BUSINESS ANALYTICS

BUS 578 MASTER OF BUSINESS ADMINISTRATION (MBA)

KINGS UNIVERSITY, HONOLULU, HAWAII,
UNITED STATES OF AMERICA

LEARNING OUTCOMES

Upon the successful completion of this course the student shall be able to understand the concepts of Business Analytics, the tools used,

Table of Contents

Business Analytics

Analytics can be defined as the science of analysis, that is, the analysis of data and methods and software tools used for the same.

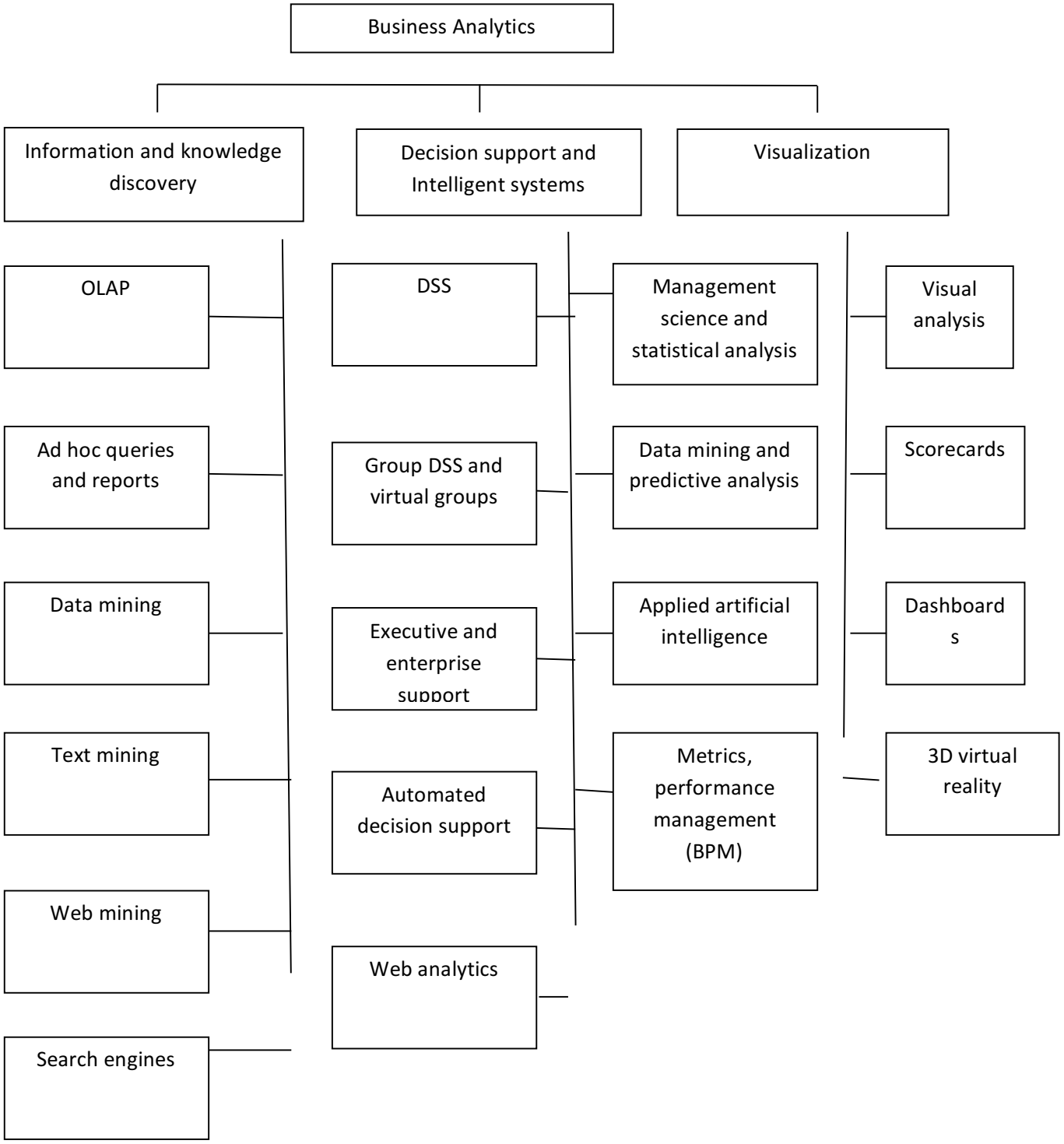
Business analytics can be defined as the applications and techniques for gathering, storing, analyzing and providing access to data to help users make better and strategic decisions (is also known as analytical processing, business intelligence tools or business intelligence applications).

Categories of analytic tools and techniques:

1. Information and knowledge discovery,

2. Decision support and Intelligent systems,

3. Visualization



MicroStrategy's Classification of Business Analytic tools (styles of Business Intelligence)

1. Enterprise reporting products – generate highly formatted static reports,

2. Cube analysis – provide Online Analytical Processing (OLAP) multidimensional slice-and-dice analytical capabilities,

3. Ad hoc querying and analysis – relational OLAB tools are used to allow power users to query a database for any answer, slice-and-dice the entire database, and drill down to the lowest level of transactional information,

4. Statistical analysis and data mining – perform predictive analysis and discover the cause-and-effect correlation between two metrics,

5. Report delivery and alerting – send full reports or alerts to large user populations, based on subscriptions, schedules or threshold events in the database.

Predictive Analytics Software (for better decisions): Discover the risks that could threaten the survival of the business, Don't waste money collecting data - use it to make better decisions, Predictive analytics without having to hire Stanford graduates:

Sophisticated Data Analysis The data modeling process begins with analysis. Segmenting your data, understanding what fields are important, what fields are not, selecting the right demographic profile that you know will have the greatest impact is often the most demanding task.

Drag-and-drop Simplicity It begins with a user interface that is 100% drag-and-drop enabled. Selecting your fields to analyze is as simple as a drag-and-drop operation. Or double click your fields in rapid sequence to add as many as you need to your data analysis process.

Instant Analysis The baseline statistical analysis is performed in seconds. Once your baseline analysis is done, simply drag and drop the fields you wish to analyze further onto the charting areas. Insight! instantly renders the perfect pie or bar charts for you, giving you the ability to quickly understand relationships, identify outliers and narrow your list of demographic fields down to the few that count.

Compare anything Comparing multiple fields is just as simple. Drop two or more fields into the chart areas and see how the two fields compare - giving you instant

feedback and the ability to understand how two data elements can work together quickly and intuitively.

Dynamic Filtering Dynamic filtering methods let you further reduce your data sets to only the records you need - once again giving you the power and flexibility to identify the data that makes sense and that will have the greatest impact in seconds.

Share Your Analysis Getting to the bottom of your analysis often means collaborating with others. With Insight! not only can you share your results with other Insight! users, you can also send your data to other users as Excel files - complete with data and charts - to give them an opportunity to further review and comment on your analysis - whether they have Insight! or not.

Data Modeling for the Rest of Us Understanding the key drivers for your business ultimately comes down to understanding your data. Your data stores months and years of information on what went right, and what went wrong, modeling that data reveals the reasons why.

An Easier Way Data mining and predictive analytics software has been available for years. But unfortunately you typically need a PhD just to be able to get started. With Insight! nothing could be further from the truth. That's because we have made it simpler. Simpler by using two of the most intuitive - and accurate - modeling algorithms available (decision tree analysis and rule induction), and simpler because you use them without even knowing it.

Wizard Driven It starts with the wizard driven interface that simply walks you through a series of questions about the data you are using. Set a few settings, check a few boxes and your data models will begin to appear before your eyes. Yes, it's really that simple.

Set the Boundaries Getting the most out of any data model, however, is about getting it "just" right. That's why in Insight! you can set the key drivers that build your model to suit your desires, giving you the ability to modify your models as you see fit.

Try, Test, Repeat The process of building and using data models is, by its very nature, a repetitive one. With Insight! changing your models around to account for new options, new scenarios and new assumptions is as simple as clicking on the appropriate pane, changing the fields you wish to analyze and repeating the process. Because of the speed with which Insight! executes the analysis, you can work

through any number of analyses until you arrive at the models that best suit your needs.

Share your Results Sharing your results with others is equally simple and intuitive with Insight!. Simply share the project file with other users and they will be able to take full advantage of your model in what-if analysis and target analysis regardless of whether they have access to the original data source or not. Sharing with non-Insight! users is just as simple; export your results in a formatted report or an Excel file and share your results with anyone.

Deploying Your Data Model The entire point of data mining is to be able to create business models that help you drive decision making to reduce risk and augment your ability to benefit from opportunities that present themselves to your business. Taking that step with traditional data mining software, however, can be daunting, expensive and work intensive.

An Easier Way Insight! makes the process of deploying your lessons quick and easy through a built-in "target analysis" engine. Simply load a data set you wish to use your business model on, use the built-in matching wizard to match up your model with your data and click go - Insight! will automatically show you all the records in your data sample that match the business rules you created. Using built-in matching is an invaluable method of deploying data mining in the real world. For example, let's say you have created a set of models that describe customers who are at highest risk for defaulting on their outstanding debt. Using the built-in target analysis you can load a list of your current prospects, select the business model you wish to use for analysis and in seconds Insight! will show you which prospects from within your list of hundreds or thousands of prospective customers are at highest risk for defaulting. How useful would that information be before you close the deal with your prospect?

What-if Scenario Analysis With Insight! testing your data models is simple and intuitive with a built-in What-if scenario analyzer that quickly lets you identify the statistical or algorithmic probability of your "what if" scenario according to the model you created.

Automated Matching Insight! lets you quickly apply your data models to every day business scenarios with a built-in what-if engine that lets you input sample data and automatically identifies statistical and algorithmic probabilities. Learn in seconds what element in your model a sample customer might fit, change a few parameters and watch the models realign automatically - all in just seconds.

Graphical Representation Understanding your data is always best done in graphical format. With Insight! that's exactly what you can do because it automatically shows you the matching data in a graphical representation, giving you the ability to quickly see how your data models are being used.

Classification of Strategic enterprise Management: Operational, Managerial and Strategic

1. Operational – ERP mainly support transaction processing on the operational level;

2. Managerial – access reports, arranged by functional areas – can make queries and drill down;

3. Strategic – SAP-SEM (Strategic Enterprise Management (includes BA).

BI activities evolved from two tools:

1. Executive information systems – EIS = a computer-based system – serves the information needs of the top executives: exception reporting and drill-down;

2. Executive support systems – ESS = comprehensive support system that goes beyond EIS to include analysis support, communications, office automation and intelligence support;

Online Analytical Processing (OLAP):

• refers to a variety of activities usually performed by end users in online systems;

• usually includes such activities as generating and answering queries, requesting ad hoc reports and executing them, conducting traditional or modern statistical analyses and building visual presentations; • multidimensional analysis and presentations, EIS/ESS and data mining;

• provide modeling, analysis and visualisation capabilities to large data sets, either to database management systems (DBMS), data warehouse systems, and a multidimensional conceptual view of the data.

Characteristics of OLAP tools:

• Categorical analysis – static analysis based on historical data;

• Exegetical analysis – base on historical data and drill-down analysis (query further into data to determine the detail data that were used to determine a derived value);

• Contemplative analysis – allows a user to determine a derived value;

• Formulaic analysis – permits changes to multiple variables;

Types of OLAP:

• Multidimensional (MOLAP) – cub structure the user can rotate – queries are fast;

• Relational (ROLAP) – create multidimensional views on-the-fly; a large number of attributes – it can be easily placed in a cube structure;

• Database and Web (DOLAP and WOLAP) – refers to a relational database management system (RDBMS) - is designed to host OLAP structures and perform OLAP calculations; Web OLAP refers o OLAP data that is accessible from a Web browser.

• Desktop OLAP involves low-priced, simple OLAP tools that perform local multidimensional analysis and presentation of data downloaded to client machines from relational or multidimensional database; can move desktop processing to an intermediate server which increases the scalability.

Reports and queries:

• The activities of OLAP and BI are using reports and queries. OLAP reporting must be uniform, flexible and adjustable.

Types of reports:

1. Rutine Reports – are generated automatically and distributed periodically to subscribers on mailing lists;

2. Ad Hoc (On-Demand) Reports – are created for a specific user whenever needed

Software examples:

• Business Objects' Crystal Reports – tool kit that helps in creating flexible, featurerich reports and integrates them into Web and Windows applications;

• Micro Strategy – provides monitoring and report creation tools for production and operational reports;

• Cognos & Business Intelligence – includes a complete list of serf-serve report types, adaptable to any data source;

• Hyperion – provides a full spectrum of management reporting capabilities that combine both operational and financial information;

• Microsoft – has included Report Builder – a user-friendly feature that allows report creation or modification.

Multidimensionality (for analysis and presentation):

• Factors for Multidimensional presentation: Dimensions, Measures and Time.

• Data cube: represent data along some measure of interest – each dimension represents some attribute in the database (the cells represent measure of interest); Cube analysis lets people perform queries by flipping through a series of report views.

Advanced Business Analytics

• Data mining and Predictive Analysis (multiple regression analysis, special forecasting and prediction methods)

• Data mining tools extract hidden, predictive information from database and search for the patterns in large transaction database;

• Predictive Analysis tools determine the probable future outcome for an event or the likelihood of the situation occurring, and identify relationships and patterns.

Tools for Advanced Analytics

• MicroStrategy - more than 400 statistical, mathematical and financial function for creating reports and analyzing their results.

• Hyperion's System – includes the Essbase Analytics module for quickly performing sophisticates analyses to interpret complex data;

• Cognos & Business Intelligence analytics – includes customizable time-series analysis & trends, deep competitive analysis, drill-down, forecasting and optimization;

• Microsoft – offers advanced analytics in its Microsoft Dynamics;

• Fair Isaac – offers intelligent tools for conducting risk analysis, fraud detection, profitability analysis and intelligent querying;

• ILOG – offer optimisation (maximize resource utilization, cost-benefit analysis).

Data Visualization

Visual technologies make decision support application more attractive and understandable to users.

Data visualization refers to technologies that support visualization and interpretation of data and information. It includes digital images, GIS, Graphical user interface, graphs, virtual reality, dimensional presentations, videos and animation. Visual tools can help idenftify relationships such as trends.

Graphic interface design for DSS

GRDSS (Geographic Resources Decision Support System) is an open source GIS based on GRASS (Geographic Resources Analysis Support System) that has functionalities such as raster, topological vector, image processing, graphics production, etc. It operates through a GUI developed in Tcl/Tk under LINUX. GRDSS include options such as Import / Export (of different data formats) including extraction of individual bands from the IRS (Indian Remote Sensing Satellites) data, display, digital image processing, map editing, raster analysis, vector analysis, point analysis, spatial query, geo-visualisation tools etc.

DESCRIPTIVE STATISTICS

Introduction

Statistics is concerned with the scientific method by which information is collected, organised, analysed and interpreted for the purpose of description and decision making.

Examples using statistics are: Hang Seng Index, Life or car insurance rate, Unemployment rate, Consumer Price Index, etc.

There are two subdivisions of statistical method.

(a) Descriptive Statistics - It deals with the presentation of numerical facts, or data, in either tables or graphs form, and with the methodology of analysing the data.

(b) Inferential Statistics - It involves techniques for making inferences about the whole population on the basis of observations obtained from samples.

Some Basic Definitions

(a) Population - A population is the group from which data are to be collected.

(b) Sample - A sample is a subset of a population.

(c) Variable - A variable is a feature characteristic of any member of a population differing in quality or quantity from one member to another.

(d) Quantitative variable - A variable differing in quantity is called quantitative variable, for example, the weight of a person, number of people in a car.

(e) Qualitative variable - A variable differing in quality is called a qualitative variable or attribute, for example, color, the degree of damage of a car in an accident.

(f) Discrete variable - A discrete variable is one which no value may be assumed between two given values, for example, number of children in a family.

(g) Continuous variable - A continuous variable is one which any value may be assumed between two given values, for example, the time for 100-meter run.

Measures of Central Tendency

According to Prof Bowley “Measures of central tendency (averages) are statistical constants which enable us to comprehend in a single effort the significance of the whole.”

The main objectives of Measure of Central Tendency are

1) To condense data in a single value.

2) To facilitate comparisons between data.

There are different types of averages, each has its own advantages and disadvantages.

Requisites of a Good Measure of Central Tendency:

1. It should be rigidly defined.

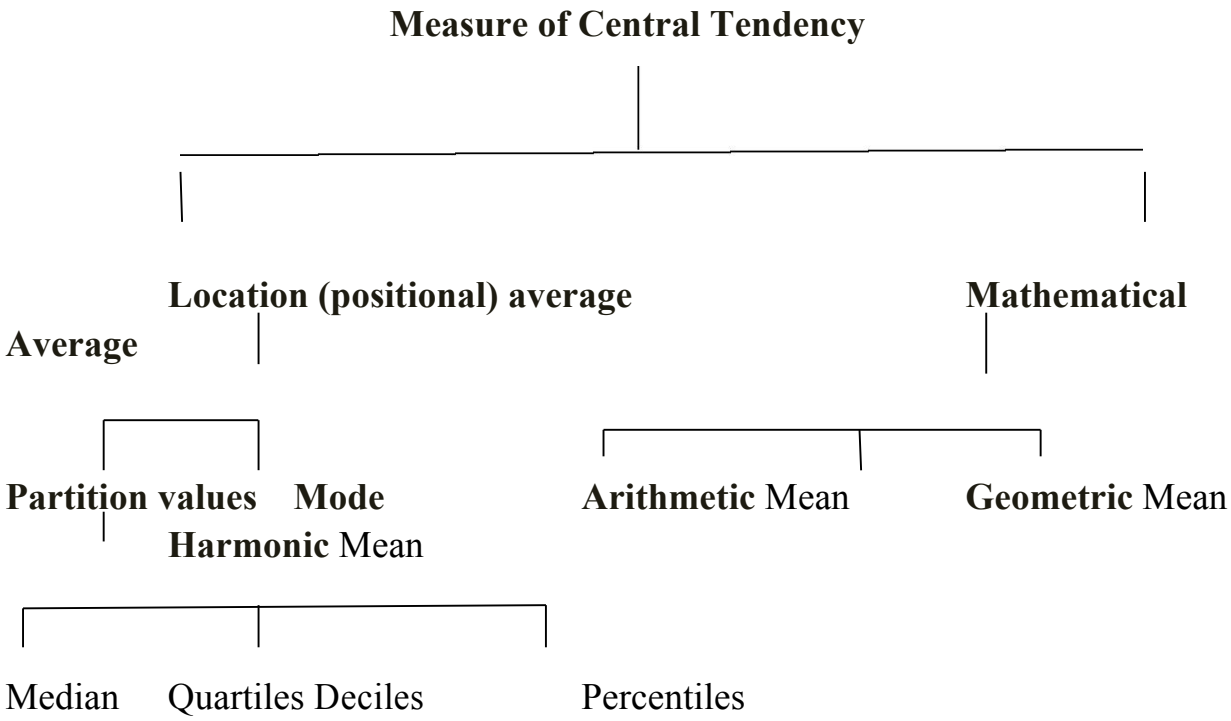
2. It should be simple to understand & easy to calculate.

3. It should be based upon all values of given data.

4. It should be capable of further mathematical treatment.

5. It should have sampling stability.

6. It should be not be unduly affected by extreme values



Partition values: The points which divide the data in to equal parts are called Partition values.

Median: The point or the value which divides the data in to two equal parts., or

when the data isarranged in numerical order

The data must be ranked (sorted in ascending order) first. The median is the number in the middle. Depending on the data size we define median as It is the middle value when data size N is odd. It is the mean of the middle two values, when data size N is even.

Ungrouped Frequency Distribution

Find the cumulative frequencies for the data. The value of the variable corresponding to which a cumulative frequency is greater than (N+1)/2 for the first time. (Where N is the total number of observations.)

Example 1: Find the median for the following frequency distribution.

X	1	2	3	4	5	6	7	8	9
Freq	8	10	11	16	20	25	15	9	6

Solution: Calculate cumulative frequencies less than type.

X	1	2	3	4	5	6	7	8	9
Freq	8	10	11	16	20	25	15	9	6
Cum freq	8	18	29	45	65	90	105	114	120

N=120, (N+1)/2=60.5 this value is first exceeded by cumulative frequency 65 , this value is corresponding to X-value 5, hence median is 5

Grouped Frequency Distribution First obtain the cumulative frequencies for the data. Then markthe class corresponding to which a cumulative frequency is greater than (N)/2 for the first time. (N is the total number of observations.) Then that class is median class. Then median is evaluated by interpolation formula.

$$median = l_1 + (l_2 - l_1)\frac{(\frac{N}{2} - cf)}{f_m}$$

Where l_1= lower limit of the median class, l_2= upper limit of the median class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the median class.

f_m= frequency of the median class.

Quartiles : The data can be divided in to four equal parts by three points. These three points areknown as quartiles. The quartiles are denoted by Q_i , i = 1,2,3
Qi is the value corresponding to $(iN/4)^{th}$ observation after arranging the data in the increasing order.

For grouped data : First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/4 for the first time. (Where N is total number of observations.). Then that class is Qi class. Then Qi is evaluated by interpolation formula.

$$Qi = l_1 + (l_2 - l_1)\frac{(\frac{iN}{4} - cf)}{f_q} \qquad i= 1, 2, 3$$

Where l_1= lower limit of the Qi class, l_2= upper limit of the Qi class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Qi class.

f_q= frequency of the Qi class.

Deciles are nine points which divided the data in to ten equal parts.

Di is the value corresponding to $(iN/10)^{th}$ observation after arranging the data in the increasing order.

For grouped data :First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/10 for the first time. (Where N is total number of observations.). Then that class is Di class. Then Di is evaluated by interpolation formula.

$$Di = l_1 + (l_2 - l_1)\frac{(\frac{iN}{10} - cf)}{f_d} \qquad i= 1, 2, \ldots\ldots\ldots\ldots 10.$$

Where l_1= lower limit of the Di class, l_2= upper limit of the Di class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Di class.

f_d= frequency of the Di class.

Percentiles are ninety-nine points which divided the data in to hundred equal parts.

Pi is the value corresponding to $(iN/100)^{th}$ observation after arranging the data in the increasing order.

For grouped data : First obtain the cumulative frequencies for the data. Then mark the class corresponding to which a cumulative frequency is greater than (iN)/100 for the first time. (Where N

is total number of observations.) Then that class is Pi class. Then Pi is evaluated by interpolation formula.

$$Pi = l_1 + (l_2 - l_1)\frac{(\frac{iN}{100} - cf)}{f_p}$$

Where l_1= lower limit of the Pi class, l_2= upper limit of the Pi class

N= Number of observations.

cf = cumulative frequency of the class proceeding to the Pi class.

f_p= frequency of the Pi class.

Graphical method for locating partition values: These partition values can be located graphicallyby using ogives. The point of intersection of both ogives is median.

To locate quartiles, mark N/4 on Y- axis, from that point draw a line parallel to X-axis, it cuts less than type ogive at Q_1 and intersects greater than or equal to curve at Q_3.

To locate Di mark iN/10 on Y-axis , from that point draw line parallel to X-axis, it intersects less than type curve at Di.

Similarly to locate Pi mark iN/100 on Y-axis , from that point draw line parallel to X-axis, it intersects less than type curve at Pi.

Example**2** . Find the median

Daily wages in Rs	100-200	200-300	300-400	400-500	500-600	600-700
No workers	4	6	20	10	5	5

Solution : To locate median class we have to calculate cumulative frequencies

Daily wages in Rs	100-200	200-300	300-400	400-500	500-600	600-700
No workers	4	6	20	10	5	5
Cum Freq	4	10	30	40	45	90

N=50 , N/2= 25 so median class is 300-400

$$median = l_1 + (l_2 - l_1)\frac{\left(\frac{N}{2} - cf\right)}{f_m} = 300 + (400 - 300)\frac{(25-10)}{20} = 300 + 100 * \frac{15}{20} = 375$$

Merits of Median

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is not affected by extreme values.
4. Even if extreme values are not known median can be calculated.
5. It can be located just by inspection in many cases.
6. It can be located graphically.
7. It is not much affected by sampling fluctuations.
8. It can be calculated for data based on ordinal scale.

Demerits of Median

1. It is not based upon all values of the given data.
2. For larger data size the arrangement of data in the increasing order is difficult process.
3. It is not capable of further mathematical treatment.
4. It is insensitive to some changes in the data values.

MODE

The mode is the most frequent data value. Mode is the value of the variable which is predominant in the given data series. Thus in case of discrete frequency distribution, mode is the value corresponding to maximum frequency. Sometimes there may be no single mode if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).

For grouped frequency distributions, the modal class is the class with the largest frequency. After identifying modal class mode is evaluated by using interpolated formula. This formula is applicable when classes are of equal width.

$$mode = l_1 + (l_2 - l_1)\frac{d1}{d1 + d2}$$

Where l_1= lower limit of the modal class,

l_2= upper limit of the modal class"

d1 =fm-f_0 and d2=fm-f_1

where fm= frequency of the modal class,

f_0 = frequency of the class preceding to the modal class,

f_1= frequency of the class succeeding to the modal class.

Mode can be located graphically by drawing histogram.
Steps:

1) Draw histogram
2) Locate modal class (highest bar of the histogram
3) Join diagonally the upper end points of the end points of the highest bar to the adjacent bars.
4) Mark the point of intersection of the diagonals.
5) Draw the perpendicular from this point on the X-axis .
6) The point where the perpendicular meets X-axis gives the modal value.

Merits of Mode

1. It is easy to understand & easy to calculate.
2. It is not affected by extreme values or sampling fluctuations.
3. Even if extreme values are not known mode can be calculated.
4. It can be located just by inspection in many cases.
5. It is always present within the data.
6. It can be located graphically.

7. It is applicable for both qualitative and quantitative data.

Demerits of Mode

1. It is not rigidly defined.
2. It is not based upon all values of the given data.
3. It is not capable of further mathematical treatment.

Arithmetic Mean

This is what people usually intend when they say "average"

Sample mean: If $X_1, X_2, \ldots\ldots\ldots\ldots\ldots X_n$ are data values then arithmetic mean is given by

$$\bar{X} = \frac{X_1 + X_2 + \cdots . + X_n}{n} = \frac{\sum_1^n Xi}{n}$$

Frequency Distribution: Let $X_1, X_2, \ldots\ldots\ldots\ldots\ldots X_n$ are class marks and the corresponding frequencies are $f_1, f_2, \ldots\ldots\ldots\ldots f_n$, then arithmetic mean is given by

$$N = \sum fi$$

$$\bar{X} = \frac{\sum_1^n fiXi}{N}$$

Example 5 : The Marks obtained in 10 class tests are 25, 20, 20, 9, 16, 10, 21, 12, 8, 13.

The mean $= \bar{X} = \frac{25+20+20+9+16+10+21+12+8+13}{10} = \frac{154}{10}$

Properties of Mean:

1) Effect of shift of origin and scale.
   If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values . New values U are obtained by shifting the origin to „a" and changing scale by „h"
   $Ui = \frac{Xi-a}{h}$ then Mean= $\bar{X} = a + hU$
2) Algebraic sum of deviations of set of values taken from their mean is zero.
   a. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values then
   b. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values with corresponding frequencies f1, f2,……fn then

3) The sum of squares of deviation of set of values about its mean is minimum.

$$\sum_1^n((Xi-\bar{X})^2) < \sum_1^n(Xi-A)^2 \qquad \text{where } A \neq \bar{X}$$

4) If $Zi = Xi \pm Yi$ i= 1,2 …..n
then $\bar{Z} = \bar{X} \pm \bar{Y}$

5) If $\overline{X1}$ *and* $\overline{X2}$ are the means of two sets of values containing n_1 and n_2 observations respectively then the mean of the combined data is given by $\bar{X} = \frac{n_1}{n_1+n_2} \overline{X1} + n_2 \overline{X2}$

This formula can be extended for k sets of data values as

$$\bar{X} = \frac{n_1\overline{X1} + n_2\overline{X2} + \cdots . + n_k\overline{Xk}}{n_1 + n_2 + \cdots \ldots + n_k}$$

Merits of Mean

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is based upon all values of the given data.
4. It is capable of further mathematical treatment.
5. It is not much affected by sampling fluctuations.

Demerits of Mean

1. It cannot be calculated if any observations are missing.
2. It cannot be calculated for the data with open end classes.
3. It is affected by extreme values.
4. It cannot be located graphically.
5. It may be number which is not present in the data.
6. It can be calculated for the data representing qualitative characteristic.

Empirical formula: For symmetric distribution Mean, Median and Mode coincide. If thedistribution is moderately asymmetrical the Mean, Median and Mode satisfy the following relationship

Mean- Mode =3(Mean-Median) Or Mode=3Median-2Mean

Weighted mean : If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$are given values with corresponding weights $W_1, W_2, \ldots\ldots W_n$ then the weighted mean is given by

$$\bar{X}_w = \frac{\sum_1^n WiXi}{\sum_1^n Wi}$$

The mean of a frequency distribution is also the weighted mean.

Use the mean to describe the middle of a set of data that *does not* have an outlier.

Geometric Mean:

a. If $X_1, X_2 \ldots\ldots\ldots\ldots\ldots X_n$ are given values then

$$GM = \sqrt[n]{X1 * X2 \ldots\ldots Xn}$$

Or GM= antilog$\left(\frac{\sum_{1}^{n} logXi}{n}\right)$

b. If X_1, X_2……X_n are given values with corresponding frequencies f1, f2,……fn then
if N=∑fi

$$GM = \sqrt[N]{X1^{f1} * X2^{f2} * \dots\dots Xn^{fn}}$$

GM= antilog$\left(\frac{\sum_{1} fi\ logXi}{N}\right)$

Merits of Geometric Mean

1. It is based upon all values of the given data.
2. It is capable of further mathematical treatment.
3. It is not much affected by sampling fluctuations.

Demerits of Geometric Mean

1. It is not easy to understand & not easy to calculate
2. It is not well defined.
3. If anyone data value is zero then GM is zero.
4. It cannot be calculated if any observations are missing.
5. It cannot be calculated for the data with open end classes.
6. It is affected by extreme values.
7. It cannot be located graphically.
8. It may be number which is not present in the data.
9. It cannot be calculated for the data representing qualitative characteristic

Harmonic Mean:

a. If X_1, X_2……X_n are given values then Harmonic Mean is given by

$$HM = \frac{n}{\sum \frac{1}{Xi}}$$

b. If X_1, X_2……X_n are given values with corresponding frequencies f1, f2,……fn then Harmonic Mean given by
if N=∑fi

$$HM = \frac{N}{\sum \frac{fi}{Xi}}$$

Merits of Harmonic Mean

1. It is rigidly defined.
2. It is easy to understand & easy to calculate.
3. It is based upon all values of the given data.
4. It is capable of further mathematical treatment.
5. It is not much affected by sampling fluctuations.

Demerits of Harmonic Mean

1. It is not easy to understand & not easy to calculate.
2. It cannot be calculated if any observations are missing.
3. It cannot be calculated for the data with open end classes.
4. It is usually not a good representative of the data.
5. It is affected by extreme values.
6. It cannot be located graphically.
7. It may be number which is not present in the data.
8. It can be calculated for the data representing qualitative characteristic.

Selection of an average:

No single average can be regarded as the best or most suitable under all circumstances. Each average has its merits and demerits and its own particular field of importance and utility. A proper selection of an average depends on the 1) nature of the data and 2) purpose of enquiry or requirement of the data.

A.M. satisfies almost all the requisites of a good average and hence can be regarded as the best average but it cannot be used

1) in case of highly skewed data.
2) in case of uneven or irregular spread of the data.
3) in open end distributions.
4) When average growth or average speed is required.
5) When there are extreme values in the data.

Except in these cases AM is widely used in practice.

Median: is the best average in open end distributions or in distributions which give highly skew orj or reverse j type frequency curves. In such cases A.M. gives unnecessarily high or low value whereas median gives a more representative

value. But in case of fairly symmetric distribution there is nothing to choose between mean, median and mode, as they are very close to each other.

Mode : is especially useful to describe qualitative data. According to Freunel and Williams,consumer preferences for different kinds of products can be compared using modal preferences as we cannot compute mean or median. Mode can best describe the average size of shoes or shirts.

G.M. is useful to average relative changes, averaging ratios and percentages. It is theoretically thebest average for construction of index number. But it should not be used for measuring absolute changes.

H.M. is useful in problems where values of a variable are compared with a constant quantity ofanother variable like time, distance travelled within a given time, quantities purchased or sold over a unit.

In general we can say that A.M. is the best of all averages and other averages may be used under special circumstances.

Karl Pearson's Coefficient of Correlation

- Pearson's 'r' is the most common correlation coefficient.
- Karl Pearson's Coefficient of Correlation denoted by- 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y.
- Karl Pearson's Coefficient of Correlation denoted by- r $-1 \leq r \geq +1$
- Degree of Correlation is expressed by a value of Coefficient
- Direction of change is Indicated by sign (- ve) or (+ ve)
- When deviation taken from actual mean: $r(x, y)= \Sigma xy / \sqrt{\Sigma x^2 \Sigma y^2}$
- When deviation taken from an assumed mean:
  $r = N \Sigma dxdy - \Sigma dx \Sigma dy \sqrt{N \Sigma dx^2-(\Sigma dx)^2} \sqrt{N \Sigma dy^2-(\Sigma dy)^2}$

Procedure for computing the correlation coefficient

1. Calculate the mean of the two series 'x' &'y'
2. Calculate the deviations 'x' &'y' in two series from their respective mean.

3. Square each deviation of ‘x’ &’y’ then obtain the sum of the squared deviation i.e. ∑ x 2 & . ∑ y 2
4. Multiply each deviation under x with each deviation under y & obtain the product of ‘xy’.Then obtain the sum of the product of x , y i.e. ∑xy
5. Substitute the value in the formula.

Interpretation of Correlation Coefficient (r)

1. The value of correlation coefficient ‘r’ ranges from -1 to +1
2. If r = +1, then the correlation between the two variables is said to be perfect and positive
3. If r = -1, then the correlation between the two variables is said to be perfect and negative
4. If r = 0, then there exists no correlation between the variables

Properties of Correlation coefficient „

- The correlation coefficient lies between -1 & +1 symbolically (- 1 ≤ r ≥ 1)
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient.
- r = √ bxy * byx
- The one regression coefficient is (+ve) other regression coefficient is also (+ve) correlation coefficient is (+ve)

Assumptions of Pearson’s Correlation Coefficient „

There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points. „

Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson’s Coefficient

„ It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson's Coefficient „

- Always assume linear relationship „
- Interpreting the value of r is difficult. „
- Value of Correlation Coefficient is affected by the extreme values. „
- Time consuming methods

Coefficient of Determination

- The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination. „
- The Coefficient of Determination = r 2. „
- Suppose: r = 0.9, r2 = 0.81 this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.
- The maximum value of r2 is 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it. „
- Coefficient of Determination = Explained variation / Total variation

Coefficient of Determination: An example „

Suppose: r = 0.60 r = 0.30 It does not mean that the first correlation is twice as strong as the second the 'r' can be understood by computing the value of r2 . When

r = 0.60

r2 = 0.36 -----(1)

r = 0.30 r2 = 0.09 -----(2)

This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained.

Spearman's Rank Coefficient of Correlation

- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson's correlation coefficient can not be used in such case Spearman Rank correlation can be used.
- R = 1- (6 ∑ D2) / N (N 2 – 1)

1. R = Rank correlation coefficient
2. D = Difference of rank between paired item in two series.
3. N = Total number of observation.

Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from -1 to +1 „ If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction „
- If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction „
- If R = 0, then there is no correlation

Rank Correlation Coefficient (R)

a) Problems where actual rank are given.

1) Calculate the difference 'D' of two Ranks i.e. (R1 – R2).

2) Square the difference & calculate the sum of the difference i.e. ∑ D 2

3) Substitute the values obtained in the formula.

Rank Correlation Coefficient

b) Problems where Ranks are not given :If the ranks are not given, then we need to assign ranks to the data series. The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1. We need to follow the same scheme of ranking for the other series.

Then calculate the rank correlation coefficient in similar way as we do when the ranks are given.

Equal Ranks or tie in Ranks: In such cases average ranks should be assigned to each individual. R = 1- (6 ∑ D2) + AF / N (N 2 – 1)

AF = 1/12(m 1 3 – m 1) + 1/12(m 2 3 – m 2) +…. 1/12(m 2 3 – m 2) m = The number of time an item is repeated

Merits Spearman's Rank Correlation „

- This method is simpler to understand and easier to apply compared to karl pearson's correlation method. „
- This method is useful where we can give the ranks and not the actual data. (qualitative term) „
- This method is to use where the initial data in the form of ranks.

Limitation Spearman's Correlation „

- Cannot be used for finding out correlation in a grouped frequency distribution. „
- This method should be applied where N exceeds 30.

Advantages of Correlation studies „

- Show the amount (strength) of relationship present „
- Can be used to make predictions about the variables under study. „
- Can be used in many places, including natural settings, libraries, etc. „
- Easier to collect co relational data

Regression Analysis „

- Regression Analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable, given the value of another variable, when those variables are related to each other.
- Regression Analysis is mathematical measure of average relationship between two or more variables.
- Regression analysis is a statistical tool used in prediction of value of unknown variable from known variable.

Advantages of Regression Analysis „

- Regression analysis provides estimates of values of the dependent variables from the values of independent variables. „
- Regression analysis also helps to obtain a measure of the error involved in using the regression line as a basis for estimations. „
- Regression analysis helps in obtaining a measure of the degree of association or correlation that exists between the two variables.

Assumptions in Regression Analysis

- Existence of actual linear relationship. „
- The regression analysis is used to estimate the values within the range for which it is valid. „
- The relationship between the dependent and independent variables remains the same till the regression equation is calculated. „
- The dependent variable takes any random value but the values of the independent variables are fixed. „

- In regression, we have only one dependant variable in our estimating equation. However, we can use more than one independent variable.

Regression line „

1. Regression line is the line which gives the best estimate of one variable from the value of any other given variable. „
2. The regression line gives the average relationship between the two variables in mathematical form. „

The Regression would have the following properties:

a) ∑(Y – Y c) = 0 and

b) ∑(Y – Y c)2 = Minimum

For two variables X and Y, there are always two lines of regression

– Regression line of X on Y : gives the best estimate for the value of X for any specific given values of Y

X = a + b Y a = X - intercept „

- b = Slope of the line
- X = Dependent variable „
- Y = Independent variable

For two variables X and Y, there are always two lines of regression – „

Regression line of Y on X : gives the best estimate for the value of Y for any specific given values of X „

Y = a + bx

- a = Y - intercept
- b = Slope of the line
- Y = Dependent variable
- x= Independent variable

The Explanation of Regression Line „

- In case of perfect correlation (positive or negative) the two line of regression coincide. „
- If the two R. line are far from each other then degree of correlation is less, & vice versa. „
- The mean values of X &Y can be obtained as the point of intersection of the two regression line. „
- The higher degree of correlation between the variables, the angle between the lines is smaller & vice versa.

Regression Equation / Line & Method of Least Squares „

Regression Equation of y on x

$Y = a + bx$

In order to obtain the values of ‘a’ & ‘b’

$\sum y = na + b \sum x$

$\sum xy = a \sum x + b \sum x^2$

Regression Equation of x on y

$X = c + dy$

In order to obtain the values of ‘c’ & ‘d’

$\sum x = nc + d \sum y$

$\sum xy = c \sum y + d \sum y^2$

Regression Equation / Line when Deviation taken from Arithmetic Mean

Regression Equation of y on x:

$Y = a + bx$ In order to obtain the values of ‘a’ & ‘b’

$a = Y - bX$

b = ∑xy / ∑ x 2

Regression Equation of x on y:

X = c + dy

c = X – dY

d = ∑xy / ∑ y 2

Regression Equation / Line when Deviation taken from Arithmetic Mean

Regression Equation of y on x:

Y – Y = byx (X –X)

byx = ∑xy / ∑ x 2

byx = r (σy / σx)

Regression Equation of x on y:

X – X = bxy (Y –Y)

bxy = ∑xy / ∑ y 2

bxy = r (σx / σy)

Properties of the Regression Coefficients

- The coefficient of correlation is geometric mean of the two regression coefficients. r = √ byx * bxy
- If byx is positive than bxy should also be positive & vice versa.
- If one regression coefficient is greater than one the other must be less than one.
- The coefficient of correlation will have the same sign as that our regression coefficient.

- Arithmetic mean of byx & bxy is equal to or greater than coefficient of correlation. byx + bxy / 2 ≥ r
- Regression coefficient are independent of origin but not of scale.

Standard Error of Estimate. „

- Standard Error of Estimate is the measure of variation around the computed regression line. „
- Standard error of estimate (SE) of Y measure the variability of the observed values of Y around the regression line. „
- Standard error of estimate gives us a measure about the line of regression. of the scatter of the observations about the line of regression.

Standard Error of Estimate of Y on X is:

S.E. of Yon X (SExy) = $\sqrt{\sum}$(Y – Ye)2 / n-2

Y = Observed value of y

Y e = Estimated values from the estimated equation that correspond to each y value

e = The error term (Y – Y e) n = Number of observation in sample.

The convenient formula:

(SExy) = $\sqrt{\sum}$ Y2 _ a ∑ Y _ b ∑YX / n – 2

X = Value of independent variable.

Y = Value of dependent variable.

a = Y intercept.

b = Slope of estimating equation.

n = Number of data points.

Probability

Basic ideas

In this chapter, we don't really answer the question 'What is probability?' Nobody has a really good answer to this question. We take a mathematical approach, writing down some basic axioms which probability must satisfy, and making deductions from these. We also look at different kinds of sampling, and examine what it means for events to be independent.

Sample space, events The general setting is: We perform an experiment which can have a number of different outcomes. The sample space is the set of all possible outcomes of the experiment. We usually call it S.

It is important to be able to list the outcomes clearly. For example, if I plant ten bean seeds and count the number that germinate, the sample space is

S = {0,1,2,3,4,5,6,7,8,9,10}.

If I toss a coin three times and record the result, the sample space is

S = {HHH,HHT,HT H,HT T,T HH,T HT,T T H,T T T},

where (for example) HT H means 'heads on the first toss, then tails, then heads again'.

Sometimes we can assume that all the outcomes are equally likely. (Don't assume this unless either you are told to, or there is some physical reason for assuming it. In the beans example, it is most unlikely. In the coins example, the assumption will hold if the coin is 'fair': this means that there is no physical reason for it to favor one side over the other.) If all outcomes are equally likely, then each has probability 1/|S|. (Remember that |S| is the number of elements in the set S).

In calculating entropy by molecular-theoretic methods, the word "probability" is often used in a sense differing from the way the word is defined in probability theory. In particular, "cases of equal probability" are often hypothetically stipulated when the theoretical methods employed are definite enough to permit a deduction rather than a stipulation.

In other words: Don’t just assume that all outcomes are equally likely, especially when you are given enough information to calculate their probabilities!

An event is a subset of S. We can specify an event by listing all the outcomes that make it up. In the above example, let A be the event ‘more heads than tails’ and B the event ‘heads on last throw’. Then

A = {HHH,HHT,HT H,T HH},

B = {HHH,HT H,T HH,T T H}.

The probability of an event is calculated by adding up the probabilities of all the outcomes comprising that event. So, if all outcomes are equally likely, we have

P(A) = |A| |S| .

In our example, both A and B have probability 4/8 = 1/2. An event is simple if it consists of just a single outcome, and is compound otherwise. In the example, A and B are compound events, while the event ‘heads on every throw’ is simple (as a set, it is {HHH}). If A = {a} is a simple event, then the probability of A is just the probability of the outcome a, and we usually write P(a), which is simpler to write than P({a}). (Note that a is an outcome, while {a} is an event, indeed a simple event.)

We can build new events from old ones:

• A∪B (read ‘A union B’) consists of all the outcomes in A or in B (or both!)

• A∩B (read ‘A intersection B’) consists of all the outcomes in both A and B;

• A\B (read ‘A minus B’) consists of all the outcomes in A but not in B;

• A 0 (read ‘A complement’) consists of all outcomes not in A (that is, S \A);

• 0/ (read ‘empty set’) for the event which doesn’t contain any outcomes.

What is probability?

There is really no answer to this question.

Some people think of it as 'limiting frequency'. That is, to say that the probability of getting heads when a coin is tossed means that, if the coin is tossed many times, it is likely to come down heads about half the time. But if you toss a coin 1000 times, you are not likely to get exactly 500 heads. You wouldn't be surprised to get only 495. But what about 450, or 100?

Some people would say that you can work out probability by physical arguments, like the one we used for a fair coin. But this argument doesn't work in all cases, and it doesn't explain what probability means.

Some people say it is subjective. You say that the probability of heads in a coin toss is 1/2 because you have no reason for thinking either heads or tails more likely; you might change your view if you knew that the owner of the coin was a magician or a con man. But we can't build a theory on something subjective.

We regard probability as a mathematical construction satisfying some axioms (devised by the Russian mathematician A. N. Kolmogorov). We develop ways of doing calculations with probability, so that (for example) we can calculate how unlikely it is to get 480 or fewer heads in 1000 tosses of a fair coin. The answer agrees well with experiment.

Kolmogorov's Axioms

Remember that an event is a subset of the sample space S. A number of events, say A1,A2,..., are called mutually disjoint or pairwise disjoint if $A_i \cap A_j = 0/$ for any two of the events Ai and Aj ; that is, no two of the events overlap.

According to Kolmogorov's axioms, each event A has a probability P(A), which is a number. These numbers satisfy three axioms:

Axiom 1: For any event A, we have $P(A) \geq 0$.

Axiom 2: $P(S) = 1$.

Axiom 3: If the events A1,A2,... are pairwise disjoint, then

$P(A1 \cup A2 \cup \cdots) = P(A1) + P(A2) + \cdots$

Note that in Axiom 3, we have the union of events and the sum of numbers. Don't mix these up; never write $P(A1) \cup P(A2)$, for example. Sometimes we separate Axiom

3 into two parts: Axiom 3a if there are only finitely many events A1,A2,...,An, so that we have

$P(A1 \cup \cdots \cup An) = n \sum i=1 P(Ai)$,

and Axiom 3b for infinitely many. We will only use Axiom 3a, but 3b is important later on.

Notice that we write

$n \sum i=1 P(Ai)$ for $P(A1) + P(A2) + \cdots + P(An)$.

Proving things from the axioms

You can prove simple properties of probability from the axioms. That means, every step must be justified by appealing to an axiom. These properties seem obvious, just as obvious as the axioms; but the point of this game is that we assume only the axioms, and build everything else from that.

Here are some examples of things proved from the axioms. There is really no difference between a theorem, a proposition, and a corollary; they all have to be proved. Usually, a theorem is a big, important statement; a proposition a rather smaller statement; and a corollary is something that follows quite easily from a theorem or proposition that came before.

Proposition 1.1 If the event A contains only a finite number of outcomes, say A = {a1,a2,...,an}, then

$P(A) = P(a1) + P(a2) + \cdots + P(an)$.

To prove the proposition, we define a new event Ai containing only the outcome ai , that is,

Ai = {ai}, for i = 1,...,n. Then A1,...,An are mutually disjoint.

(each contains only one element which is in none of the others), and $A1 \cup A2 \cup \cdots \cup An = A$; so by Axiom 3a, we have

$P(A) = P(a1) + P(a2) + \cdots + P(an)$.

Corollary 1.2 If the sample space S is finite, say S = {a1,...,an}, then

P(a1) +P(a2) +···+P(an) = 1.

For P(a1) +P(a2) +···+P(an) = P(S) by Proposition 1.1, and P(S) = 1 by Axiom 2. Notice that once we have proved something, we can use it on the same basis as an axiom to prove further facts.

Now we see that, if all the n outcomes are equally likely, and their probabilities sum to 1, then each has probability 1/n, that is, 1/|S|. Now going back to Proposition 1.1, we see that, if all outcomes are equally likely, then

P(A) = |A|/ |S|

for any event A, justifying the principle we used earlier.

Proposition 1.3 P(A 0) = 1−P(A) for any event A.

Let A1 = A and A2 = A 0 (the complement of A). Then A1 ∩A2 = 0/ (that is, the events A1 and A2 are disjoint), and A1 ∪A2 = S. So

P(A1) +P(A2) = P(A1 ∪A2) (Axiom 3)

= P(S) = 1 (Axiom 2).

So P(A) = P(A1) = 1−P(A2).

Corollary 1.4 P(A) ≤ 1 for any event A.

For 1−P(A) = P(A 0) by Proposition 1.3, and P(A 0) ≥ 0 by Axiom 1; so 1− P(A) ≥ 0, from which we get P(A) ≤ 1.

Remember that if you ever calculate a probability to be less than 0 or more than 1, you have made a mistake!

Corollary 1.5 P(0/) = 0.

For 0/ = S 0 , so P(0/) = 1−P(S) by Proposition 1.3; and P(S) = 1 by Axiom 2, so P(0/) = 0.

Here is another result. The notation A ⊆ B means that A is contained in B, that is, every outcome in A also belongs to B.

Proposition 1.6 If $A \subseteq B$, then $P(A) \leq P(B)$.

This time, take $A1 = A$, $A2 = B \setminus A$. Again we have $A1 \cap A2 = 0/$ (since the elements of $B\setminus A$ are, by definition, not in A), and $A1 \cup A2 = B$. So by Axiom 3,

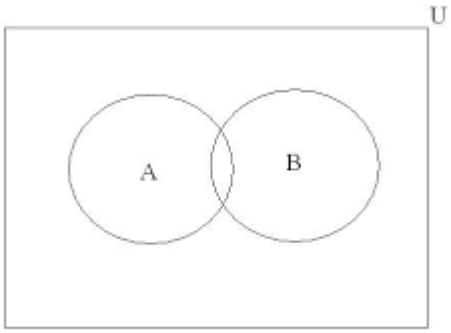
$P(A1) + P(A2) = P(A1 \cup A2) = P(B)$.

In other words, $P(A) + P(B\setminus A) = P(B)$. Now $P(B\setminus A) \geq 0$ by Axiom 1; so

$P(A) \leq P(B)$,

as we had to show.

Inclusion-Exclusion Principle



A Venn diagram for two sets A and B suggests that, to find the size of $A\cup B$, we add the size of A and the size of B, but then we have included the size of $A\cap B$ twice, so we have to take it off. In terms of probability:

$P(A\cup B) = P(A) + P(B) - P(A\cap B)$.

We now prove this from the axioms, using the Venn diagram as a guide. We see that $A\cup B$ is made up of three parts, namely

$A1 = A\cap B$,

$A2 = A\setminus B$,

$A3 = B\setminus A$.

Indeed we do have A∪B = A1 ∪A2 ∪A3, since anything in A∪B is in both these sets or just the first or just the second. Similarly we have A1∪A2 = A and A1∪A3 = B.

The sets A1,A2,A3 are mutually disjoint. (We have three pairs of sets to check. Now A1 ∩A2 = 0/, since all elements of A1 belong to B but no elements of A2 do. The arguments for the other two pairs are similar – you should do them yourself.)

SAMPLING

Now there is another issue, depending on whether we care about the order in which the pens are chosen. We will only consider this in the case of sampling without replacement. It doesn't really matter in this case whether we choose the pens one at a time or simply take two pens out of the drawer; and we are not interested in which pen was chosen first. So in this case the sample space is

{{R,G},{R,B},{R,P},{G,B},{G,P},{B,P}},

containing six elements. (Each element is written as a set since, in a set, we don't care which element is first, only which elements are actually present. So the sample space is a set of sets!) The event 'at least one red pen' is {{R,G},{R,B},{R,P}}, with probability 3/6 = 1/2. We should not be surprised that this is the same as in the previous case.

There are formulae for the sample space size in these three cases. These involve the following functions:

$n! = n(n-1)(n-2)\cdots 1$

$nPk = n(n-1)(n-2)\cdots(n-k+1)$

$nCk = nPk/k!$

Note that n! is the product of all the whole numbers from 1 to n; and

$nPk = n!/(n-k)!$,

so that

$nCk = n!/k!(n-k)!$

Theorem 1.10 The number of selections of k objects from a set of n objects is given in the following table.

	with replacement	without replacement
ordered sample	n k	nPk
unordered sample		nCk

In fact the number that goes in the empty box is n+k−1Ck, but this is much harder to prove than the others, and you are very unlikely to need it.

Here are the proofs of the other three cases. First, for sampling with replacement and ordered sample, there are n choices for the first object, and n choices for the second, and so on; we multiply the choices for different objects. (Think of the choices as being described by a branching tree.) The product of k factors each equal to n is n k .

For sampling without replacement and ordered sample, there are still n choices for the first object, but now only n − 1 choices for the second (since we do not replace the first), and n−2 for the third, and so on; there are n−k+1 choices for the kth object, since k −1 have previously been removed and n−(k −1) remain. As before, we multiply. This product is the formula for nPk.

For sampling without replacement and unordered sample, think first of choosing an ordered sample, which we can do in nPk ways. But each unordered sample could be obtained by drawing it in k! different orders. So we divide by k!, obtaining nPk/k! = nCk choices.

In our example with the pens, the numbers in the three boxes are 42 = 16, 4P2 = 12, and 4C2 = 6, in agreement with what we got when we wrote them all out.

Note that, if we use the phrase ‘sampling without replacement, ordered sample’, or any other combination, we are assuming that all outcomes are equally likely.

Example The names of the seven days of the week are placed in a hat. Three names are drawn out; these will be the days of the Probability I lectures. What is the probability that no lecture is scheduled at the weekend?

Here the sampling is without replacement, and we can take it to be either ordered or unordered; the answers will be the same. For ordered samples, the size of the sample

space is $7P3 = 7 \cdot 6 \cdot 5 = 210$. If A is the event 'no lectures at weekends', then A occurs precisely when all three days drawn are weekdays; so $|A| = 5P3 = 5 \cdot 4 \cdot 3 = 60$. Thus, $P(A) = 60/210 = 2/7$.

If we decided to use unordered samples instead, the answer would be 5C3/ 7C3, which is once again 2/7.

Stopping rules

Suppose that you take a typing proficiency test. You are allowed to take the test up to three times. Of course, if you pass the test, you don't need to take it again. So the sample space is

S = {p, f p, f f p, f f f },

where for example f f p denotes the outcome that you fail twice and pass on your third attempt.

If all outcomes were equally likely, then your chance of eventually passing the test and getting the certificate would be 3/4.

But it is unreasonable here to assume that all the outcomes are equally likely. For example, you may be very likely to pass on the first attempt. Let us assume that the probability that you pass the test is 0.8. (By Proposition 3, your chance of failing is 0.2.) Let us further assume that, no matter how many times you have failed, your chance of passing at the next attempt is still 0.8. Then we have

$P(p) = 0.8$,

$P(f\ p) = 0.2 \cdot 0.8 = 0.16$,

$P(f\ f\ p) = 0.2\ 2 \cdot 0.8 = 0.032$,

$P(f\ f\ f) = 0.2\ 3 = 0.008$.

Thus the probability that you eventually get the certificate is $P(\{p, f\ p, f\ f\ p\}) = 0.8+0.16+0.032 = 0.992$. Alternatively, you eventually get the certificate unless you fail three times, so the probability is $1-0.008 = 0.992$.

A stopping rule is a rule of the type described here, namely, continue the experiment until some specified occurrence happens. The experiment may potentially be infinite.

Question A couple are planning to have a family. They decide to stop having children either when they have two boys or when they have four children. Suppose that they are successful in their plan.

(a) Write down the sample space.

(b) Assume that, each time that they have a child, the probability that it is a boy is 1/2, independent of all other times. Find P(E) and P(F) where E = “there are at least two girls”, F = “there are more girls than boys”.

Solution

(a) S = {BB,BGB,GBB,BGGB,GBGB,GGBB,BGGG,GBGG, GGBG,GGGB,GGGG}.

(b) E = {BGGB,GBGB,GGBB,BGGG,GBGG,GGBG,GGGB,GGGG},

F = {BGGG,GBGG,GGBG,GGGB,GGGG}.

Now we have

P(BB) = 1/4, P(BGB) = 1/8,

P(BGGB) = 1/16,

and similarly for the other outcomes. So P(E) = 8/16 = 1/2, P(F) = 5/16.

Conditional probability

Alice and Bob are going out to dinner. They toss a fair coin 'best of three' to decide who pays: if there are more heads than tails in the three tosses then Alice pays, otherwise Bob pays.

Clearly each has a 50% chance of paying. The sample space is

$$S = \{HHH,HHT,HT\ H,HT\ T,T\ HH,T\ HT,T\ T\ H,T\ T\ T\},$$

and the events 'Alice pays' and 'Bob pays' are respectively

$$A = \{HHH,HHT,HT\ H,T\ HH\},$$

$$B = \{HT\ T,T\ HT,T\ T\ H,T\ T\ T\}.$$

They toss the coin once and the result is heads; call this event E. How should we now reassess their chances? We have

$$E = \{HHH,HHT,HT\ H,HT\ T\},$$

and if we are given the information that the result of the first toss is heads, then E now becomes the sample space of the experiment, since the outcomes not in E are no longer possible. In the new experiment, the outcomes 'Alice pays' and 'Bob pays' are

$$A \cap E = \{HHH,HHT,HT\ H\},$$

$$B \cap E = \{HT\ T\}.$$

Thus the new probabilities that Alice and Bob pay for dinner are 3/4 and 1/4 respectively.

In general, suppose that we are given that an event E has occurred, and we want to compute the probability that another event A occurs. In general, we can no longer count, since the outcomes may not be equally likely. The correct definition is as follows.

Let E be an event with non-zero probability, and let A be any event. The conditional probability of A given E is defined as

$$P(A \mid E) = P(A \cap E)\ P(E)\ .$$

Again I emphasise that this is the definition. If you are asked for the definition of conditional probability, it is not enough to say “the probability of A given that E has occurred”, although this is the best way to understand it. There is no reason why event E should occur before event A!

Note the vertical bar in the notation. This is P(A | E), not P(A/E) or P(A\E).

Note also that the definition only applies in the case where P(E) is not equal to zero, since we have to divide by it, and this would make no sense if P(E) = 0.

To check the formula in our example:

P(A | E) = P(A∩E) P(E) = 3/8 1/2 = 3 4 ,

P(B | E) = P(B∩E) P(E) = 1/8 1/2 = 1 4 .

It may seem like a small matter, but you should be familiar enough with this formula that you can write it down without stopping to think about the names of the events. Thus, for example,

P(A | B) = P(A∩B) P(B) if P(B) 6= 0.

Example

A random car is chosen among all those passing through Trafalgar Square on a certain day. The probability that the car is yellow is 3/100: the probability that the driver is blonde is 1/5; and the probability that the car is yellow and the driver is blonde is 1/50.

Find the conditional probability that the driver is blonde given that the car is yellow.

GENETICS

Solution: If Y is the event ‘the car is yellow’ and B the event ‘the driver is blonde’, then we are given that P(Y) = 0.03, P(B) = 0.2, and P(Y ∩B) = 0.02. So

P(B | Y) = P(B∩Y) /P(Y) = 0.02/ 0.03 = 0.667

to 3 d.p. Note that we haven’t used all the information given.

There is a connection between conditional probability and independence:

Proposition 2.1 Let A and B be events with P(B) 6= 0. Then A and B are independent if and only if

P(A | B) = P(A).

Proof The words 'if and only if' tell us that we have two jobs to do: we have to show that if A and B are independent, then P(A | B) = P(A); and that if P(A | B) = P(A), then A and B are independent.

So first suppose that A and B are independent. Remember that this means that $P(A \cap B) = P(A) \cdot P(B)$. Then

$P(A \mid B) = P(A \cap B) / P(B) = P(A) \cdot P(B) / P(B) = P(A)$,

that is, P(A | B) = P(A), as we had to prove.

Now suppose that P(A | B) = P(A). In other words,

$P(A \cap B) / P(B) = P(A)$,

using the definition of conditional probability. Now clearing fractions gives

$P(A \cap B) = P(A) \cdot P(B)$,

which is just what the statement 'A and B are independent' means.

This proposition is most likely what people have in mind when they say 'A and B are independent means that B has no effect on A'.

BAYES' THEOREM

By the Theorem of Total Probability,

P(B) = P(B | A1)P(A1) +P(B | A2)P(A2) +P(B | A3)P(A3)

= (1/3)×(1/2) + (2/3)×(1/4) + (2/3)×(1/4)

= 1/2.

We have reached by a roundabout argument a conclusion which you might think to be obvious. If we have no information about the first pen, then the second pen is equally likely to be any one of the four, and the probability should be 1/2, just as for the first pen. This argument happens to be correct. But, until your ability to

distinguish between correct arguments and plausible-looking false ones is very well developed, you may be safer to stick to the calculation that we did. Beware of obvious-looking arguments in probability! Many clever people have been caught out.

Bayes' Theorem

There is a very big difference between P(A | B) and P(B | A).

Suppose that a new test is developed to identify people who are liable to suffer from some genetic disease in later life. Of course, no test is perfect; there will be some carriers of the defective gene who test negative, and some non-carriers who test positive. So, for example, let A be the event 'the patient is a carrier', and B the event 'the test result is positive'.

The scientists who develop the test are concerned with the probabilities that the test result is wrong, that is, with P(B | A 0) and P(B 0 | A). However, a patient who has taken the test has different concerns. If I tested positive, what is the chance that I have the disease? If I tested negative, how sure can I be that I am not a carrier? In other words, P(A | B) and P(A 0 | B 0).

These conditional probabilities are related by Bayes' Theorem:

Theorem 2.4

Let A and B be events with non-zero probability. Then

P(A | B) = P(B | A)·P(A) P(B) .

The proof is not hard. We have

P(A | B)·P(B) = P(A∩B) = P(B | A)·P(A),

using the definition of conditional probability twice. (Note that we need both A and B to have non-zero probability here.) Now divide this equation by P(B) to get the result.

If P(A) 6= 0,1 and P(B) 6= 0, then we can use Corollary 17 to write this as

P(A | B) = P(B | A)·P(A) /P(B | A)·P(A) +P(B | A0)·P(A0) .

Bayes' Theorem is often stated in this form.

Example Consider the ice-cream salesman from Section 2.3. Given that he sold all his stock of ice-cream, what is the probability that the weather was sunny? (This question might be asked by the warehouse manager who doesn't know what the weather was actually like.) Using the same notation that we used before, A1 is the event 'it is sunny' and B the event 'the salesman sells all his stock'. We are asked for P(A1 | B). We were given that P(B | A1) = 0.9 and that P(A1) = 0.3, and we calculated that P(B) = 0.59. So by Bayes' Theorem,

P(A1 | B) = P(B | A1)P(A1) /P(B) = 0.9×0.3 /0.59 = 0.46

to 2 d.p.

ITERATED CONDITIONAL PROBABILITY

= 0.01×0.001 (0.01×0.001) + (0.95×0.999)

= 0.00001 0.94095

= 0.00001.

So a patient with a negative test result can be reassured; but a patient with a positive test result still has less than 2% chance of being a carrier, so is likely to worry unnecessarily.

Of course, these calculations assume that the patient has been selected at random from the population. If the patient has a family history of the disease, the calculations would be quite different.

Example 2% of the population have a certain blood disease in a serious form; 10% have it in a mild form; and 88% don't have it at all. A new blood test is developed; the probability of testing positive is 9/10 if the subject has the serious form, 6/10 if the subject has the mild form, and 1/10 if the subject doesn't have the disease.

I have just tested positive. What is the probability that I have the serious form of the disease?

Let A1 be 'has disease in serious form', A2 be 'has disease in mild form', and A3 be 'doesn't have disease'. Let B be 'test positive'. Then we are given that A1, A2, A3 form a partition and

$P(A1) = 0.02$

$P(A2) = 0.1$

$P(A3) = 0.88$

$P(B \mid A1) = 0.9$

$P(B \mid A2) = 0.6$

$P(B \mid A3) = 0.1$

Thus, by the Theorem of Total Probability,

$P(B) = 0.9\times0.02+0.6\times0.1+0.1\times0.88 = 0.166,$

and then by Bayes' Theorem,

$P(A1 \mid B) = P(B \mid A1)P(A1)\ P(B) = 0.9\times0.02\ 0.166 = 0.108$

to 3 d.p.

Random variables

The Holy Roman Empire was, in the words of the historian Voltaire, "neither holy, nor Roman, nor an empire". Similarly, a random variable is neither random nor a variable:

A random variable is a function defined on a sample space.

The values of the function can be anything at all, but for us they will always be numbers. The standard abbreviation for 'random variable' is r.v.

Example I select at random a student from the class and measure his or her height in centimetres.

Here, the sample space is the set of students; the random variable is 'height', which is a function from the set of students to the real numbers: h(S) is the height of student S in centimetres. (Remember that a function is nothing but a rule for associating with each element of its domain set an element of its target or range set. Here the domain set is the sample space S, the set of students in the class, and the target space is the set of real numbers.)

Example I throw a six-sided die twice; I am interested in the sum of the two numbers. Here the sample space is

$$S = \{(i, j) : 1 \leq i, j \leq 6\},$$

and the random variable F is given by $F(i, j) = i + j$. The target set is the set $\{2,3,...,12\}$.

The two random variables in the above examples are representatives of the two types of random variables that we will consider. These definitions are not quite precise, but more examples should make the idea clearer.

A random variable F is discrete if the values it can take are separated by gaps. For example, F is discrete if it can take only finitely many values (as in the second example above, where the values are the integers from 2 to 12), or if the values of F are integers (for example, the number of nuclear decays which take place in a second in a sample of radioactive material – the number is an integer but we can't easily put an upper limit on it.)

A random variable is continuous if there are no gaps between its possible values. In the first example, the height of a student could in principle be any real number between certain extreme limits. A random variable whose values range over an interval of real numbers, or even over all real numbers, is continuous.

One could concoct random variables which are neither discrete nor continuous (e.g. the possible, values could be 1, 2, 3, or any real number between 4 and 5), but we will not consider such random variables. We begin by considering discrete random variables.

Probability mass function

Let F be a discrete random variable. The most basic question we can ask is: given any value a in the target set of F, what is the probability that F takes the value a? In other words, if we consider the event

$$A = \{x \in S : F(x) = a\}$$

what is P(A)? (Remember that an event is a subset of the sample space.) Since events of this kind are so important, we simplify the notation: we write

$P(F = a)$

in place of

$P(\{x \in S : F(x) = a\})$.

(There is a fairly common convention in probability and statistics that random variables are denoted by capital letters and their values by lower-case letters. In fact, it is quite common to use the same letter in lower case for a value of the random variable; thus, we would write $P(F = f)$ in the above example. But remember that this is only a convention, and you are not bound to it.)

EXPECTED VALUE AND VARIANCE

The probability mass function of a discrete random variable F is the function, formula or table which gives the value of P(F = a) for each element a in the target set of F. If F takes only a few values, it is convenient to list it in a table; otherwise we should give a formula if possible. The standard abbreviation for 'probability mass function' is p.m.f.

Example I toss a fair coin three times. The random variable X gives the number of heads recorded. The possible values of X are 0,1,2,3, and its p.m.f. is

a	0 1 2 3
P(X = a)	1/ 8 3 /8 3/ 8 1/ 8

For the sample space is {HHH,HHT,HT H,HT T,T HH,T HT,T T H,T T T}, and each outcome is equally likely. The event X = 1, for example, when written as a set of outcomes, is equal to {HT T,T HT,T T H}, and has probability 3/8.

Two random variables X and Y are said to have the same distribution if the values they take and their probability mass functions are equal. We write X ∼ Y in this case.

In the above example, if Y is the number of tails recorded during the experiment, then X and Y have the same distribution, even though their actual values are different (indeed, Y = 3−X).

Expected value and variance

Let X be a discrete random variable which takes the values a1,...,an. The expected value or mean of X is the number E(X) given by the formula

$$E(X) = n \sum i=1 \; aiP(X = ai).$$

That is, we multiply each value of X by the probability that X takes that value, and sum these terms. The expected value is a kind of 'generalised average': if each of the values is equally likely, so that each has probability 1/n, then E(X) = (a1 +···+an)/n, which is just the average of the values.

There is an interpretation of the expected value in terms of mechanics. If we put a mass pi on the axis at position ai for i = 1,...,n, where pi = P(X = ai), then the centre of mass of all these masses is at the point E(X).

If the random variable X takes infinitely many values, say a1,a2,a3,..., then we define the expected value of X to be the infinite sum

$$E(X) = \infty \sum i=1 \; aiP(X = ai).$$

Of course, now we have to worry about whether this means anything, that is, whether this infinite series is convergent. This is a question which is discussed at great length in analysis. We won't worry about it too much. Usually, discrete random variables will only have finitely many values; in the few examples we consider where there are infinitely many values, the series will usually be a geometric series or something similar, which we know how to sum. In the proofs below, we assume that the number of values is finite.

The variance of X is the number Var(X) given by

$$Var(X) = E(X\,2\,) - E(X)\,2\,.$$

Here, X 2 is just the random variable whose values are the squares of the values of X. Thus

$$E(X\,2\,) = n \sum i=1 \; a\,2\,i\,P(X = ai)$$

(or an infinite sum, if necessary). The next theorem shows that, if E(X) is a kind of average of the values of X, then Var(X) is a measure of how spread-out the values are around their average.

Proposition 3.1 Let X be a discrete random variable with $E(X) = \mu$. Then

$$Var(X) = E((X - \mu)\,2\,) = n \sum i=1 \; (ai - \mu)\,2P(X = ai).$$

For the second term is equal to the third by definition, and the third is

$$n \sum i=1 \; (ai - \mu)\,2P(X = ai)$$

$$= n \sum i=1 \; (a\,2\,i - 2\mu ai + \mu\,2\,)P(X = ai)$$

$$= n \sum i=1 \; a\,2\,i\,P(X = ai)\,! - 2\mu \; n \sum i=1 \; aiP(X = ai)\,! + \mu\,2 \; n \sum i=1 \; P(X = ai)\,!\,.$$

(What is happening here is that the entire sum consists of n rows with three terms in each row. We add it up by columns instead of by rows, getting three parts with n terms in each part.) Continuing, we find

$$E((X-\mu)^2) = E(X^2) - 2\mu E(X) + \mu^2$$

$$= E(X^2) - E(X)^2,$$

and we are done. (Remember that $E(X) = \mu$, and that $\sum_{i=1}^{n} P(X = a_i) = 1$ since the events $X = a_i$ form a partition.)

JOINT P.M.F. OF TWO RANDOM VARIABLES

Some people take the conclusion of this proposition as the definition of variance.

Example I toss a fair coin three times; X is the number of heads. What are the expected value and variance of X?

$E(X) = 0\times(1/8) +1\times(3/8) +2\times(3/8) +3\times(1/8) = 3/2,$

$Var(X) = 0\ 2 \times(1/8) +1\ 2 \times(3/8) +2\ 2 \times(3/8) +3\ 2 \times(1/8)-(3/2)\ 2 = 3/4.$

If we calculate the variance using Proposition 3.1, we get

$Var(X) = -3\ 2\ 2 \times 1\ 8 + -1\ 2\ 2 \times 3\ 8 + 1\ 2\ 2 \times 3\ 8 + 3\ 2\ 2 \times 1\ 8 = 3\ 4\ .$

Two properties of expected value and variance can be used as a check on your calculations.

• The expected value of X always lies between the smallest and largest values of X.

• The variance of X is never negative. (For the formula in Proposition 3.1 is a sum of terms, each of the form $(a_i - \mu)\ 2$ (a square, hence non-negative) times $P(X = a_i)$ (a probability, hence non-negative)

Joint p.m.f. of two random variables

Let X be a random variable taking the values $a_1,...,a_n$, and let Y be a random variable taking the values $b_1,...,b_m$. We say that X and Y are independent if, for any possible values i and j, we have

$P(X = a_i, Y = b_j) = P(X = a_i)\cdot P(Y = b_j).$

Here $P(X = a_i, Y = b_j)$ means the probability of the event that X takes the value a_i and Y takes the value b_j . So we could re-state the definition as follows:

The random variables X and Y are independent if, for any value a_i of X and any value b_j of Y, the events $X = a_i$ and $Y = b_j$ are independent (events).

Note the difference between 'independent events' and 'independent random variables'.

Binomial random variable Bin(n, p)

Remember that for a Bernoulli random variable, we describe the event X = 1 as a 'success'. Now a binomial random variable counts the number of successes in n independent trials each associated with a Bernoulli(p) random variable.

For example, suppose that we have a biased coin for which the probability of heads is p. We toss the coin n times and count the number of heads obtained. This number is a Bin(n, p) random variable.

A Bin(n, p) random variable X takes the values 0,1,2,...,n, and the p.m.f. of X is given by

P(X = k) = nCkq n−k p k

for k = 0,1,2,...,n, where q = 1− p. This is because there are nCk different ways of obtaining k heads in a sequence of n throws (the number of choices of the k positions in which the heads occur), and the probability of getting k heads and n−k tails in a particular order is q n−k p k .

Note that we have given a formula rather than a table here. For small values we could tabulate the results; for example, for Bin(4, p):

k 0 1 2 3 4 P(X = k) q 4 4q 3 p 6q 2 p 2 4qp3 p 4

Note: when we add up all the probabilities in the table, we get n

∑ k=0 nCkq n−k p k = (q+ p) n = 1,

as it should be: here we used the binomial theorem

(x+y) n = n ∑ k=0 nCkx n−k y k .

(This argument explains the name of the binomial random variable!) If X ∼ Bin(n, p), then

E(X) = np, Var(X) = npq.

There are two ways to prove this, an easy way and a harder way. The easy way only works for the binomial, but the harder way is useful for many random variables. However, you can skip it if you wish: I have set it in smaller type for this reason.

Here is the easy method. We have a coin with probability p of coming down heads, and we toss it n times and count the number X of heads. Then X is our Bin(n, p) random variable. Let Xk be the random variable defined by

Xk = 1 if we get heads on the kth toss,

0 if we get tails on the kth toss.

In other words, Xi is the indicator variable of the event 'heads on the kth toss'. Now we have

$$X = X1 + X2 + \cdots + Xn$$

(can you see why?), and X1,...,Xn are independent Bernoulli(p) random variables (since they are defined by different tosses of a coin). So, as we saw earlier, E(Xi) = p, Var(Xi) = pq. Then, by Theorem 21, since the variables are independent, we have

$$E(X) = p + p + \cdots + p = np,$$

$$Var(X) = pq + pq + \cdots + pq = npq.$$

The other method uses a gadget called the probability generating function. We only use it here for calculating expected values and variances, but if you learn more probability theory you will see other uses for it. Let X be a random variable whose values are non-negative integers. (We don't insist that it takes all possible values; this method is fine for the binomial Bin(n, p), which takes values between 0 and n. To save space, we write pk for the probability P(X = k). Now the probability generating function of X is the power series

$$G_X(x) = \sum p_k x^k .$$

(The sum is over all values k taken by X.)

We use the notation $[F(x)]_{x=1}$ for the result of substituting x = 1 in the series F(x).

Proposition 3.4 Let $G_X(x)$ be the probability generating function of a random variable X. Then

(a) $[G_X(x)]_{x=1} = 1$;

(b) $E(X) = \frac{d}{dx} G_X(x)\big|_{x=1}$;

(c) $\mathrm{Var}(X) = \left[\frac{d^2}{dx^2} G_X(x)\right]_{x=1} + E(X) - E(X)^2$.

Part (a) is just the statement that probabilities add up to 1: when we substitute $x = 1$ in the power series for $G_X(x)$ we just get $\sum p_k$.

For part (b), when we differentiate the series term-by-term (you will learn later in Analysis that this is OK), we get

$d/dx\, G_X(x) = \sum k p_k x^{k-1}$.

Now putting $x = 1$ in this series we get

$\sum k p_k = E(X)$.

For part (c), differentiating twice gives

$d^2/dx^2\, G_X(x) = \sum k(k-1) p_k x^{k-2}$.

Now putting $x = 1$ in this series we get

$\sum k(k-1) p_k = \sum k^2 p_k - \sum k p_k = E(X^2) - E(X)$.

Adding $E(X)$ and subtracting $E(X)^2$ gives $E(X^2) - E(X)^2$, which by definition is $\mathrm{Var}(X)$.

Bernoulli random variable Bernoulli(p)

• Occurs when there is a single trial with a fixed probability p of success.

• Takes only the values 0 and 1.

• p.m.f. $P(X = 0) = q$, $P(X = 1) = p$, where $q = 1 - p$.

• $E(X) = p$, $\mathrm{Var}(X) = pq$

Binomial random variable Bin(n, p)

• Occurs when we are counting the number of successes in n independent trials with fixed probability p of success in each trial, e.g. the number of heads in n coin tosses. Also, sampling with replacement from a population with a proportion p of distinguished elements.

• The sum of n independent Bernoulli(p) random variables.

• Values 0,1,2,...,n.

• p.m.f. P(X = k) = nCkq n−k p k for 0 ≤ k ≤ n, where q = 1− p.

• E(X) = np, Var(X) = npq.

Hypergeometric random variable Hg(n,M,N)

• Occurs when we are sampling n elements without replacement from a population of N elements of which M are distinguished.

• Values 0,1,2,...,n.

• p.m.f. P(X = k) = (MCk · N−MCn−k)/ NCn.

• E(X) = n M N , Var(X) = n M N N −M N N −n N −1 .

• Approximately Bin(n,M/N) if n is small compared to N,M,N −M.

Geometric random variable Geom(p)

• Describes the number of trials up to and including the first success in a sequence of independent Bernoulli trials, e.g. number of tosses until the first head when tossing a coin.

• Values 1,2,... (any positive integer).

• p.m.f. P(X = k) = q k−1 p, where q = 1− p.

• E(X) = 1/p, Var(X) = q/p 2 .

Poisson random variable Poisson(λ)

• Describes the number of occurrences of a random event in a fixed time interval, e.g. the number of fish caught in a day.

• Values 0,1,2,... (any non-negative integer)

• p.m.f. P(X = k) = e −λλ k/k! .

• E(X) = λ, Var(X) = λ.

• If n is large, p is small, and np = λ, then Bin(n, p) is approximately equal to Poisson(λ) (in the sense that the p.m.f.s are approximately equal).

Uniform random variable U[a,b]

• Occurs when a number is chosen at random from the interval [a,b], with all values equally likely.

• p.d.f. f(x) = (0 if x < a, 1/(b−a) if a ≤ x ≤ b, 0 if x > b.

• c.d.f. F(x) = (0 if x < a, (x−a)/(b−a) if a ≤ x ≤ b, 1 if x > b.

• E(X) = (a+b)/2, Var(X) = (b−a) 2/12.

Exponential random variable Exp(λ)

• Occurs in the same situations as the Poisson random variable, but measures the time from now until the first occurrence of the event.

• p.d.f. f(x) = 0 if x < 0, λ e −λx if x ≥ 0.

• c.d.f. F(x) = 0 if x < 0, 1−e −λx if x ≥ 0.

• E(X) = 1/λ, Var(X) = 1/λ 2 .

• However long you wait, the time until the next occurrence has the same distribution.

Normal random variable N(μ,σ 2)

• The limit of the sum (or average) of many independent Bernoulli random variables. This also works for many other types of random variables: this statement is known as the Central Limit Theorem.

• p.d.f. f(x) = 1 σ √ 2π e −(x−μ) 2/2σ 2

. • No simple formula for c.d.f.; use tables.

• E(X) = μ, Var(X) = σ 2 .

• For large n, Bin(n, p) is approximately N(np,npq).

• Standard normal N(0,1) is given in the table. If $X \sim N(\mu,\sigma^2)$, then $(X-\mu)/\sigma \sim N(0,1)$. The c.d.f.s of the Binomial, Poisson, and Standard Normal random variables are tabulated in the New Cambridge Statistical Tables, Tables 1, 2 and 4.

DECISION MAKING UNDER THE CONDITIONS OF RISK AND UNCERTAINTY

Introduction

Modeling for decision making involves two distinct parties—one is the decision maker and the other is the model builder known as the analyst. The analyst is to assist the decision maker in his/her decision making process. Therefore, the analyst must be equipped with more than a set of analytical methods. Specialists in model building are often tempted to study a problem, and then go off in isolation to develop an elaborate mathematical model for use by the manager (i.e., the decision maker). Unfortunately the manager may not understand this model and may either use it blindly or reject it entirely. The specialist may feel that the manager is too ignorant and unsophisticated to appreciate the model, while the manager may feel that the specialist lives in a dream world of unrealistic assumptions and irrelevant mathematical language. Such miscommunication can be avoided if the manager works with the specialist to develop first a simple model that provides a crude but understandable analysis. After the manager has built up confidence in this model, additional detail and sophistication can be added, perhaps progressively only a bit at a time. This process requires an investment of time on the part of the manager and sincere interest on the part of the specialist in solving the manager's real problem, rather than in creating and trying to explain sophisticated models. This progressive model building is often referred to as the bootstrapping approach and is the most important factor in determining successful implementation of a decision model. Moreover the bootstrapping approach simplifies the otherwise difficult task of model validating and verification processes.

In deterministic models, a good decision is judged by the outcome alone. However, in probabilistic models, the decision maker is concerned not only with the outcome value but also with the amount of risk each decision

carries. As an example of deterministic versus probabilistic models, consider the past and the future. Nothing we can do can change the past, but everything we do influences and changes the future, although the future has an element of uncertainty. Managers are captivated much more by shaping the future than the history of the past.

Uncertainty is the fact of life and business. Probability is the guide for a "good" life and successful business. The concept of probability occupies an important place in the decision making process, whether the problem is one faced in business, in government, in the social sciences, or just in one's own everyday personal life. In very few decision making situations is perfect information—all the needed facts—available. Most decisions are made in the face of uncertainty. Probability enters into the process by playing the role of a substitute for certainty—a substitute for complete knowledge .

Probabilistic modeling is largely based on application of statistics for probability assessment of uncontrollable events (or factors), as well as risk assessment of your decision. The original idea of statistics was the collection of information about and for the state. The word statistics is not derived from any classical Greek or Latin roots, but from the Italian word for state. Probability has a much longer history. Probability is derived from the verb to probe meaning to "find out" what is not too easily accessible or understandable. The word "proof" has the same origin that provides necessary details to understand what is claimed to be true. Probabilistic models are viewed as similar to that of a game; actions are based on expected outcomes. The center of interest moves from the deterministic to probabilistic models using subjective statistical techniques for estimation, testing and predictions. In probabilistic modeling, risk means uncertainty for which the probability distribution is 3 known. Therefore risk assessment means a study to determine the outcomes of decisions along with their probabilities.

Decision makers often face a severe lack of information. Probability assessment quantifies the information gap between what is known, and what needs to be known for an optimal decision. The probabilistic models are used for protection against adverse uncertainty, and exploitation of propitious uncertainty. Difficulty in probability assessment arises from information that is scarce, vague, inconsistent or incomplete. A statement

such as “the probability of a power outage is between 0.3 and 0.4” is more natural and realistic than its “exact” counterpart, such as “the probability of a power outage is 0.36342”.
It is a challenging task to compare several courses of action and then select one action to be implemented. At times, the task may prove too challenging. Difficulties in decision making arise through complexities in decision alternatives. The limited informationprocessing capacity of a decision-maker can be strained when considering the consequences of only one course of action. Yet, choice requires that the implications of various courses of action be visualized and compared. In addition, unknown factors always intrude upon the problem situation and seldom are outcomes known with certainty. Almost always, an outcome depends upon the reactions of other people who may be undecided themselves. It is no wonder that decision makers sometimes postpone choices for as long as possible. Then, when they finally decide, they neglect to consider all the implications of their decision [6, 7].

Business decision making is almost always accompanied by conditions of uncertainty. Clearly, the more information the decision maker has, the better the decision will be. Treating decisions as if they were gambles is the basis of decision theory. This means that we have to trade off the value of a certain outcome against its probability. To operate according to the canons of decision theory, we must compute the value of a certain outcome and its probabilities; hence, determining the consequences of our choices. The origin of decision theory is derived from economics by using the utility function of payoffs. It suggests that decisions be made by computing the utility and probability, the ranges of options, and also lays down strategies for good decisions .
Objectives are important both in identifying problems and in evaluating alternative solutions. Evaluating alternatives requires that a decision maker’s objectives be expressed as criteria that reflect the attributes of the alternatives relevant to the choice. The systematic study of decision making provides a framework for choosing courses of action in a complex, uncertain or conflicting situation. The choices of possible actions, and the prediction of expected outcomes, derive from a logical analysis of the

decision situation. A possible drawback in the decision analysis approach: You might have already noticed that the above criteria always result in selection of only one course of action. However, in many decision problems, the decision maker might wish to consider a combination of some actions. For example, in the investment problem, the investor might wish to distribute the assets among a mixture of the choices in such a way to optimize the portfolio's return [2-4].

Relevant information and knowledge used to solve a decision problem sharpens our flat probability. Useful information moves the location of a problem from the pure uncertain "pole" towards the deterministic "pole." Probability assessment is nothing more than the quantification of uncertainty. In other words, quantification of uncertainty allows for the communication of uncertainty between persons. There can be uncertainties regarding events, states of the world, beliefs and so on. Probability is the tool for both communicating 4 uncertainty and managing it. There are different types of decision models that help to analyze the different scenarios. Depending on the amount and degree of knowledge we have, the three most widely used types are:

• Decision making under pure uncertainty
• Decision making under risk
• Decision making by buying information (pushing the problem towards the deterministic "pole")

In decision making under pure uncertainty, the decision maker has absolutely no knowledge, not even about the likelihood of occurrence for any state of nature. In such situations, the decision maker's behavior is purely based on his/her attitude toward the unknown. Some of these behaviors are optimistic, pessimistic and least regret, among others. Consider three following known ideas about a glass of water and a captain in a rough sea:

Figure 1 Known Ideas about a Glass of Water and a Captain in a Rough Sea

A glass of water:
Optimist: The glass is half-full.
Pessimist: The glass is half-empty.
Manager: The glass is twice as large as it needs to be.

A captain in a rough sea:
The optimist expects it to change.
The pessimist complains about the wind.
The realist adjusts the sails.

Optimists are right; so are the pessimists. It is up to you to choose which you will be [8, 9]. The optimist sees opportunity in every problem; the pessimist sees problem in every opportunity. Both optimists and pessimists contribute to our society. The optimist invents the airplane and the pessimist the parachute. Whenever the decision maker has some knowledge regarding the states of nature, he/she may be able to assign subjective probability for the occurrence of each state of nature. By doing so, the problem is then classified as decision making under risk. In many cases, the decision maker may need an expert's judgment to sharpen his/her uncertainties with respect to the likelihood of each state of nature. In such a case, the decision maker may buy the expert's relevant knowledge in order to make a better decision [10, 14]. The procedure used to incorporate the expert's advice with the decision maker's probabilities assessment is known as the Bayesian approach.

This paper presents the decision analysis process both for public and private decision making under different decision criteria, type and quality of available information. Basic elements in the analysis of decision alternatives and choice are described as well as the goals and objectives that guide decision making. In the subsequent sections, we will examine key issues related to a decision maker's preferences regarding alternatives, criteria for choice and choice modes.

Decision Making Under Pure Uncertainty

Decision making

Decision taking is a multidimensional process and it is not simply to make one choice. Decision taking as an integral part of management is one of determining characteristics of leadership. The quality of taken decisions has main impact on the success of non-success of organisation. The decision taking is considered managerial function and organisation process because in most cases this action requires not only quality but also readiness to face with effects, those it passes individual and becomes concern of determined group of people. Decision taking is the process ofproblem introduction, formulation of alternatives, analyse of alternatives and selection of best

alternative, than continues with the implementation of alternative and its control. The decision making process is constituted from eight steps:

1. Identification of problem: This process starts with existence of problem and the difference between existing and desired state. The managers are good if they are able to understand three main characteristics of problem: to be in flow about the problem, to be in pressure to act

2. Identification of criteria for decision making: After identification of problem it should be identified the criteria for solution of problem. The criteria should be based on the importance and weight depending from issue or problem for which is needed decision.

3. Distribution of importance / weight of criteria: The decision maker should weight the importance of criteria and classify them by giving priority according to its importance.

4. Development of alternatives: Decision maker must be creative, thus in cooperation with team should list the alternatives based on which certain problem could be solved.

5. Analyse of alternatives: The selected alternatives are put into analyse in this step. There will be carried out the investigation of information and additional material in order to identify the priorities and weaknesses for each presented alternative.

6. Selection of alternative: After carrying the weight of presented alternatives, in this step is chosen the best alternative which generates the highest amount calculated at previous step.

7. Execution of decision: In this step is placed the decision on action, and the decision has to be followed to the persons of concern as well as their engagement is accepted for the work which follows. In the cases people who execute decision participate in the process they are enthusiast to support the implementation of decision.

8. Evaluation of effectiveness of decision: It is evaluation of result where may be seen if the problem is solved. In the cases where the problem still exist than the manager have to see what was going wrong and to return to previous steps.

The perspective of decision making

The decision making process involves evaluating a scenario from different angles, or perspectives, in order to identify solutions that will lead to the desired outcome. 87 Three main perspectives on decision making are rationality, limited rationality and intuition. Rational decision making means managers make sustainable solution that maximizes value under the conditions of specific limitations. When managers take rational decisions but are limited from their ability to process information, this presents rational limited decision taking. Whereas intuitive decision making is when the decisions are taken based on experience, feeling and accumulated judgment. The right solving of problems is mainly intuitive and decision taking results from the unclear mixture of experience, immagination, intellegence and feeling joined with conscience.

Depending from the nature of problem, the manager may take different type of decisions. When we have to deal with structural problems the decision making is programmed while when we have to deal with unstructured problems it is not programmed. Managers try to take good decisions because they will be judged based on the results of those decisions. It is known that managers often disregard regulations and norms when they have to take decisions under risk and uncertainty. An attention also must be paid to the context of decisions that means to analyse the structure of organisation and its organisation culture.

Attitude to risk on decision making

Risk management is difficult process and its purpose is to save property, capital or profit of organisation by decreasing the potential of loses. Risk management may be defined as making decision for the activities by which would be decreased the probability and the consequences of unpleasant effects that may be done only through identifying and assessing risk. Researches tried to study the role of risk in their field of interest. According to researches P. Slivic89 (2000) and P. More (1983)90 people percept the risk in different ways depending from the field of work. One valuable definition for risk in the decision making field introduced

authors H. Raiffa and R.D Luce91 , who make the distinction of three conditions that managers are faced with while taking decisions:

1. Security -accurate decision making because results of each alternative are known therefore managers may take secure decisions.

2. Risk – each action leads to one of specific results, where results of alternatives cannot be evaluated. The ability of determining probability may be the result of previous experience or of secondary information.

3. Unsecurity -actions may lead to a group of consequences, but where the probability of result is completely unknown. For decision maker the security and justification for settling an alternative is missing.

In generally managers are not inclined to accept risk. Some studies showed that managers do not accept that risk by which they are faced with is inseparable with situation and they simply move its acceptance by considering it as controllable issue.

Hypothesis testing

A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

Hypothesis testing is formulated in terms of two hypotheses:

• H0: the null hypothesis;

• H1: the alternate hypothesis

The hypothesis we want to test is if H1 is “likely” true. So, there are two possible outcomes:

• Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1;

• Do not reject H0 because of insufficient evidence to support H1.

Very important!!

Note that failure to reject H0 does not mean the null hypothesis is true. There is no formal outcome that says "accept H0." It only means that we do not have sufficient evidence to support H1.

Example In a jury trial the hypotheses are:

• H0: defendant is innocent;

• H1: defendant is guilty. H0 (innocent) is rejected if H1 (guilty) is supported by evidence beyond "reasonable doubt." Failure to reject H0 (prove guilty) does not imply innocence, only that the evidence is insufficient to reject it.

Case study

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the true defective probability. We want to test if:

• H0 : $p = 0.05$

• H1 : $p > 0.05$

We are going to use a sample of 100 chips from the production to test.

Let X denote the number of defective in the sample of 100. Reject H0 if $X \geq 10$ (chosen "arbitrarily" in this case). X is called the test statistic.

p = 0.05 Reject H0, p > 0.05 0 10 100 Do not reject H0 critical value

Design of experiments

Design of experiments (DOE) is a systematic method to determine the relationship between factors affecting a process and the output of that process. In other words, it is used to find cause-and-effect relationships. This information is needed to manage process inputs in order to optimize the output.

Simple Comparative Experiments

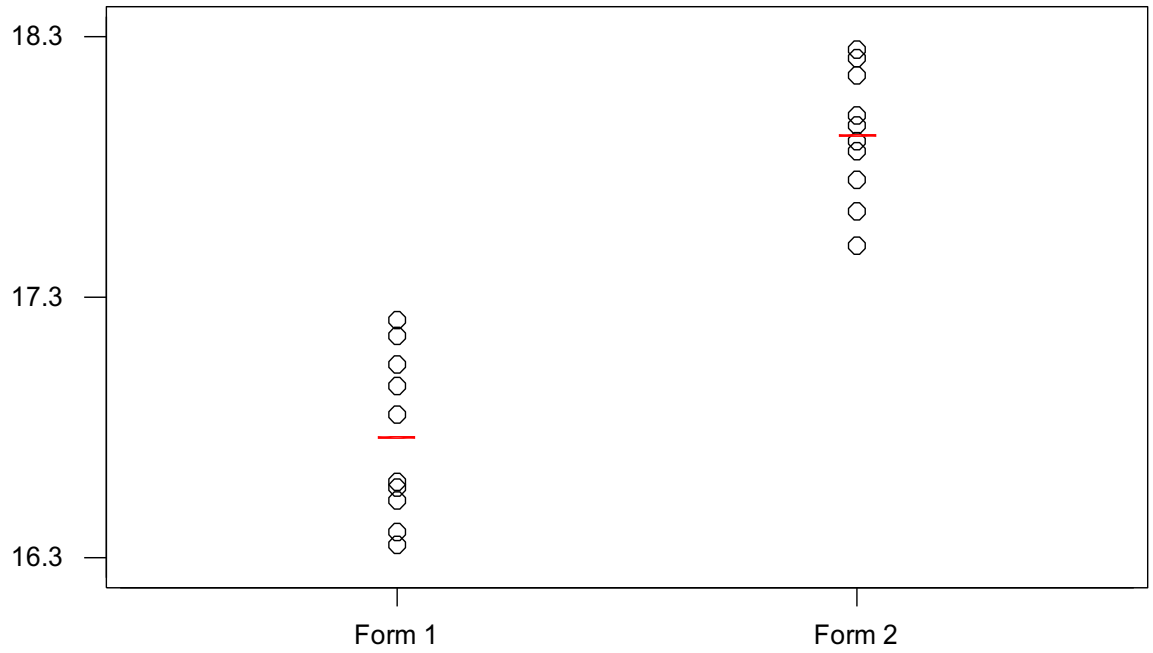
- Consider experiments to compare two conditions

- Simple comparative experiments
- Example:
  - The strength of portland cement mortar
  - Two different formulations: modified v.s. unmodified
  - Collect 10 observations for each formulations
  - Formulations = Treatments (levels)
- The data (Table 2.1)

Observation (sample), *j*	Modified Mortar (Formulation 1)	Unmodified Mortar (Formulation 2)
1	16.85	17.50
2	16.40	17.63
3	17.21	18.25
4	16.35	18.00
5	16.52	17.86
6	17.04	17.75
7	16.96	18.22
8	17.15	17.90
9	16.59	17.96
10	16.57	18.15

- Dot

- diagram: Form 1 (modified) v.s. Form 2 (unmodified)
- unmodified (17.92) > modified (16.76)

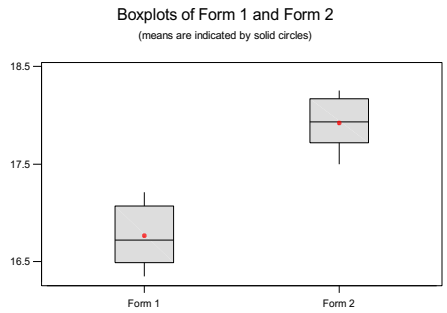


- Hypothesis testing (significance testing): a technique to assist the experiment in comparing these two formulations.

Basic Statistical Concepts

- Run = each observations in the experiment
- Error = random variable
- Graphical Description of Variability
  - Dot diagram: the general location or central tendency of observations
  - Histogram: central tendency, spread and general shape of the distribution of the data (Fig. 2-2)

– Box-plot: minimum, maximum, the lower and upper quartiles and the median



- Probability Distributions
- Mean, Variance and Expected Values

  Sampling and Sampling Distribution
- Random sampling
- Statistic: any function of the observations in a sample that does not contain unknown parameters
- Sample mean and sample variance
- Properties of sample mean and sample variance
  - Estimator and estimate
  - Unbiased and minimum variance
- Degree of freedom:
  - Random variable y has v degree of freedom if $E(SS/v) = \sigma^2$
  - The number of independent elements in the sum of squares
- The normal and other sampling distribution:
  - Sampling distribution
  - Normal distribution: The Central Limit Theorem

- Chi-square distribution: the distribution of SS
- t distribution
- F distribution

Inferences about the Differences in Means, Randomized Designs

- Use hypothesis testing and confidence interval procedures for comparing two treatment means.
- Assume a completely randomized experimental design is used. (a random sample from a normal distribution)

Hypothesis Testing

- Compare the strength of two different formulations: unmodified v.s. modified
- Two levels of the factor
- y_{ij} : the the jth observation from the ith factor level, i=1, 2, and j = 1,2,…, n_i

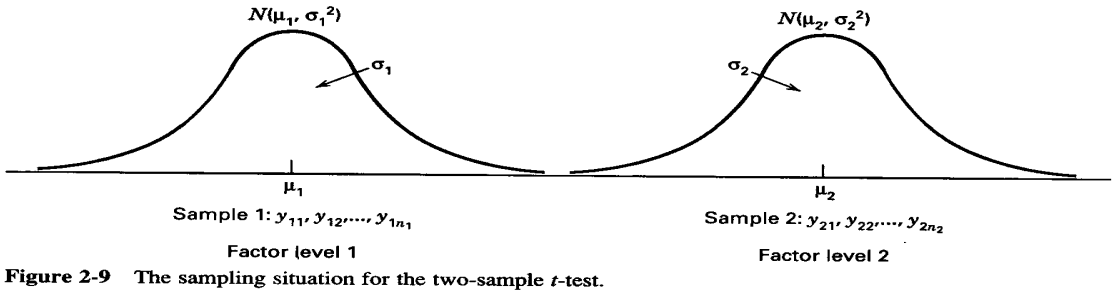


Figure 2-9 The sampling situation for the two-sample *t*-test.

- Model: $y_{ij} = \mu_i + \varepsilon_{ij}$
- $y_{ij} \sim N(\mu_i, \sigma_i^2)$
- Statistical hypotheses:
- Test statistic, critical region (rejection region)
- Type I error, Type II error and Power
- The two-sample t-test:

$$\bar{y}_1 - \bar{y}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

If $\sigma^2 = \sigma_1^2 = \sigma_2^2$ and is known, then the testing statistic is

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\quad}}$$

Use S_1^2 and S_2^2 to estimate σ_1^2 and σ_2^2

The previous ratio becomes $\dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

However, we have the case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Pool the individual sample variances:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The test statistic is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

- Values of t_0 that are near zero are consistent with the null hypothesis
- Values of t_0 that are very different from zero are consistent with the alternative hypothesis
- t_0 is a "distance" measure-how far apart the averages are expressed in standard deviation units
- Notice the interpretation of t_0 as a **signal-to-noise** ratio

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{9(0.100) + 9(0.061)}{10 + 10 - 2} = 0.081$$

$$S_p = 0.284$$

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{16.76 - 17.92}{0.284\sqrt{\dfrac{1}{10} + \dfrac{1}{10}}} = -9.13$$

The two sample means are about 9 standard deviations apart

- So far, we haven't really done any "statistics"
- We need an **objective** basis for deciding how large the test statistic t_0 really is
- In 1908, W. S. Gosset derived the **referencedistribution** for t_0 … called the *t* distribution
- Tables of the *t* distribution - text, page 640

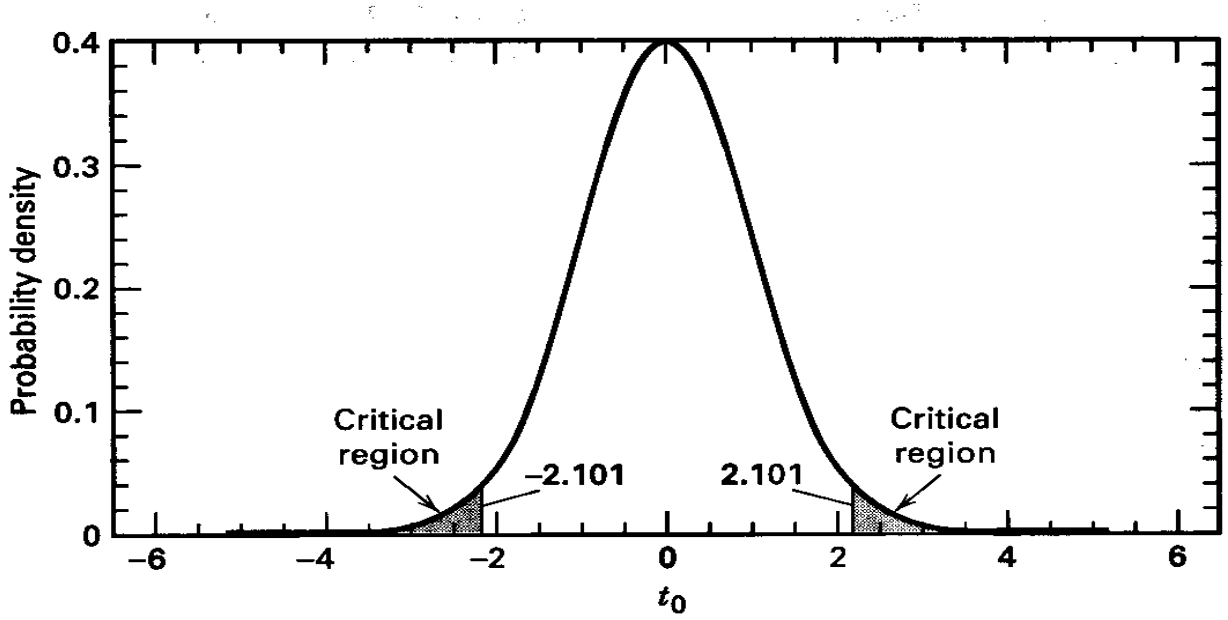


Figure 2-10 The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

- A value of t_0 between –2.101 and 2.101 is consistent with equality of means
- It is possible for the means to be equal and t_0 to exceed either 2.101 or –2.101, but it would be a "**rareevent**" … leads to the conclusion that the means are different

- Could also use the ***P*-value** approach

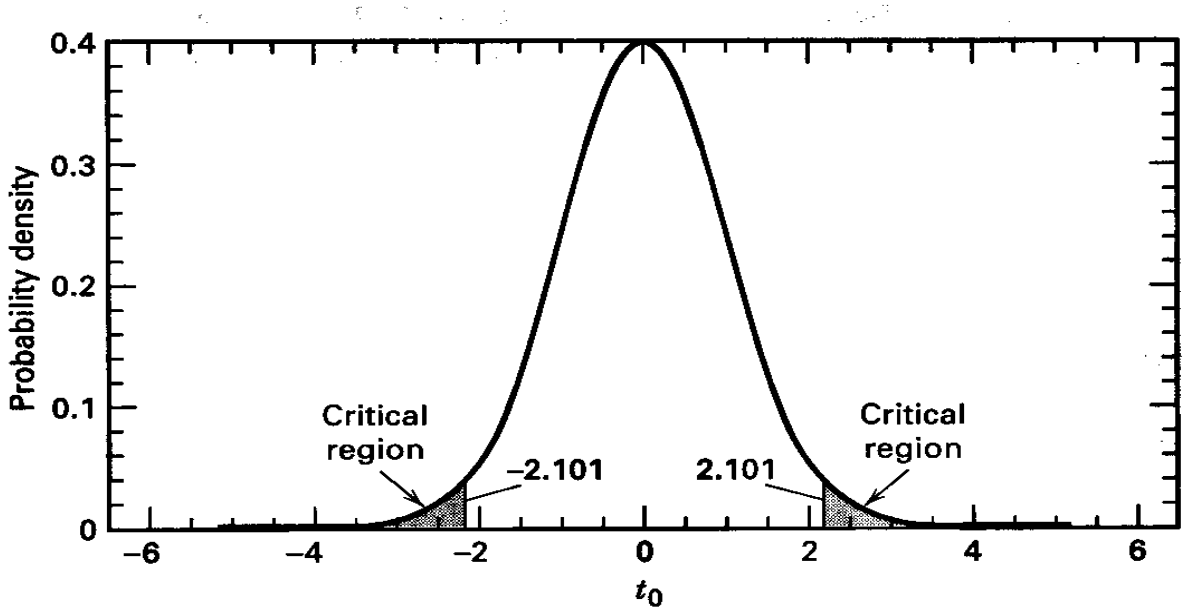


Figure 2-10 The t distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

- The ***P*-value** is the risk of **wrongly rejecting** the null hypothesis of equal means (it measures rareness of the event)
- The *P*-value in our problem is *P* = 3.68E-8

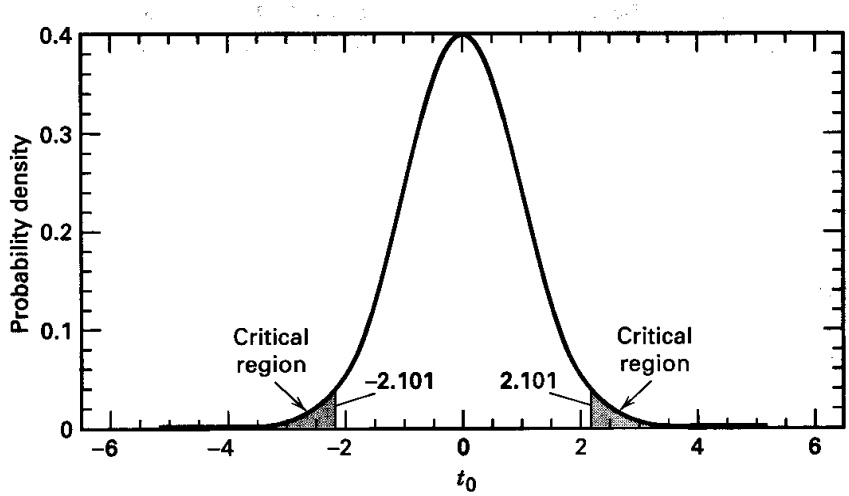
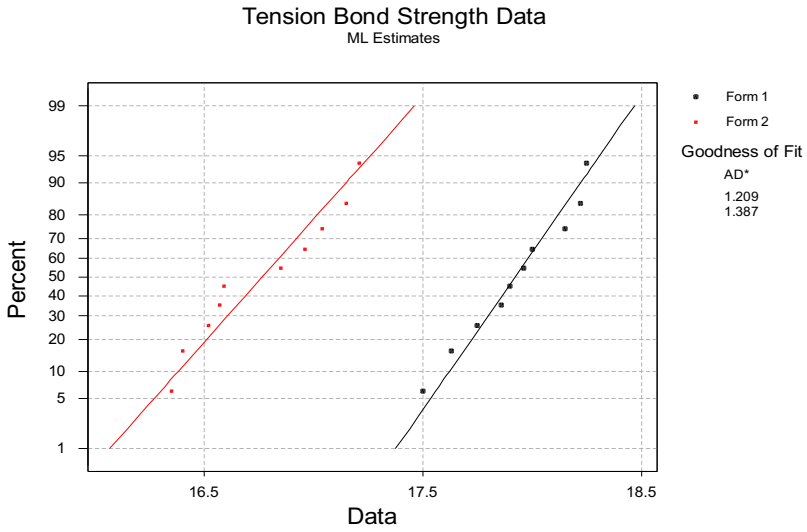


Figure 2-10 The t distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$.

- Checking Assumptions in the t-test:
  - Equal-variance assumption
  - Normality assumption
- Normal Probability Plot: $y_{(j)}$ v.s. $(j - 0.5)/n$



- Estimate mean and variance from normal probability plot:
  - Mean: 50 percentile
  - Variance: the difference between 84^{th} and 50^{th} percentile
- Transformations

Choice of Sample Size

- Type II error in the hypothesis testing
- Operating Characteristic curve (O.C. curve)
  - Assume two population have the same variance (unknown) and sample size.

– For a specified sample size and α, larger differences are more easily detected

– To detect a specified difference δ, the more powerful test, the more sample size we need.

Confidence Intervals

- The confidence interval on the difference in means
- General form of a confidence interval:
- The 100(1-α) percent confidence interval on the difference in two means

$$\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2} S_p \sqrt{(1/n_1)+(1/n_2)} \le \mu_1 - \mu_2 \le \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2} S_p \sqrt{(1/n_1)+(1/n_2)}$$

Inferences about the Differences in Means, Paired Comparison Designs

- Example: Two different tips for a hardness testing machine
- 20 metal specimens
- Completely randomized design (10 for tip 1 and 10 for tip 2)
- Lack of homogeneity between specimens
- An alternative experiment design: 10 specimens and divide each specimen into two parts.
- The statistical model:

$$y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}, i = 1,2, j = 1,2,\cdots,10$$

  - μ_i is the true mean hardness of the ith tip,
  - β_j is an effect due to the jth specimen,

$$d_j = y_{1j} - y_{2j}$$

  - ε_{ij} is a random error with mean zero and variance σ_i^2

- The difference in the jth specimen:
- The expected value of this difference is

$$\mu_d = E(d_j) = E(y_{1j} - y_{2j}) = \mu_1 - \mu_2$$

- Testing $\mu_1 = \mu_2$ <=> testing $\mu_d = 0$
- The test statistic for H_0: $\mu_d = 0$ v.s. H_1: $\mu_d \neq 0$

$$t_0 = \frac{\bar{d}}{S_d / \sqrt{n}}$$

- Under H_0, $t_0 \sim t_{n-1}$ (paired t-test)
- Paired comparison design
- Block (the metal specimens)
- Several points:
  - Only n-1 degree of freedom (2n observations)
  - Reduce the variance and narrow the C.I. (the noise reduction property of blocking)

Inferences about the Variances of Normal Distributions

- Test the variances
- Normal distribution
- Hypothesis: H_0: $\sigma^2 = \sigma_0^2$ v.s. H_1: $\sigma^2 \neq \sigma_0^2$
- The test statistic is

$$\chi_0^2 = \frac{SS}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

- Under H_0,
- The 100(1-α) C.I.: $\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2),n-1}}$
- Hypothesis: H_0: $\sigma_1^2 = \sigma_2^2$ v.s. H_1: $\sigma_1^2 \neq \sigma_2^2$

- The test statistic is $F_0 = S_1^2 / S_2^2$, and under H_0, $F_0 = S_1^2 / S_2^2 \sim F_{n1-1,n2-1}$
- The 100(1-α) C.I.:

$$\frac{S_1^2}{S_2^2} F_{1-\alpha/2,n_2-1,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} F_{\alpha/2,n_2-1,n_1-1}$$

DOE Types

The following is a summary of some of the most common DOE types

1 One Factor Designs

These are the designs where only one factor is under investigation, and the objective is to determine whether the response is significantly different at different factor levels. The factor can be *qualitative* or *quantitative*. In the case of qualitative factors (*e.g.* different suppliers, different materials, etc.), no extrapolations (*i.e.* predictions) can be performed outside the tested levels, and only the effect of the factor on the response can be determined. On the other hand, data from tests where the factor is quantitative (*e.g.* temperature, voltage, load, etc.) can be used for both effect investigation and prediction, provided that sufficient data are available.

2 Factorial Designs

In factorial designs, multiple factors are investigated simultaneously during the test. As in one factor designs, qualitative and/or quantitative factors can be considered. The objective of these designs is to identify the factors that have a significant effect on the response, as well as investigate the effect of interactions (depending on the experiment design used). Predictions can also be performed when quantitative factors are present, but care must be taken since certain designs are very limited in the choice of the predictive model. For example, in two level designs only a linear relationship between the response and the factors can be used, which may not be realistic.

General Full Factorial Designs

In general full factorial designs, each factor can have a different number of levels, and the factors can be quantitative, qualitative or both.

Two Level Full Factorial Designs

These are factorial designs where the number of levels for each factor is restricted to two. Restricting the levels to two and running a full factorial experiment reduces the number of treatments (compared to a general full factorial experiment) and allows for the investigation of all the factors and all their interactions. If all factors are quantitative, then the data from such experiments can be used for predictive purposes, provided a linear model is appropriate for modeling the response (since only two levels are used, curvature cannot be modeled).

Two Level Fractional Factorial Designs

This is a special category of two level designs where not all factor level combinations are considered and the experimenter can choose which combinations are to be excluded. Based on the excluded combinations, certain interactions cannot be determined.

Plackett-Burman Designs

This is a special category of two level fractional factorial designs, proposed by R. L.

Plackett and J. P. Burman, where only a few specifically chosen runs are performed to investigate just the main effects (*i.e.* no interactions).

Taguchis Orthogonal Arrays

Taguchis orthogonal arrays are highly fractional designs, used to estimate main effects using only a few experimental runs. These designs are not only applicable to two level factorial experiments, but also can investigate main effects when factors have more than two levels. Designs are also available to investigate main effects for certain mixed level experiments where the factors included do not have the same number of levels.

3 Response Surface Method Designs

These are special designs that are used to determine the settings of the factors to achieve an optimum value of the response.

4 Reliability DOE

This is a special category of DOE where traditional designs, such as the two level designs, are combined with reliability methods to investigate effects of different factors on the life of a unit. In Reliability DOE, the response is a life metric (*e.g.* age, miles, cycles, etc.), and the data may contain censored observations (suspensions, interval data). One factor designs and two level factorial designs (full, fractional, and Plackett-Burman) are available in **DOE++** to conduct a Reliability DOE analysis.

Introduction to Factorial Designs

Basic Definitions and Principles

- Study the effects of two or more factors.
- Factorial designs
- Crossed: factors are arranged in a factorial design
- Main effect: the change in response produced by a change in the level of the factor

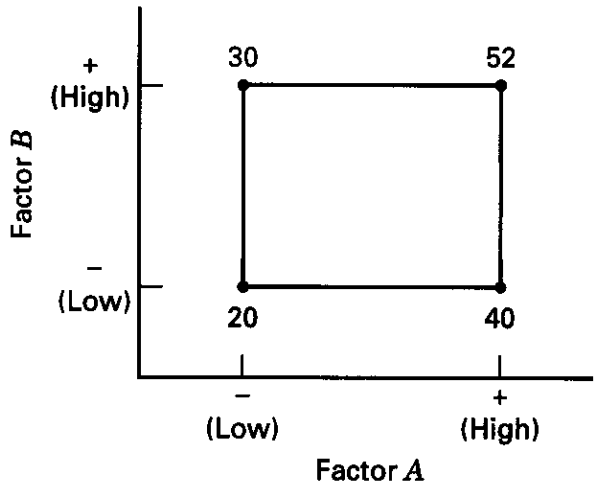


Figure 5-1 A two-factor factorial experiment, with the response (y) shown at the corners.

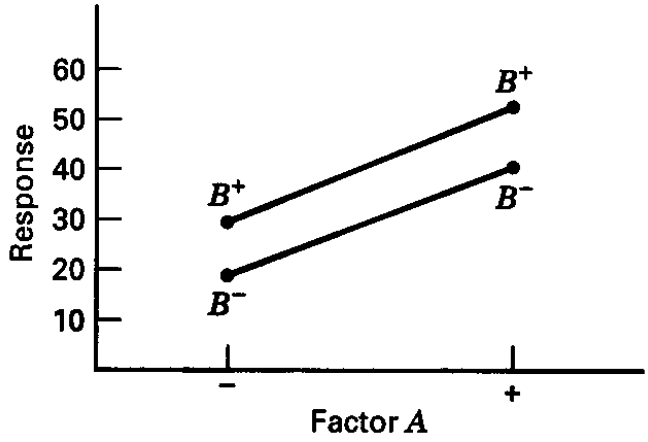


Figure 5-3 A factorial experiment without interaction.

Definition of a factor effect: The change in the mean response when the factor is changed from low to high

$$A = \bar{y}_{A^+} - \bar{y}_{A^-} = \frac{40+52}{2} - \frac{20+30}{2} = 21$$

$$B = \bar{y}_{B^+} - \bar{y}_{B^-} = \frac{30+52}{2} - \frac{20+40}{2} = 11$$

$$AB = \frac{52+20}{2} - \frac{30+40}{2} = -1$$

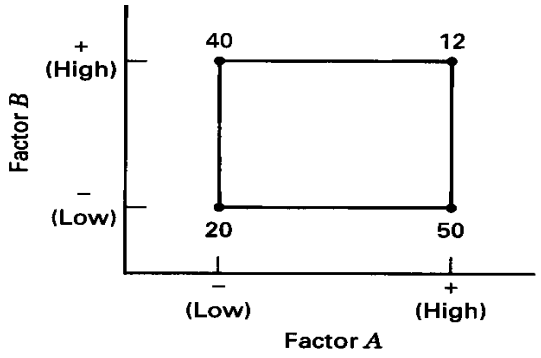


Figure 5-2 A two-factor factorial experiment with interaction.

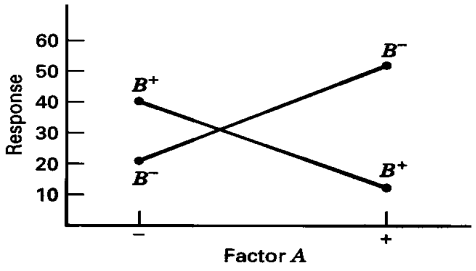


Figure 5-4 A factorial experiment with interaction.

$$A = \bar{y}_{A^+} - \bar{y}_{A^-} = \frac{50+12}{2} - \frac{20+40}{2} = 1$$

$$B = \bar{y}_{B^+} - \bar{y}_{B^-} = \frac{40+12}{2} - \frac{20+50}{2} = -9$$

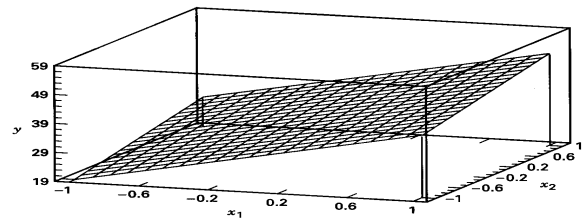
Regression Model & The Associated Response Surface

$$AB = \frac{12+20}{2} - \frac{40+50}{2} = -29$$

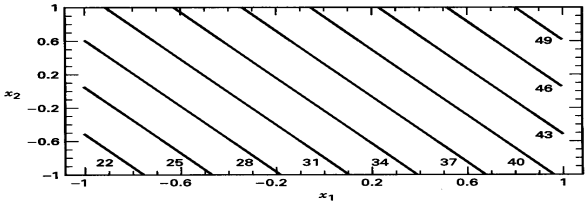
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

The least squares fit is

$$\hat{y} = 35.5 + 10.5 x_1 + 5.5 x_2 + 0.5 x_1 x_2 \cong 35.5 + 10.5 x_1 + 5.5 x_2$$



(a) The response surface



(b) The contour plot

Figure 5-5 Response surface and contour plot for the model $\hat{y} = 35.5 + 10.5x_1 + 5.5x_2$.

- When an interaction is large, the corresponding main effects have little practical meaning.
- A significant interaction will often mask the significance of main effects.

The Two-Factor Factorial Design

An Example

- *a* levels for factor A, *b* levels for factor B and *n* replicates
- Design a battery: the plate materials (3 levels) v.s. temperatures (3 levels), and $n = 4$
- The data for the Battery Design:

Table 5-1 Life (in hours) Data for the Battery Design Example

Material Type	Temperature (°F)					
	15		70		125	
1	130	155	34	40	20	70
	74	180	80	75	82	58
2	150	188	136	122	25	70
	159	126	106	115	58	45
3	138	110	174	120	96	104
	168	160	150	139	82	60

- Completely randomized design: a levels of factor A, b levels of factor B, n replicates

Table 5-2 General Arrangement for a Two-Factor Factorial Design

		Factor B			
		1	2	⋯	b
Factor A	1	$y_{111}, y_{112}, \cdots, y_{11n}$	$y_{121}, y_{122}, \cdots, y_{12n}$		$y_{1b1}, y_{1b2}, \cdots, y_{1bn}$
	2	$y_{211}, y_{212}, \cdots, y_{21n}$	$y_{221}, y_{222}, \cdots, y_{22n}$		$y_{2b1}, y_{2b2}, \cdots, y_{2bn}$
	⋮				
	a	$y_{a11}, y_{a12}, \cdots, y_{a1n}$	$y_{a21}, y_{a22}, \cdots, y_{a2n}$		$y_{ab1}, y_{ab2}, \cdots, y_{abn}$

Statistical (effects) model:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., b \\ k = 1, 2, ..., n \end{cases}$$

Testing hypotheses

$$H_0 : \tau_1 = \cdots = \tau_a = 0 \text{ v.s. } H_1 : \text{at least one } \tau_i \neq 0$$

$$H_0 : \beta_1 = \cdots = \beta_b = 0 \text{ v.s. } H_1 : \text{at least one } \beta_j \neq 0$$

]

Statistical Analysis of the Fixed Effects Model

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2$$
$$+ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

$SS_T = SS_A + SS_B + SS_{AB} + SS_E$

df breakdown:

$abn-1 = a-1+b-1+(a-1)(b-1)+ab(n-1)$

- Mean squares

$$E(MS_A) = E(SS_A/(a-1)) = \sigma^2 + \frac{bn\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$E(MS_B) = E(SS_B/(b-1)) = \sigma^2 + \frac{an\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E(MS_{AB}) = E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{ij}^2}{(a-1)(b-1)}$$

$$E(MS_E) = E\left(\frac{SS_E}{ab(n-1)}\right) = \sigma^2$$

- The ANOVA table:

Table 5-3 The Analysis of Variance Table for the Two-Factor Factorial, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A treatments	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_0 = \frac{MS_A}{MS_E}$
B treatments	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_0 = \frac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_0 = \frac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n-1)$	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	SS_T	$abn-1$		

- Multiple Comparisons:
  - Use the methods in Chapter 3.
  - Since the interaction is significant, fix the factor B at a specific level and apply Turkey's test to the means of factor A at this level.
  - See Pages 182, 183
  - Compare all *ab* cells means to determine which one differ significantly

Estimating the Model Parameters

- The model is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

- The normal equations:

$$\mu : abn\mu + bn\sum_{i=1}^{a}\tau_i + an\sum_{j=1}^{b}\beta_j + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\tau\beta)_{ij} = y_{...}$$

$$\tau_i : bn\mu + bn\tau_i + n\sum_{j=1}^{b}\beta_j + n\sum_{j=1}^{b}(\tau\beta)_{ij} = y_{i..}$$

$$\beta_j : an\mu + n\sum_{i=1}^{a}\tau_i + an\beta_j + n\sum_{i=1}^{a}(\tau\beta)_{ij} = y_{.j.}$$

$$(\tau\beta)_{ij} : n\mu + n\tau_i + n\beta_j + n(\tau\beta)_{ij} = y_{ij.}$$

- Constraints:

$$\sum_{i=1}^{a}\tau_i = 0, \sum_{j=1}^{b}\beta_j = 0, \sum_{i=1}^{a}(\tau\beta)_{ij} = \sum_{j=1}^{b}(\tau\beta)_{ij} = 0$$

- Estimations:

$$\hat{\mu} = \bar{y}_{...}$$
$$\hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{...}$$
$$\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$$
$$(\tau\beta)_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

The fitted value:

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + (\tau\beta)_{ij} = \bar{y}_{ij\cdot}$$

- Choice of sample size: Use OC curves to choose the proper sample size.
- Consider a two-factor model without interaction:
    - Table 5.8
    - The fitted values:
    - Figure 5.15
- One observation per cell:
    - The error variance is not estimable because the two-factor interaction and the error can not be separated.
    - Assume no interaction. (Table 5.9) $\hat{y}_{ijk} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + (\tau\beta)_{ij} = \bar{y}_{ij\cdot}$
    - Tukey (1949): assume $(\tau\beta)_{ij} = r\tau_i\beta_j$ (Page 192)
    - Example 5.2

The General Factorial Design

- More than two factors: *a* levels of factor A, *b* levels of factor B, *c* levels of factor C, …, and *n* replicates.
- Total *abc ... n* observations.
- For a fixed effects model, test statistics for each main effect and interaction may be constructed by dividing the corresponding mean square for effect or interaction by the mean square error.
- Degree of freedom:
    - Main effect: # of levels – 1
    - Interaction: the product of the # of degrees of freedom associated with the individual components of the interaction.

- The three factor analysis of variance model:
  - $$y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij}$$
  - $$+ (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$
  - The ANOVA table (see Table 5.12)
  - Computing formulas for the sums of squares (see Page 196)
  - Example 5.3

Foundations of analytics

The importance of Business Analytics

The importance of business analytics (BA) is revealed in a recent Bloomberg survey, which found that BA has been effective in decision making for three out of four enterprises. Among the improvements are increased profitability, reduced cost, faster decision making and critical performance improvements.

Meaning can be found in data

BA refers to technologies, applications, skills and practices for the investigation of past business performance to improve insight into this past performance. BA is much more than merely providing simple data to a business. By using BA tools, meaning can be found in data which results in a business improving its business intelligence. Some examples of Business Analytics include statistical analysis, decision processes, web analytics etc.

Creating novel insights through Business Analytics

An important distinction needs to be made between BA and Business Intelligence (BI). BA mostly focuses on creating novel insights and the understanding of business performance based on statistical methods, data, quantitative analysis, explanatory and predictive modeling and fact-based management to *drive decision making*. Business Intelligence also uses data and statistical methods, yet it rather focuses on using a set of metrics to measure both past performance and *guide business planning*. Also it looks at querying, reporting, OLAP and alerts. Examples of questions asked by Business Analytics are “why is this happening”, “what if these trends continue”, “what will happen next” and “what is the optimal outcome”. Examples of questions

asked by BI are “what happened”, “how many times did it happen”, “where the problem is” and “what are the solutions to the problem”.

Big Data Analytics

With big data only going to increase, the use of business analytics has increased with a greater need to understand this data and the realization that Business Intelligence has many advantages. In the past three years, the percentage of respondents benefitting from business analytics has risen from 90% to 97%, clearly emphasising business’s increasing thirst to make sense of its data.

Embrace of analytics by business executives

The most important characteristics of business analytics for businesses are a top-down embrace of analytics by senior leaders and managers. This allows them to make better decisions by accessing data themselves which might only have previously been accessible to the IT guys. In this way Business Analytics that informs business decisions can be used to optimize business processes. Also, data-driven companies use their data as a strategic asset by enhancing competitiveness.

Create an environment which is receptive to innovation

A recent paper by IBM shows the three ways in which the most successful companies are operating when they deploy business analytics. The high performing businesses are firstly *challenging* in that they create an environment which is receptive to innovation by disrupting the status quo, and they are more ambitious in applying new insights. Secondly, they’re good at *anticipating and predicting* outcomes and trade-offs proactively. Thirdly, they’re *empowering* by allowing employees to use information with confidence which helps them act on their insights to improve the business.

Applications of Business Analytics

Some specific applications of Business Analytics, which allow for many novel avenues of opportunity for businesses to optimize and adapt their business model, are:

- critical product analysis
- improved customer service
- up-selling opportunities
- simplified inventory management
- and competitive price insights.

Critical product analysis allows minor alterations to be made of a location specific product, including aiding the study in trends associated with those locations. Improved customer service keep track of frequent customer queries which prevents businesses from repeating mistakes and improving customer satisfaction. Up-selling opportunities identify the most prominent needs of a business's customer base. Simplified inventory management is supported by Business Analytics as gathered data can help predict which products are on the verge of becoming outdated, minimising losses. Competitive price insights can help businesses make their prices competitive by tracking the customer trends and price ranges which suite the customers.

The quality of the data can be a huge headache for managers

Start up businesses without any data of their own need to either purchase data or focus on gathering it over time. Purchasing data collected by third parties is a solution, however these are also prone to inaccuracies or defunct records. Business Analytics can therefore be a time-consuming and expensive affair for start up businesses. The quality of the data can also be a huge headache for managers, as some records are either duplicated or misleading, making analysis difficult. However, working with data deviations will allow for fairly accurate analysis.

Project management

Project management is the application of rocesses, methods, knowledge, skills and experience to achieve the project objectives. General. A project is a unique, transient endeavour, undertaken to achieve planned objectives, which could be defined in terms of outputs, outcomes or benefits.

Definition of PERT

PERT is an acronym for Program (Project) Evaluation and Review Technique, in which planning, scheduling, organizing, coordinating and controlling of uncertain activities take place. The technique studies and represents the tasks undertaken to complete a project, to identify the least time for completing a task and the minimum time required to complete the whole project. It was developed in the late 1950s. It is aimed to reduce the time and cost of the project.

PERT uses time as a variable which represents the planned resource application along with performance specification. In this technique, first of all the project is divided into activities and events. Thereafter proper sequence is ascertained and a network is constructed. After that time needed in each activity is calculated and the critical path (longest path connecting all the events) is determined.

Definition of CPM

Developed in the late 1950's, Critical Path Method or CPM is an algorithm used for planning, scheduling, coordination and control of activities in a project. Here, it is assumed that the activity duration are fixed and certain. CPM is used to compute the earliest and latest possible start time for each activity.

The process differentiates the critical and non-critical activities in order to reduce time and avoid the queue generation in the process. The reason behind the identification of critical activities is that, if any activity is delayed, it will cause the whole process to suffer. That is why, it is named as Critical Path Method.

In this method, first of all, a list is prepared consisting of all the activities needed to complete a project, followed by the computation of time required to complete each activity. Thereafter, the dependency between the activities is determined. Here, 'path' is defined as a sequence of activities in a network. The critical path is the path with highest length.

Comparison Chart

BASIS FOR COMPARISON	PERT	CPM
Meaning	PERT is a project management technique, used to manage uncertain activities of a project.	CPM is a statistical technique of project management that manages well defined activities of a project.
What is it?	A technique of planning and control of time.	A method to control cost and time.
Focus on	Event	Activity
Model	Deterministic Model	Probabilistic Model
Estimates	Three time estimates	One time estimate
Appropriate for	High precision time estimate	Reasonable time estimate
Management of	Unpredictable Activities	Predictable activities
Nature of jobs	Non-repetitive nature	Repetitive nature
Critical and Non-critical activities	No differentiation	Differentiated

BASIS FOR COMPARISON	PERT	CPM
Suitable for	Research and Development Project	Non-research projects like civil construction, ship building etc.
Crashing concept	Not Applicable	Applicable