many classes. The x-axis corresponds to the classes and the y-axis gives the frequency of the observations. Note that this is bar graph for numerical categories. The shape of the diagram provides an idea of the true population distribution. The construction of histogram depends on the number of classes. Usually one looks at the range of the data and then with the number of classes known, determines the width of the interval. Once, the frequency for each class is determined, plotting is done very easily.

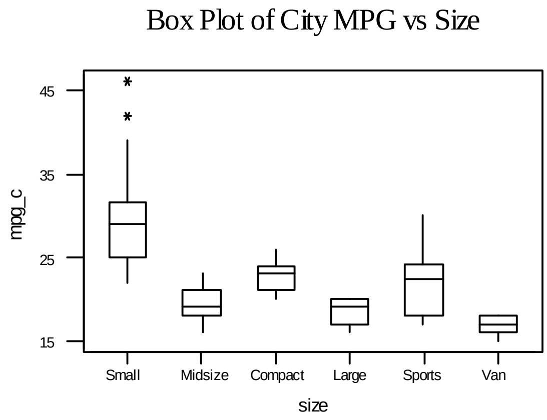


SCATTER PLOT

Construct x-axis and y-axis that cover the ranges of two variables. Plot (x_i, y_i) points for each observation in the data set.

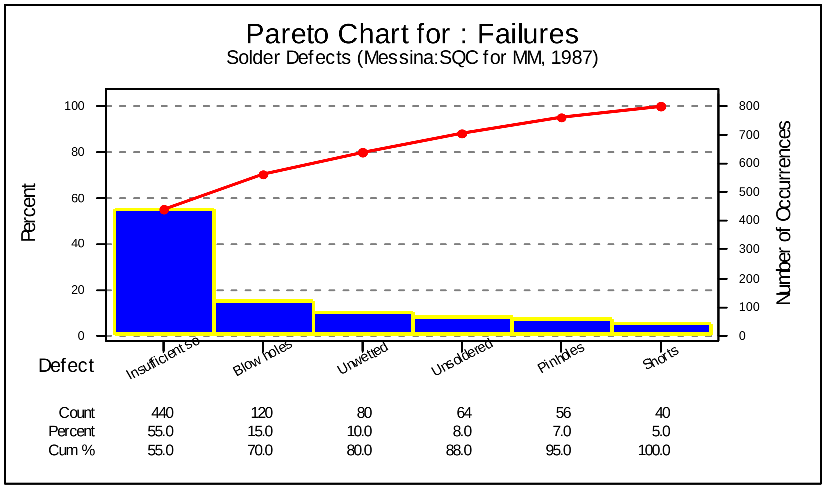


BOX PLOT



This is due to J. Tukey and provides a great deal of information about the data. This plot describes the data by a rectangle whose lower and upper limits are the first and third quartiles, respectively. The median (50th percentile or second quartile) is designated by a horizontal line segment inside the rectangle box. The average value is marked by a symbol such as "x" or "+". The width of the box is usually arbitrary. The main steps for a box plot are: (a) Calculate the mean and the quartiles of the data; (b) Calculate the inter-quartile range(IQR); (c) Draw a box with a convenient width such that the upper edge is at the third quartile, lower edge is the first quartile, a horizontal line within the box identifying the median, and put an "x" to identify the mean; (d) Draw vertical lines from the center of the upper and lower edges of the box to the most extreme data values (that are no farther than 3 IQR); (e) Plot all points that are more extreme than the vertical lines.

PARETO CHART



MEASURES OF LOCATION

Suppose that X_i, 1 # i # n, denotes the i-th observation of the sample taken from the population under study.

Mean: is used very often in analyzing the data. Although this is a common measure, if the data vary greatly the average may take a non-typical value and could be misleading.

Median: is the halfway point of the data and tells us something about the location of the distribution of the data.

Mode: if exists, gives the data point that occur most frequently. It is possible for a set of data to have 0, 1 or more modes.

- Mean and median always exist.
- Mode need not exist.
- Median and mode are less sensitive to extreme observations.
- Mean is most widely used.
- There are some data set for which median or mode may be more appropriate than mean.

Percentiles: The 100p-th percentile of a set of data is the value below which a proportion p of the data points will lie.

- Percentiles convey more information and are very useful in setting up warranty or guarantee periods for manufactured items.
- Also referred to as quantiles.

  The shape of the frequency data can be classified into several classes.
- Symmetric: mean = median = mode
- Positively skewed: tail to the right; mean > median
- Negatively skewed:tail to the right; median > mean
- In problems, such as waiting time problems one is interested in the tails of the distributions.
- For skewed data median is preferred to the mean.

MEASURES OF SPREAD

As pointed out before, one should not solely rely on mean or median or mode. Also two or more sets of data may have the same mean but they may be qualitatively different. In order to make a meaningful study, we need to rely on other measures. For example, we may be interested to see how the data is spread.

Range: is the difference between the largest and the smallest observations.

- Quick estimate on the standard deviation.
- Plays an important role in SPC.

Standard deviation: describes how the data is spread around its mean.

Coefficient of variation: The measures we have seen so far depend on the unit of measurements. It is sometimes necessary and convenient to have a measure that is independent of the unit and such a useful and common measure is given by the ratio of the standard deviation to the mean called the coefficient of variation.

Interquartile range: is the difference between the 75th and 25th percentiles.

- Gives the interval which contains the central 50 % of the observations.
- Avoids the total dependence on extreme data

HOW TO TAKE SAMPLES?

So far we discussed the computation of some basic numerical measures given a set of data. Most of the times we are interested in making inferences about the population(s) under study.

- The inference is very much dependent on the sample(s) drawn from the population(s).
- Much care should be devoted to the sampling.
- There is always going to be some error involved in making inferences about the populations based on the samples.
- The goal is to minimize this error as much as possible.
- There are many ways of bringing in systematic bias (consistently misrepresent the population).
- This can be avoided by taking random samples.

Simple random sample: all units are equally likely to be selected.

Multi-stage sample: units are selected in several stages.

Cluster sample: is used when there is no list of all the elements in the population and the elements are clustered in larger units.

Stratified sample: In cases where population under study may be viewed as comprising different groups (stratas) and where elements in each group are more or less homogeneous, we randomly select elements from every one of the strata.

Convenience sample: samples are taken based on convenience of the experimenter.

Systematic sample: units are taken in a systematic way such as selecting every 10th item after selecting the first item at random.

HOW TO USE SAMPLES?

- Samples should represent the population.
- Random sample obtained will not always be an exact copy of the population.
- Thus, there is bound to be some error:

1. **Random or unbiased error:** This is due to the random selection of the sample and the mean of such error will be 0 as positive deviation and negative deviation cancel out. This random error is also referred to as random deviation and is measured by the standard deviation of the estimator.
2. **Non-random or biased error:** this occurs due to several sources such as human, machines, mistakes due to copying or punching, recording and so on. Through careful planning we should try to avoid or minimize this error.

EXPERIMENTS USING MINITAB

We will illustrate the concepts of sample, sampling error, etc with practical data using MINITAB. See class lecture for details.