

BRM Presentation

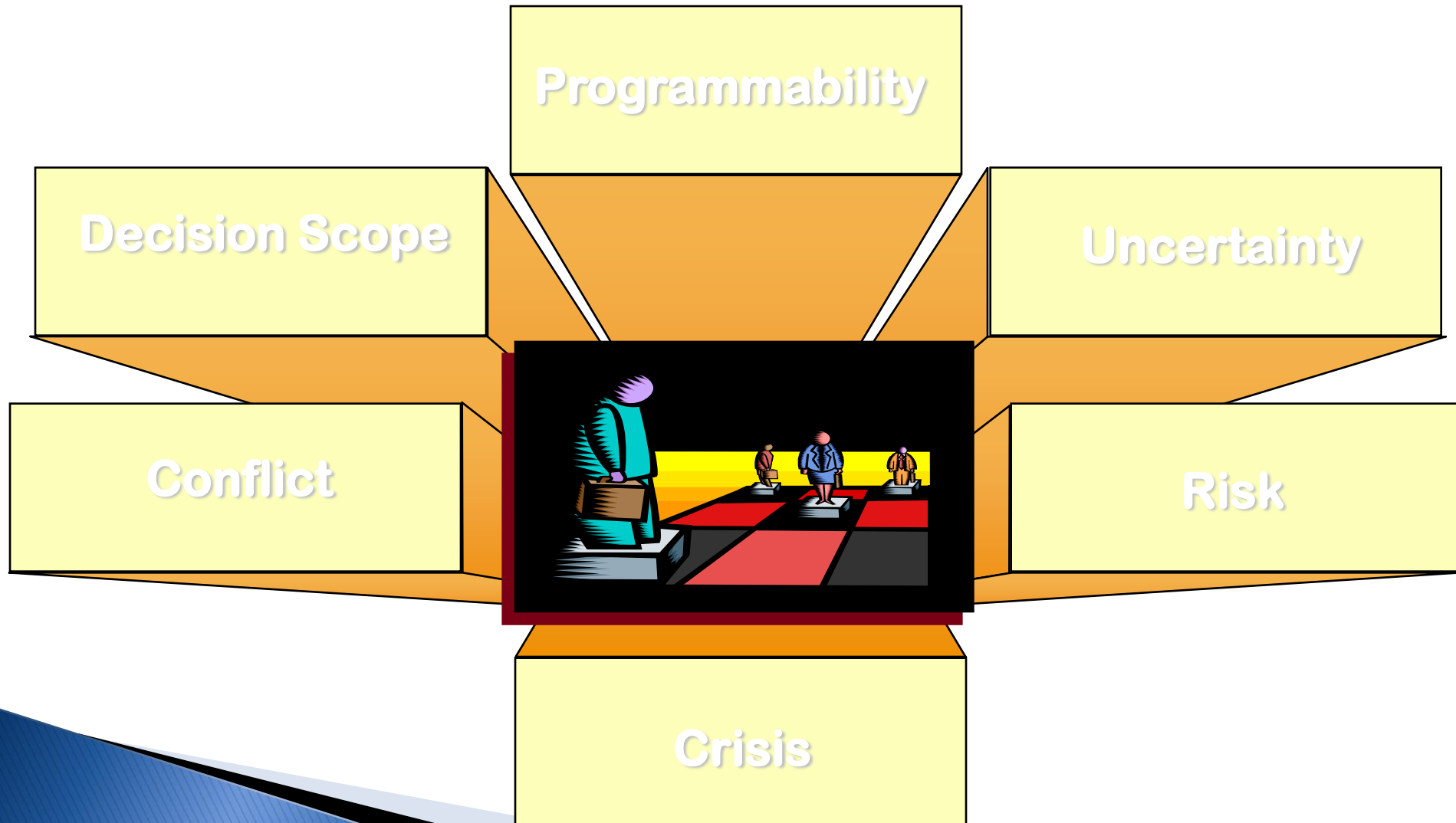
Dr. Hemal Pandya

Decision Making

- ▶ The process of identifying problems and opportunities and resolving them.
- ▶ Management decisions can be made by managers, teams, or individual employees, depending on:
 - The scope of the decision, and
 - The design and structure of the organization.



Characteristics of Management Decision Making



The Nature of Management Decisions

- ▶ There are differences between management and scientific decisions
- ▶ Management decisions usually affect many more people and need to be made in a changing and uncertain environment
- ▶ Anyhow, the process used for scientific decision making is similar to management decision making

Management Decision

- ▶ The steps are:
 - definition of the problem
 - gathering facts related to the problem
 - comparing these with right and wrong criteria based on knowledge and experience
 - and then taking the best course of action
- ▶ Management decision making, however, is often an art rather than a science. The conventional theories of decision making do not always apply

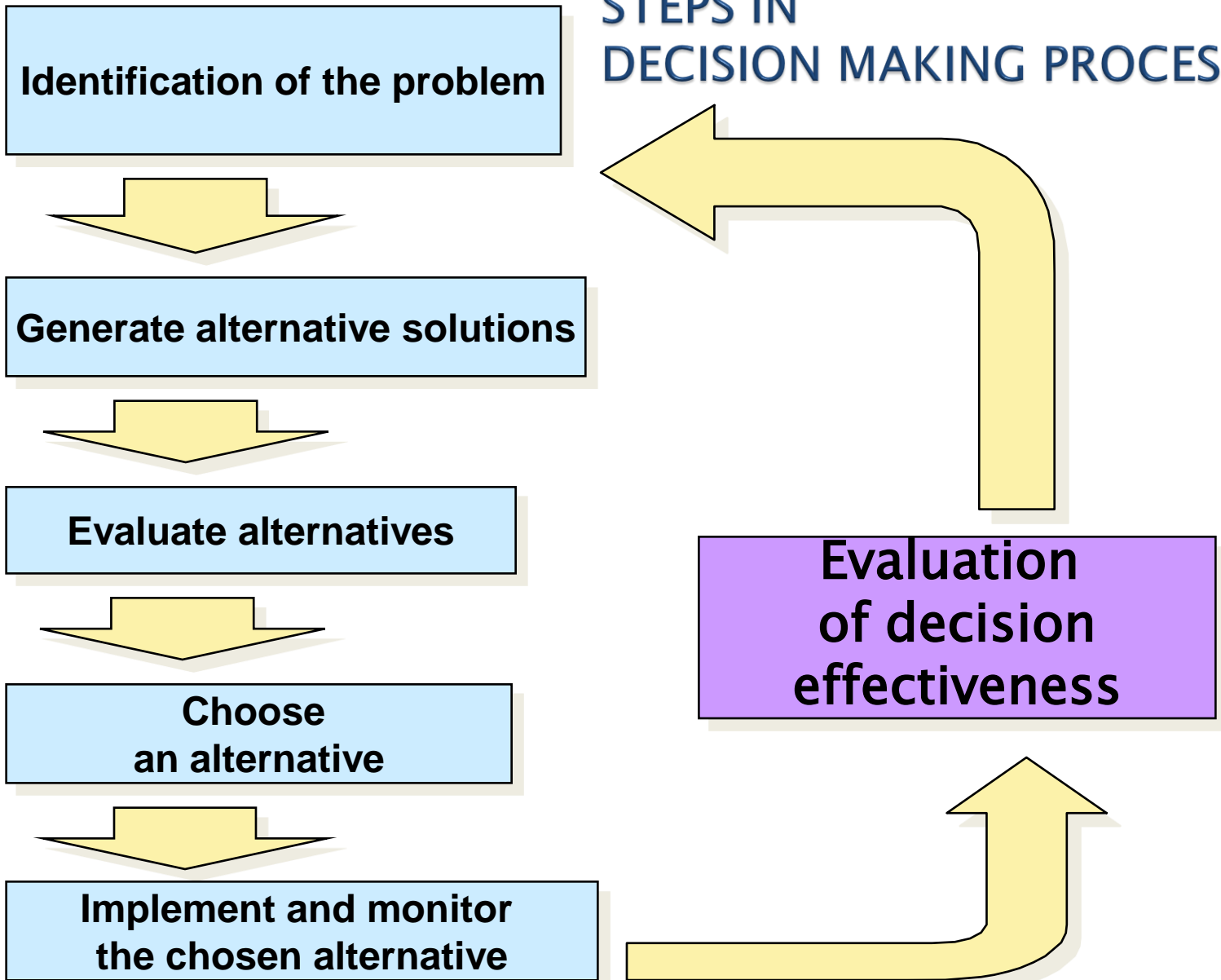
Main characteristics of management decision making

- ▶ The problem is often ill defined or unpredictable because of being related to human behaviour
- ▶ Management problems are usually much wider in scope than technical problems, and affect many more people and functions
- ▶ Gathering information on which to base a management decision is often a hazardous process. The problem is continually changing, due to market or people considerations
- ▶ The management decision, like the scientific one, is essentially one of choosing between several alternatives. This may not be easy to do as the consequences of each alternative are often difficult to predict

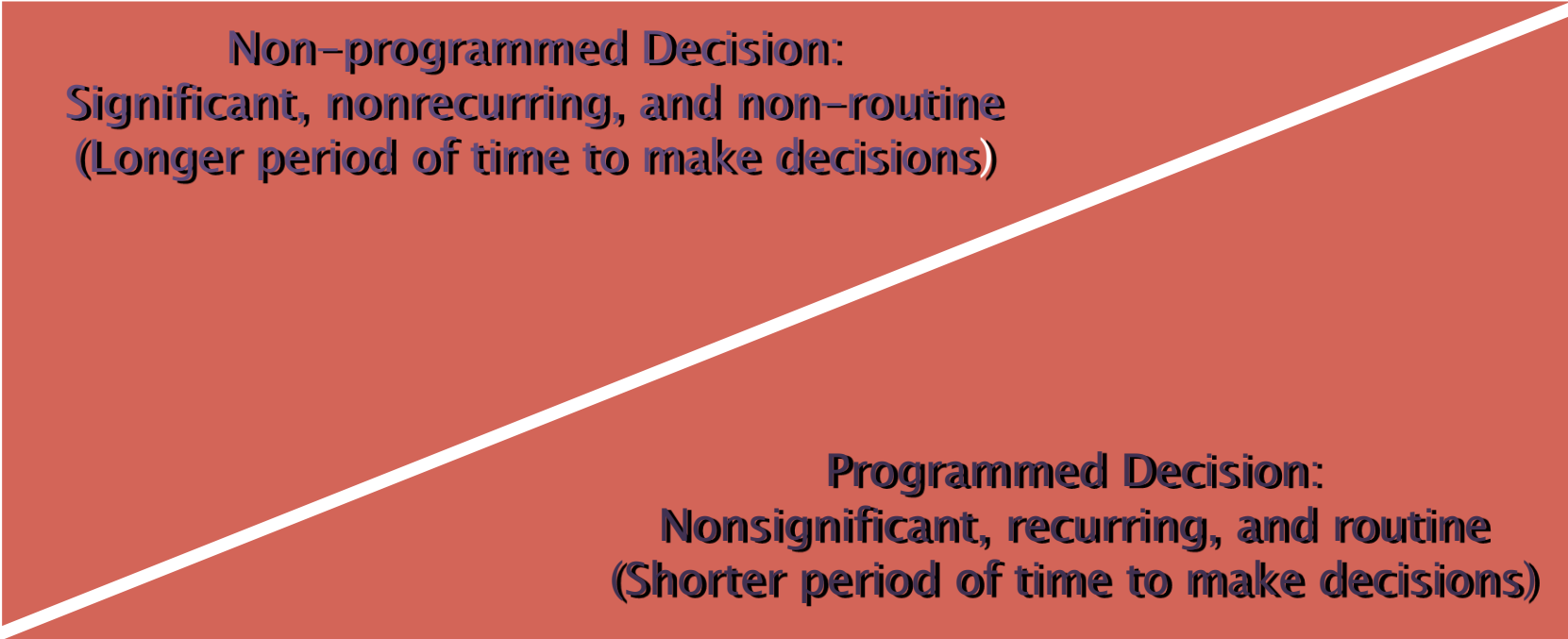
Main characteristics of management decision making (contd)

- ▶ There is *rarely* a 'best' solution in management problems
- ▶ It is highly probable that what is considered to be the best solution today may turn out to be the worst choice in the future
- ▶ Once a decision has been made it requires consensus and commitment from the people who are affected and who are to implement the solutions
- ▶ Following implementation the decision must be continually monitored to see whether it is still valid in a changing environment.
- ▶ The process of managements decision making cannot be learned, except the experience and judgment
- ▶ Decision making styles are often affected by the organization in which a manager operates

STEPS IN DECISION MAKING PROCESS



Decision Structure Continuum



Non-programmed Decision:
Significant, nonrecurring, and non-routine
(Longer period of time to make decisions)

Programmed Decision:
Nonsignificant, recurring, and routine
(Shorter period of time to make decisions)

Types of Decisions

- ▶ Programmed Decisions
 - Recurring or routine situations in which the decision-maker should use decision rules or organizational policies and procedures to make the decision. **Programmed decision** is a repetitive decision that can be handle by a routine approach. (Procedure, rule, policy)
- ▶ Non-programmed Decisions
 - Significant and non-recurring and non-routine situations in which the decision-maker should use the decision-making model. **Non-programmed decision** is a decision that must be custom-made to solve unique and non-recurring problems.

Continua for Classifying a Problem (cont'd)

Which Decision Model to Use

Decision-Making Model

Bounded Rationality Model



(Optimizing)

(Satisficing)

When to Use Group or Individual Decision Making

Group Decisions

Individual Decisions



Models of Decision Making

- ▶ Rational–Economic Model



- ▶ Behavioral Decision Model

Rational–Economic Model

- ▶ A framework that assumes managers have completely **accurate information**.
- ▶ Concentrates on how decisions **should** be made, **not** on how they are **actually** made.

Assumptions of Rational–Economic Model

- ▶ Managers have “**perfect information.**”
- ▶ **Objectives are known** and agreed to.
- ▶ **Managers are rational**, systematic, and logical.
- ▶ Managers work in **the best interests** of their organizations.
- ▶ **Ethical decisions** do not arise in the decision–making process.

Drawbacks of the Rational–Economic Model

- ▶ In practice, the model **may not always be realistic.**
- ▶ Leaders rarely have access to **perfect information.**
 - Even if available, decision makers are limited in their ability to **comprehend vast amounts of information.**

Drawbacks of the Rational–Economic Model

- Decision makers seldom can accurately **forecast future consequences**.
- Fatigue, emotions, attitudes, **motives of behaviors all intervene** to prevent a rational decision making.
- Individual culture and **ethical values** influence the decision process.

Behavioral Decision Model

- ▶ The behavioral model acknowledges **human limitations** that make rational **decisions difficult**.
- ▶ The behavioral model suggests that cognitive **ability to process** information **is limited**.

Behavioral Decision Model

- ▶ Managers usually attempt to behave rationally within their **limited perception of a situation.**
- ▶ Most organizational situations are so complex that managers view problems within **sharply restricted bounds.**
- ▶ Manager's behavior can be considered **rational**, in **terms** of their **simplified view** of the problem.

Bounded Rationality

- ▶ This is the behavior that people construct simplified models that extract the essential features from problems without capturing all of their complexities in order to decide rationally.

What is research?

Management research is an unbiased, structured, and *sequential method of enquiry*, directed towards a clear implicit or explicit business objective. This enquiry might lead to *validating existing postulates* or *arriving at new theories and models*.

Scientific Approach to Research

The scientific method is based on certain “articles of faith” which are:

- ▶ Logical reasoning process
- ▶ Reliance on empirical evidence
- ▶ Use of relevant concepts
- ▶ Commitment to objectivity
- ▶ Ethical neutrality
- ▶ Generalization
- ▶ Verifiability

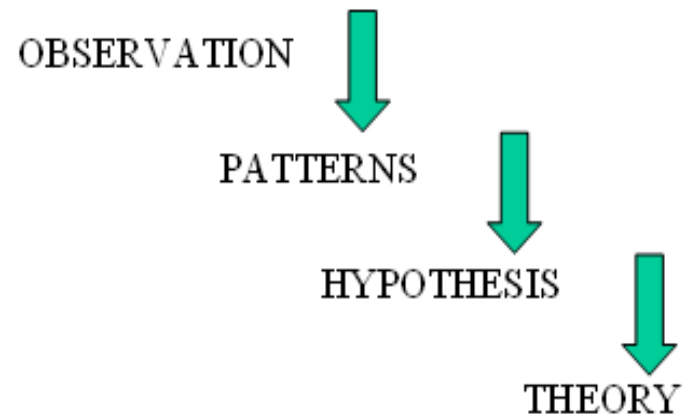
Logical Reasoning Process:

- ▶ Induction
- ▶ Deduction

INDUCTION

Figure-8: Inductive Reasoning

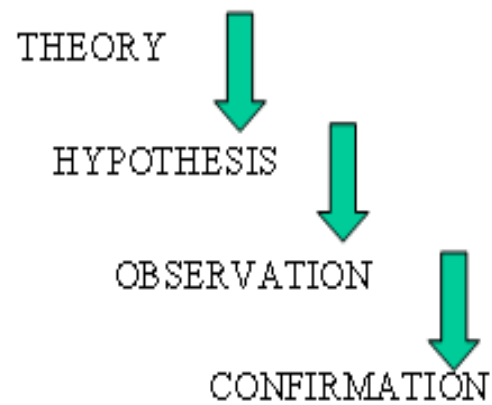
Inductive thinking (Qualitative)



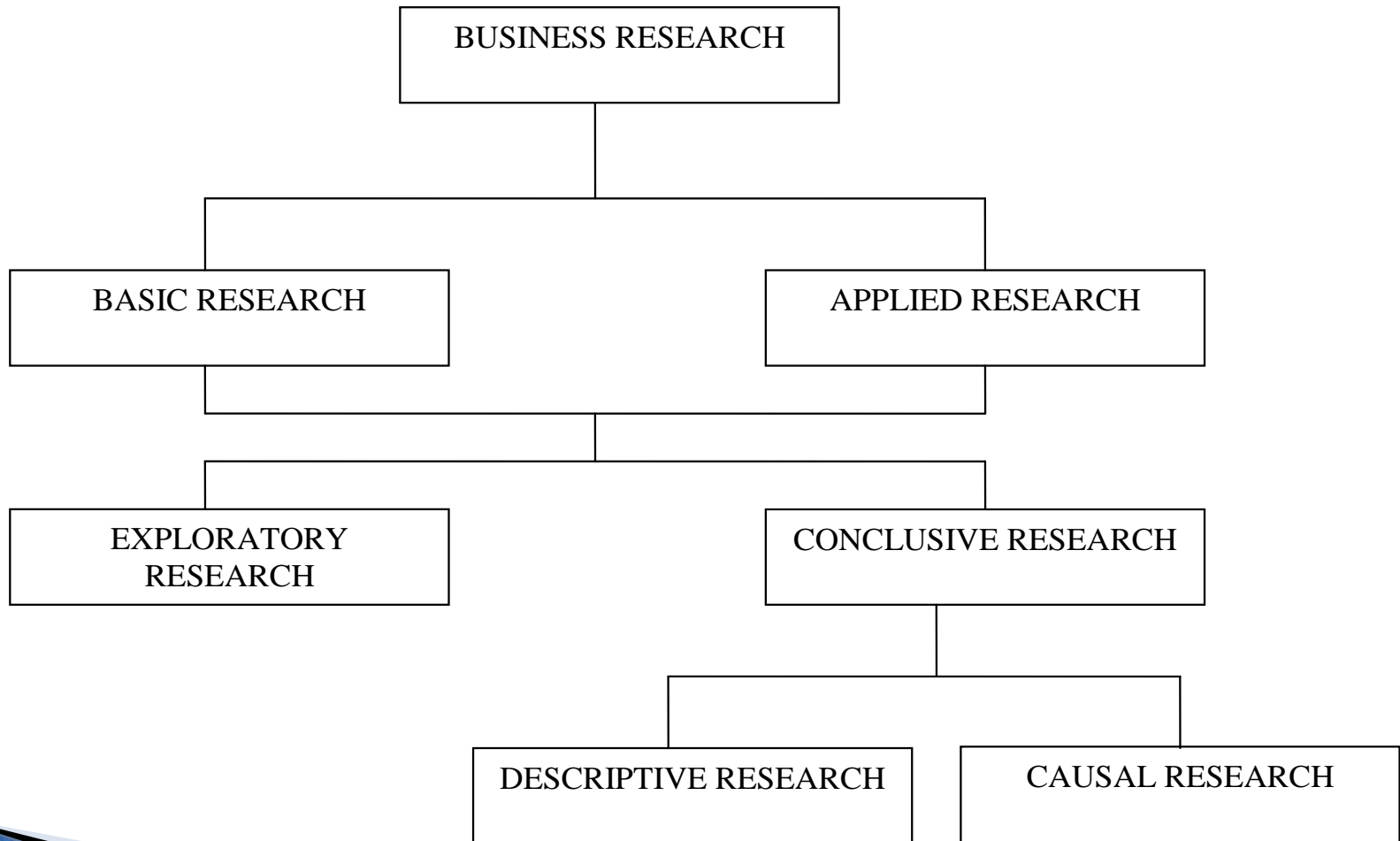
DEDUCTION

Figure-9: Deductive Reasoning

Deductive thinking (Quantitative)



Types of Research

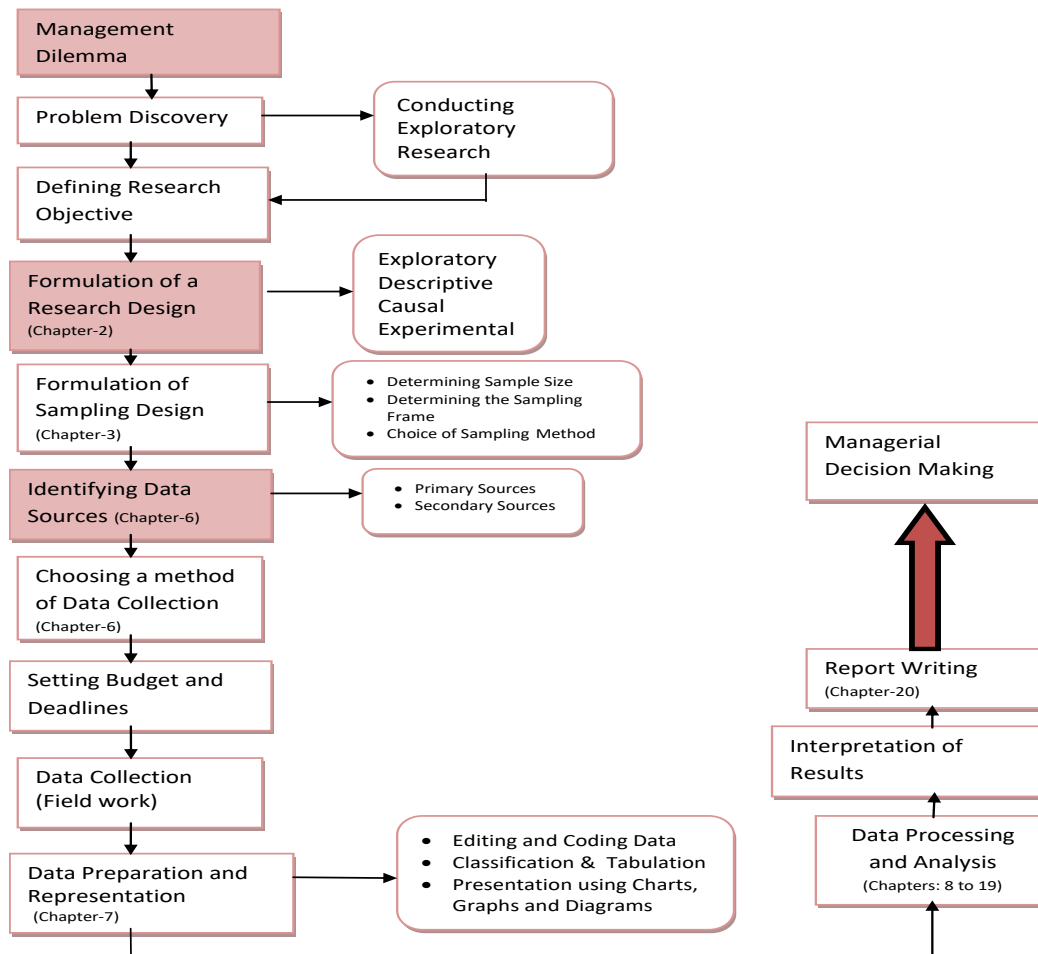


Types of Research

- ▶ Basic or Pure Research
- ▶ Applied Research
- ▶ Diagnostic Study
- ▶ Evaluation Study
- ▶ Action Research
- ▶ Historical Research
- ▶ Experimental Research
- ▶ Literature Review Research
- ▶ Library Research
- ▶ Paradigm Research
- ▶ Pre-Paradigm Research

Business Research Process

Business Research Process



Objectives of Business Research

- ▶ To identify and define opportunities and problems.
- ▶ To define, monitor, and refine strategies.
- ▶ To define, monitor, and refine tactics.
- ▶ To frame the operational action plans in order to successfully achieve the short-term goals and in turn the long-term objectives leading to corporate growth.

Factors Aiding Business Research

Business and management have become distinctive focus areas for research due to:

- ▶ Competition – local, domestic and global
- ▶ More complex business organizations
- ▶ Stakeholders demanding greater role
- ▶ Shortened length of Product life cycle and Process life cycle
- ▶ Growth of Internet
- ▶ Growth of technology and greater computing power
- ▶ Advanced analytical techniques
- ▶ Availability of analytical soft wares

These factors not only aid the business research but also instigate a strong need for business research in the managerial decision-making process.

Scope of Business Research

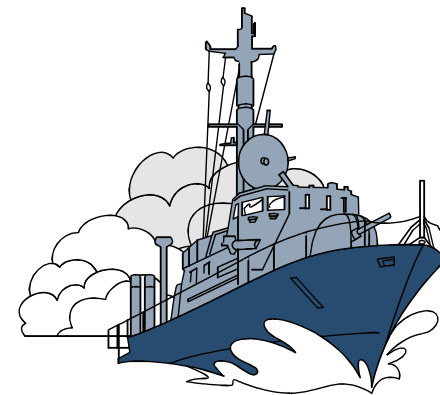
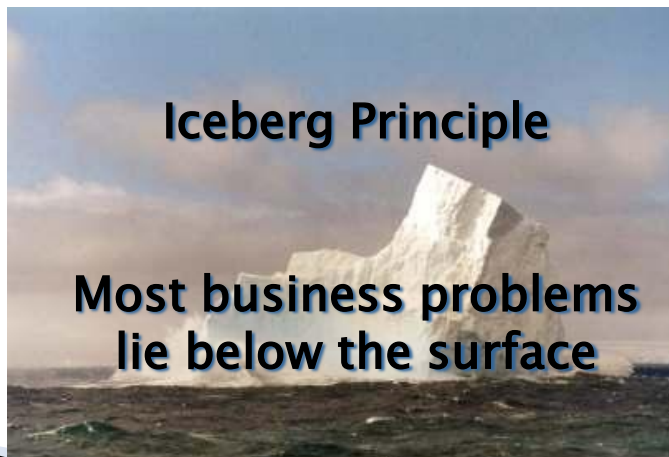
Scope of Business Research	
<p>General Business Conditions and Corporate Research</p> <ul style="list-style-type: none"> • Short Range & Long-Range Forecasting • Business and Industry Trends • Global Environments • Inflation and Pricing • Plant and Warehouse Location • Acquisitions <p>Management and Organizational Behaviour Research</p> <ul style="list-style-type: none"> • Total Quality Management • Morale and Job Satisfaction • Leadership Style • Employee Productivity • Organizational Effectiveness • Structural issues • Absenteeism and turnover • Organizational Climate <p>Financial and Accounting Research</p> <ul style="list-style-type: none"> • Forecasts of financial interest rate trends, • Stock, bond and commodity value predictions • capital formation alternatives • mergers and acquisitions • risk-return trade-offs • portfolio analysis • impact of taxes • research on financial institutions • expected rate of return • capital asset pricing models • credit risk • cost analysis 	<p>Sales and Marketing Research</p> <ul style="list-style-type: none"> • Market Share • Market segmentation • Market characteristics • Sales Analysis • Establishment of sales quotas • Distribution channels • New product concepts • Test markets • Advertising research • Buyer behaviour • Customer satisfaction • Website visitation rates <p>Information Systems Research</p> <ul style="list-style-type: none"> • Knowledge and information needs assessment • Computer information system use and evaluation • Technical support satisfaction • Database analysis • Data mining • Enterprise resource planning systems • Customer relationship management systems <p>Corporate Responsibility Research</p> <ul style="list-style-type: none"> • Ecological Impact • Legal Constraints on advertising and promotion • Sex, age and racial discrimination / worker equity • Social values and ethics • Corporate Governance

Defining the Research Problem

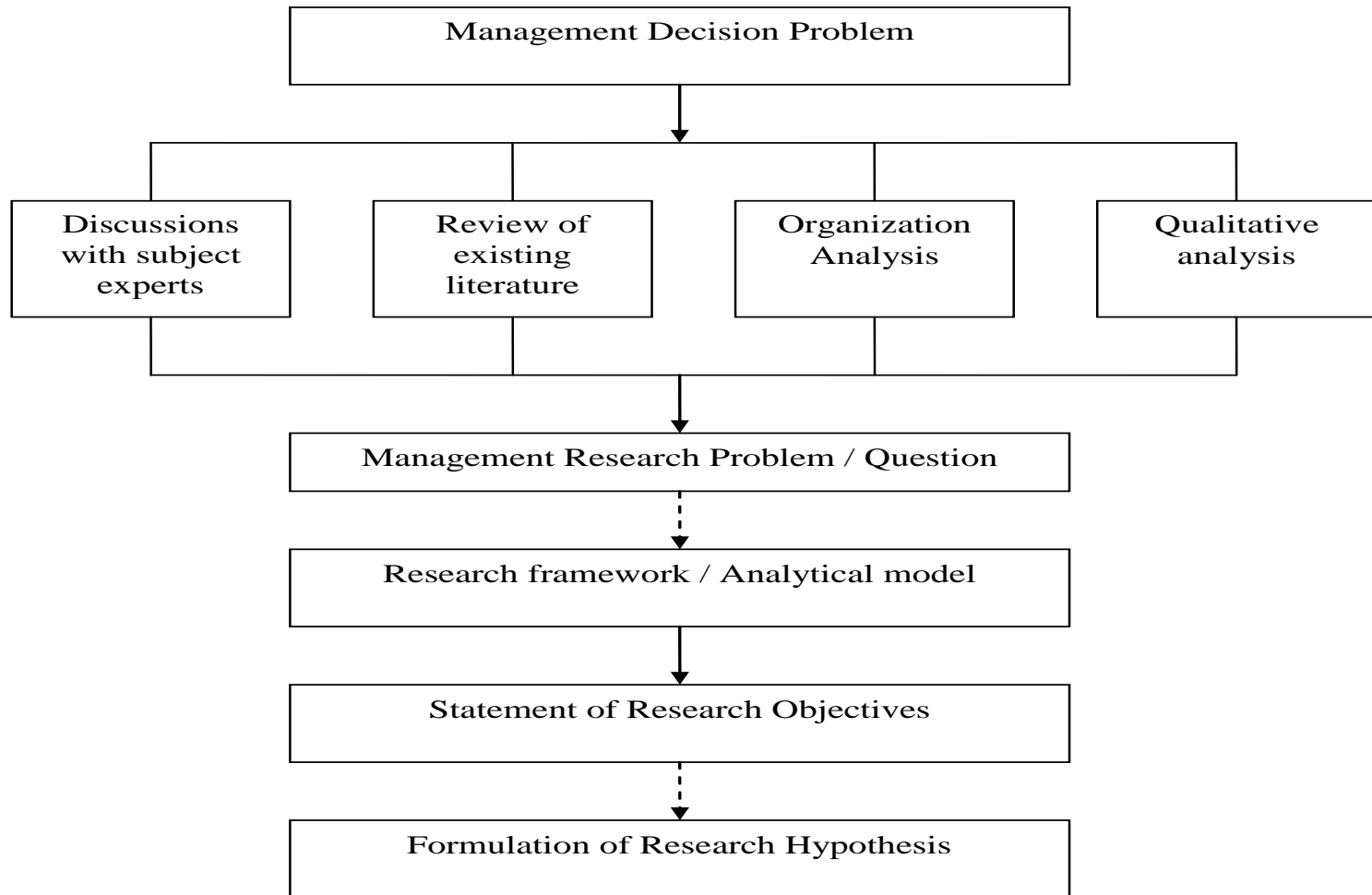
A research problem can be defined as a *gap* or uncertainty in the decision makers' *existing body of knowledge* which inhibits efficient decision making. The gap could be *academic & theoretical* (basic) or *real time and action oriented* (applied).

The Iceberg Principle

- ▶ The principle indicating that the dangerous part of many business problems is neither visible to nor understood by managers.



Problem Identification Process



Elements of a Research Problem

- Unit of analysis
- Independent variable
- Dependent variable
- Extraneous independent variable
- Intervening variables
- Moderating variables

The Research Hypothesis

Hypothesis is any assumption/presupposition that the researcher makes about the probable direction of the results that might be obtained on the completion of the research process

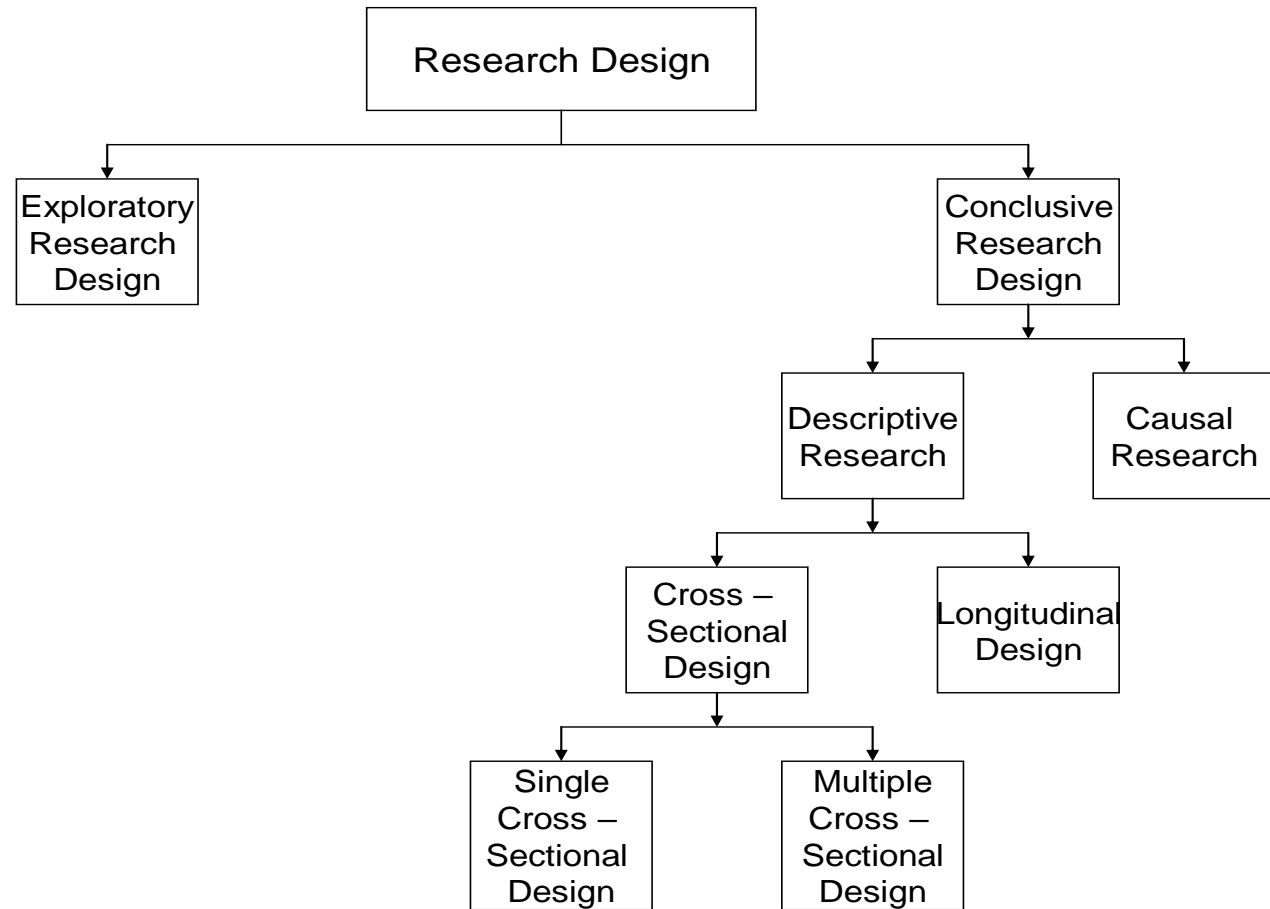
Descriptive hypothesis: This is simply a statement about the magnitude, trend, or behaviour of a population under study.

- ▶ **Relational hypothesis:** These are the typical kind of hypotheses which state the expected relationship between two variables.

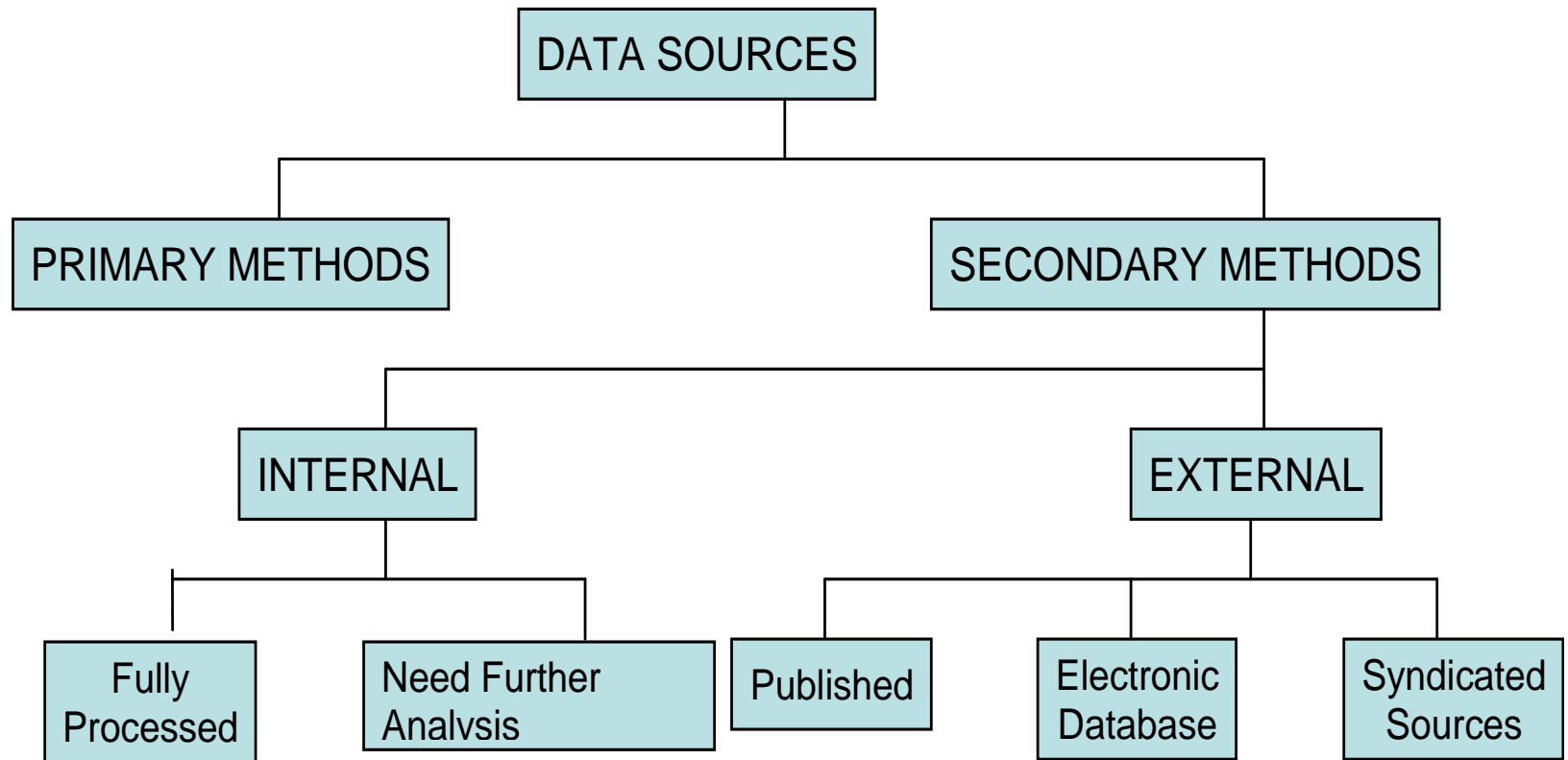
Criteria for Hypothesis Formulation

- ▶ In simple, declarative statement form
- ▶ Measurable and quantifiable
- ▶ Is essentially a conjectural statement
- ▶ Has underlying assumptions on the testing of the stated relationships

Classification of Research Designs



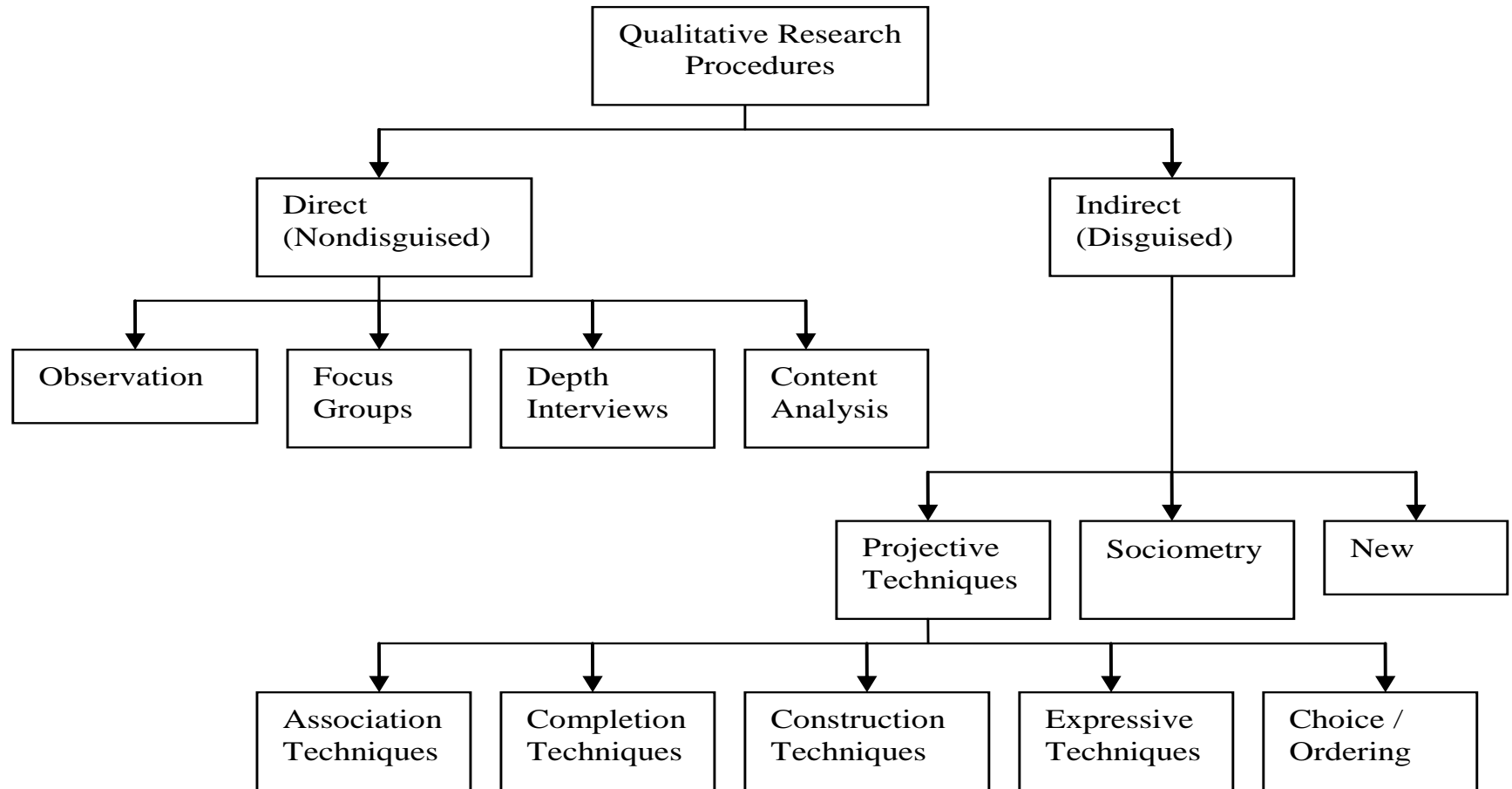
Sources of Data



A Comparison of Qualitative and Quantitative Research

	Qualitative Research	Quantitative Research
Basic research objective	To gain a broad qualitative understanding of the underlying reasons and motivations; As a first step in multistage research	To quantify the data and generalize the results from the sample to the population of interest; Recommend a final course of action
Type of sample used	Small numbers of non-representative cases	Large number of representative cases
Data collection Method	Unstructured	Structured
Nature of data analysis	Non-statistical	Statistical

Classification of Qualitative Methods



Qualitative research = Observation technique

Classification of Observation

▶ Direct vs indirect:

- *Direct* >> *observing behavior as it occurs*
- *Indirect* >> *observing the effects of behavior*

▶ Disguised vs nondisguised

- *Nondisguised* >> *Direct*
- *Disguised* >> *Indirect*

▶ Structured vs unstructured

- *Structured* >> *predetermine what to observe*
- *Unstructured* >> *monitor all behavior*

▶ Human vs Mechanical

- *Human* >> *observation done by human beings*
- *Mechanical* >> *observation by machine*

Observation I

▶ **Appropriate Conditions**

- The event must occur in a short time interval,
 - *avoid lag affect*
- Must occur in a setting where the researcher can observe the behavior
 - *Praying, cooking are not suitable things to observe*
- Necessary under situations of faulty recall
 - *Faulty recall>>remembering things such as how many times one looked at his wristwatch.*

Observation:

Advantages and Limitations

▶ **Advantages**

- *Greater data accuracy than direct questioning, in natural settings people behave naturally,*
- *Problems of refusal, not at home, false response, non-cooperation etc. are absent,*
- *No recall error,*
- *In some situations, only way*
 - *Number of customers visiting a store*
 - *Studying children's behavior*

▶ **Limitations**

- *Time consuming,* *-- too many things to*
observe,
- *may not be representative,*
- *difficulty in determining root cause of the behavior.*

Focus Group I

- ▶ *An interview conducted by a trained moderator in a non-structured and natural manner with a small group of respondents.*

Group size	8-12
Group composition prescreened	Homogenous, respondents
Physical setting	Relaxed, informal setting
Time duration	1 - 3 hours
Recording	Use of audio and video cassettes
Moderator	Observational, interpersonal, good communication skills needed.

Focus Groups II

▶ Objectives:

- Generate new product or service ideas
- Understand consumer vocabulary
 - *Useful for ad campaigns*
- Reveal consumer needs, motives, perceptions and attitudes,
 - *Generating future research objectives*
- Facilitate understanding of the quantitative studies

Focus Group Procedure

Determine the objectives of the Marketing Research Project and define the problem

Specify the objective of qualitative research

State the objectives/questions to be answered by the focus group

Write a screening questionnaire

Develop a moderator's outline

Conduct the focus group interview

Review tapes and analyze data

Summarize the findings and plan follow-up research

The Focus Group Moderator

- ▶ **The person who conducts the focus group session.**
 - Success of focus groups depend on him/her,
 - He/she must strive for generating a stimulating natural discussion without losing sight of the focus,
 - Must take initiative, but should not dominate the discussion unduly,
 - Should have feeling of urgency,
 - Should participate in the research from the beginning,
 - Must add value beyond just conducting the session.

Traits of a Good Focus Group Moderator

A Good Focus Group Moderator...

1. *must have experienced in conducting focus group research;*
 2. *should participate in conceptualizing the focus group research design, rather than simply executing the groups exactly as asked and take personal and take responsibility for the recruitment, screening, and selecting of participants.*
 3. *must engage in advance preparation to improve overall knowledge of the area being discussed and prepare a detailed guide to moderate the focus group..*
 4. *must demonstrate the enthusiasm and exhibit the energy necessary to keep the group interested yet maintain control of the group without leading or influencing the participants;*
 5. *should be open to modern techniques (e.g., attitude scaling, conceptual mapping, visual stimulation, or role-playing) which can be used to delve deeper into the minds of participants;*
 6. *must share in the feeling of urgency to complete the focus group while desiring to achieve an excellent total research project; and*
 7. *must provide some “added value” to the project beyond just conducting the session.*
-

Focus Group: Advantages and Disadvantages

▶ Major Advantages:

- *Synergism, Snowballing, Stimulation, Security, Spontaneity, Speed and Cost savings.*

▶ Major Disadvantages:

- *Lack of representativeness, Misuse, Misjudge, Moderation problem, and Difficulty of analysis*
- ▶ *A very promising technique.*

Seven Advantages of Focus Group I

1. **Synergism.** When a group of people with similar interests discuss an issue together, they are likely to produce a richer insight, wider range of information, and innovative ideas than will individual responses obtained privately.
2. **Snowballing.** In a group discussion, one person's comment often triggers a chain reaction from the other participants and generates more views.
3. **Stimulation.** Once the focus group discussion is underway, general level of excitement over the topic increases, and a large number of respondents want to express their ideas and expose their feelings.

Seven Advantages of Focus Group II

4. Security. Because of homogeneity of composition, focus group participants have similar feelings. This enables them to feel comfortable and uninhibited to express their ideas/feelings.

5. Spontaneity. In focus groups participants are not required to answer specific questions. Their responses can be spontaneous and unconventional reflecting an accurate idea of their views

6. Speed. Because people discuss issues simultaneously, data collection and analysis in focus group proceed relatively quickly.

7. Inexpensive. Considering the richness of output, it is a relatively inexpensive method of data collection.

Five Disadvantages of Focus Group

1. **Lack of representativeness.** Focus groups are not representative of the general population. Hence, results of focus group discussions are not projectable and should not be the only basis for decision making.
2. **Misjudgement.** Focus groups are generally susceptible to client and researcher biases. As such, compared to the results of other data collection techniques, focus group results could be easily misjudged.
3. **Misuse.** Focus groups can be misused and abused by considering the results as conclusive rather than exploratory.
4. **Moderation.** Skills of the moderator is a major determinant of focus group success and the quality of their results. But moderators with desirable skills are rare.
5. **Difficult to analyze.** The unstructured nature of the responses in focus group discussions makes coding, analysis, and interpretation difficult.

Other Qualitative Techniques

Depth Interview: *An unstructured interview that seeks opinions of respondents on a one-to-one basis. Useful for sensitive issues, politics etc.*

Protocol Analysis: *Involves placing a person in a decision making situation and asking him/her to state everything he/she considers in making a decision. Useful in 1. Purchasing involving a long time frame (car, house) and 2. Where the decision process is too short (greeting card).*

Projective technique: *Involve situations in which participants are placed in simulated activities hoping that they will divulge information about themselves that are unlikely to be revealed under direct questioning.*

Projective Techniques

- ▶ These are indirect interviewing methods which enable sampled respondents to project their views, beliefs and feelings onto a third-party or into some task situation.
- ▶ The researcher sets up a situation for the respondents asking them to express their own views, or to complete/interpret some ambiguous stimulus presented to them.
- ▶ **Various types. More common ones are:**
 - Free Word Association
 - Sentence Completion
 - Unfinished scenario/story completion
 - Cartoon completion test

FREE WORD ASSOCIATION

In this technique, a list of carefully selected stimulus words or phrases related to the topic of research are read out, one at a time, to a respondent. The respondent is asked to respond with the first word or phrase that comes to his/her mind. The list of words should contain a mixture of test words and neutral words.

*In the example shown here, the researchers seems to be interested in studying high-tech banking (words with *).*

However, analyzing and interpreting test results are rather difficult.

<u>Stimulus Word</u>	<u>Response</u>
Postman	_____
Bank Teller*	_____
Networking	_____
Automatic teller machine*	_____
Persian Carpet	_____
Driver	_____
Bank by Phone*	_____
Transitlink	_____

SENTENCE COMPLETION

This technique is an extension of the free-word association test. In this technique, the respondent is presented with some sentences containing incomplete stimuli and is asked to complete them. Like the free-word association method, interpreting and analysing data obtained from this technique is also difficult.

Automatic teller machine users are

Automatic teller machines may be convenient, but they

My major concern about automatic teller machines is

UNFINISHED SCENARIO COMPLETION

This technique is similar to the sentence completion test. However, in this technique, the respondent is presented with a specific scenario containing incomplete stimuli [*see example below*] and is asked to complete the scenario. Interpreting and analyzing data obtained from this technique is also difficult.

Since Mr. Albert Lee had received a large commission by check just before leaving home for a holiday trip, he wanted to deposit it in an automatic teller machine, because _____, but his friend Mr. Wong told him that he should _____, because _____.

CARTOON COMPLETION TEST

In the cartoon technique, the respondent is shown a comic-strip like cartoon with two characters in a conversation. While the speech of one character is shown in his/her balloon, the other balloon is empty. The respondent is asked to assume the role of the other person and fill the empty balloon with a speech.



Government publications

	Sub-type	Sources	Data	Uses
1.	Census data conducted every ten years throughout the country	Registrar General of India conducting census survey http://censusindia.gov.in/	Size of population and its distribution by age, sex, occupation and income levels. 2010 census is taking many more variables to get a better picture of the population	Population information is significant as forecasts of purchase, estimates of growth and development, as well as policy decisions can be made on this base.
2.	Statistical Abstract India – annually	CSO (Central Statistical Organization) for the past 5 years http://www.mospi.gov.in/cso_test1.htm	Education; health; residential information at the state level is part of this document	Making demand estimations and state level assessment of Government support and policy changes can be made
3.	White paper on national income	CSO http://www.mospi.gov.in/cso_test1.htm	Estimates of national income, savings and consumption.	Significant indication of the financial trends; investment forecasts and monetary policy formulation
4.	Annual Survey of Industries – all industries	CSO No. of units, persons employed, capital output ratio, turnover, etc. http://www.mospi.gov.in/cso_test1.htm		Information on existing units give perspective on the Industrial development and helps in creating the employee profile
5.	Monthly survey of selected industries	CSO http://www.mospi.gov.in/cso_test1.htm	Production statistics in detail	Demand –supply estimations.

Government publications (contd.)

	Sub-type	Sources	Data	Uses
6.	Foreign Trade of India Monthly Statistics	Director General of Commercial Intelligence http://www.dgciskol.nic.in/	Exports & Imports countrywise and productwise	Forecast manufacturing and trade estimations
7.	Wholesale price index weekly all India Consumer Price Index	Ministry of Commerce and Industry http://india.gov.in/sectors/commerce/ministry_commerce.php	Reporting of prices of Products like food articles, foodgrains, minerals, fuel, power, lights, lubricants, textiles, chemicals, metal, machinery & transport	Establishing price bands of product categories; pricing estimations for new products; determining consumer spend
8.	Economic Survey – annual publication.	Dept. of Economic Affairs, Ministry of Finance, patterns, currency and finance http://finmin.nic.in/the_ministry/dept_eco_affairs/	Descriptive reporting of the current economic status	Estimations of the future and evaluation of policy decisions and extraneous factors in that period
9.	National Sample Survey (NSS)	Ministry of Planning http://www.planningcommission.gov.in/	Social, economic, demographic, industrial and agricultural statistics.	Significant for making policy decisions as well as studying sociological patterns

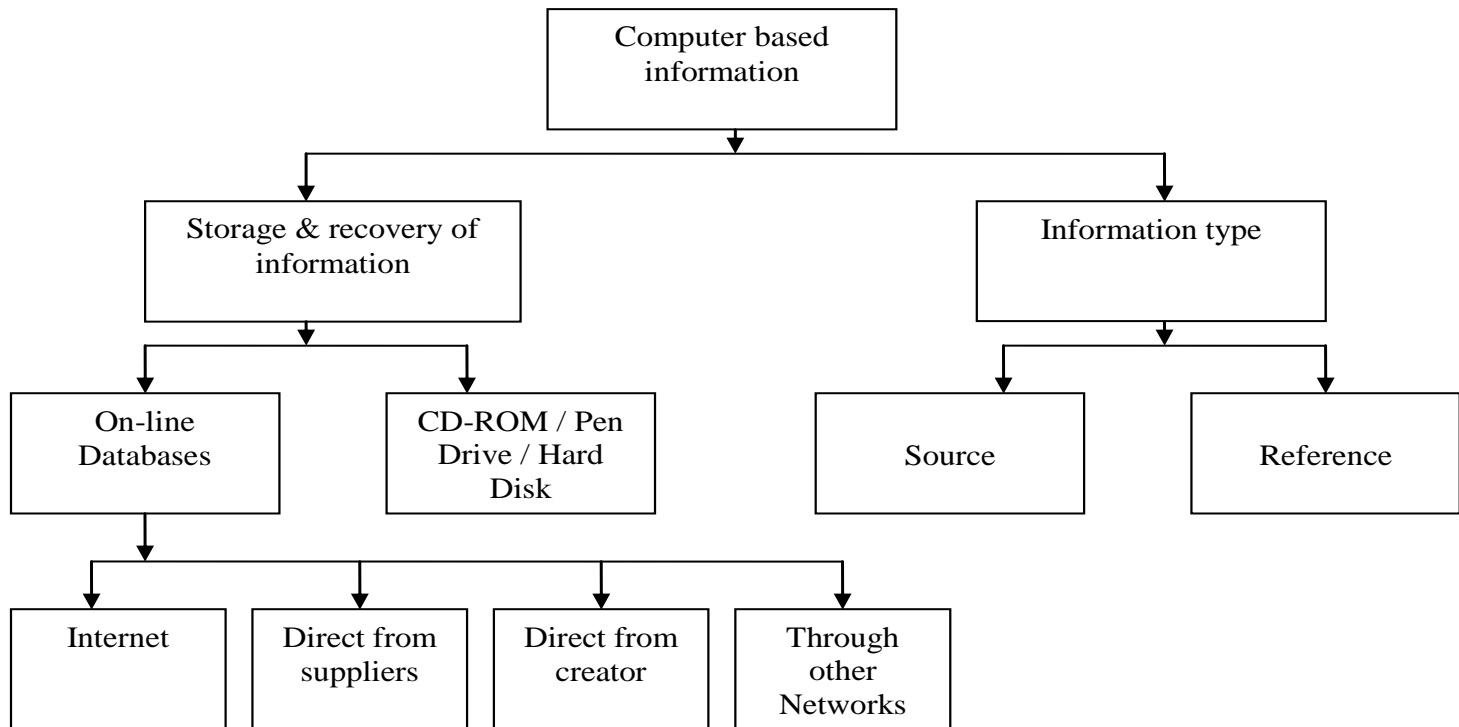
Non-government publications

	Sub-type	Sources	Data	Uses
1.	Company Working Results – Stock Exchange Directory	Bombay Stock Exchange http://www.bseindia.com/	A complete database of the companies registered with the Stock exchange and comprehensive details about stock policies and current share prices	Significant in determining the financial health various sectors as well as assessment of corporate funding and predictions of outcomes
2.	Status reports by various commodity boards	The Commodity Board or the Industry Associations like Jute Board, Cotton Industry, Sugar Association, Pulses Board, Metal Board, Chemicals, Spices, Fertilizers, Coir, Pesticides, Rubber, Handicrafts, Plantation Boards etc.	Detailed information on current assets-in terms of units current production figures and market condition	These are useful for individual sectors in working out their plans as well as evaluating causes of success or failure.
3.	Industry Associations on problems faced by private sector, etc.	FICCI, ASSOCHAM, AIMA, Association of Chartered Accountants & Financial Analysts, Indo-American Chamber of Commerce, etc. http://www.ficci.com/ http://www.assochem.org/ http://www.aima-ind.org/ www.iaccindia.com/	Cases/ comprehensive reports by the supplier or user or any other section associated with the sector	Cognizance of the gaps and problems in the effective functioning of the organization ; trouble shooting
4.	Export related data – commodity wise.	Leather Exports Promotion Council, Apparel Export Promotion Council, Handicrafts, Spices Tea, etc., Exim Bank etc. http://www.leatherindia.org/ http://www.aepcindia.com/	Product and country wise data on the export figures as well as information on existing policies related to the sector.	To estimate the demand; gauge opportunities for trade and impetus required in terms of manufacturing and policy changes

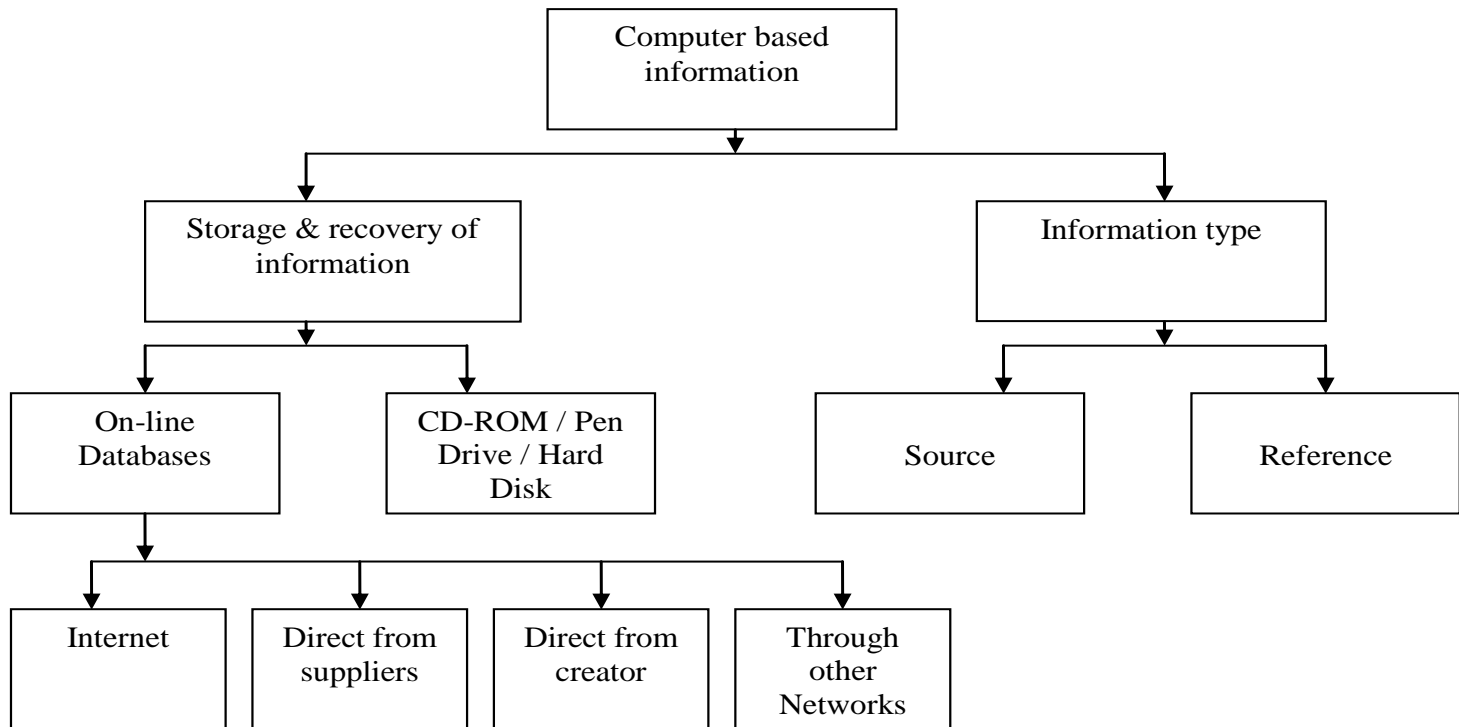
Non-government publications (contd)

	Sub-type	Sources	Data	Uses
5.	Retail Store Audit on pharmaceutical, veterinary, consumer products.	ORG (Operations Research Group); Monthly reports on urban sector. Quarterly reports on rural sector	The touch point for this data is the retailer, who provides the figures related to product sales; the data is very comprehensive and covers most brands .the data is region specific and covers both inventory and goods sold.	Market analysis and market structure mapping with estimations of market share of leading brands. The audit can also be used to study consumption trends at different time periods or subsequent to sales promotion or other activities.
6.	National Readership Surveys (NRS)	IMRB- survey of reading behavior for different segments as well as different products. http://www.imrbint.com/	Today these surveys are done by various bodies with different sample bases. Today the survey base has become younger, with the age of the reader is lowered to 12+.	Media planning and measuring exposure as well as reach for product categories.
7.	THOMPSON INDICES: Urban market index, Rural market index	Hindustan Thompson Associates	All towns with population of more than one lakh are covered and information of demographic and socio-economic variables are given for each city with Bombay as base. The rural index similarly covers about 400 districts with socio-economic indicators like value of agriculture output, etc.	The inclinations to purchase consumer products are directly related to socio-economic development of communities in general. The indices provide barometers to measure such potentials for each city and has implications for the researcher in terms of data collection sources.

Computerized Databases



Computerized Databases



Syndicate Data Sources

Household/ individual data—conducted on individual consumers. They are usually of the following types

- ***Surveys***: are usually one-time assessments conducted on a large representative respondent base.
- ***Product purchase panels***: These specially selected respondents groups periodically record certain identified purchases, generally related to household products and groceries.
- ***Media-specific panels***: media panels are created for collecting information related to promotion and advertising. They generally make use of different kinds of electronic equipments to automatically record consumer viewing behaviour.

Syndicate Data Sources

Institutional syndicated data—the second group of syndicated sources collect information from organizations and institutions.

- ▶ *Retailer audits*: for various product/service categories periodically recorded data is available to track the movement of stocks at the retail end.
- ▶ *Wholesaler audits*: these measure warehouse movement. Participating operators include wholesalers, super and hyper markets and frozen-food warehouses.

Meaning of Measurement and Scaling

- **Measurement:** The term ‘measurement’ means assigning numbers or some other symbols to the characteristics of certain objects. When numbers are used, the researcher must have a rule for assigning a number to an observation in a way that provides an accurate description.
- **Scaling:** Scaling is an extension of measurement. Scaling involves creating a continuum on which measurements on objects are located.

Types of Measurement Scale

Nominal scale: This is the lowest level of measurement. Here, numbers are assigned for the purpose of identification of the objects. Any object which is assigned a higher number is in no way superior to the one which is assigned a lower number.

Example:

- Are you married?

(a) Yes

(b) No

- Married person may be assigned a no. 1.
- Unmarried person may be assigned a no. 2.

The assigned numbers cannot be added, subtracted, multiplied or divided. The only arithmetic operations that can be carried out are the count of each category. Therefore, a frequency distribution table can be prepared for the nominal scale variables and mode of the distribution can be worked out.

Types of Measurement Scale

Ordinal scale: This is the next higher level of measurement. One of the limitations of the nominal scale measurements is that we cannot say whether the assigned number to an object is higher or lower than the one assigned to another option. The ordinal scale measurement takes care of this limitation. An ordinal scale measurement tells whether an object has more or less of characteristics than some other objects.

Types of Measurement Scale

Example:

Rank the following attributes while choosing a restaurant for dinner. The most important attribute may be ranked one, the next important may be assigned a rank of 2 and so on.

Attribute	Rank
Food quality	
Prices	
Menu variety	
Ambience	
Service	

In the ordinal scale, the assigned ranks cannot be added, multiplied, subtracted or divided. One can compute median, percentiles and quartiles of the distribution. The other major statistical analysis which can be carried out is the rank order correlation coefficient, sign test.

Types of Measurement Scale

Interval scale: The interval scale measurement is the next higher level of measurement.

- It takes care of the limitation of the ordinal scale measurement where the difference between the score on the ordinal scale does not have any meaningful interpretation.
- In the interval scale the difference of the score on the scale has meaningful interpretation.
- It is assumed that the respondent is able to answer the questions on a continuum scale.
- The mathematical form of the data on the interval scale may be written as

$$Y = a + bX \quad \text{where } a \neq 0$$

- Ratio of the score on this scale does not have a meaningful interpretation.

Types of Measurement Scale

Example:

- The counter-clerks at ICICI Bank, (Vasant Kunj Branch) are very friendly.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

The numbers on this scale can be added, subtracted, multiplied or divided. One can compute arithmetic mean, standard deviation, correlation coefficient and conduct a t-test, Z-test, regression analysis and factor analysis.

Types of Measurement Scale

Ratio scale: This is the highest level of measurement and takes care of the limitations of the interval scale measurement, where the ratio of the measurements on the scale does not have a meaningful interpretation.

- The mathematical form of the ratio scale data is given by $Y = b X$.
- In ratio scale, there is a natural zero (origin).

Example:

How many chemist shops are there in your locality?

How many students are there in the MBA programme at IIFT?

- All mathematical and statistical operations can be carried out using the ratio scale data.

Definition of Attitude

- An attitude is viewed as an enduring disposition to respond consistently in a given manner to various aspects of the world, including persons, events and objects.

Components of Attitude:

- Cognitive component
- Affective component
- Intention or action component

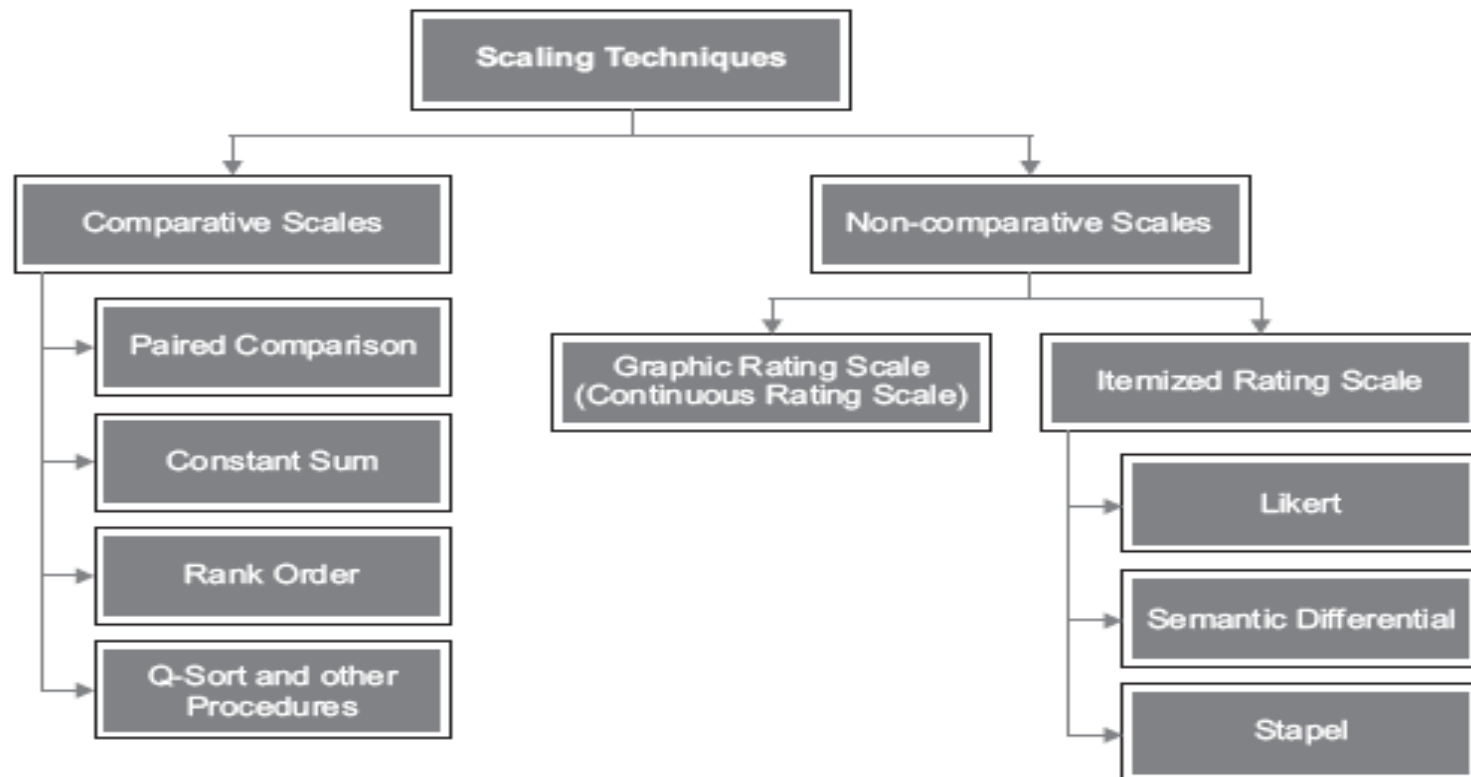
Classification of Scales

Single item vs multiple item scale:

- ▶ In the single item scale, there is only one item to measure a given construct.
- ▶ In multiple item scale, there are many items that play a role in forming the underlying construct that the researcher is trying to measure. This is because each of the item forms some part of the construct which the researcher is trying to measure.

Classification of Scales

Comparative vs non-comparative scales



Classification of Scales

- **Comparative scales** - In comparative scales it is assumed that respondents make use of a standard frame of reference before answering the question.

Example:-

- Please rate Domino's in comparison to Pizza Hut on the basis of your satisfaction level on an 11-point scale, based on the following parameters: (1 = Extremely poor, 6 = Average, 11 = Extremely good). Circle your response:

a.	Variety of menu options	1	2	3	4	5	6	7	8	9	10	11
b.	Value for money	1	2	3	4	5	6	7	8	9	10	11
c.	Speed of service (delivery time)	1	2	3	4	5	6	7	8	9	10	11
d.	Promotional offers	1	2	3	4	5	6	7	8	9	10	11
e.	Food quality	1	2	3	4	5	6	7	8	9	10	11
f.	Brand name	1	2	3	4	5	6	7	8	9	10	11
g.	Quality of service	1	2	3	4	5	6	7	8	9	10	11
h.	Convenience in terms of takeaway location	1	2	3	4	5	6	7	8	9	10	11
i.	Friendliness of the salesperson on the phone	1	2	3	4	5	6	7	8	9	10	11
j.	Quality of packaging	1	2	3	4	5	6	7	8	9	10	11
k.	Adaptation of Indian taste	1	2	3	4	5	6	7	8	9	10	11
l.	Side orders/appetizers	1	2	3	4	5	6	7	8	9	10	11

Classification of Scales

Formats of Comparative Scales –

- Paired comparison scales
- Rank order scale
- Constant sum rating scale
- Q-sort technique
- **Non-Comparative Scales** – In the non-comparative scales, the respondents do not make use of any frame of reference before answering the questions.

More on Analysis of Data

Calculating summarized rank order

The rankings of attributes while choosing a restaurant for dinner for 32 respondents can be presented in the form of frequency distribution in the table below.

Attribute	Rank				
	1	2	3	4	5
Ambience	4	5	13	5	5
Food Quality	16	13	2	1	0
Menu Variety	7	2	2	9	12
Service	3	8	11	6	4
Location	2	4	4	11	11
Total	32	32	32	32	32

More on Analysis of Data

To calculate a summary rank ordering, the attribute with the first rank is given the lowest number (1) and the least preferred attribute is given the highest number (5). The summarized rank order is obtained with the following computations as:

Ambience	:	$(4 \times 1) + (5 \times 2) + (13 \times 3) + (5 \times 4) + (5 \times 5)$	=	98
Food Quality	:	$(16 \times 1) + (13 \times 2) + (2 \times 3) + (1 \times 4) + (0 \times 5)$	=	52
Menu Variety	:	$(7 \times 1) + (2 \times 2) + (2 \times 3) + (9 \times 4) + (12 \times 5)$	=	113
Service	:	$(3 \times 1) + (8 \times 2) + (11 \times 3) + (6 \times 4) + (4 \times 5)$	=	96
Location	:	$(2 \times 1) + (4 \times 2) + (4 \times 3) + (11 \times 4) + (11 \times 5)$	=	121

The total lowest score indicates the first preference ranking. The results show the following rank ordering:

- (1) Food quality
- (2) Service
- (3) Ambience
- (4) Menu variety
- (5) Location

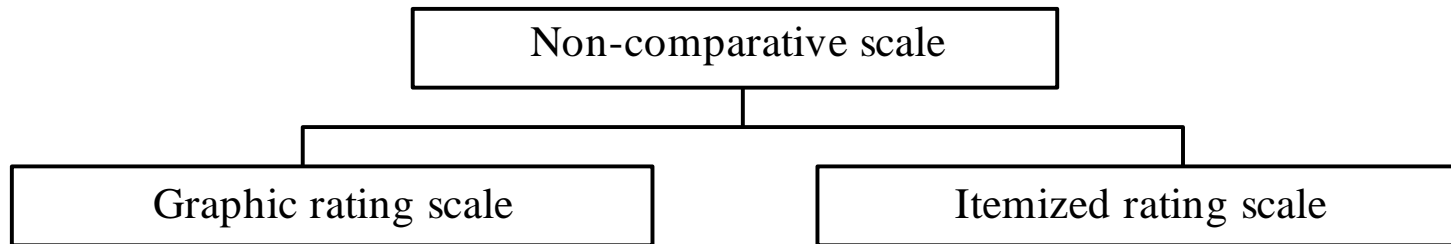
More on Analysis of Data

- ▶ The researcher could create new variables by re-specifying the data with numeric or logical transformation. Suppose a multiple-item Likert scale designed to measure the perception of a customer towards the bank has 10 items. The total score of a respondent can be computed as:

Total score of i^{th} respondent = Score of i^{th} respondent on item 1 + Score of i^{th} respondent on item 2 + ... + Score of i^{th} respondent on item 10.

Once the total score for each of the respondent is computed, the average score can be obtained by dividing it by the number of items. It can be further categorized as favourable, neutral and unfavourable perception that could be related to various demographic variables depending upon the objectives of research.

Classification of Scales



Graphic Rating Scale – This is a continuous scale and the respondent is asked to tick his preference on a graph.

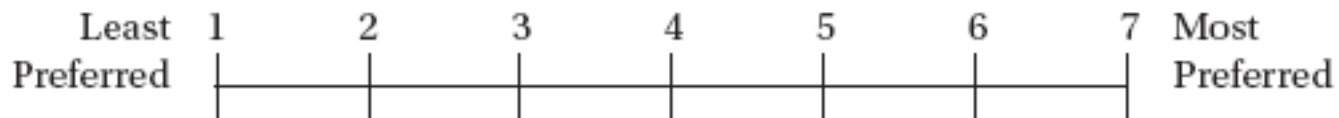
Examples:

- Please put a tick mark (✓) on the following line to indicate your preference for fast food.



Classification of Scales

Please put a tick mark (·) on the following line to indicate your preference for fast food.



Alternative Presentation of Graphic Rating Scale –

Please indicate how much do you like fast food by pointing to the face that best shows your attitude and taste. If you do not prefer it at all, you would point to face one. In case you prefer it the most, you would point to face seven.



Classification of Scales

Itemized rating scale – In the itemized rating scale, the respondents are provided with a scale that has a number of brief descriptions associated with each of the response categories. There are certain issues that should be kept in mind while designing the itemized rating scale.

- Number of categories to be used
- Odd or even number of categories
- Balanced versus unbalanced scales
- Nature and degree of verbal description
- Forced versus non–forced scales
- Physical form

Classification of Scales

Examples of Itemized Rating Scales:

Likert scale

- The respondents are given a certain number of items (statements) on which they are asked to express their degree of agreement/disagreement.
- This is also called a summated scale because the scores on individual items can be added together to produce a total score for the respondent.
- An assumption of the Likert scale is that each of the items (statements) measures some aspect of a single common factor, otherwise the scores on the items cannot legitimately be summed up.
- In a typical research study, there are generally 25 to 30 items on a Likert scale.

Classification of Scales

Example of a Likert Scale:

No.	Statement	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1.	The company makes quality products			✓		
2.	It is a leader in technology					✓
3.	It doesn't care about the general public.		✓			
4.	The company leads in R&D to improve products				✓	
5.	The company is not a good paymaster.	✓				
6.	The products of the company go through stringent quality tests.				✓	
7.	The company has not done anything to curb pollution.		✓			
8.	It does not care about the community near its plant.	✓				
9.	The company's stocks are good to buy or own.				✓	
10.	The company does not have good labour relations.		✓			

Classification of Scales

Semantic Differential Scale

- This scale is widely used to compare the images of competing brands, companies or services.
- Here the respondent is required to rate each attitude or object on a number of five–or seven–point rating scales.
- This scale is bounded at each end by bipolar adjectives or phrases.
- The difference between Likert and Semantic differential scale is that in Likert scale, a number of statements (items) are presented to the respondents to express their degree of agreement/disagreement. However, in the semantic differential scale, bipolar adjectives or phrases are used.

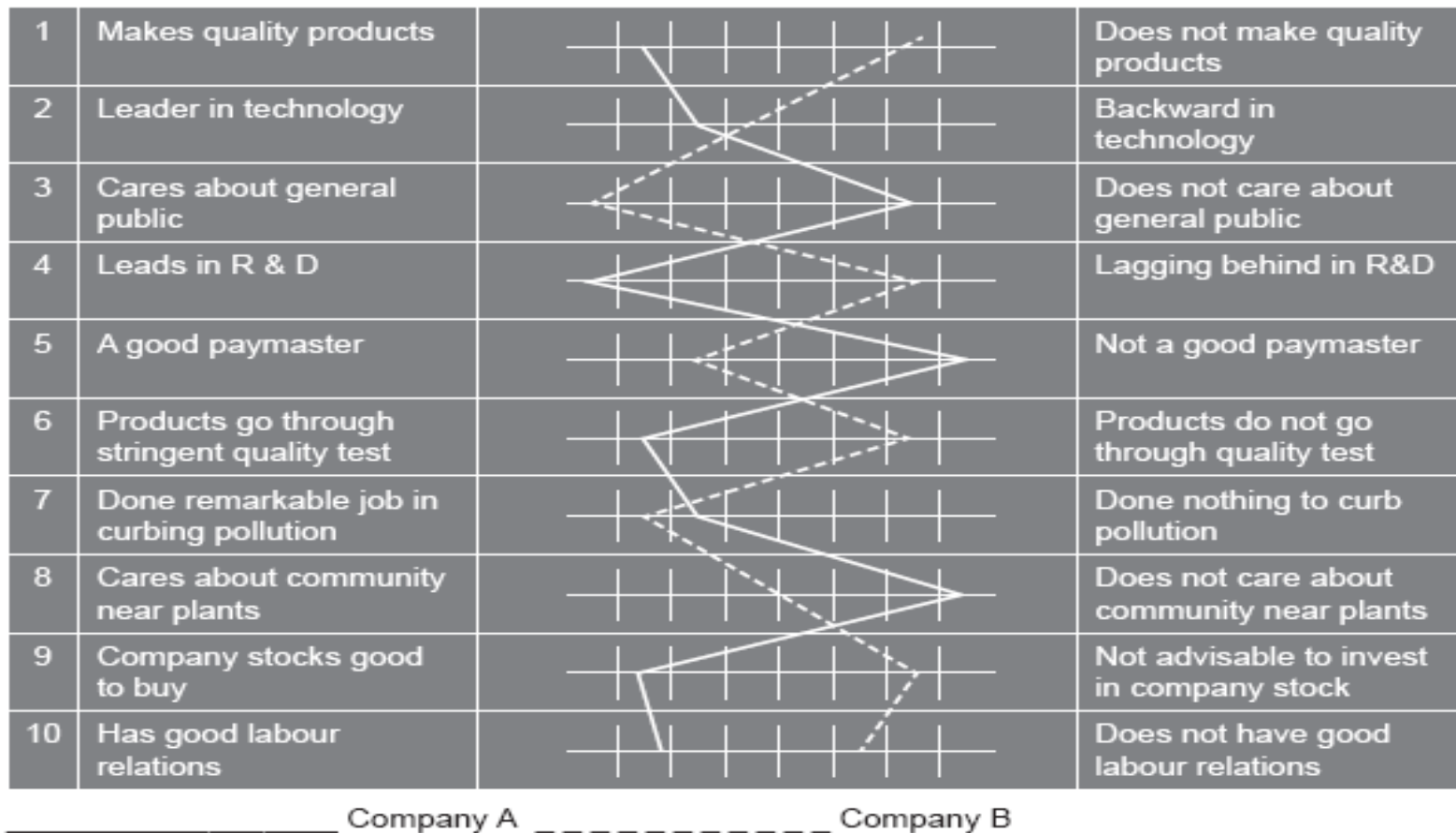
Classification of Scales

Example of Semantic Differential Scale:

1	Makes quality products	□ □ □ □ □ □ □	Does not make quality products
2	Leader in technology	□ □ □ □ □ □ □	Backward in technology
3	Does not care about general public	□ □ □ □ □ □ □	Cares about general public
4	Leads in R & D	□ □ □ □ □ □ □	Lagging behind in R&D
5	Not a good paymaster	□ □ □ □ □ □ □	A good paymaster
6	Products go through stringent quality test	□ □ □ □ □ □ □	Products don't go through quality test
7	Does nothing to curb pollution	□ □ □ □ □ □ □	Does a remarkable job in curbing pollution
8	Does not care about community near plants	□ □ □ □ □ □ □	Cares about community near plants
9	Company stocks good to buy	□ □ □ □ □ □ □	Not advisable to invest in company stock
10	Does not have good labour relations	□ □ □ □ □ □ □	Has good labour relations

Classification of Scales

Example of Semantic Differential Scale: (Pictorial Profile)



Classification of Scales

- ▶ Stapel Scale

RESTAURANT	
+5	+5
+4	+4
+3	+3
+2*	+2
+1	+1
<i>Quality of Food</i>	<i>Quality of Service</i>
-1	-1
-2	-2
-3	-3
-4	-4
-5	-5*

Measurement Error

This occurs when the observed measurement on a construct or concept deviates from its true values.

Reasons

- Mood, fatigue and health of the respondent
- Variations in the environment in which measurements are taken
- A respondent may not understand the question being asked and the interviewer may have to rephrase the same. While rephrasing the question the interviewer's bias may get into the responses.
- Some of the questions in the questionnaire may be ambiguous errors may be committed at the time of coding, entering of data from questionnaire to the spreadsheet

Criteria for Good Measurement

Reliability

Reliability is concerned with consistency, accuracy and predictability of the scale.

Methods to measures Reliability

- Test-retest reliability
- Split-half reliability
- Cronbach's Alpha

Criteria for good measurement

Validity

The validity of a scale refers to the question whether we are measuring what we want to measure.

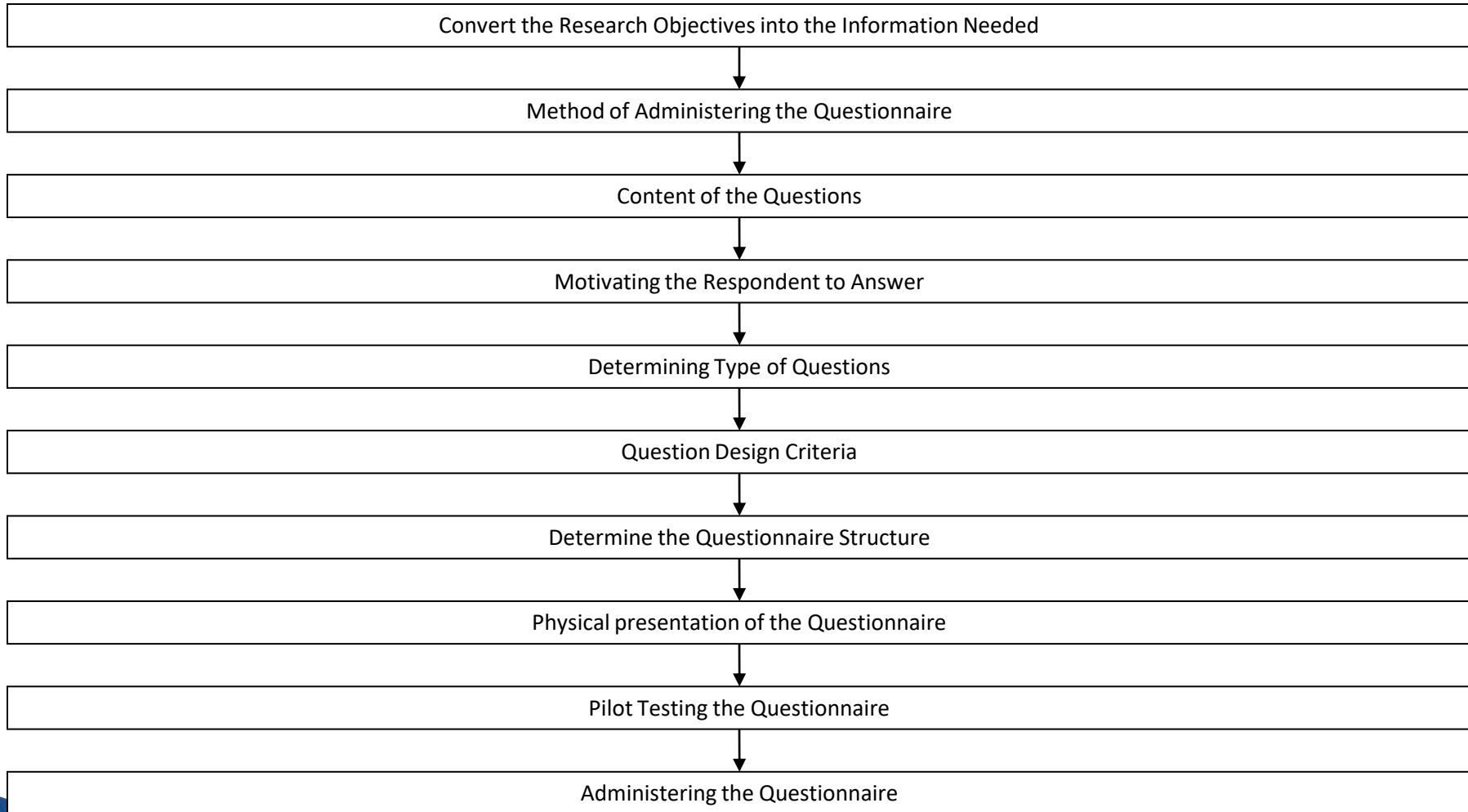
Different ways to measure Validity

- Content validity
- Concurrent validity
- Predictive validity

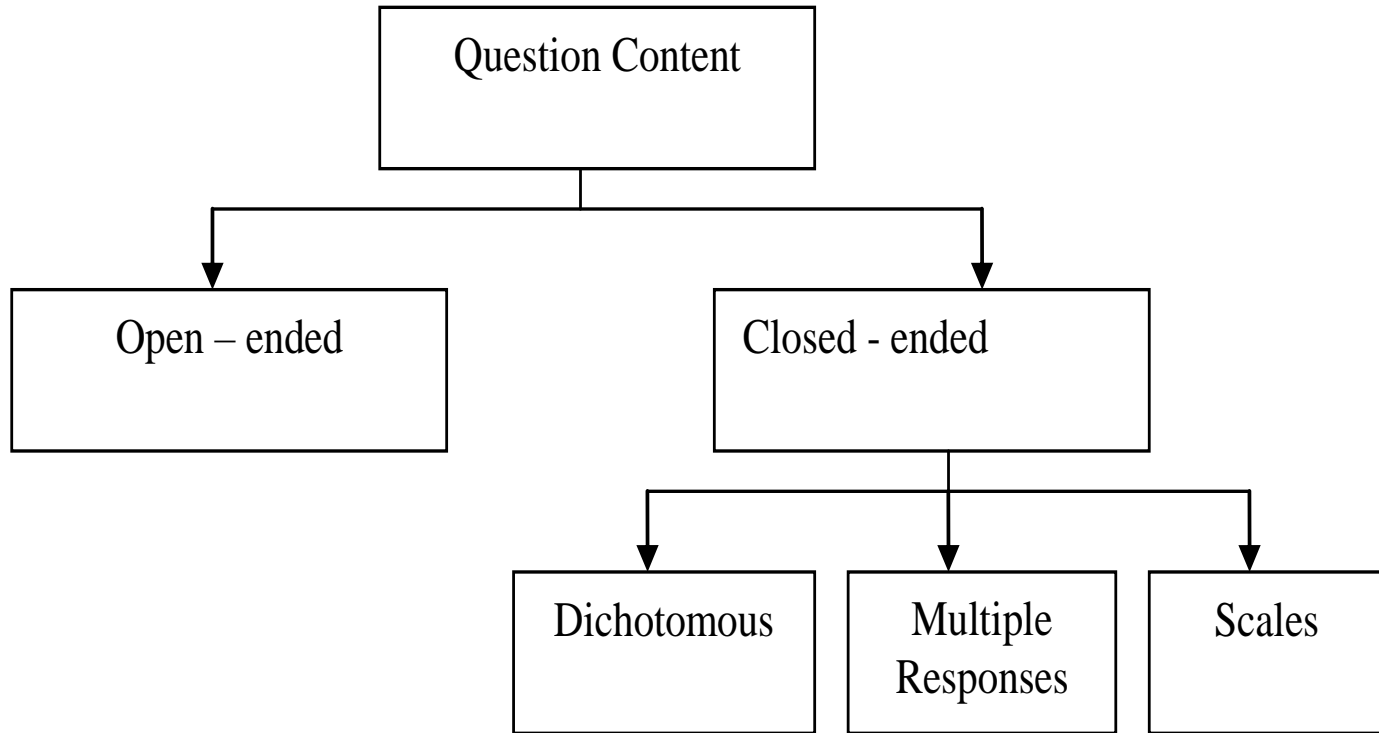
Sensitivity

Sensitivity refers to an instrument's ability to accurately measure the variability in a concept.

The Questionnaire Design Process



Type of Questions



Questions: Common Pitfalls

▶ Incomprehensible

- When you consider a new stadium, is it possible that part of your determining factors might rest on the fact that you sometimes choose to forgo entertainment in favor of more pedestrian activities like walking your dog?

▶ Unanswerable

- Will building a new stadium in Boise cost too much?

▶ Leading

- Wouldn't you agree that it is a great idea to build a new multiuse facility in Boise?

▶ Double-barreled

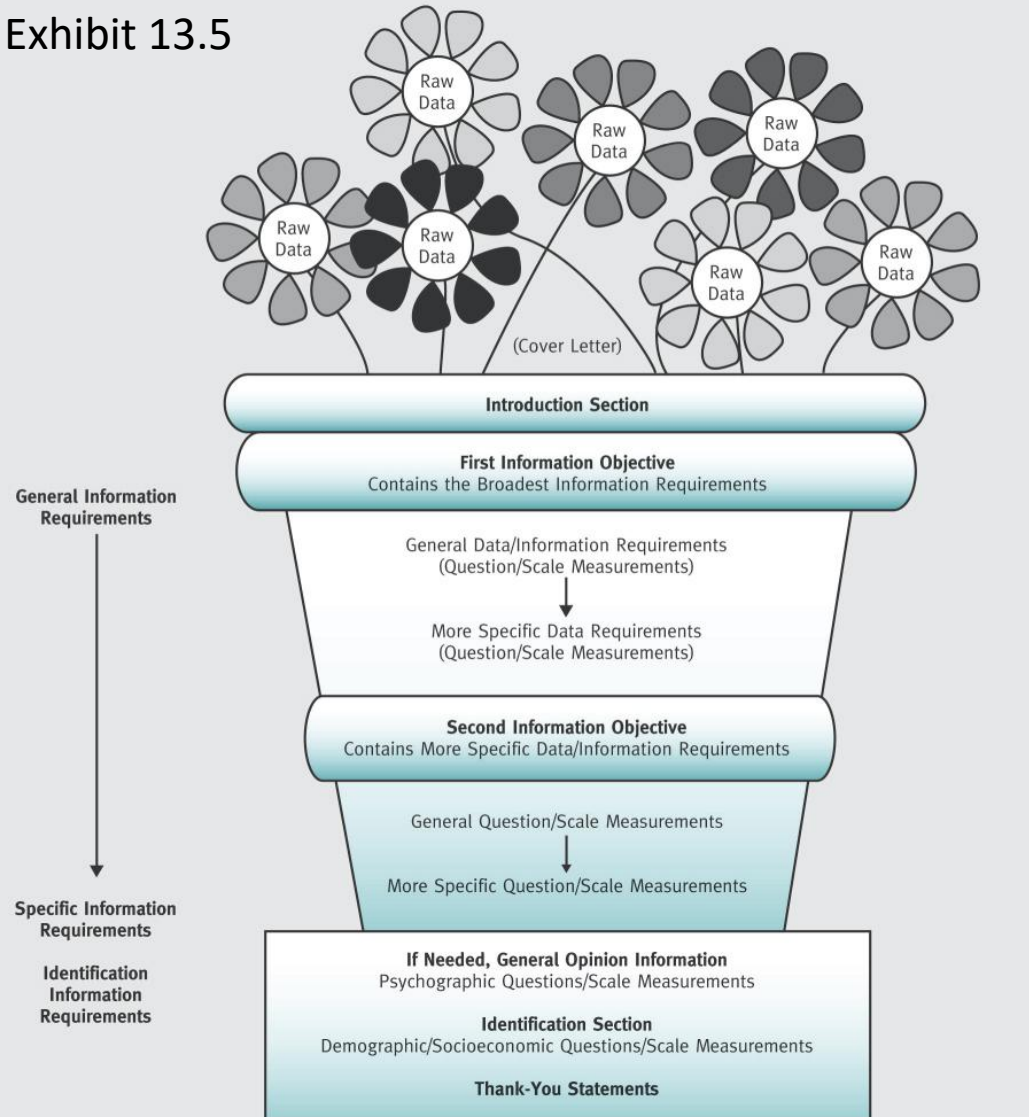
- Do you think the old stadium needs to be replaced and a new stadium should be built downtown?

Purpose of Questions

- ▶ **Descriptive**
 - Describe characteristics of your sample
- ▶ **Predictive (Hypothesis Testing)**
 - What demographic factors are associated with support for the stadium?
- ▶ **Every question should be designed to provide useful information**
 - What is our primary goal in the stadium study?
 - What questions are we trying to address?
- ▶ **How do our questions stack up?**
 - Good, bad questions?
Do they each answer a question? Descriptive?
Predictive?

Arrangement of Questions: Flowerpot Approach

Exhibit 13.5

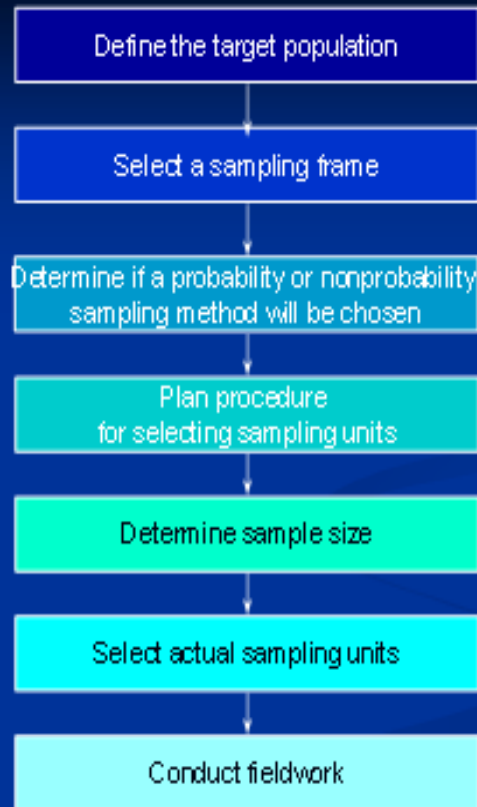


Sampling Concepts

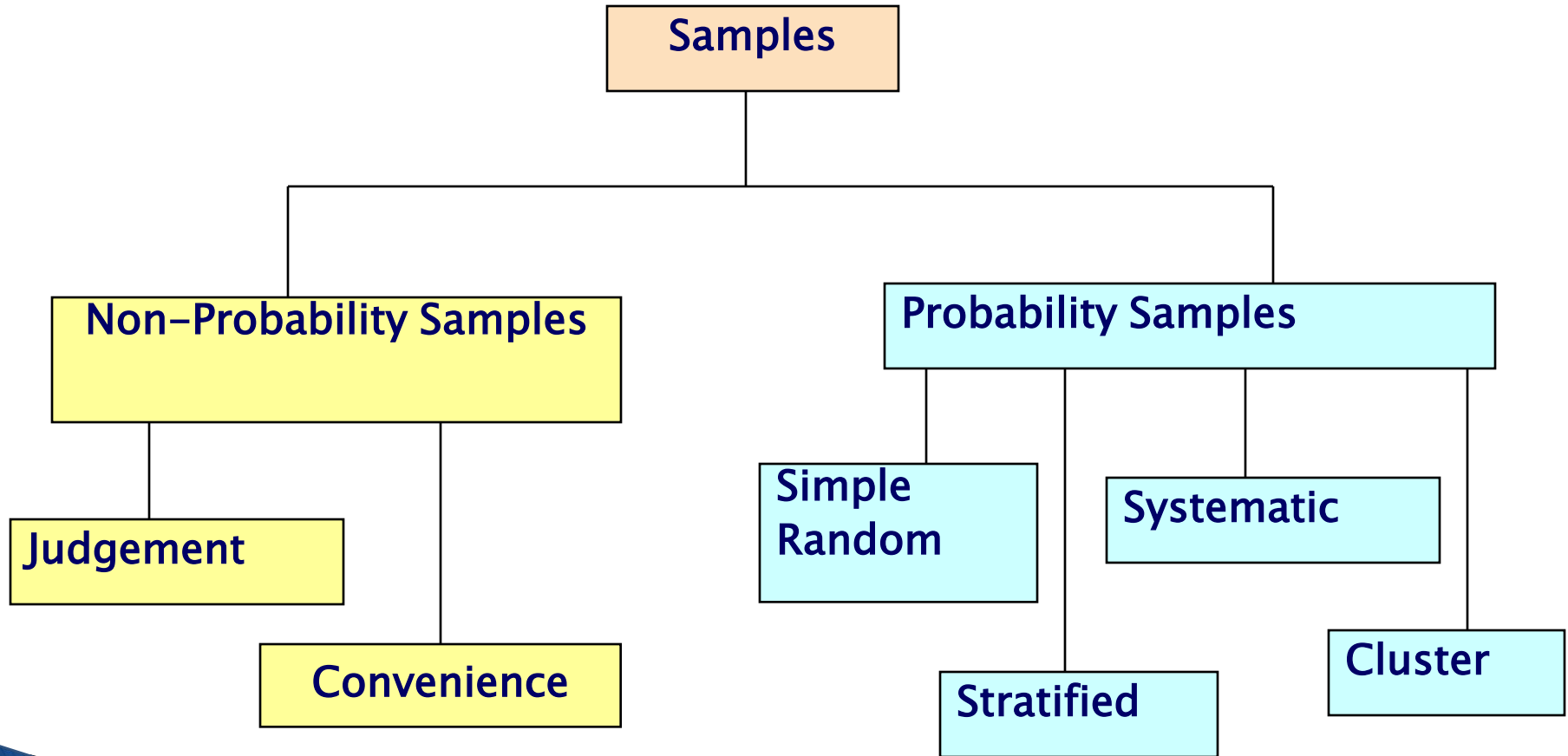
- ▶ **Population:** Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways.
- ▶ **Element:** An element comprises a single member of the population.
- ▶ **Sampling frame:** Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling.
- ▶ **Sample:** It is a subset of the population. It comprises only some elements of the population.
- ▶ **Sampling unit:** A sampling unit is a single member of the sample.
- ▶ **Sampling:** It is a process of selecting an adequate number of elements from the population so that the study of the sample will not only help in understanding the characteristics of the population but will also enable us to generalize the results.
- ▶ **Census (or complete enumeration):** An examination of each and every element of the population is called census or complete enumeration.

Stages in Sample Selection

Stages in the Selection of a Sample



Sampling Techniques



Sampling Design

Probability Sampling Design – Probability sampling designs are used in conclusive research. In a probability sampling design, each and every element of the population has a known chance of being selected in the sample.

Types of Probability Sampling Design

- Simple random sampling with replacement
- Simple random sampling without replacement
- Systematic sampling
- Stratified random sampling
- Cluster sampling
- Two Stage Sampling
- Multi-stage Sampling
- Probability Proportional to size sampling
- Area Sampling

Sampling Design

Non-probability Sampling Designs – In case of non-probability sampling design, the elements of the population do not have any known chance of being selected in the sample.

Types of Non-Probability Sampling Design

- Convenience sampling
- Judgemental sampling
- Snowball sampling
- Quota sampling

Determination of Sample Size

Sample size for estimating population mean
– The formula for determining sample size is given as:

$$n = \frac{Z^2 \sigma^2}{e^2}$$

Where

n = Sample size

σ = Population standard deviation

e = Margin of error

Z = The value for the given confidence interval

Determination of Sample Size

Sample size for estimating population proportion –

1. When population proportion p is known

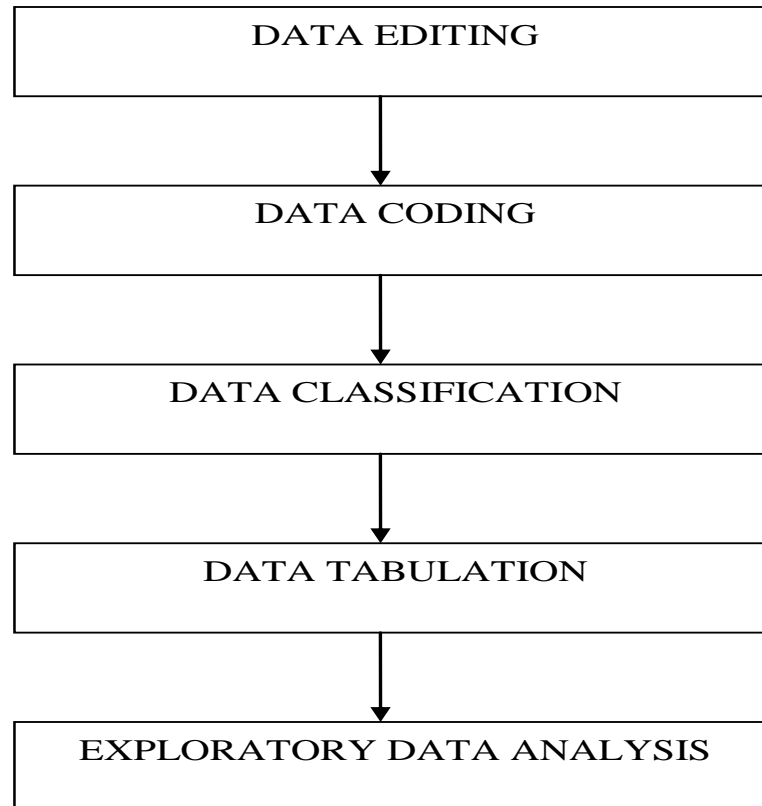
$$n = \frac{Z^2 pq}{e^2}$$

2. When population proportion p is not known

$$n = \frac{1}{4} \frac{Z^2}{e^2}$$

DATA PROCESSING

The Data Preparation Process



Data Editing

Field editing: usually done by the field investigators at the end of every field day the investigator(s) who must review the filled forms for any inconsistencies, non-response, illegible responses or incomplete questionnaires.

Centralized in-house editing: usually done at the researcher's end.

- Backtracking
- Allocating missing values
- Plug value
- Discarding unsatisfactory responses

Data Coding

The process of identifying and denoting a numeral to the responses given by the respondent is called coding

- ▶ Field
- ▶ Record
- ▶ File
- ▶ Data matrix

Pre-Coding closed-ended questions

▶ *Dichotomous questions:*

Do you eat ready-to-eat food? Yes=1; no=0 (X-1)

▶ *Ranking questions*

Q.NO.	Variable name	Coding instructions	Variable name
1.	Balika Badhu	Number from 1-10	X 10a
2.	Sathiya	Number from 1-10	X 10b
3.	Sasural Genda Phool	Number from 1-10	X 10c
4.	Bidai	Number from 1-10	X 10d
5.	Pathshala	Number from 1-10	X 10e
6.	Bandini	Number from 1-10	X 10f
7.	Laptaganj	Number from 1-10	X 10g
8.	Sajan Ghar Jaaana Hai	Number from 1-10	X 10h
9.	Tere Liye	Number from 1-10	X 10i
10.	Uttaran	Number from 1-10	X 10j

Pre-Coding

Closed-ended questions

▶ Checklists/multiple responses

How many columns will you make for the following question?

Which of the following newspapers do you read (tick all that you read)

Times of India	-----
Hindustan Times	-----
Mail Today	-----
Indian Express	-----
Deccan Chronicle	-----
Asian Age	-----
Mint	-----

Pre-Coding

Closed-ended questions

Scaled questions

Col.no.	Variable name	Coding instructions	Variable name
1.	Individual shops more	A number from 1 to 5 SA = 5, A = 4, N = 3, D = 2, SD = 1	X 1a
2.	Well informed	- do -	X 1b
3.	Knows what to buy	- do -	X 1c
4.	More spending money	- do -	X 1d
5.	More shopping options	- do -	X 1e

Data coding

Sample record: Excel sheet for two-wheeler owners

Unit Column 1	occupation Column 2	Vehicle Column 3	Km/day Column 4	Marital status Column 5	Family size Column 6
1	4	1	20	1	3
2	3	2	25	2	1
3	5	1	25	1	4
4	2	1	15	2	2
5	4	2	20	2	4
6	5	2	35	2	6
7	1	1	40	1	3
8	5	2	20	2	4

Data Coding

Sample record: Excel sheet for two-wheeler owners

Unit Column 1	occupation Column 2	Vehicle Column 3	Km/day Column 4	Marital status Column 5	Family size Column 6
1	4	1	20	1	3
2	3	2	25	2	1
3	5	1	25	1	4
4	2	1	15	2	2
5	4	2	20	2	4
6	5	2	35	2	6
7	1	1	40	1	3
8	5	2	20	2	4

Code Book Formulation

- ▶ Appropriate to the research objective
- ▶ Comprehensive
- ▶ Mutually exclusive
- ▶ Single variable entry

Sample Code Book Extract

Question No.	Variable Name	Coding Instruction	Symbol used for variable name
1.	Buy ready to eat food products	Yes = 1 No = 0	X1
2.	Use ready to eat food products	Yes = 1 No = 0	X2
22.	Age	Less than 20 yrs = 1, 21 to 26 years = 2, 27 to 35 years = 3, 36 to 45 years = 4, More than 45 years = 5	X22
23.	Gender	Male = 1 Female = 2	X23
24.	Marital status	Single = 1 Married = 2 Divorced/widow = 3	X24
25.	No. of children	Exact no. to be written	X25
26.	Family size	One to two = 1, Three to five = 2, Six & more = 3	X26
27.	Monthly household income	Rs.20000 to Rs.34999 = 1, Rs.35000 to Rs.50000 = 2, Rs.50001 to Rs.74999 = 3 Rs.75000 & above = 4	X27
28.	Education	Less than graduation = 1 Graduation = 2 Post graduation & above = 3	X28
29.	Occupation	Student = 1 Businessman = 2 Professional = 3 Service = 4 Housewife = 5 Others = 6	X29

Post-Coding

Open-ended Questions

If you think Lean was a success so far, please specify three most significant reasons that have contributed to its success in your opinion?

Col.no.	Variable name	Coding instructions	Variable name
63.	Improvement at work place by eliminating waste.	Yes = 1 No = 0	X 63a
64	To meet increasing demands of customers	Yes = 1 No = 0	X 63b
65	To improve quality	Yes = 1 No = 0	X 63c
66	To achieve corporate goal	Yes = 1 No = 0	X 63d
67	It reduces cycle time of the manufacturing & production.	Yes = 1 No = 0	X 63e
68	Reduced response time	Yes = 1 No = 0	X 63f
69	Enhanced innovation and creativity	Yes = 1 No = 0	X 63g

Types of Statistical Analysis

Five types of statistical analysis

Descriptive

What are the characteristics of the respondents?

Inferential

What are the characteristics of the population?

Differences

Are two or more groups the same or different?

Associative

Are two or more variables related in a systematic way?

Predictive

Can we predict one variable if we know one or more other variables?

Descriptive vs. Inferential Analysis

Descriptive analysis – Descriptive analysis deals with summary measures relating to the sample data. The common ways of summarizing data are by calculating average, range, standard deviation, frequency and percentage distribution. The first thing to do when data analysis is taken up is to describe the sample.

Examples of Descriptive Analysis:

- ▶ What is the average income of the sample?
- ▶ What is the average age of the sample?
- ▶ What is the standard deviation of ages in the sample?
- ▶ What is the standard deviation of incomes in the sample?
- ▶ What percentage of sample respondents are married?
- ▶ What is the median age of the sample respondents?
- ▶ Is there any association between the frequency of purchase of product and income level of the consumers?

Descriptive vs. Inferential Analysis

- ▶ Is the level of job satisfaction related with the age of the employees?
- ▶ Which TV channel is viewed by the majority of viewers in the age group 20–30 years?
- ▶ **Types of Descriptive Analysis** – The table below presents the type of descriptive analysis that is applicable under each form of measurement.

Type of Measurement	Type of Descriptive Analysis
Nominal	Frequency table, Proportion percentages, Mode
Ordinal	Median, Quartiles, Percentiles, Rank order correlation
Interval	Arithmetic mean, Correlation coefficient
Ratio	Index numbers, Geometric mean, Harmonic mean

Descriptive vs. Inferential Analysis

Inferential Analysis – Under inferential statistics, inferences are drawn on population parameters based on sample results. The researcher tries to generalize the results to the population based on sample results.

Examples of Inferential Analysis:

- ▶ Is the average age of the population significantly different from 35?
- ▶ Is the average income of population significantly greater than 25,000 per month?
- ▶ Is the job satisfaction of unskilled workers significantly related with their pay packet?
- ▶ Do the users and non-users of a brand vary significantly with respect to age?
- ▶ Is the growth in the sales of the company statistically significant?

Descriptive vs. Inferential Analysis

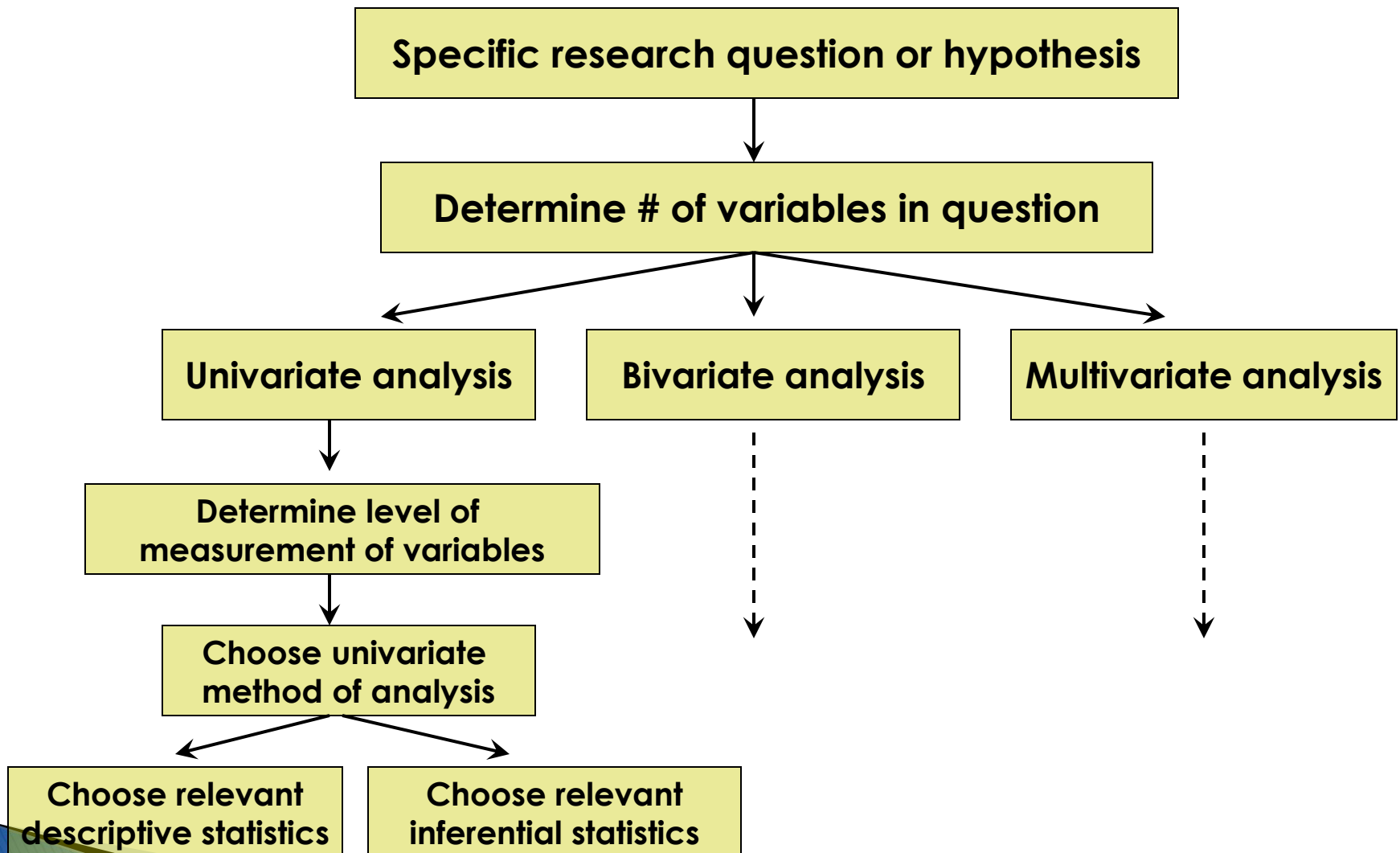
- ▶ Does the advertisement expenditure influence sale significantly?
- ▶ Are consumption expenditure and disposable income of households significantly correlated?
- ▶ Is the proportion of satisfied workers significantly more for skilled workers than for unskilled works?
- ▶ Do urban and rural households differ significantly in terms of average monthly expenditure on food?
- ▶ Is the variability in the starting salaries of fresh MBA different with respect to marketing and finance specialization?

Meaning of Univariate, Bivariate & Multivariate Analysis of Data

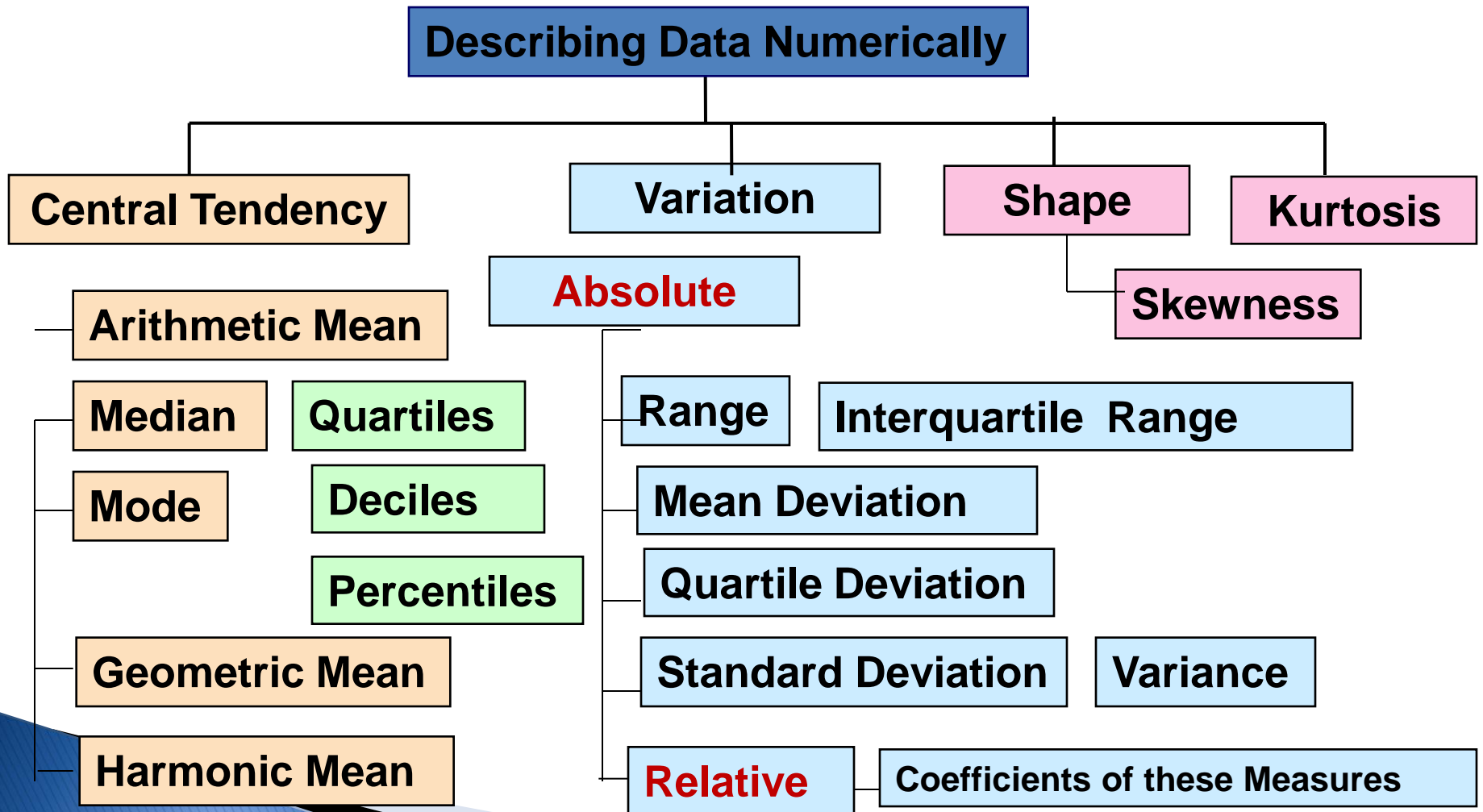
- ▶ **Univariate Analysis** – In univariate analysis, one variable is analyzed at a time.
- ▶ **Bivariate Analysis** – In bivariate analysis two variables are analyzed together and examined for any possible association between them.
- ▶ **Multivariate Analysis** – In multivariate analysis, the concern is to analyze more than two variables at a time.

The type of statistical techniques used for analyzing univariate and bivariate data depends upon the level of measurements of the questions pertaining to those variables. Further, the data analysis could be of two types, namely, descriptive and inferential.

Choosing the Statistical Technique*



Univariate Analysis: Summary Measures



Descriptive Analysis of Univariate Data

Measures of Central Tendency

- ▶ Arithmetic mean (appropriate for Interval and Ratio scale data)
- ▶ Geometric Mean, Harmonic Mean Weighted Mean
- ▶ Median, Quartiles, Deciles, Percentiles (appropriate for Ordinal, Interval and Ratio scale data)
- ▶ Mode (appropriate for Ordinal, Interval and Ratio scale data)

Descriptive Analysis of Univariate Data

Measures of Central Tendency: The Geometric Mean

- *Geometric mean*

- $\bar{X}_G = (X_1 \times X_2 \times \dots \times X_n)^{1/n}$ *of a variable over time*

- Geometric mean rate of return

- Measures the status of an investment over time

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_n)]^{1/n} - 1$$

- Where R_i is the rate of return in time period i

Descriptive Analysis of Univariate Data

- ▶ Measures of Central Tendency : The Harmonic Mean

$$\bar{X}_H = n / \sum (1/X_i)$$

Descriptive Analysis of Univariate Data

- ▶ Measures of Central Tendency: Weighted Mean

$$\bar{X}_w = (\sum W_i X_i) / (\sum W_i)$$

Descriptive Analysis of Univariate Data

Measures of Dispersion

Absolute Measures

- ▶ Range (appropriate for Interval and Ratio scale data)
- ▶ Mean Deviation
- ▶ Quartile Deviation
- ▶ Variance and Standard Deviation (appropriate for interval and ratio scale data)

Relative Measures

- ▶ Coefficient of Range (appropriate for Interval and Ratio scale data)
- ▶ Coefficient of Mean Deviation
- ▶ Coefficient of Quartile Deviation (appropriate for Ordinal, Interval and Ratio scale data)
- ▶ Coefficient of variation (appropriate for Ratio scale data)
- ▶ Relative and absolute frequencies (appropriate for Nominal scale data)

Descriptive Analysis of Univariate Data

Measures of Skewness

Types of Skewness

- ▶ Positive
- ▶ Negative

Coefficients of Skewness

- ▶ Karl–Pearson's Measure = $(\text{Mean} - \text{Mode}) / \text{S.D.}$
- ▶ Bowley's Measure = $(Q_3 + Q_1 - 2Md) / (Q_3 - Q_1)$
- ▶ Kelly's Measure = $(D_9 + D_1 - 2Md) / (D_9 - D_1)$
- ▶ Moments based coefficient (Beta-1)

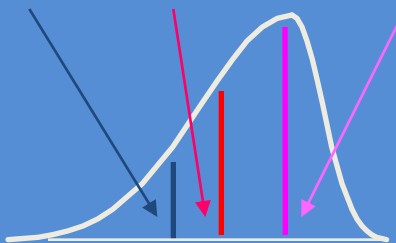
Descriptive Analysis of Univariate Data

Shape of a Distribution

- ▶ Describes how data is distributed
- ▶ Measures of shape
 - Symmetric or skewed

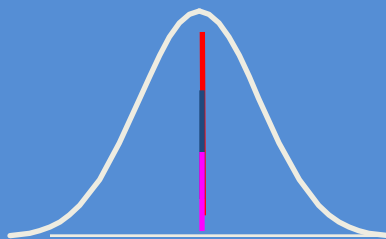
Left-Skewed

Mean < Median < Mode



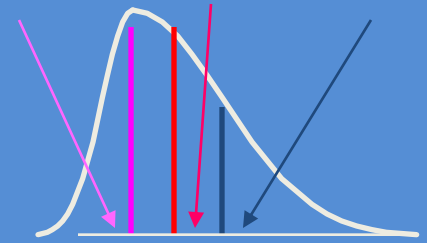
Symmetric

Mean = Median = Mode



Right-Skewed

Mode < Median < Mean



© 2002 Prentice-Hall, Inc.

Descriptive Analysis of Univariate Data

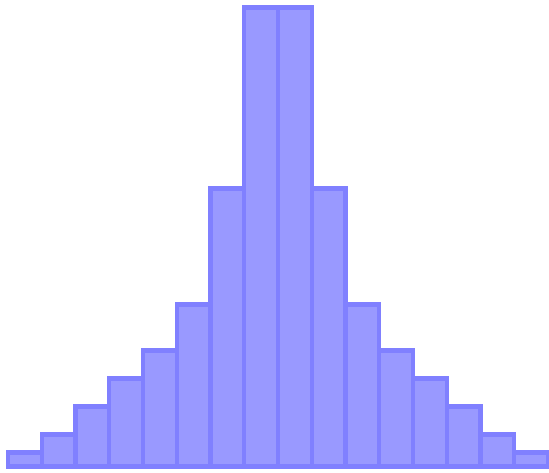
Measures of Kurtosis

Types of Kurtosis

- ▶ Platykurtic
- ▶ Mesokurtic
- ▶ Leptokurtic

Measure based on Moments (Beta-2)

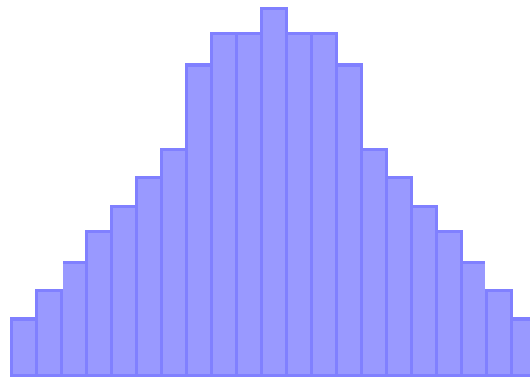
Kurtosis



Leptokurtic

(high peak)

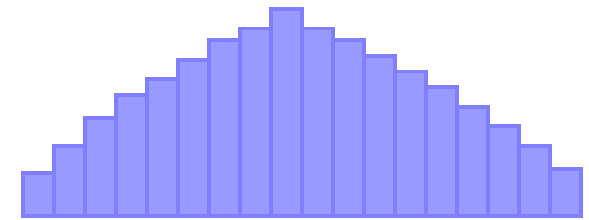
(+ve kurtosis)



Mesokurtic

(normal)

(zero kurtosis)



Platykurtic

(low peak)

(-ve kurtosis)

Mesokurtic distribution...kurtosis = 3

Leptokurtic distribution...kurtosis < 3

Platykurtic distribution...kurtosis > 3

BIVARIATE ANALYSIS: RELATIONSHIPS BETWEEN VARIABLES AND MEASURES OF ASSOCIATION

Handout #9

Bivariate Analysis

- ▶ Recall Problem Set #3A on Identifying Variables.
 - In each sentence you were asked to identify *two* variables pertaining to the same unit of analysis.
 - This was because each sentence claimed or implied that there is a *relationship* or *association* between these two variables.
 - For example, that a case with a “high” value on one variable is likely to have a “high” value on the other variable also, and likewise that a case with a “low” value on one variable is likely to have a “low” value on the other variable also.
 - Moreover, many of the sentences also claimed or implied that this association exists because there is a *cause and effect relationship* between the two variables
 - That is, that having a “high” (or “low”) value on one variable causes a case to have a “high” (or “low”) value on the other variable.
 - Thus, Problem Set #3A anticipated that we would move beyond *univariate analysis* to *bivariate analysis*.

Bivariate Data Analysis

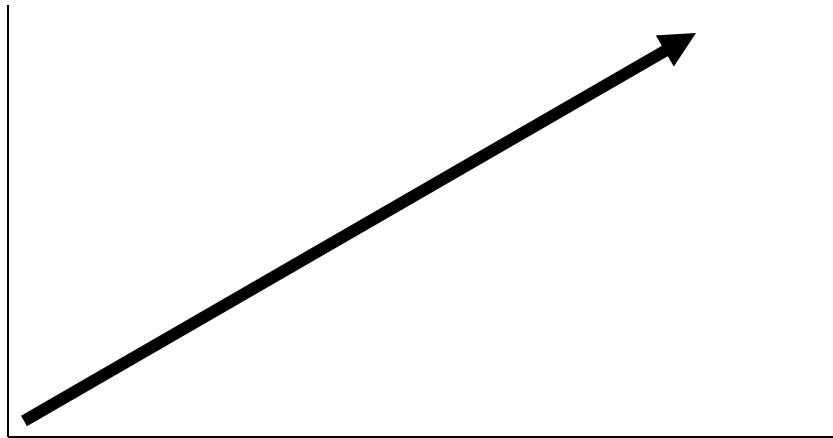
- ▶ Bivariate data analysis occurs when you look at the relationship between two variables.
- ▶ Research questions will ask “is there a relationship” between two variables.
- ▶ Hypotheses will predict a directional relationship between an independent and dependent variable.
- ▶ Descriptive statistics: frequencies, measures of central tendency and measures of dispersion/variation are all ways of conducting Univariate data analysis (e.g., the focus is on one variable).

Questions Re: Bivariate Relationships

- ▶ Does there appear to be a relationship?
 - Yes if the distribution of the dependent variable is different depending on the values of the independent variable.
- ▶ How strong is it?
 - The larger the distribution differences, the stronger the relationship.
- ▶ What is the direction of the relationship?
 - Used with ordinal or interval–ratio level variables.

Directional Relationships

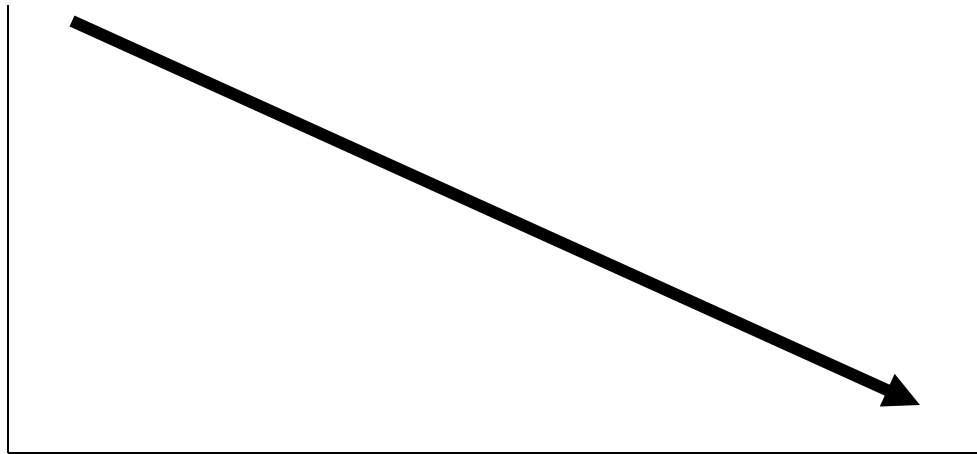
- ▶ Positive:
 - As the values of one variable increases, the values of the other variable increase



Directional Relationships

- ▶ Negative

- As the values of one variable increase, the values of the other variable decrease.

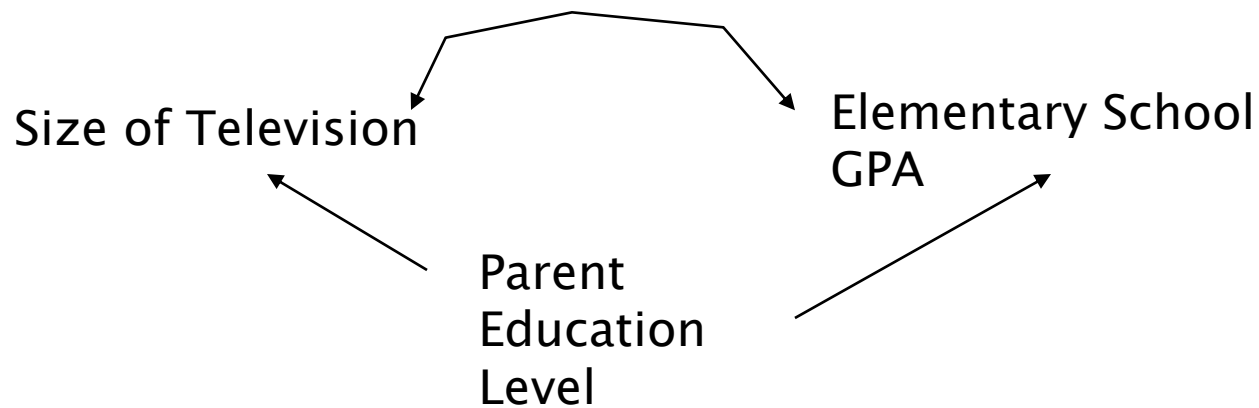


Elaboration

- ▶ A process to consider other factors that may influence a bivariate relationship
- ▶ Control variables: Used in multi-variate analysis to take into account the potential influence of one variable while looking at the relationship between at least two other variables.
- ▶ Example: What would you want to control for while looking at the relationship between Type of Family (single parent vs two parent) and school success?

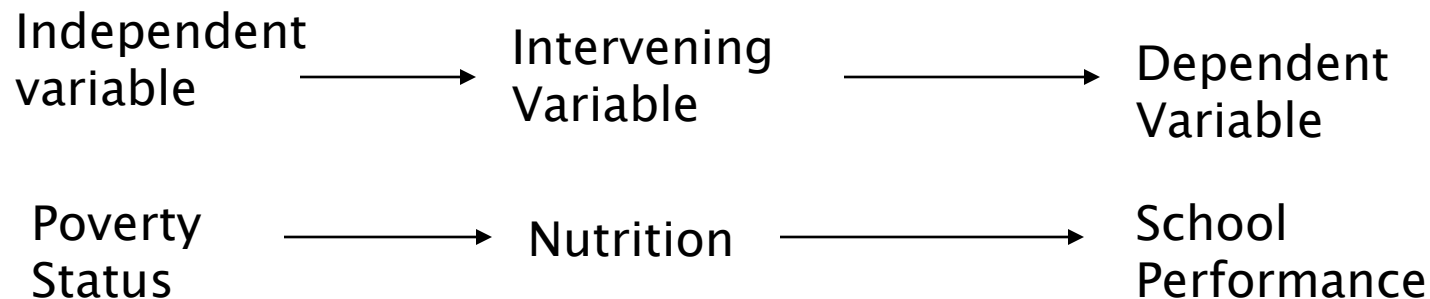
Spurious Relationship

- ▶ The apparent relationship between two variables is related to their causal connection to the third variable.



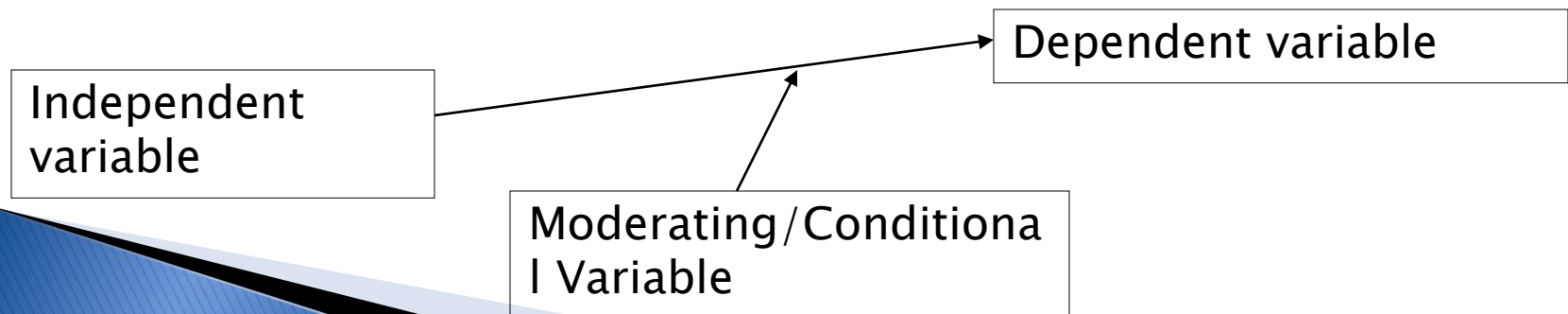
Intervening Relationship

- ▶ An intervening variable falls in time between the independent and dependent variable and has a causal relationship with the dependent variable and independent variable.



Conditional Relationship

- ▶ Also called a Moderating Variable.
- ▶ It moderates the influence between the independent and dependent variable. The relationship between the IV and DV will perform differently depending on the values of the moderating variable.



An example of a conditional relationship

Psychotropic
Treatment for
PTSD

(Meds or No
Meds)



PTSD
Symptoms

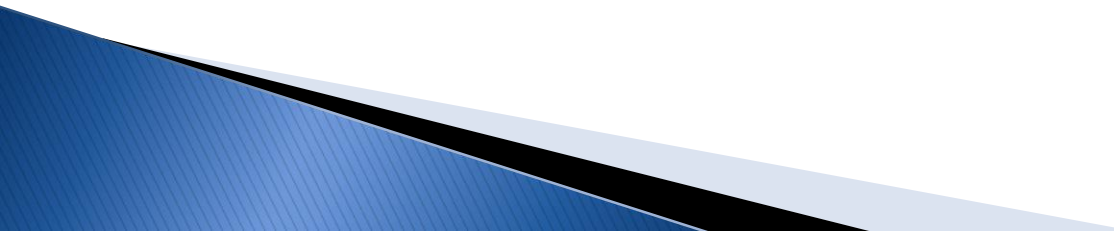
Type of Trauma

(Veteran or Crime
Victim)

Determining if a Relationship Exists

- ▶ Compute percentages for each value of X (down each column)
 - Base = marginal for each column
- ▶ Read the table by comparing values of X for each value of Y
 - Read table across each row
- ▶ Terminology
 - strong/ weak; positive/ negative; linear/ curvilinear

Bivariate Statistics

- Association between two variables.
 - Choose statistic based on the type of variables you have.
- 

Association

Association in bivariate data means that certain values of one variable tend to occur more often with some values of the second variable than with other variables of that variable (Moore p.242)

Cross Tabulation

***Correlation
Coefficient***

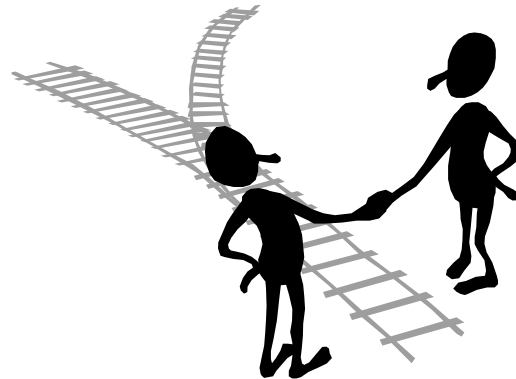
Describing association



Direction

Positive - Negative

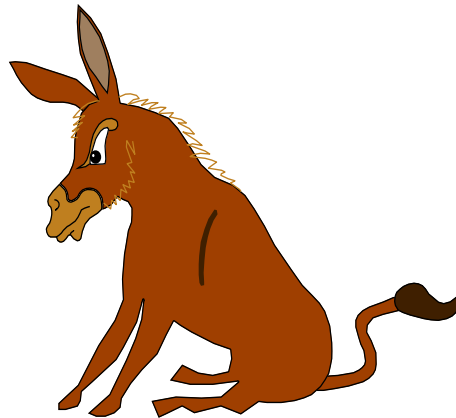
Strong - Weak



Strength

Nature

Linear - Curvilinear



Describing association

Two variables are *positively* associated when larger values of one tend to be accompanied by larger values of the other

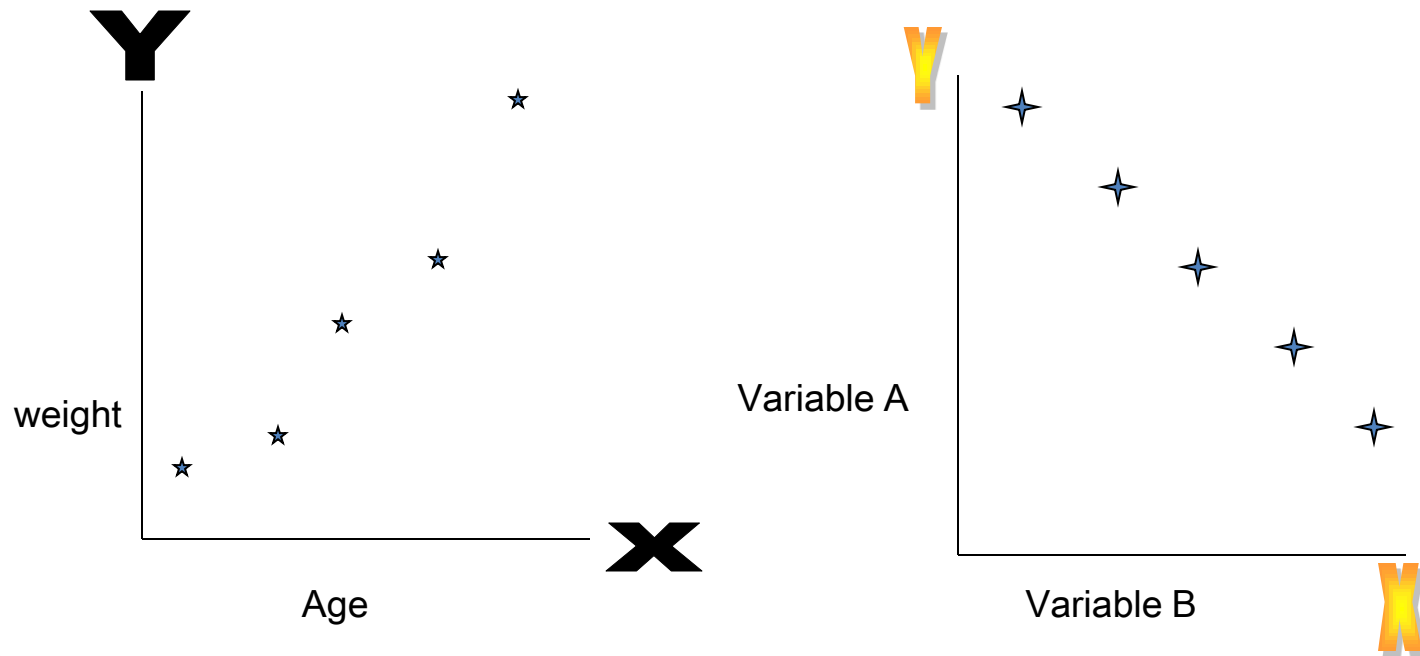
The variables are *negatively* associated when larger values of one tend to be accompanied by smaller values of the other

(Moore, p. 254)

Describing association

Scattergram or scatterplot

Graph that can be used to show how two interval level variables are related to one another



Association

- ▶ Example: gender and voting
 - Are gender and party supported *associated* (related)?
 - Are gender and party supported *independent* (unrelated)?
 - Are women more likely than men to vote republican?
Are men more likely to vote democrat?

Negative Association Between Variables

- ▶ However, the *direction of association* can be negative as well as positive, that is,
 - “high” may go with “low” and “low” may go with “high” or
 - as one variable increases, the other variable decreases.
- ▶ For example, sentence #2 in PS #3 claims that there is a *negative* (or *inverse*) relationship between two variables, which we can diagram in this way:

#2 LEVEL OF EDUCATION — DEGREE OF RELIGIOSITY
[inds]
(Low to High) ===== (Low to High)

- ▶ A negative relationship is sometimes called an *inverse* relationship, as in the *Inverse Square Root Law* of random sampling (Handout #2):

SIZE — MARGIN OF ERROR [samples]
(Low to High) ===== (Low to High)

Association That Cannot be Characterized as Positive or Negative

- ▶ All the variables in the examples presented thus far have values that run from “Low” to “High” and we made use of this fact in defining positive vs. negative association.
 - All **quantitative** (**interval** or **ratio**) variables have (numerical) values that run from “Low” (e.g., 0% TURNOUT) to “High” (e.g., 100% TURNOUT) and, as we have seen, many **ordinal** variables also have values that run from “Low” to “High.”
- ▶ However, some **ordinal** variables have values that, while they run in a natural order, do not run from “Low” to “High,” but rather (for example) from
 - “Liberal” to “Conservative” (IDEOLOGY variables),
 - “Strongly Agree” to “Strongly Disagree” (many ISSUE OPINION variables)
 - from “Never” to “Always” (ABORTION OPINION), and so forth.
- ▶ Moreover, **nominal** variables by definition have values that do not “run” in any natural ordering at all.

Variables with “Matching Values”

- ▶ In general, when both of two variables have “matching values” with a *common natural ordering*, an association between them may be characterized as positive or negative.
- ▶ Here are some additional examples involving variables that have matching values but are not of the “Low to High” type (and in which we probably expect positive associations):

PARENTS’ PARTY ID + CHILD’S PARTY ID [*parent-child pairs*]
(Dem. to Rep.) ===== (Dem. to Rep.)

PRES ECONOMIC POLICY + PRES FOREIGN POLICY [*inds*]
(Strong Ap. to Strong Disap.) === (Strong Ap. to Strong Disap.)

Variables with “Non-Matching Values”

- ▶ On the other hand, pairs of ordinal and nominal variables may have “non-matching” values, as in these examples:

#14 DIRECTION OF IDEOLOGY ===== DIRECTION OF VOTE [*inds*]
(Liberal to Conservative) (Dem. vs. Rep.)

RELIGIOUS AFFILIATION ===== PRESIDENTIAL VOTE [*inds*]
(Protestant vs. Catholic) (Dem. vs. Rep.)

- ▶ We certainly would expect IDEOLOGY and VOTE to be quite closely associated, but we will have to spell out the direction verbally (e.g., as in sentence #14: “liberals generally vote Democratic and conservatives generally vote Republican”), rather than relying on the semi-mathematical shorthand of “positive vs. negative.”
- ▶ Likewise, in the first half of the twentieth century (and outside of the South), there was a strong association between RELIGION and VOTE in the U.S., but again the direction of association has to be spelled out in words, namely that Protestants tended to vote Republican and Catholics tended to vote Democratic.
- ▶ In neither of these examples can we meaningfully specify the direction of the relationship in terms of the positive vs. negative shortcut, because the variables have non-matching values; rather we need to spell out its direction in words.

Broadening the Scope of Positive vs. Negative Association

- ▶ Given a dichotomous variable with “yes” and “no” values, “no” is conventionally considered to be “low” and “yes” to be “high,” so that we can meaningfully say (for example) that LEVEL OF POLITICAL INTEREST is *positively* associated with WHETHER OR NOT VOTED IN ELECTION.
- ▶ One may be able to rename and recode variables so that they have values that run from low to high, with the result that a relationship that previously could not be summarized in positive vs. negative terms now can be so characterized.

DIRECTION OF IDEOLOGY === LEVEL OF SUPPORT FOR U.N. [inds]
(Liberal to Conservative) (Low to High)

- ▶ We can rename and recode the first variable in a way that allows us to summarize the association in positive vs. negative terms.

DEGREE OF LIBERALISM + LEVEL OF SUPPORT FOR U.N. [inds]
(Low to High) ===== (Low to High)

“Polarity” of Variables

- ▶ Reversing the “polarity” of either variable reverses the (positive or negative) sign of the association.

DEGREE OF LIBERALISM — LEVEL OF OPPOSITION TO U.N. [inds]
(Low to High) ===== (Low to High)

DEGREE OF CONSERVATISM + LEVEL OF OPP TO U.N. [inds]
(Low to High) ===== (Low to High)

DEGREE OF CONSERVATISM — LEVEL OF SUPPORT FOR U.N. [inds]
(Low to High) ===== (Low to High)

- ▶ You should be able to see that all four of the diagrams above make the same substantive empirical claim.

Strength of Association Between Variables

- ▶ Beyond the question of the *direction* of an association (if any) between two variables, there is the question of the *strength* of the association between them. For example:
 - If almost all liberals vote Democratic and almost all conservatives vote Republican, there is a *strong association* between IDEOLOGY and VOTE.
 - But if liberals vote Democratic only slightly more than conservatives do, and conservatives vote Republican only slightly more than liberals do, there is only a *weak association* between IDEOLOGY and VOTE.

Measures of Association

- ▶ A great number of different of *bivariate summary statistics* called *measures of association* are used in quantitative research, each with somewhat different properties.
 - Different measures of association are appropriate depending on whether the variables are nominal, ordinal, or interval.
 - Many are defined and discussed in Weisberg, Chapter 12; however, you are asked only to skim this chapter and you are *not* responsible for knowing the different types of measures of association, let alone how to calculate them.
 - We will later study one measure of association that you will be responsible for: the *correlation coefficient* that measures strength and direction of association between two *interval* variables.
- ▶ But in the meantime, you should understand two general properties of all measures of association.
- ▶ We'll use the symbol *a* to designate a generic measure of association.

Measures of Association (cont.)

- ▶ Measures of association are *standardized*:
 - that is, such measures take on values between 0 and 1.
- ▶ If bivariate analysis shows that there is *no relationship or association* between two variables (e.g., if liberals vote Democratic and Republican in the same proportions as conservatives do), then $a = 0$.
- ▶ If it shows that there is *perfect relationship or association* between two variables (e.g., if all liberals vote Democratic and all conservatives vote Republican [*or vice versa*]), then $a = 1$.
- ▶ As the strength of association ranges between these (empirically unlikely) extremes, the value of a ranges between 0 and 1.

Measures of Association (cont.)

- ▶ If the variables have matching values such that an association can be characterized as *positive* or *negative*, the measure of association carries the appropriate (+ or -) sign.
- ▶ In this event, a measure of association takes on values that extend from -1 through 0 to +1. For example:
 - if all Democratic parents have Democratic children and all Republican parents have Republican children (the “pure socialization hypothesis”), then $a = +1$;
 - if Democratic and Republican parents are equally likely to have children of either partisanship (the “no impact hypothesis”), then $a = 0$; and
 - if all Democratic parents have Republican children and all Republican parents have Democratic children (the “pure rebellion hypothesis”), then $a = -1$.

Independent vs. Dependent Variables

- ▶ Association between variables is *symmetric*.
 - If PARENTS' PARTY ID is associated with CHILD'S PARTY ID, then it is equally true that CHILD'S PARTY ID is associated with PARENTS' PARTY ID.
 - If the association between variables X and Y is $a = +0.7$, then the association between Y and X is also $a = +0.7$.
- ▶ However, if there is an association between variables X and Y , it *may* (or *may not*) be due to the fact that X *influences* (or has *causal impact*) on Y or that Y influences X .
- ▶ Such a *cause and effect* relationship clearly is *not symmetric*.
 - Saying that variable X influences variable Y is different from saying that Y influences X .

Independent vs. Dependent Variables (cont.)

- ▶ Consider sentence #6 in PS #3A, which says that “hard studying makes for good grades.”
- ▶ In terms of the points made here, this sentence is saying *three* distinct things:
 - there is a *relationship* or *association* between LEVEL OF STUDY EFFORT and LEVEL OF GRADES, i.e., $a \neq 0$;
 - the relationship or association between the two variables is *positive*, i.e., $a > 0$; and
 - the reason [or at least part of the reason] this positive association exists is because LEVEL OF STUDY EFFORT has (positive) *causal impact* on (“makes for”) or *influences* LEVEL OF GRADES.
 - Note, however, the sentence does not suggest how strong this association is, i.e., how much a deviates from 0.

Independent vs. Dependent Variables (cont.)

- ▶ The variable that is the (hypothesized) cause is called the *independent variable*, and the variable that is the (hypothesized) effect is called the *dependent variable*.
- ▶ In diagrams such as we have been using, it is conventional to draw an arrow *from the independent to the dependent variable* in this manner:

DEGREE OF STUDYING + LEVEL OF GRADES [students]
(Low to High) =====> (Low to high)

In general: INDEPENDENT VAR =====> DEPENDENT VAR

Use this format in PS #9.

Independent vs. Dependent Variables (cont.)

In this context (but not in some others), the independent variable is conventionally put on the left side and the dependent variable on the right, so the arrow points from left to right.

Here are several ways of characterizing independent and dependent variables:

(1) The independent variable is the *cause*, the dependent variable is the *effect*.

(2) The independent variable *influences* (or *has an impact on*) the dependent variable.

(3) We want to *explain* why cases have particular values on the dependent variable; we use their values on the independent variable as (part of) the *explanation*.

Independent vs. Dependent Variables (cont.)

- ▶ Some of the sentences in PS #3A (and PS #9) contain words and phrases that clearly indicate the direction of (claimed) causality (and whether the causal effect is positive or negative), for example:
 - LEVEL OF EDUCATION *undermines* DEGREE OF RELIGIOSITY
 - DEGREE OF STUDYING *makes for* LEVEL OF GRADES
- ▶ In other sentences, the direction of causality is (at best) implicit only, e.g., #1 and #14.

Independent vs. Dependent Variables (cont.)

- ▶ In some contexts, we may have good reason to study the association between two variables even if we don't regard one as independent and the other as dependent. For example, we would expect the following to be true (with a quite high measure of association):

SCORE ON PART I OF TEST + SCORE ON PART II OF TEST [*students*]
(Low to High) ===== (Low to High)

- ▶ If this were not true, we might conclude that *at least* one part of the test is not an especially *valid* measure of students' mastery of the material.
 - In fact, on the first POLI 300 test last semester, the correlation (association) between these two variables was +0.779.
- ▶ However, it doesn't make much sense to regard one of these variables as independent and the other dependent.

Independent vs. Dependent Variables (cont.)

- ▶ Variables are not *intrinsically* independent or dependent, but rather they assume one or other role in different hypotheses, theories, or research projects.
- ▶ For example, once we have developed a concept of PARTY IDENTIFICATION, have figured out how to measure it, have collected appropriate data, and have completed basic univariate analysis, we may turn to further *bivariate* research questions pertaining to PARTY ID.

Independent vs. Dependent Variables (cont.)

- ▶ One set of such questions concerns the *causes of* or, or *influences on*, or *explanations for* PARTY ID.
 - Why do some people think of themselves as partisans and others as Independents?
 - Among partisans, why do some people think of themselves as Democrats and others as Republicans?
 - Is PARTY ID influenced by PARENTS' PARTY ID, LEVEL OF EDUCATION, LEVEL OF INCOME, RELIGION AFFILIATION, IDEOLOGY, etc?
- ▶ Here we are treating (children's) PARTY ID as the *dependent* variable and are looking for *independent* variables that may affect the direction and strength of PARTY ID.

Independent vs. Dependent Variables (cont.)

- ▶ Another set of questions concerns the *consequences* or *effects* of PARTY ID.
 - Does PARTY ID affect how likely people are to turn out and vote?
 - Does PARTY ID effect how people vote?
 - Does PARENTS' PARTY ID affect the party identification and other political attitudes of their children?
- ▶ Here we are treating (parents') PARTY ID as the *independent* variable and are looking for *dependent* variables that may be influenced by PARTY ID.

Methods of analysis (De Vaus, 134)

Univariate
methods

Frequency distributions



Bivariate
methods

Cross tabulations



Scattergrams



Regression



Correlation



Comparison of means



Descriptive Analysis of Bivariate Data

Preparation of cross-tables

Interpretation of cross-tables – For interpretation of cross-tables, it is required to identify dependent and independent variable.

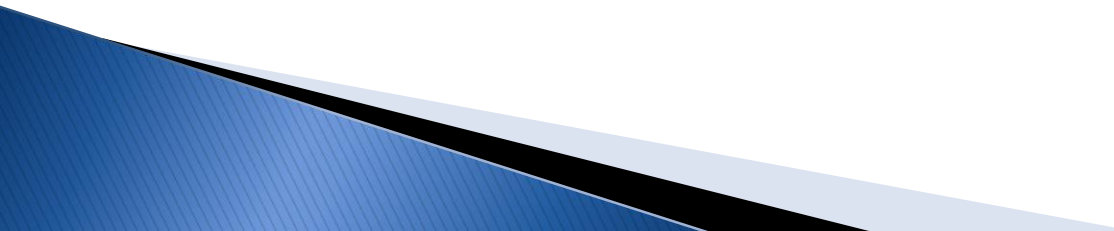
Percentages should be computed in the direction of independent variable.

There is no hard and fast rule as to where the dependent or independent variables are to be taken. They can be taken either in rows or in columns.

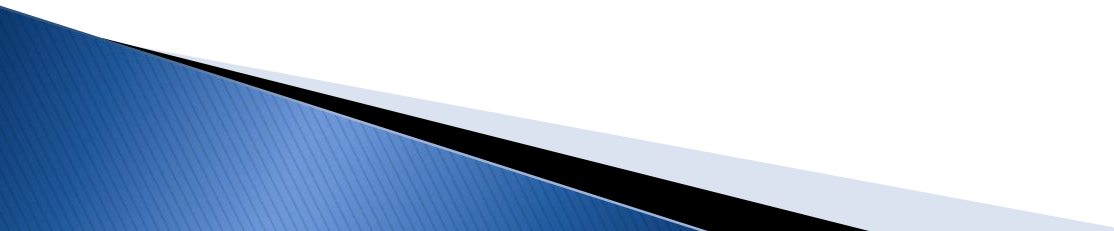
Cross Tabulation

- ▶ A cross tabulation table shows the intersection between two categorical variables (nominal or ordinal). You can construct a cross tabulation table with interval–ratio level variables, but it would be too complex and difficult to analyze, especially if there is a lot of variability.
- ▶ We are often asked to produce cross tabulation tables for social service agencies (e.g., types of services by gender, type of problem, etc...)

Setting Up a Cross Tabulation Table

- ▶ The independent variable is usually placed in the columns
 - ▶ The dependent variable is usually placed in the rows.
 - ▶ The columns represent the attributes of the IV. The rows represent the attributes of the DV.
 - ▶ Each CELL represents the intersection of one IV attribute and one DV attribute.
- 

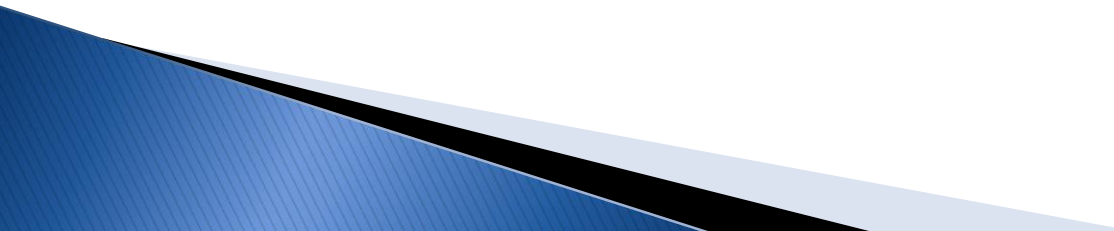
Setting Up a Cross tabulation Table

- ▶ At the base of each column, sum the frequencies in that column.
 - ▶ At the end of each row, sum the frequencies in that row.
 - ▶ The summary row and column are called marginals.
 - ▶ The lower right hand cell represents the total sample size.
- 

Hypothesis: Homeless people with criminal backgrounds will be less likely to report being employed.

		Have you ever been convicted of a crime?		
Do you have a regular job or do day labor?		NO	YES	TOTAL
	NO	314	380	694
	YES	200	196	396
	TOTAL	514	576	1090

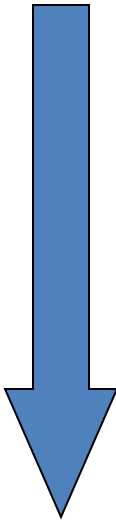
Cross Tabulation Percentages


- ▶ Column Percent: Show the distribution of each row attribute within each column.
 - ▶ Row Percent: Show the distribution of each column attribute within each row.
 - ▶ If the independent variable is in the column, select column percent.
 - ▶ If the IV is in the rows, select row percent.
 - ▶ For non-directional questions, calculate both column and row percents.
- 

Cross Tabulation Tables

- ▶ Designate the X variable and the Y variable
- ▶ Place the values of X across the table
- ▶ Draw a *column* for each X value
- ▶ Place the values of Y *down* the table
- ▶ Draw a *row* for each Y value
- ▶ Insert frequencies into each CELL
- ▶ Compute totals (MARGINALS) for each column and row

Cross tabulation tables

		<i>Occupation</i>				Total	Calculate percent
		White collar		Blue collar			
		Freq	%	Freq	%		
<i>Vote</i>	Democrat	270	27%	810	81%	1080	
	Republican	730	73%	190	19%	920	
	Totals	1000	100%	1000	100%	2000	

Read
Table 

(De Vaus pp 158-160)

Cross tabulation

- ▶ Use column percentages and compare these across the table
- ▶ Where there is a difference this indicates some association

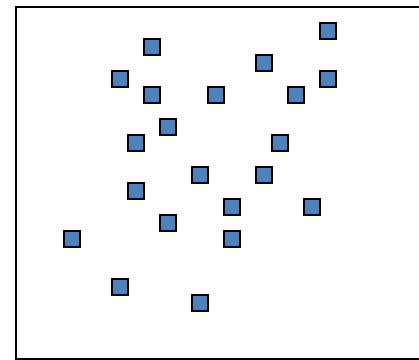
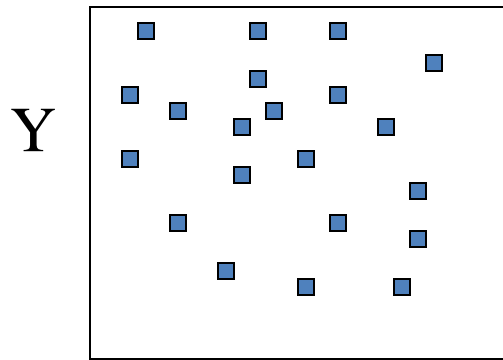
Hypothesis: Homeless people with criminal backgrounds will be less likely to report being employed.

		Have you ever been convicted of a crime?		
Do you have a regular job or do day labor?		NO	YES	TOTAL
	NO	314 (61%)	380 (66%)	694 (64%)
	YES	200 (39%)	196 (34%)	396 (36%)
	TOTAL	514 (100%)	576 (100%)	1090 (100%)

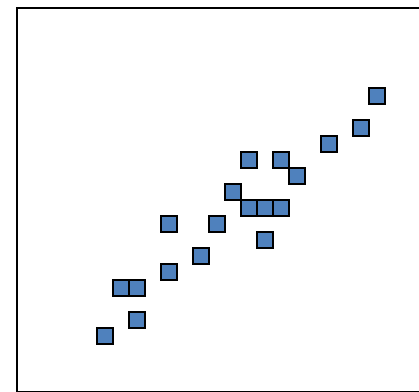
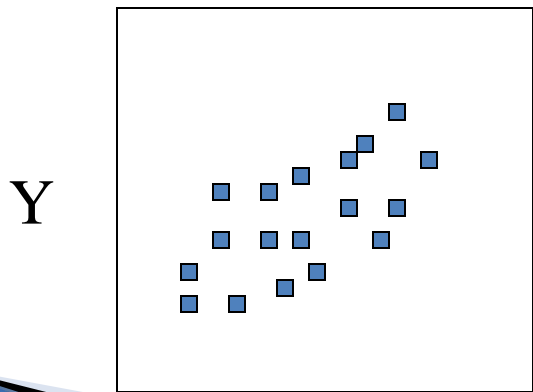
Description of Scattergrams

- Strength of Relationship
 - Strong
 - Moderate
 - Low
- Linearity of Relationship
 - Linear
 - Curvilinear
- Direction
 - Positive
 - Negative

Description of scatterplots



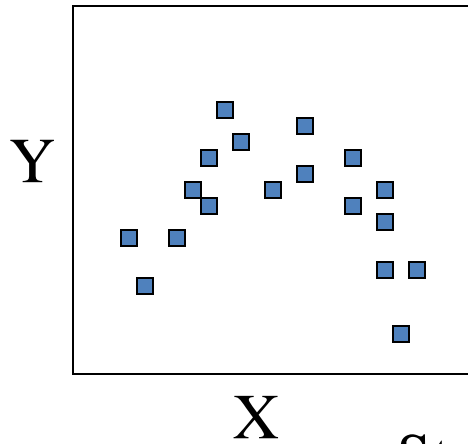
Strength and direction



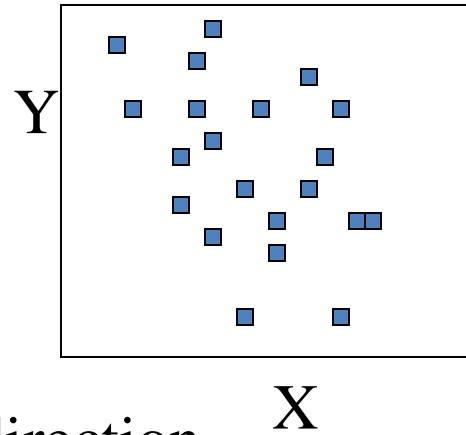
Y

X

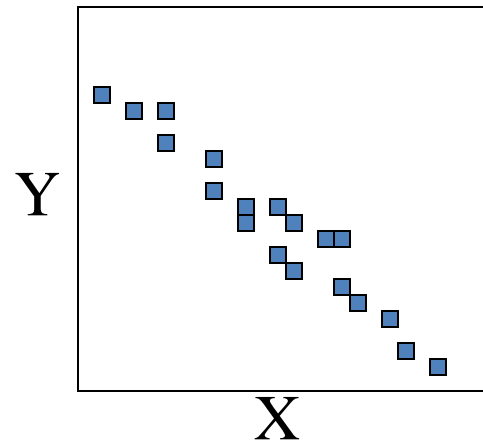
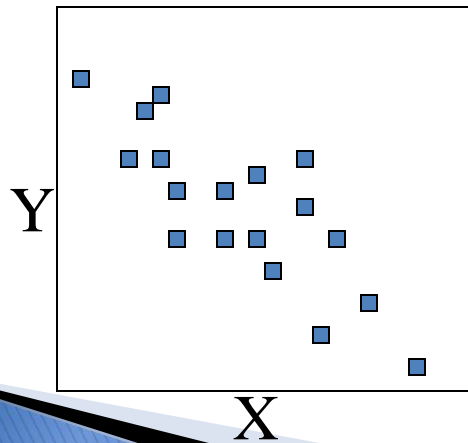
Description of scatterplots



Nature



Strength and direction

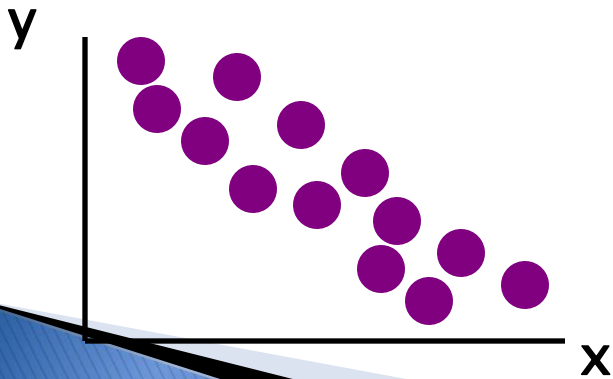
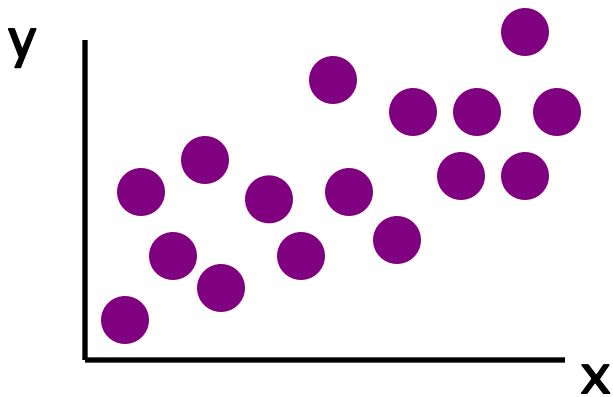


Scatter Plots and Correlation

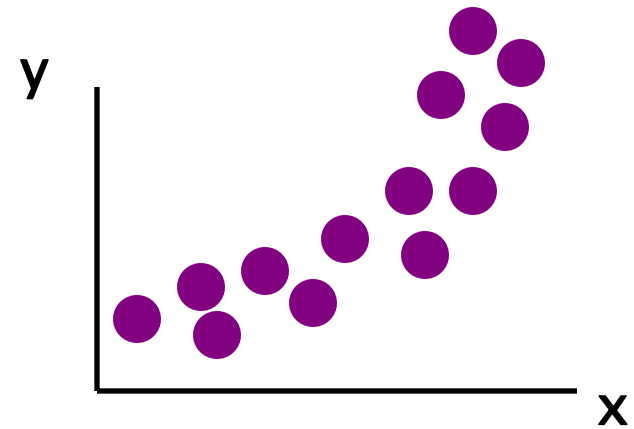
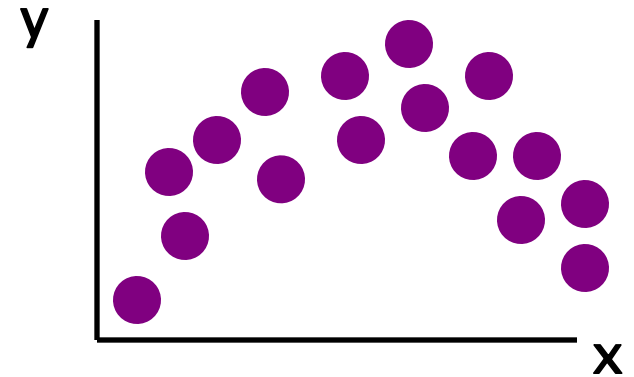
- ▶ A **scatter plot** (or scatter diagram) is used to show the relationship between two variables
- ▶ **Correlation** analysis is used to measure strength of the association (linear relationship) between two variables
 - Only concerned with strength of the relationship
 - No causal effect is implied

Scatter Plot Examples

Linear relationships



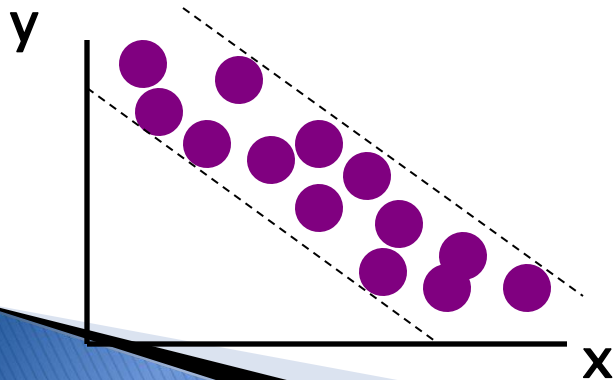
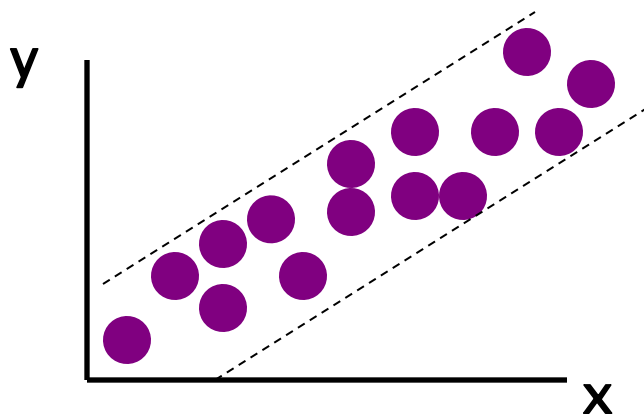
Curvilinear relationships



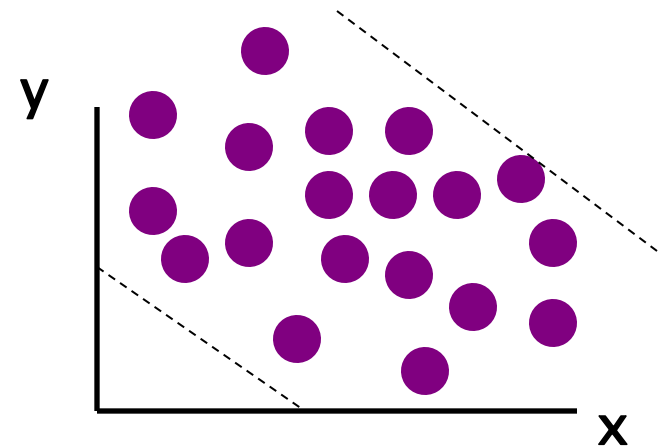
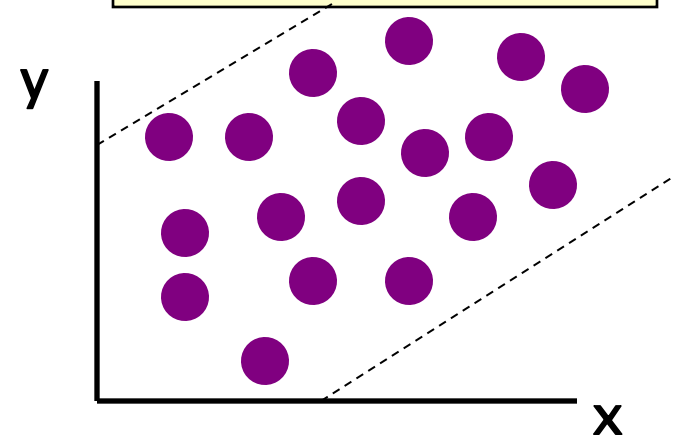
Scatter Plot Examples

(continued)

Strong relationships



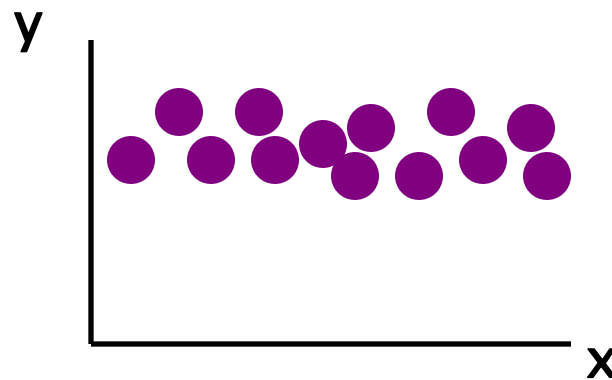
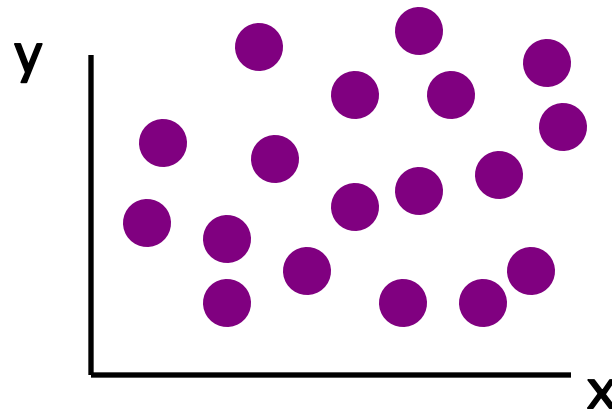
Weak relationships



Scatter Plot Examples

(continued)

No relationship



The Sample Covariance

- ▶ The covariance measures the strength of the linear relationship between **two variables**
- ▶ The **population covariance**:

$$\text{Cov}(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

- ▶ The **sample covariance**:

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Only concerned with the strength of the relationship
- No causal effect is implied

Interpreting Covariance

► Covariance between two variables:

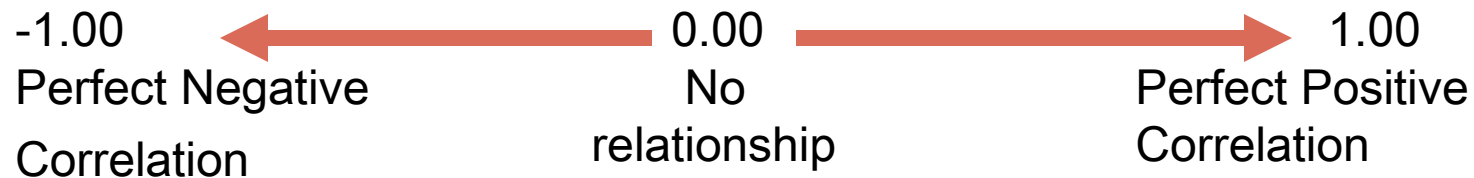
$\text{Cov}(x,y) > 0$ → x and y tend to move in the **same**
direction

$\text{Cov}(x,y) < 0$ → x and y tend to move in **opposite**
directions

$\text{Cov}(x,y) = 0$ x and y are independent

Correlation

- ▶ Correlation coefficient—number used to describe the strength and direction of association between variables
 - Very strong = .80 through 1
 - Moderately strong = .60 through .79
 - Moderate = .50 through .59
 - Moderately weak = .30 through .49
 - Very weak to no relationship 0 to .29



Correlation Coefficients

- Nominal
 - Phi
 - Cramer's V
- Ordinal (linear)
 - Gamma
- Nominal and Interval
 - Eta

<http://www.nyu.edu/its/socsci/Docs/correlate.html>

Correlation: Pearson's r

- Interval and/or ratio variables
 - Pearson product moment coefficient (r)
 - two interval variables, normally distributed
 - assumes a linear relationship
 - Can be any number from
 - 0 to -1 : 0 to 1 (+1)
 - Sign (+ or -) shows direction
 - Number shows strength
 - Linearity cannot be determined from the coefficient
- e.g.: $r = .8913$

Correlation Coefficient

(continued)

- ▶ The **population correlation coefficient ρ** (rho) measures the strength of the association between the variables
- ▶ The **sample correlation coefficient r** is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Quantitative Estimate of a Linear Correlation

- ▶ A quantitative estimate of a linear correlation between two variables X and Y is given by Karl Pearson as:

$$r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$
$$r_{xy} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \sqrt{\sum_{i=1}^n Y_i^2 - n\bar{Y}^2}}$$

Where, r_{xy} = Correlation coefficient between X and Y

\bar{X} = Mean of the variable X

\bar{Y} = Mean of the variable Y

n = Size of the sample

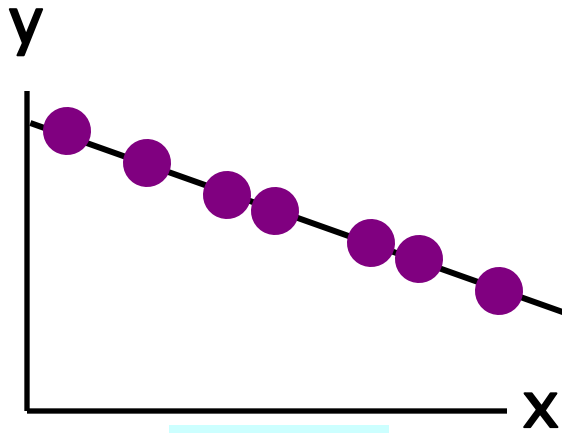
Quantitative Estimate of a Linear Correlation

- ▶ The linear correlation coefficient takes a value between -1 and $+1$ (both values inclusive).
- ▶ If the value of the correlation coefficient is equal to 1 , the two variables are perfectly positively correlated and the scatter of the points of the variables X and Y will lie on a positively sloped straight line.
- ▶ Similarly, if the correlation coefficient between the two variables X and Y is -1 , the scatter of the points of these variables will lie on a negatively sloped straight line and such a correlation will be called a perfectly negative correlation.

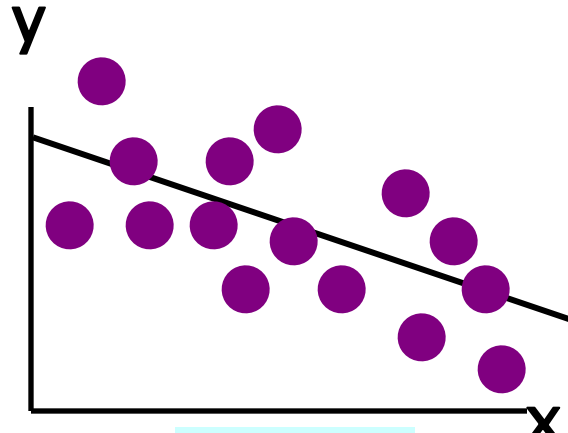
Features of ρ and r

- ▶ Unit free
- ▶ Range between -1 and 1
- ▶ The closer to -1 , the stronger the negative linear relationship
- ▶ The closer to 1 , the stronger the positive linear relationship
- ▶ The closer to 0 , the weaker the linear relationship

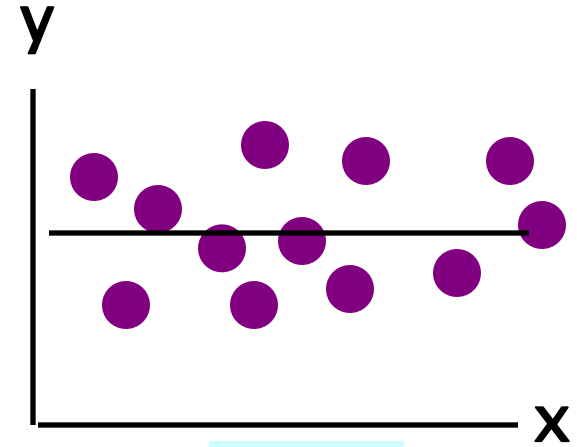
Examples of Approximate r Values



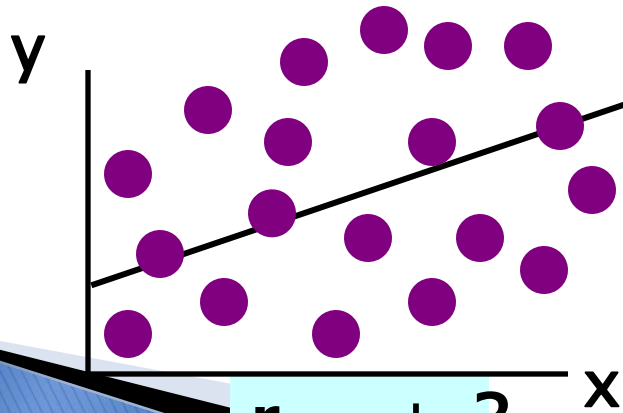
$r = -1$



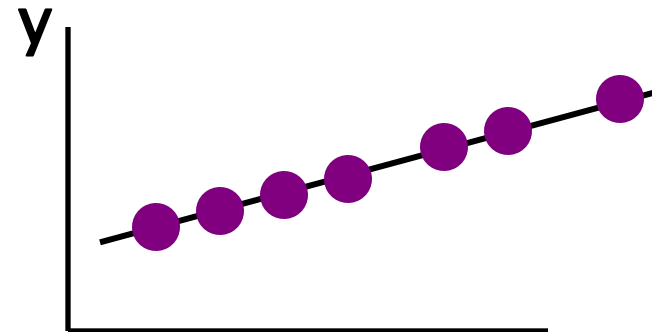
$r = -.6$



$r = 0$



$r = +.3$



$r = +1$

Calculating the Correlation Coefficient

Sample Correlation Coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$r = \hat{\rho} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right) = \frac{S_{XY}}{S_X S_Y}$$

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

Calculation Example

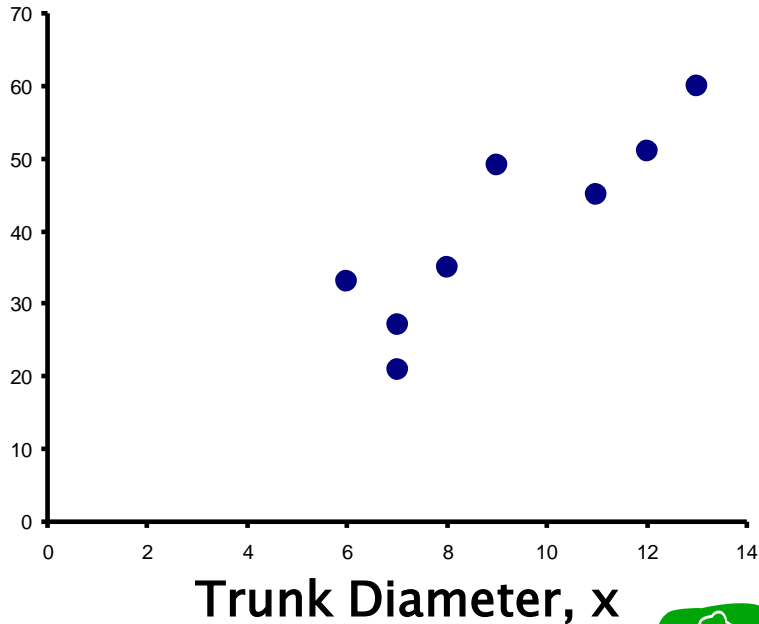
Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\Sigma=321$	$\Sigma=73$	$\Sigma=3142$	$\Sigma=14111$	$\Sigma=713$



Calculation Example

(continued)

Tree
Height,
y



$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$
$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$
$$= 0.886$$

$r = 0.886 \rightarrow$ relatively strong positive linear association between x and y

Excel Output

Excel Correlation Output

Tools / data analysis / correlation...

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between
Tree Height and Trunk
Diameter



Descriptive Analysis of Bivariate Data

Spearman's rank correlation coefficient is given by

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Where, r_s = Spearman's rank correlation coefficient

n = Sample size

d_i = Difference in the ranking for the i^{th} contestant

The rank correlation coefficient takes a value between -1 and $+1$.

Introduction to Regression Analysis

- ▶ **Regression analysis** is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable

Regression Analysis

Regression analysis examines associative relationships between a metric dependent variable and one or more independent variables in the following ways:

- ▶ Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
- ▶ Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
- ▶ Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
- ▶ Predict the values of the dependent variable.
- ▶ Control for other independent variables when evaluating the contributions of a specific variable or set of variables.
- ▶ Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.

Coefficient of Correlation

- ▶ Measures the relative strength of the linear relationship between two variables
- ▶ Population correlation coefficient:

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

- ▶ Sample correlation coefficient:

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Statistical Relationship

The descriptive statistics that measures the degree of relation between 2 variables are called **correlation coefficients**.

Three measures for statistical relationship are:

Scale <i>data</i>	Pearson's r	<ul style="list-style-type: none">•Normal Distribution•Linearity
Ordinal (<i>or above data</i>)	Kendall's Tau-b	<ul style="list-style-type: none">•Distribution free•Monotonicity
Nominal (<i>or above data</i>)	Chi-Square Test	<ul style="list-style-type: none">•Raw frequency > 5

Statistical Relationship (Cont.)

- ▶ **Pearson Correlation coefficient** (ρ , r) measures the strength of linear relationship between two variables (X and Y) assuming normal distribution. *{significance!}*

$$\rho_{x,y} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

$$r_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x}_x)(y_i - \bar{y}_y)}{(n-1)s_x s_y}$$

Statistical Relationship (Cont.)

- ▶ Correlation coefficient will **range from -1 to $+1$**
- ▶ **A correlation of 0 indicates that there is no linear relationship between two variables**
- ▶ Even a high **correlation could be observed just by chance**; to be sure we need to run a statistical test.
- ▶ Correlation between two variables **does not mean causal relationship** between them
- ▶ **Correlation Matrix** provides pair-wise correlation between more than two variables.

Statistical Relationship (Cont.)

- ▶ Square of Pearson's r (r^2) can be interpreted as explained variance if there is a Dependent Variable (DV) Independent Variable (IV) relationship exists.
- ▶ For example if $r = 0.933$ than $r^2 = 0.87049$ that means IV explains 87.05% of the variations in the DV.

Correlation Test Assumptions

Parametric Correlation Test

- ▶ Pearson's r:
 - Interval data
 - Normality
 - Equal Variance (*not needed if $n \geq 30$*)
 - Linearity

Statistical Relationship (Cont.)

▶ Kendall's tau-b

- A distribution-free (nonparametric) measure of association for ordinal (or ranked) variables that take ties into account.
- ▶ The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships.
- ▶ Possible values range from -1 to 1 .

Statistical Relationship (Cont.)

- ▶ **Spearman's rho** Commonly used distribution-free (nonparametric) measure of correlation between two ordinal variables.
- ▶ For all of the cases, the values of each of the variables are ranked from smallest to largest, and the Pearson correlation coefficient is computed on the ranks.

Correlation Test Assumptions

Non-Parametric Correlation Test

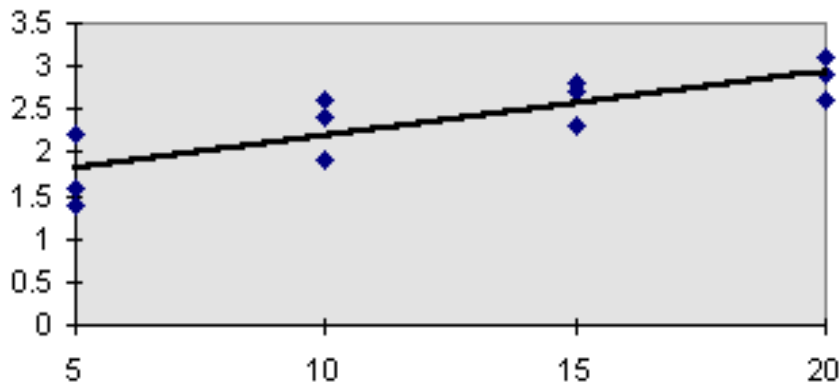
- ▶ Kendall's tau-b & Spearman's rho:
 - Ordinal data
 - Monotonicity

Simple Linear Regression Model

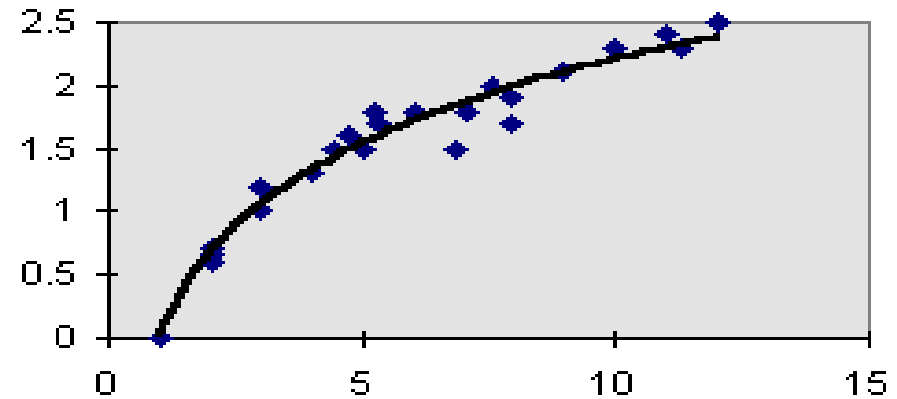
- ▶ Only **one independent variable**, x
- ▶ Relationship between x and y is described by a linear function
- ▶ Changes in y are assumed to be caused by changes in x

Types of Regression Models

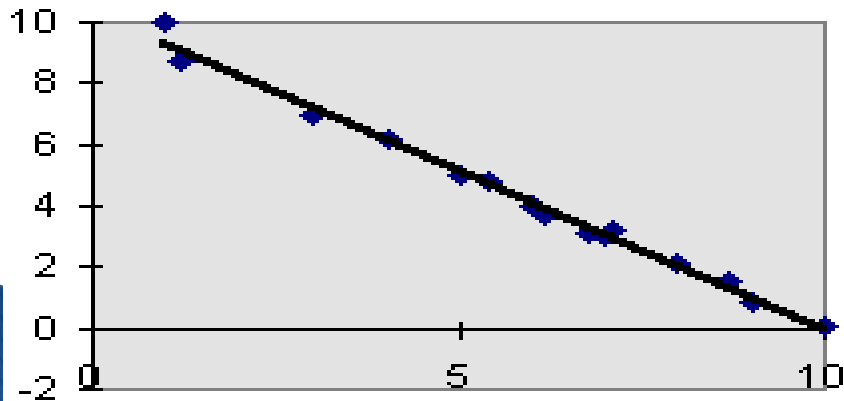
Positive Linear Relationship



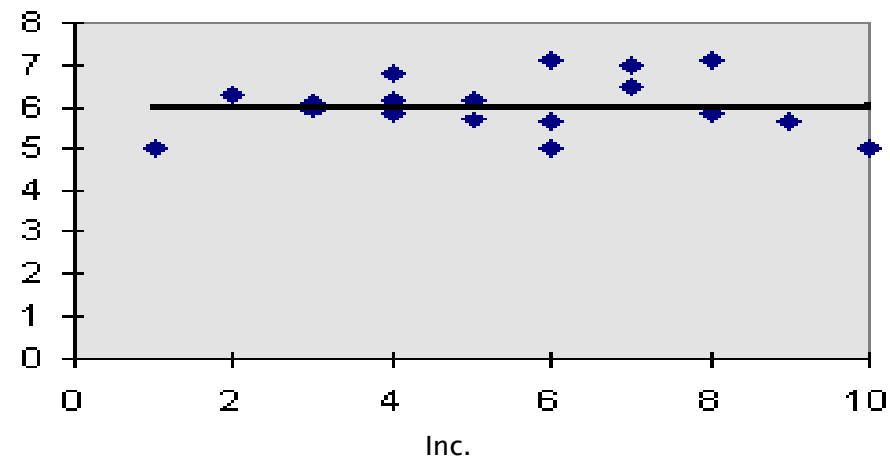
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Conducting Bivariate Regression Analysis

Plot the Scatter Diagram



Formulate the General Model



Estimate the Parameters



Estimate Standardized Regression Coefficients



Test for Significance



Determine the Strength and Significance of Association



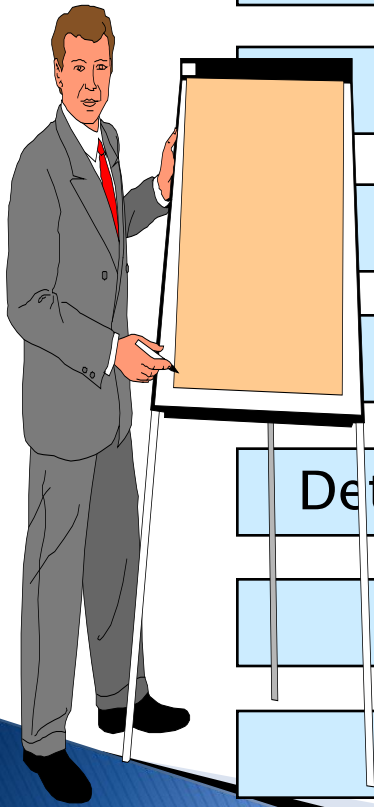
Check Prediction Accuracy



Examine the Residuals



Cross-Validate the Model



Population Linear Regression

The population regression model:

The diagram illustrates the population linear regression model, $y = \beta_0 + \beta_1 X + \epsilon$, with the following components and labels:

- Dependent Variable:** y
- Population y intercept:** β_0
- Population Slope Coefficient:** β_1
- Independent Variable:** X
- Random Error term, or residual:** ϵ

The model is composed of two main parts:

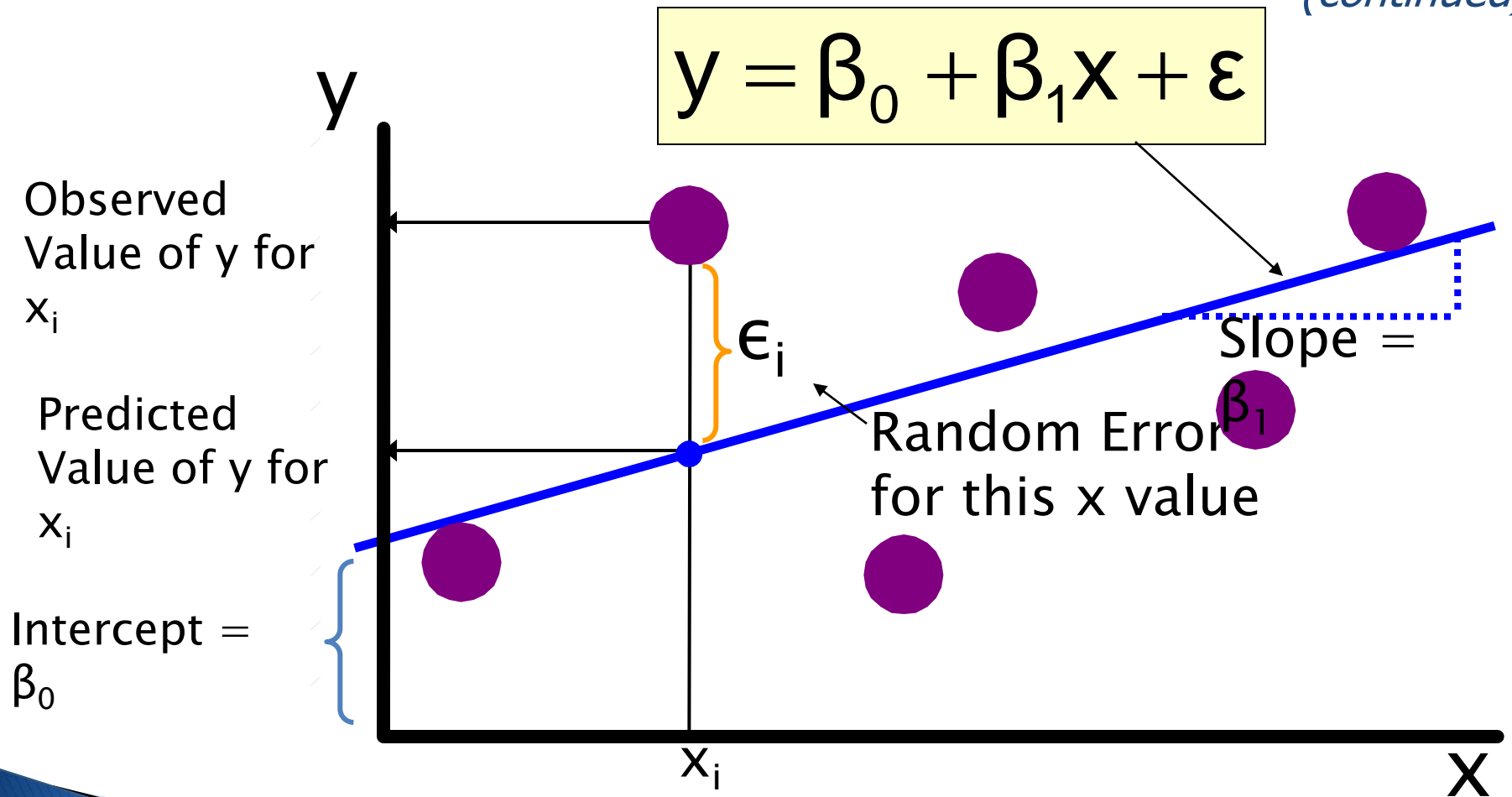
- Linear component:** $\beta_0 + \beta_1 X$
- Random Error component:** ϵ

Linear Regression Assumptions

- ▶ Error values (ϵ) are statistically independent
- ▶ Error values are normally distributed for any given value of x
- ▶ The probability distribution of the errors is normal
- ▶ The probability distribution of the errors has constant variance
- ▶ The underlying relationship between the x variable and the y variable is linear

Population Linear Regression

(continued)



Estimated Regression Model

The sample regression line provides an **estimate** of the population regression line

Estimated
(or
predicted) y
value

Estimate of
the
regression
intercept

Estimate of the
regression
slope

Independent
variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Least Squares Criterion

- ▶ b_0 and b_1 are obtained by finding the values of b_0 and b_1 that **minimize the sum of the squared residuals**

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

The Least Squares Equation

- ▶ The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b_1 = r \frac{s_y}{s_x}$$

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Interpretation of the Slope and the Intercept

- ▶ b_0 is the estimated average value of y when the value of x is zero
- ▶ b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- ▶ A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- ▶ A random sample of 10 houses is selected
 - Dependent variable (y) = house price in \$1000s
 - Independent variable (x) = square feet



Sample Data for House Price Model

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Excel Output

Regression Statistics

Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation

is: $\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$

ANOVA

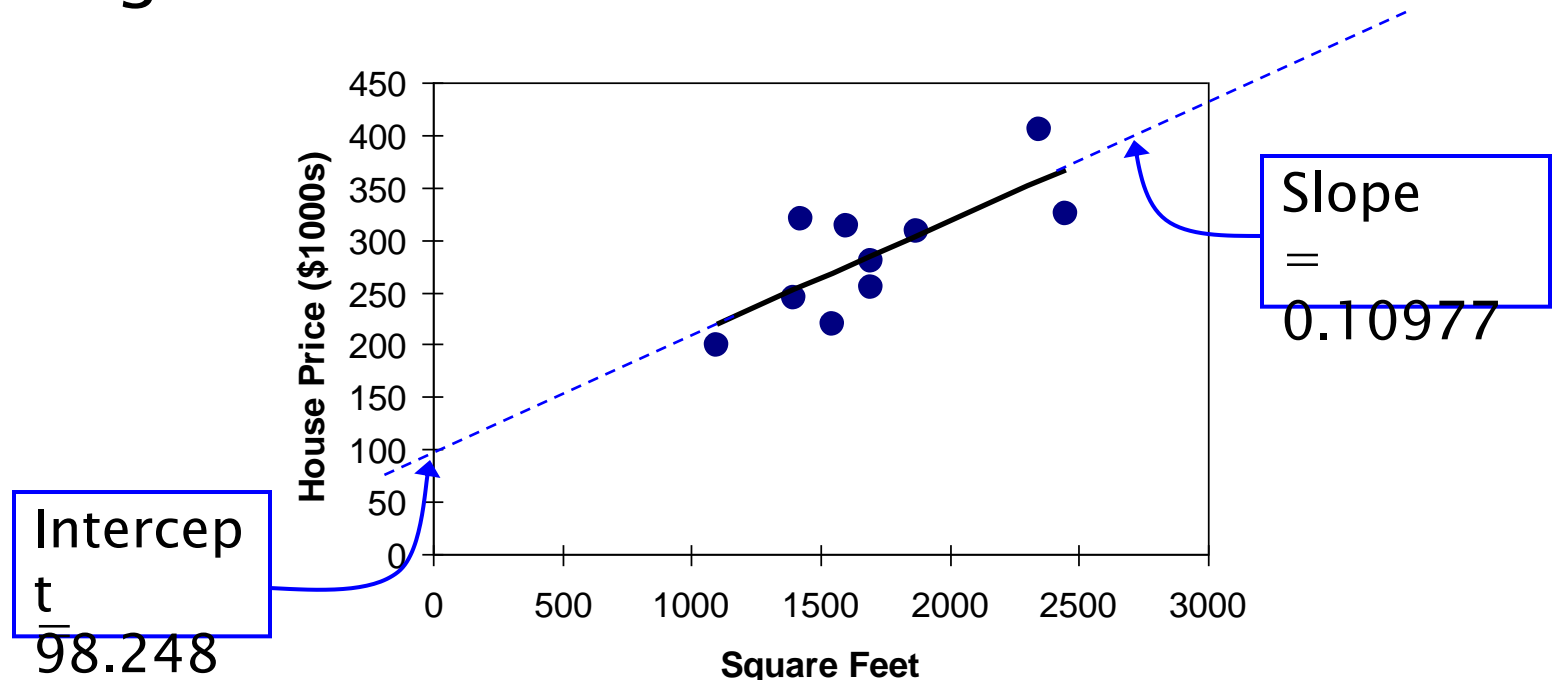
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			


	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580



Graphical Presentation

- ▶ House price model: scatter plot and regression line




$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

- ▶ b_0 is the estimated average value of Y when the value of X is zero (if $x = 0$ is in the range of observed x values)
 - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet



Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

- ▶ b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
 - Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(\$1000) = \109.77 , on average, for each additional one square foot of size



Least Squares Regression Properties

- ▶ The sum of the residuals from the least squares regression line is 0 $\sum (y - \hat{y}) = 0$
- ▶ The sum of the squared residuals is a minimum (minimized $\sum (y - \hat{y})^2$)
- ▶ The simple regression line always passes through the mean of the y variable and the mean of the x variable
- ▶ The least squares coefficients are unbiased estimates of β_0 and β_1

Coefficient of Determination, R^2

- ▶ The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- ▶ The coefficient of determination is also called **R-squared** and is denoted as R^2

$$R^2 = \frac{SSR}{SST}$$

where

$$0 \leq R^2 \leq 1$$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Note: In the single independent variable case, the coefficient of determination is

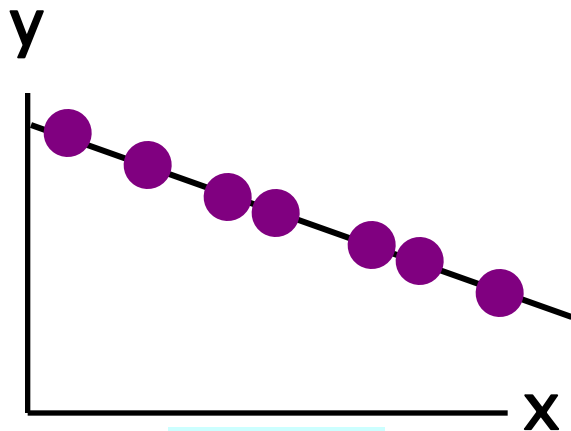
$$R^2 = r^2$$

where:

R^2 = Coefficient of determination

r = Simple correlation coefficient

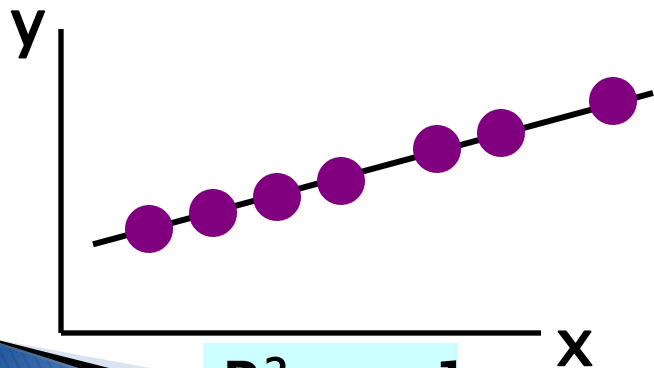
Examples of Approximate R^2 Values



$$R^2 = 1$$

$$R^2 = 1$$

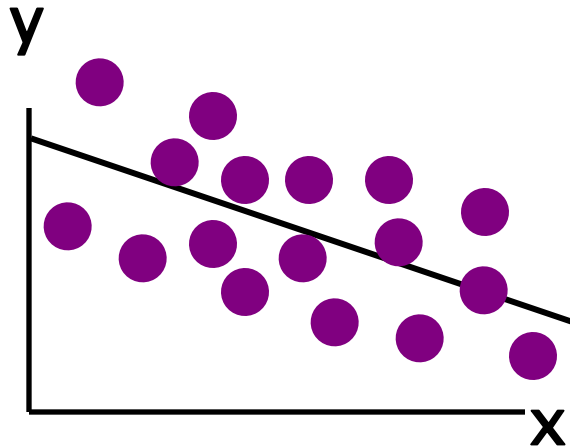
Perfect linear relationship between x and y :



$$R^2 = +1$$

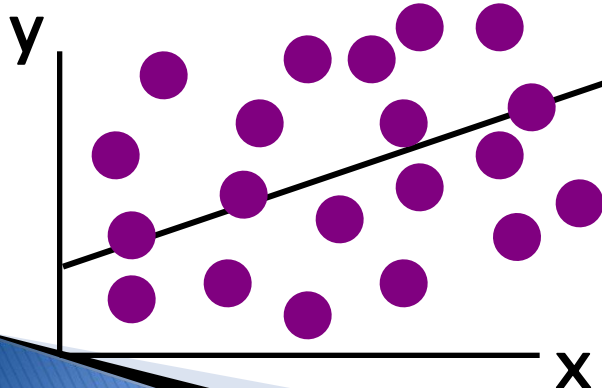
100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



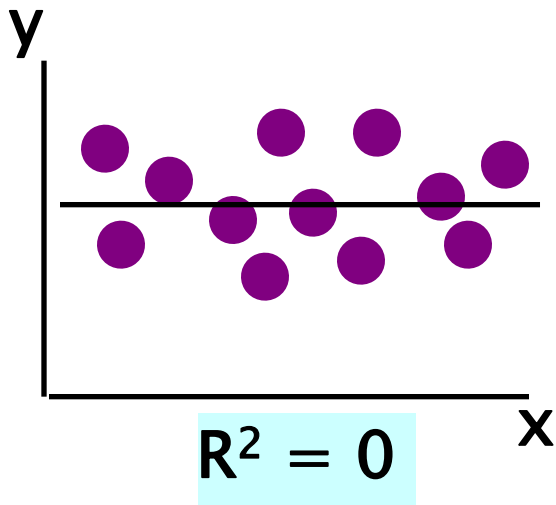
$$0 < R^2 < 1$$

Weaker linear relationship between x and y:



Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



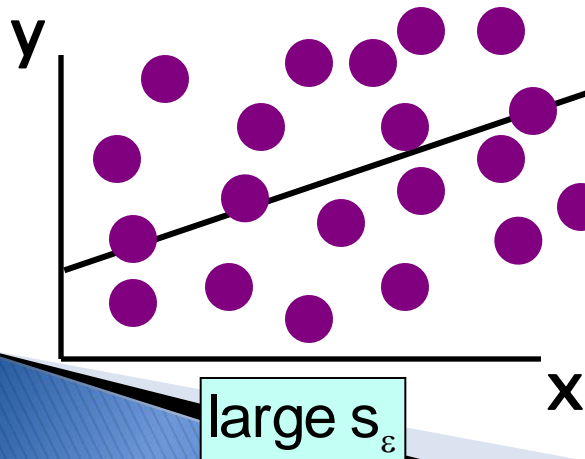
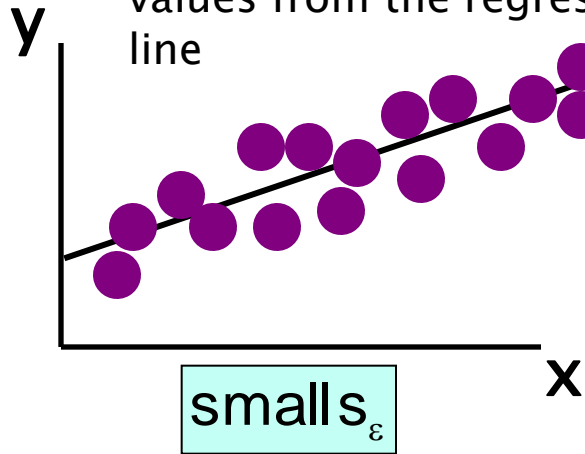
$$R^2 = 0$$

No linear relationship between x and y :

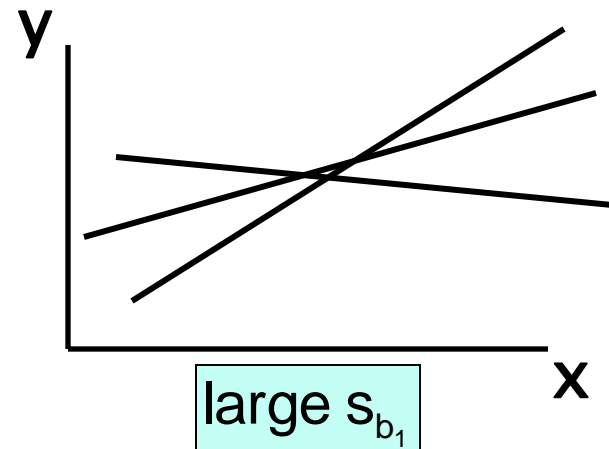
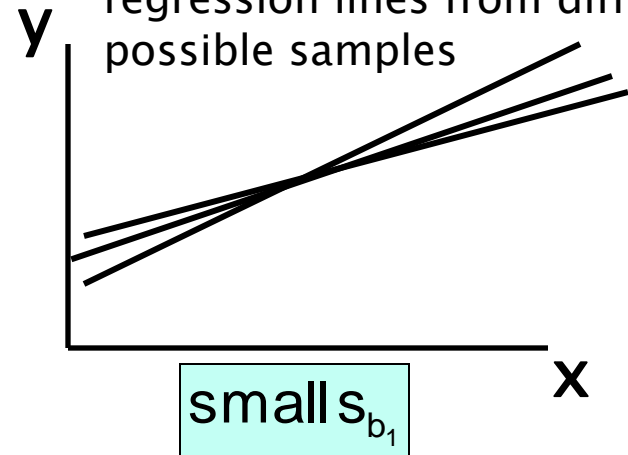
The value of Y does not depend on x . (None of the variation in y is explained by variation in x)

Comparing Standard Errors

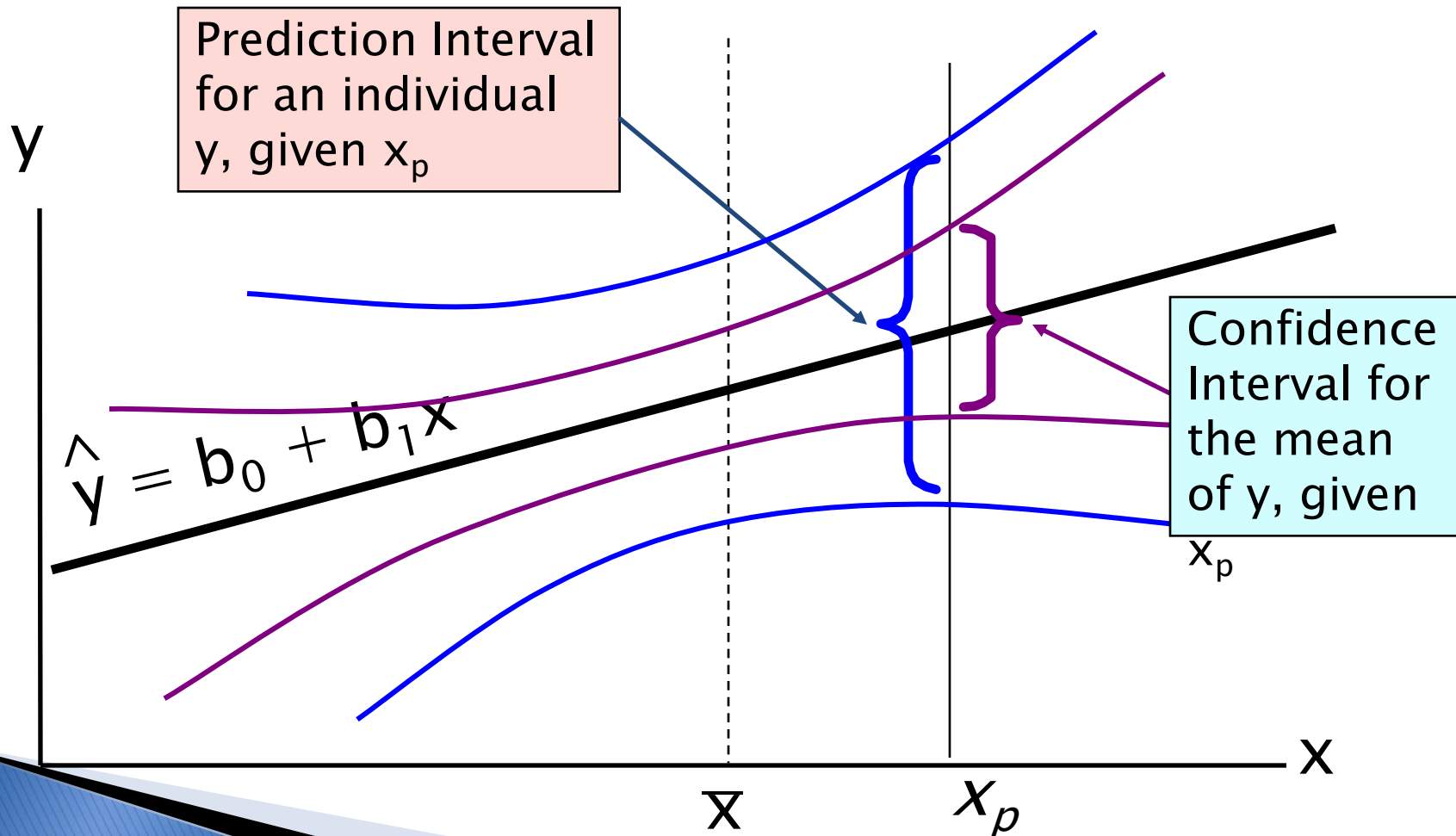
Variation of observed y values from the regression line



Variation in the slope of regression lines from different possible samples



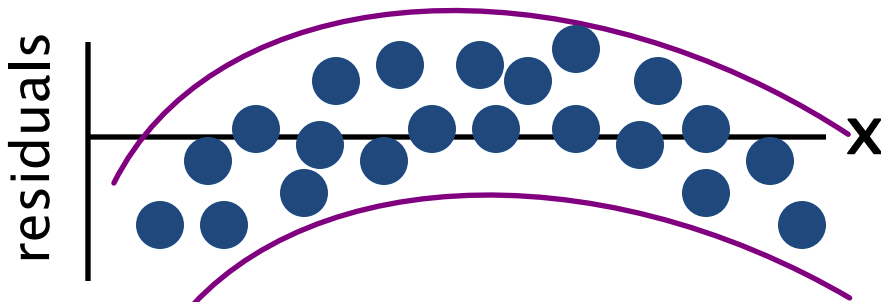
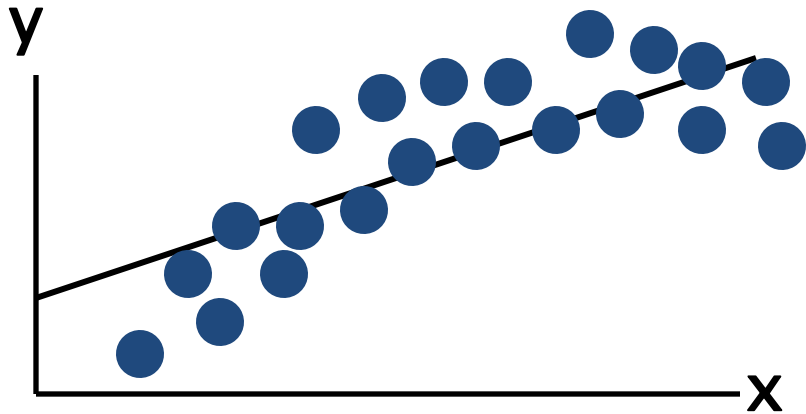
Interval Estimates for Different Values of x



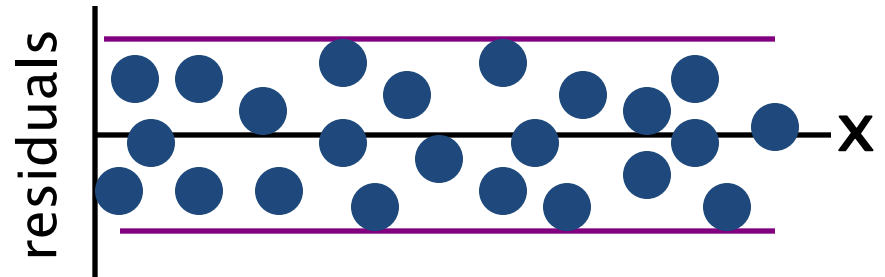
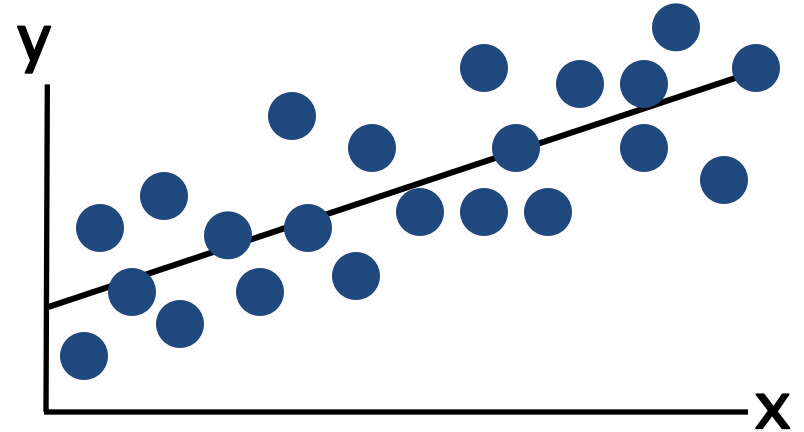
Residual Analysis

- ▶ Purposes
 - Examine for linearity assumption
 - Examine for constant variance for all levels of x
 - Evaluate normal distribution assumption
- ▶ Graphical Analysis of Residuals
 - Can plot residuals vs. x
 - Can create histogram of residuals to check for normality

Residual Analysis for Linearity

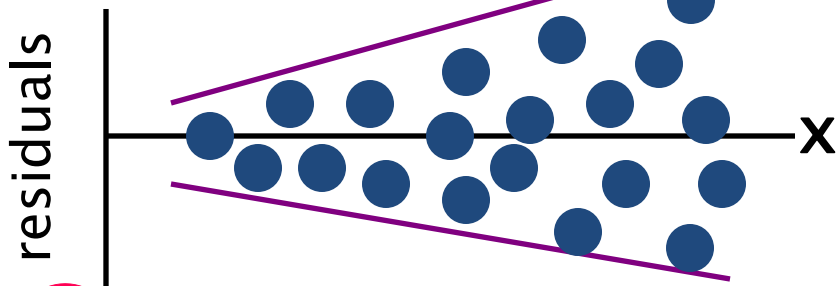
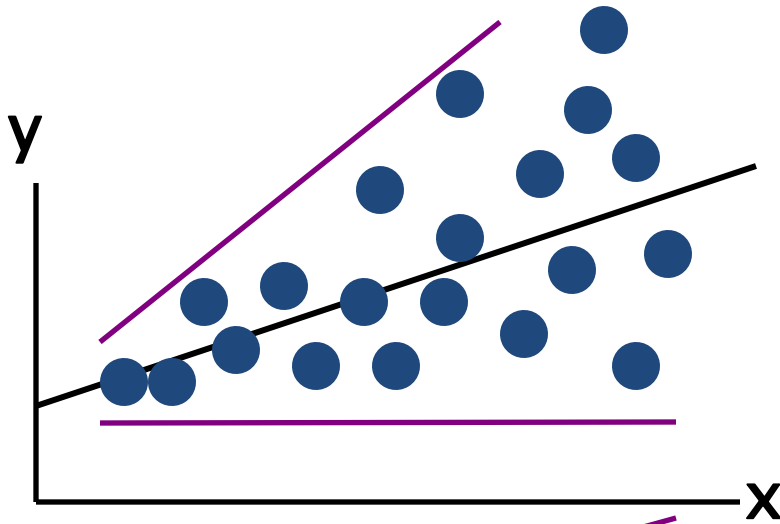


Not Linear

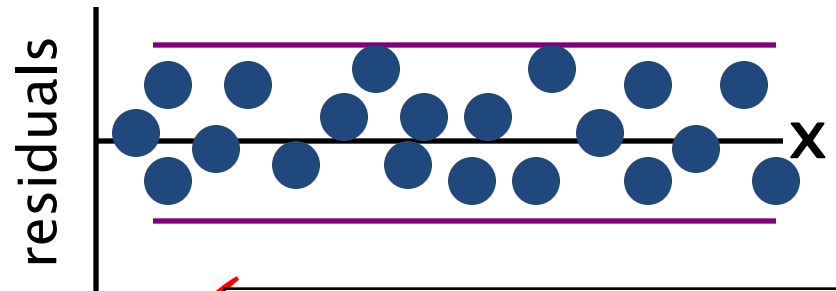
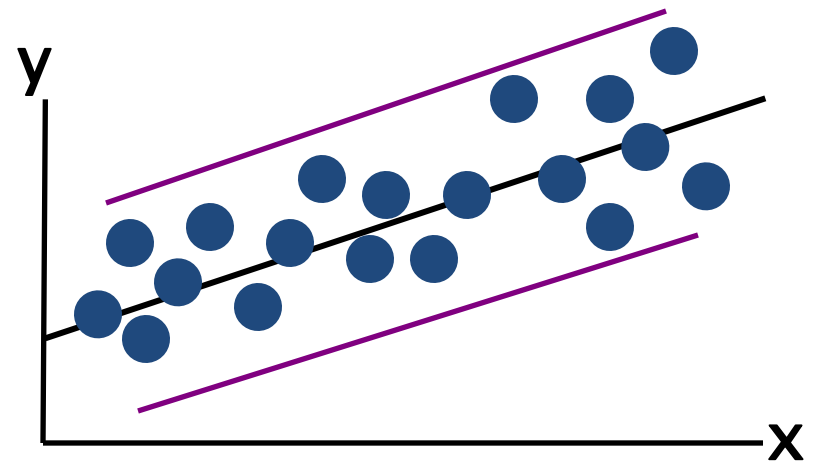


Linear

Residual Analysis for Constant Variance



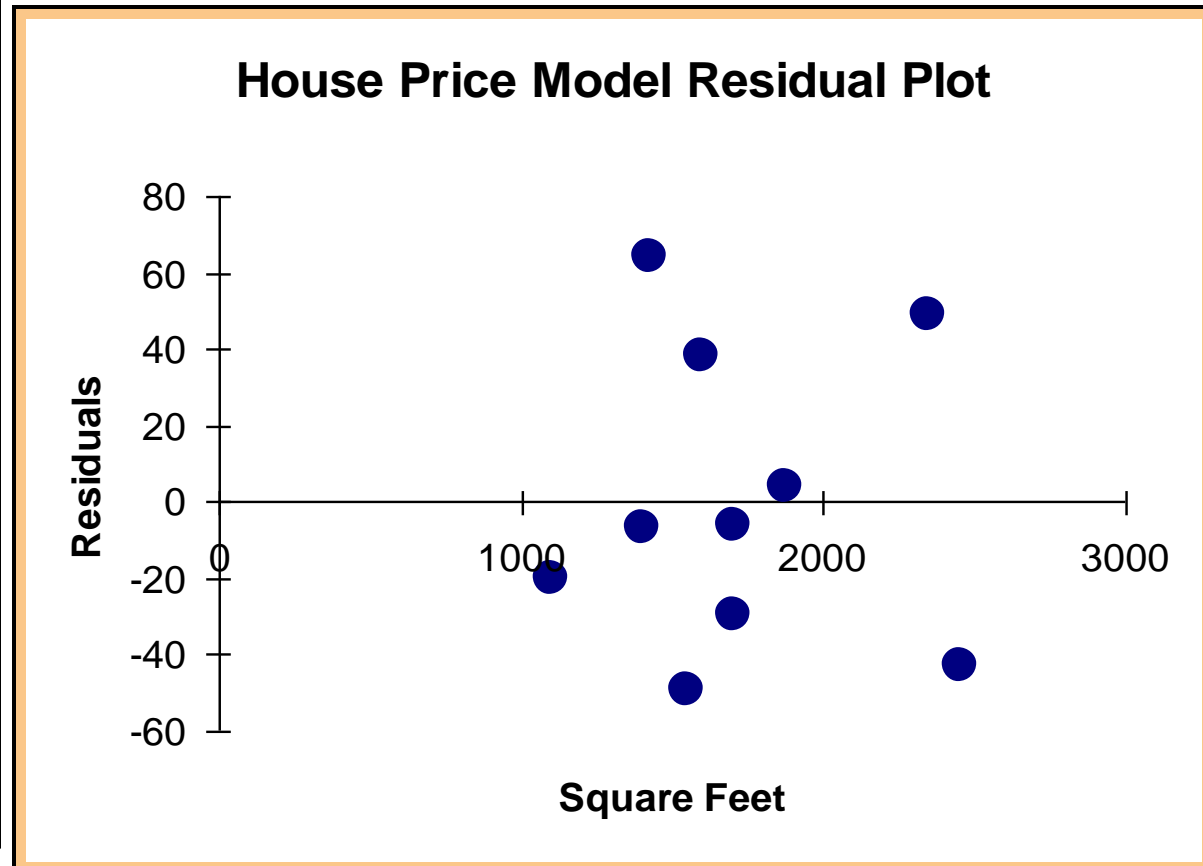
Non-constant
variance



Constant variance

Excel Output

RESIDUAL OUTPUT		
	<i>Predicted House Price</i>	<i>Residuals</i>
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Simple Regression Analysis

Simple linear regression equation can be presented as

$$Y = \alpha + \beta X + U$$

Where,

U = Stochastic error term

α , β = Parameters to be estimated

- ▶ The equation is estimated using the ordinary least squares (OLS) method of estimation.
- ▶ The OLS method of estimation states that the regression line should be drawn in such a way so as to minimize the error sum of squares.

Simple Regression Analysis

The OLS method of estimation would result in the following two normal equations:

$$\sum Y = n\hat{\alpha} + \hat{\beta}\sum X$$

$$\sum XY = \hat{\alpha}\sum X + \hat{\beta}\sum X^2$$

Solving the above normal equations results in:

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

Once $\hat{\beta}$ is estimated, the value of α may be computed as,

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Simple Regression Analysis

- ▶ The estimate of error (residual) term is obtained as:

$$\hat{U} = Y - \hat{Y}$$

where \hat{Y} = estimated value of the dependent variable

where Y = observed value of the dependent variable

The estimate of the variance of the error term is given by:

$$V(\hat{U}) = \hat{\sigma}_U^2 = \frac{\sum_{i=1}^n \hat{U}_i^2}{n - k}$$

$$\text{Standard error of estimate} = \hat{\sigma}_U = \sqrt{\frac{\sum_{i=1}^n \hat{U}_i^2}{n - k}}$$

- n and k denote the sample size and number of parameters to be estimated respectively.

Statistics for assessing an association between two variables, unpaired data

Risk factor (independent variable, exposure, group assignment)	Outcome (dependent variable)					
	Dichotomous	Nominal	Interval, normal distribution	Interval non-normal	Ordinal	Time to event, censored data
Dichotomous	Chi-squared, Fisher's exact test, risk ratio, odds ratio	Chi-squared	<i>t</i> -test	Mann-Whitney test	Chi-squared for trend, Mann- Whitney test	Log-rank, Wilcoxon, rate ratio
Nominal	Chi-squared, exact test	Chi-squared	ANOVA	Kruskal-Wallis test	Kruskal-Wallis test	Log-rank, Wilcoxon
Interval, normal distribution	<i>t</i> -test	ANOVA	Linear regression, Pearson's correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–
Interval, non-normal	Mann-Whitney test	Kruskal-Wallis test	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–
Ordinal	Chi-squared for trend, Mann- Whitney test	Kruskal-Wallis test	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–

Comparison of Bivariate tests for unpaired and paired data

	Independent observations (2 groups)	Paired observations (2 observations)	Independent observations (≥ 3 groups)	Repeated observations (≥ 3 observations)
Dichotomous variable	Chi-squared Fisher's exact	McNemar's test	Chi-squared	Cochran's Q
Normally distributed interval variable	<i>t</i> -test	Paired <i>t</i> -test	ANOVA	Repeated-measures ANOVA
Non-normally distributed interval variable	Mann-Whitney test	Wilcoxon signed rank test	Kruskal–Wallis test	Friedman statistic
Ordinal variable	Mann-Whitney test	Wilcoxon signed rank test	Kruskal–Wallis test	Friedman statistic

CHAPTER-12

TESTING OF HYPOTHESES

What is a Hypothesis?

- ▶ A hypothesis is an assumption or a statement that may or may not be true.
- ▶ The hypothesis is tested on the basis of information obtained from a sample.
- ▶ Hypothesis tests are widely used in business and industry for making decisions.
- ▶ Instead of asking, for example, what the mean assessed value of an apartment in a multistoried building is, one may be interested in knowing whether or not the assessed value equals some particular value, say Rs 80 lakh.
- ▶ Some other examples could be whether a new drug is more effective than the existing drug based on the sample data, and whether the proportion of smokers in a class is different from 0.30.

Concepts in Testing of Hypothesis

- ▶ **Null hypothesis:** The hypotheses that are proposed with the intent of receiving a rejection for them are called null hypotheses. This requires that we hypothesize the opposite of what is desired to be proved. For example, if we want to show that sales and advertisement expenditure are related, we formulate the null hypothesis that they are not related. Null hypothesis is denoted by H_0 .
- ▶ **Alternative hypothesis:** Rejection of null hypotheses leads to the acceptance of alternative hypotheses. The rejection of null hypothesis indicates that the relationship between variables (e.g., sales and advertisement expenditure) or the difference between means (e.g., wages of skilled workers in town 1 and town 2) or the difference between proportions have statistical significance and the acceptance of the null hypotheses indicates that these differences are due to chance. Alternative hypothesis is denoted by H_1 .

Concepts in Testing of Hypothesis

- ▶ **One-tailed and two-tailed tests:** A test is called one-sided (or one-tailed) only if the null hypothesis gets rejected when a value of the test statistic falls in one specified tail of the distribution. Further, the test is called two-sided (or two-tailed) if null hypothesis gets rejected when a value of the test statistic falls in either one or the other of the two tails of its sampling distribution.
- ▶ **Type I and type II error:** if the hypothesis H_0 is rejected when it is actually true, the researcher is committing what is called a type I error. The probability of committing a type I error is denoted by alpha (α). This is termed as the level of significance. Similarly, if the null hypothesis H_0 when false is accepted, the researcher is committing an error called Type II error. The probability of committing a type II error is denoted by beta (β). The expression $1 - \beta$ is called power of test.

Steps in Testing of Hypothesis Exercise

- ▶ Setting up of a hypothesis
- ▶ Setting up of a suitable significance level
- ▶ Determination of a test statistic
- ▶ Determination of critical region
- ▶ Computing the value of test-statistic
- ▶ Making decision

Test statistic for testing hypothesis about population mean

The table below summarizes the test statistic for testing hypothesis about population mean.

Sample Size	Knowledge of Population Standard Deviation (σ)	
	Known	Not Known
Large ($n > 30$)	Z	Z
Small ($n \leq 30$)	Z	t

Test Concerning Means – Case of Single Population

Case of large sample – In case the sample size n is large or small but the value of the population standard deviation is known, a Z test is appropriate. The test statistic is given by

$$Z = \frac{\bar{X} - \mu_{H_0}}{\frac{\sigma}{\sqrt{n}}}$$

Where,

\bar{X} = Sample mean

σ = Population standard deviation

μ_{H_0} = The value of μ under the assumption that the null hypothesis is true.

n = Size of sample.

Test Concerning Means – Case of Single Population

If the population standard deviation σ is unknown, the sample standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum (X - \bar{X})^2}$$

is used as an estimate of σ . There can be alternate cases of two-tailed and one-tailed tests of hypotheses. Corresponding to the null hypothesis $H_0 : \mu = \mu_0$, the following criteria could be formulated as shown in the table below.

S.No.	Alternative Hypothesis	Reject the Null Hypothesis if	Accept the Null Hypothesis if
1.	$\mu < \mu_0$	$Z < -Z_\alpha$	$Z \geq -Z_\alpha$
2.	$\mu > \mu_0$	$Z > Z_\alpha$	$Z \leq Z_\alpha$
3.	$\mu \neq \mu_0$	$Z < -Z_{\alpha/2}$ Or $Z > Z_{\alpha/2}$	$-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$

Test Concerning Means – Case of Single Population

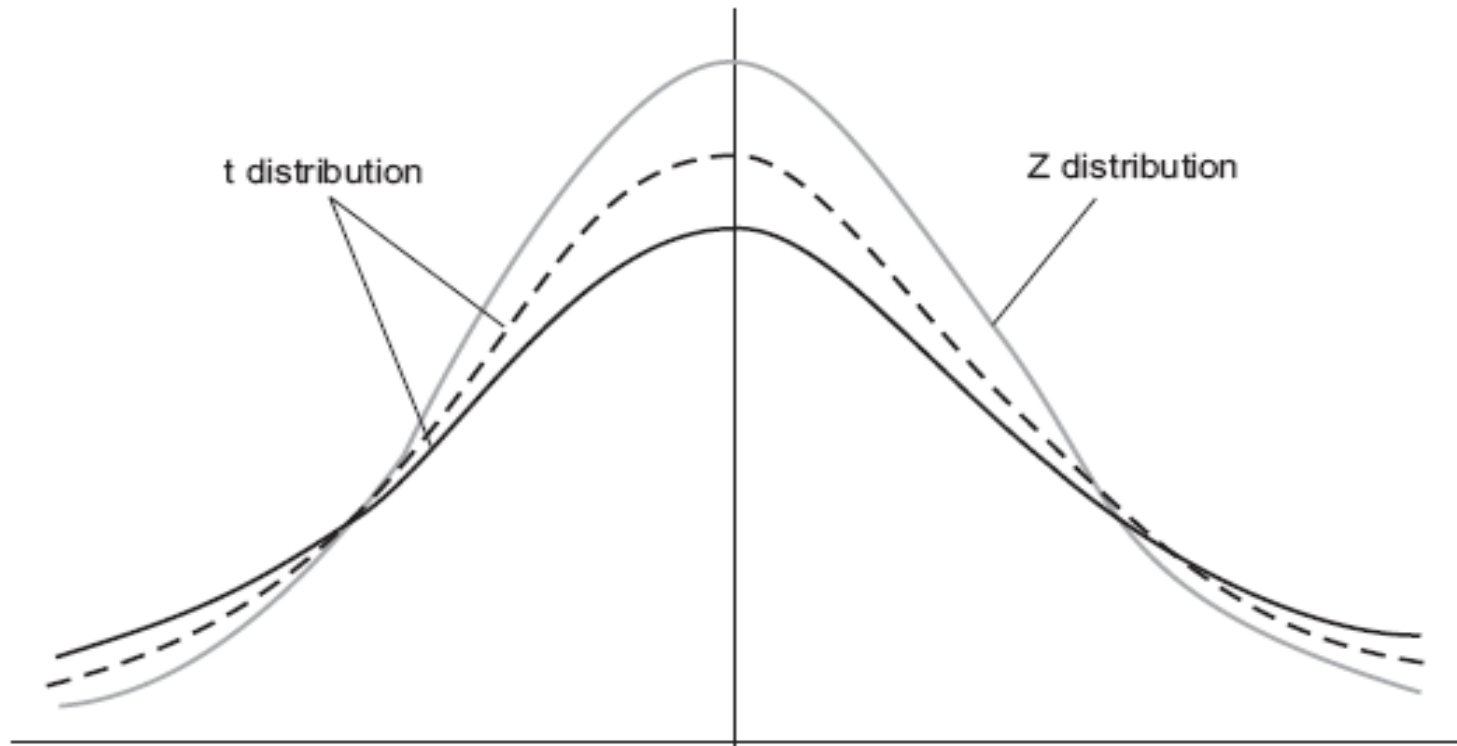
- ▶ It may be noted that Z_α and $Z_{\alpha/2}$ are Z values such that the area to the right under the standard normal distribution is α and $\alpha/2$ respectively.

Case of small sample:

- ▶ In case the sample size is small ($n \leq 30$) and is drawn from a population having a normal population with unknown standard deviation σ , a t test is used to conduct the hypothesis for the test of mean.
- ▶ The t distribution is a symmetrical distribution just like the normal one.
- ▶ However, t distribution is higher at the tail and lower at the peak. The t distribution is flatter than the normal distribution.
- ▶ With an increase in the sample size (and hence degrees of freedom), t distribution loses its flatness and approaches the normal distribution whenever $n > 30$.

Test Concerning Means – Case of Single Population

- ▶ A comparative shape of t and normal distribution is given in the figure below:



Test Concerning Means – Case of Single Population

The null hypothesis to be tested is:

$$H_0 : \mu = \mu_0$$

The alternative hypothesis could be one-tailed or two-tailed test. The test statistics used in this case is:

$$t_{n-1} = \frac{\bar{X} - \mu_{H0}}{\frac{\hat{\sigma}}{\bar{X}}}$$

$$\text{Where, } \frac{\hat{\sigma}}{\bar{X}} = \frac{s}{\sqrt{n}} \quad (\text{where } s = \text{Sample standard deviation})$$

$n-1$ = degrees of freedom

The procedure for testing the hypothesis of a mean is identical to the case of large sample.

Tests for Difference Between Two Population Means

Case of large sample – In case both the sample sizes are greater than 30, a Z test is used. The hypothesis to be tested may be written as:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where,

μ_1 = mean of population 1

μ_2 = mean of population 2

The above is a case of two-tailed test. The test statistic used is:

Tests for Difference Between Two Population Means

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)H_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

\bar{X}_1 = Mean of sample drawn from population 1

\bar{X}_2 = Mean of sample drawn from population 2

n_1 = size of sample drawn from population 1

n_2 = size of sample drawn from population 2

If σ_1 and σ_2 are unknown, their estimates given by $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are used.

$$\hat{\sigma}_1 = s_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2}$$

$$\hat{\sigma}_2 = s_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}$$

The Z value for the problem can be computed using the above formula and compared with the table value to either accept or reject the hypothesis.

Tests for Difference Between Two Population Means

Case of small sample – If the size of both the samples is less than 30 and the population standard deviation is unknown, the procedure described above to discuss the equality of two population means is not applicable in the sense that a t test would be applicable under the assumptions:

- a) Two population variances are equal.
- b) Two population variances are not equal.

Tests for Difference Between Two Population Means

Population variances are equal – If the two population variances are equal, it implies that their respective unbiased estimates are also equal. In such a case, the expression becomes:

$$\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}} = \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

(Assuming $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}^2$)

To get an estimate of σ^2 , a weighted average of s_1^2 and s_2^2 is used, where the weights are the number of degrees of freedom of each sample. The weighted average is called a ‘pooled estimate’ of σ^2 . This pooled estimate is given by the expression:

$$\hat{\sigma}^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

Tests for Difference Between Two Population Means

The testing procedure could be explained as under:

$$H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 \neq \mu_2 \Rightarrow \mu_1 - \mu_2 \neq 0$$

In this case, the test statistic t is given by the expression:

$$t_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) H_0}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{Where } \hat{\sigma} = \sqrt{\frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}$$

Once the value of t statistic is computed from the sample data, it is compared with the tabulated value at a level of significance α to arrive at a decision regarding the acceptance or rejection of hypothesis.

Tests for Difference Between Two Population Means

Population variances are not equal – In case population variances are not equal, the test statistic for testing the equality of two population means when the size of samples are small is given by:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)H_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

The degrees of freedom in such a case is given by the expression:

$$\text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

The procedure for testing of hypothesis remains the same as was discussed when the variances of two populations were assumed to be same.

Tests Concerning Population Proportion

The case of single population proportion – Suppose we want to test the hypothesis,

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

For large sample, the appropriate test statistic would be:

$$Z = \frac{\bar{p} - P_{H_0}}{\frac{\sigma}{\bar{p}}}$$

Where, \bar{p} = sample proportion

P_{H_0} = the value of p under the assumption that null hypothesis is true

$\frac{\sigma}{\bar{p}}$ = Standard error of sample proportion

Tests Concerning Population Proportion

The value of $\sigma_{\bar{p}}$ is computed by using the following formula:

$$\sigma_{\bar{p}} = \sqrt{\frac{P_{H_0} Q_{H_0}}{n}}$$

Where, $Q_{H_0} = 1 - P_{H_0}$
 $n =$ Sample size

For a given level of significance α , the computed value of Z is compared with the corresponding critical values, i.e. $Z_{\alpha/2}$ or $-Z_{\alpha/2}$ to accept or reject the null hypothesis.

Tests Concerning Population Proportion

Two Population Proportions – Here, the interest is to test whether the two population proportions are equal or not. The hypothesis under investigation is:

$$H_0 : p_1 = p_2 \Rightarrow p_1 - p_2 = 0$$

$$H_1 : p_1 \neq p_2 \Rightarrow p_1 - p_2 \neq 0$$

The alternative hypothesis assumed is two sided. It could as well have been one sided. The test statistic is given by:

$$Z = \frac{\bar{p}_1 - \bar{p}_2 - (p_1 - p_2) H_0}{\sigma_{\bar{p}_1 - \bar{p}_2}}$$

Tests Concerning Population Proportion

Where, \bar{p}_1 = Sample proportion possessing a particular attribute from population 1

\bar{p}_2 = Sample proportion possessing a particular attribute from population 2

$\sigma_{\bar{p}_1 - \bar{p}_2}$ = Standard error of difference between proportions.

$(p_1 - p_2)_{H_0}$ = Value of difference between population proportion under the assumption that the null hypothesis is true.

The formula for $\sigma_{\bar{p}_1 - \bar{p}_2}$ is given by:

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Tests Concerning Population Proportion

We do not know the value of p_1, p_2 , etc., but under the null hypothesis $p_1 = p_2 = p$.

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The best estimate of p is given by:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Where, x_1 = Number of successes in sample 1
 x_2 = Number of successes in sample 2
 n_1 = Size of sample taken from population 1
 n_2 = Size of sample taken from population 2

Tests Concerning Population Proportion

It is known that $\bar{p}_1 = \frac{x_1}{n_1}$ and $\bar{p}_2 = \frac{x_2}{n_2}$. Therefore $x_1 = n_1\bar{p}_1$, and $x_2 = n_2\bar{p}_2$.

Therefore,
$$\hat{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$$

Therefore, the estimate of standard error of difference between the two proportions is given by:

$$\hat{\sigma}_{\bar{p}_1 - \bar{p}_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where \hat{p} is as defined above and $\hat{q} = 1 - \hat{p}$. Now, the test statistic may be rewritten as:

Tests Concerning Population Proportion

$$Z = \frac{\bar{p}_1 - \bar{p}_2 - (p_1 - p_2)_{H_0}}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Now, for a given level of significance α , the sample Z value is compared with the critical Z value to accept or reject the null hypothesis.

Testing the significance of the correlation coefficient

The hypothesis to be tested is mentioned below:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Test statistic is given by,

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where,

ρ = Population correlation coefficient between the variables X and Y

r = Sample correlation coefficient between the variables X and Y

$n - 2$ = The degrees of freedom

Now for a given level of significance, if computed $|t|$ is greater than tabulated $|t|$ with $n - 2$ degrees of freedom, the null hypothesis of no correlation between X and Y is rejected.

Test of Significance of Regression Parameters

The hypothesis to be tested for the slope coefficient is mentioned below as:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

The test statistic to be used to test the significance of the slope coefficient is given by:

$$t_{n-k} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

$$V(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum(X - \bar{X})^2}$$

$$SE(\hat{\beta}) = \frac{\hat{\sigma}_u}{\sqrt{\sum(X - \bar{X})^2}}$$

Where, $\hat{\beta}$ = estimated value of beta (β)

$SE(\hat{\beta})$ = standard error of estimate of β

At a given level of significance, computed t is compared with absolute t to accept or reject the null hypothesis.

Goodness of fit of regression equation

- ▶ r^2 is used to measure the goodness of fit of regression equation. This measure is also called the coefficient of determination of a regression equation and it takes value between 0 and 1. Higher the value of r^2 , higher is the goodness of fit.

$$\begin{aligned}r^2 &= 1 - \frac{\sum \hat{U}^2}{\sum (Y - \bar{Y})^2} \\ &= r_{xy}^2 \\ &= r_{y\hat{y}}^2 \\ &= \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}\end{aligned}$$

Testing the Significance of r^2

The hypothesis to be tested is:

$$H_0 : r^2 = 0$$

$$H_1 : r^2 > 0$$

The test statistic F is given by the expression:

$$F = \frac{r^2 / (k-1)}{(1-r)^2 / (n-k)}$$

For a given level of significance α , the computed value of the F statistic is compared with the tabulated value of F with $k - 1$ degrees of freedom in the numerator and $n - k$ degrees of freedom in the denominator. If the computed F exceeds the tabulated F, the null hypothesis is rejected in favour of the alternative hypothesis.

Multiple Regression Model

In the multiple regression model, there are at least two independent variables. The linear multiple regression model with two independent variables would look like:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + U$$

b_0 , b_1 and b_2 are the parameters to be estimated.

The estimation is carried out using the OLS method which results in the following:

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2$$

Where,

$$x_1 = X_1 - \bar{X}_1$$

$$x_2 = X_2 - \bar{X}_2$$

Multiple Regression Model

- ▶ In case of multiple regression model, we have the concept of the multiple correlation squared given by $R^2_{Y.X_1X_2}$ which indicates the explanatory power of the model. The various formulae for R^2 are given as under:

$$R^2_{Y.X_1X_2} = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = 1 - \frac{\sum \hat{u}^2}{\sum y^2} = \frac{\sum y^2 - \sum \hat{u}^2}{\sum y^2}$$
$$= \frac{\hat{b}_1 \sum yx_1 + \hat{b}_2 \sum yx_2}{\sum y^2} = (r_{Y\hat{Y}})^2$$

Where, $y = Y - \bar{Y}$

The test of significance of the individual parameters is conducted using the t statistic. To be able to use the t statistic we need the estimates of the variance of the estimated coefficients of the regression equation.

Multiple Regression Model

- ▶ The estimates of the variance of estimated coefficients are presented below:

$$\text{var}(\hat{b}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}_1^2 \sum x_2^2 + \bar{X}_2^2 \sum x_1^2 - 2\bar{X}_1\bar{X}_2 \sum x_1x_2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \right]$$

$$\text{var}(\hat{b}_1) = \hat{\sigma}^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}$$

$$\text{var}(\hat{b}_2) = \hat{\sigma}^2 \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}$$

Where,

$$\hat{\sigma}^2 = \frac{\sum \hat{u}^2}{n - k}$$
$$\hat{u} = Y - \hat{Y}$$

Multiple Regression Model

Let us assume that we want to test the significance of the slope coefficient of the variable X_1 . We can write the null and alternative hypothesis as:

$$H_0 : b_1 = 0$$

$$H_1 : b_1 \neq 0$$

The test statistic may be written as:

$$t = \frac{\hat{b}_1 - b_{1H_0}}{\sqrt{V(\hat{b}_1)}}$$

The value of the test statistic t is computed and compared with the table value of t for a given level of significance. If the computed value of $|t|$ is greater than table value of $|t|$, we reject H_0 in favour of the alternative hypothesis H_1 .

Testing the Significance of R^2

- ▶ The test for the significance of R^2 is carried out using the F statistic, which is already explained in the case of the two variable linear regression model. The hypothesis to be tested is listed as under:

$$H_0 : b_0 = b_1 = b_2 = 0 \Rightarrow R^2 = 0$$

$$H_1 : \text{All } b\text{'s are not zero} \Rightarrow R^2 > 0$$

- ▶ If R^2 is equal to 0 that means all the coefficients are equal to zero since none of the independent variables would explain any variations in Y.
- ▶ The test for the significance of R^2 is shown through the analysis of variance (ANOVA) table.

Testing the Significance of R^2

Source	Sum of Squares	d.f.	Mean Square	F
Due to Regression	$R^2 \Sigma y^2$	$K - 1$	$\frac{R^2 \Sigma y^2}{K - 1}$	$\frac{R^2 (n - K)}{(1 - R^2)(K - 1)}$
Due to Residual	$(1 - R^2) \Sigma y^2$	$n - K$	$\frac{(1 - R^2) \Sigma y^2}{n - K}$	
Total	Σy^2	$n - 1$		

Dummy Variables in Regression Analysis

- ▶ In regression analysis, the dependent variable is generally metric in nature and it is most often influenced by other metric variables.
- ▶ However, there could be situations where the dependent variable may be influenced by the qualitative variables like gender, marital status, profession, geographical region, colour, or religion.
- ▶ The question arises how to quantify qualitative variables.
- ▶ In situations like this, the dummy variables come to our rescue. They are used to quantify the qualitative variables.
- ▶ The number of dummy variables required in the regression model is equal to the number of categories of data less one.
- ▶ Dummy variables usually assume two values 0 and 1.

Example of a Dummy Variable Regression

Suppose the starting salary of a college lecturer is influenced not only by years of teaching experience but also by gender. Therefore, the model could be specified as:

$$Y = f(X, D)$$

Where,

Y = Starting salary of a college lecturer in thousands ` per month

X = No. of years of work experience

D is a dummy variable which takes values

D = 1 (if the respondent is a male)

= 0 (if the respondent is a female)

The model could be written as,

$$Y = \alpha + \beta X + \gamma D + U$$

Example of a Dummy Variable Regression

This can be estimated by using ordinary least squares (OLS) techniques. Suppose the estimated regression equation looks like:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\gamma}D$$

Now, for the male respondents, the salary equation would look like:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\gamma}$$

$$\hat{Y} = (\hat{\alpha} + \hat{\gamma}) + \hat{\beta}X$$

For the female respondents, the salary equation would look like:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

The above two equations differ by the amount $\hat{\gamma}$. It is known that $\hat{\gamma}$ can be positive or negative. If $\hat{\gamma}$ is positive it would imply that the average salary of a male lecturer is more than that of a female lecturer by the amount $\hat{\gamma}$ while keeping the number of years of experience constant.

Difference between Parametric & Non-parametric Tests

	Parametric Tests	Non-Parametric Tests
Assumptions:	Normality assumption is required.	Normality assumption is not required.
	Uses the metric data.	Ordinal or interval scale data is used.
	Can be applied for both small and large samples.	Can be applied for small samples.
Applications:	One sample using Z or t statistics.	One sample using the sign test.
	Two independent samples using a t or z test.	Two independent samples using the Mann-Whitney U statistics.
	Two paired samples using a t or z test.	Two paired samples using the sign test and Wilcoxon matched pair rank test.
	Randomness – no test in parametric is available.	Randomness – using runs test.
	Several independent samples using F test in ANOVA.	Several independent samples using Kruskal-Wallis test.

Types of Non-Parametric Tests

Chi-square Tests – For the use of a chi-square test, the data is required in the form of frequencies. The majority of the applications of chi-square are with the discrete data. The test could also be applied to continuous data, provided it is reduced to certain categories and tabulated in such a way that the chi-square may be applied. Some of the important properties of the chi-square distribution are:

- ▶ Unlike the normal and t distribution, the chi-square distribution is not symmetric.
- ▶ The values of a chi-square are greater than or equal to zero.
- ▶ The shape of a chi-square distribution depends upon the degrees of freedom. With the increase in degrees of freedom, the distribution tends to normal

χ^2 -Test

Chi-square test is suitable for analyzing nominal and ordinal data. (*Interval and ratio data should be grouped first*)

Chi-square test is used for

- Goodness-of-fit (*1-Way classification; 1-DV, 1-IV*)
- Test for independence (*2-Way classification; 1-DV, 2+IV*)

Categorical data in Rows

Ordinal data in Columns

χ^2 -Test Assumptions

- Categorical data
- Any cell raw frequency > 5
- Random Sampling

Applications of Chi-square

1. Chi-square test for the goodness of fit
2. Chi-square test for the independence of variables
3. Chi-square test for the equality of more than two population proportions.

Common principles of all the chi-square tests are as under:

- ▶ State the null and the alternative hypothesis about a population.
- ▶ Specify a level of significance.
- ▶ Compute the expected frequencies of the occurrence of certain events under the assumption that the null hypothesis is true.
- ▶ Make a note of the observed counts of the data points falling in different cells
- ▶ Compute the chi-square value given by the formula.

Applications of Chi-square

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Where,

O_i = Observed frequency of i^{th} cell

E_i = Expected frequency of i^{th} cell

k = Total number of cells

$k-1$ = degrees of freedom

Compare the sample value of the statistic as obtained in previous step with the critical value at a given level of significance and make the decision.

Applications of Chi-square

Chi-square test for goodness of fit

▶ The hypothesis to be tested in this case is:

H_0 : Probabilities of the occurrence of events E_1, E_2, \dots, E_k are given by the specified probabilities p_1, p_2, \dots, p_k

H_1 : Probabilities of the k events are not the p_i stated in the null hypothesis.

The procedure has already been explained.

Applications of Chi-square

Chi-square test for independence of variables

The chi-square test can be used to test the independence of two variables each having at least two categories. The test makes a use of contingency tables also referred to as cross-tabs with the cells corresponding to a cross classification of attributes or events. A contingency table with three rows and four columns (as an example) is as shown below.

Second Classification Category	First Classification Category				Total
	1	2	3	4	
1	O_{11}	O_{12}	O_{13}	O_{14}	R_1
2	O_{21}	O_{22}	O_{23}	O_{24}	R_2
3	O_{31}	O_{32}	O_{33}	O_{34}	R_3
Total	C_1	C_2	C_3	C_4	n

Applications of Chi-square

Assuming that there are r rows and c columns, the count in the cell corresponding to the i^{th} row and the j^{th} column is denoted by O_{ij} , where $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$. The total for row i is denoted by R_i whereas that corresponding to column j is denoted by C_j . The total sample size is given by n , which is also the sum of all the r row totals or the sum of all the c column totals.

The hypothesis test for independence is:

H_0 : Row and column variables are independent of each other.

H_1 : Row and column variables are not independent.

The hypothesis is tested using a chi-square test statistic for independence given by:

Applications of Chi-square

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The degrees of freedom for the chi-square statistic are given by $(r - 1)(c - 1)$.

The expected frequency in the cell corresponding to the i^{th} row and the j^{th} column is given by:

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Where, R_i = Total for the i^{th} row
 C_j = Total for the j^{th} column
 n = Total sample size.

For a given level of significance α , the sample value of the chi-square is compared with the critical value for the degree of freedom $(r - 1)(c - 1)$ to make a decision.

Applications of Chi-square

Chi-square test for the equality of more than two population proportions

The analysis is carried out exactly in the same way as was done for the other two cases. The formula for a chi-square analysis remains the same. However, two important assumptions here are different.

- (i) We identify our population (e.g., age groups or various class employees) and the sample directly from these populations.
- (ii) As we identify the populations of interest and the sample from them directly, the sizes of the sample from different populations of interest are fixed. This is also called a chi-square analysis with fixed marginal totals. The hypothesis to be tested is as under:

H₀ : The proportion of people satisfying a particular characteristic is the same in population.

H₁ : The proportion of people satisfying a particular characteristic is not the same in all populations.

The expected frequency for each cell could also be obtained by using the formula as explained early. The decision procedure remains the same.

Applications of Chi-square

Examining strength of relationship between two nominal scale variables

1. **Contingency coefficient** – Applicable when number of rows equal the number of columns in a contingency table.

The value of the contingency coefficient is given by:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

The lower limit of C equals zero when χ^2 is zero. The upper limit of C when the number of rows is equal to the number of columns is given by the expression:

$$\sqrt{\frac{r-1}{r}}$$

Where, r = number of rows

Applications of Chi-square

2. **Phi coefficient (Φ)** – Can be applied when the number of rows and columns in a contingency table are two. The phi-coefficient like the correlation coefficient can assume any value between -1 and 1. In a 2x2 table given below phi coefficient can be computed as:

	Column 1	Column 2	Total
Row 1	a	b	a + b
Row 2	c	d	c + d
Total	a + c	b + d	(a+b+c+d)

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

Applications of Chi-square

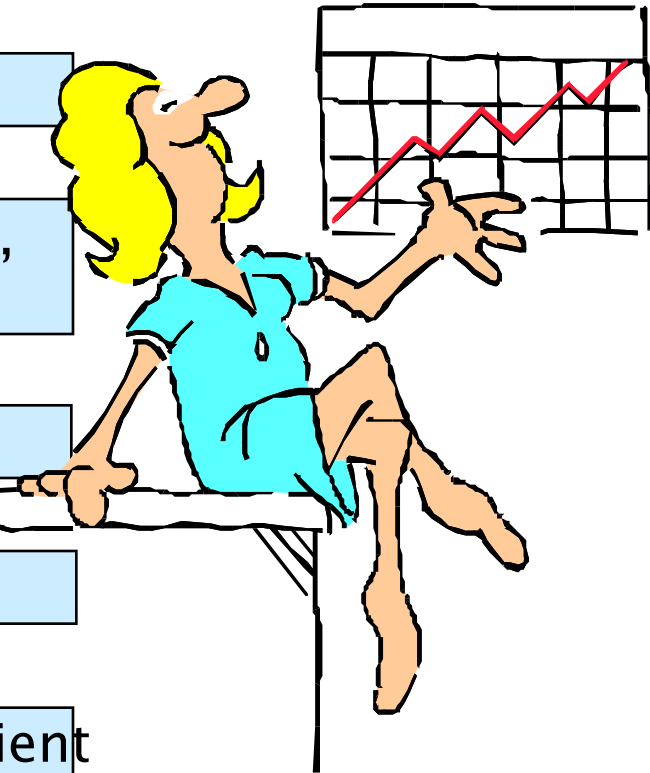
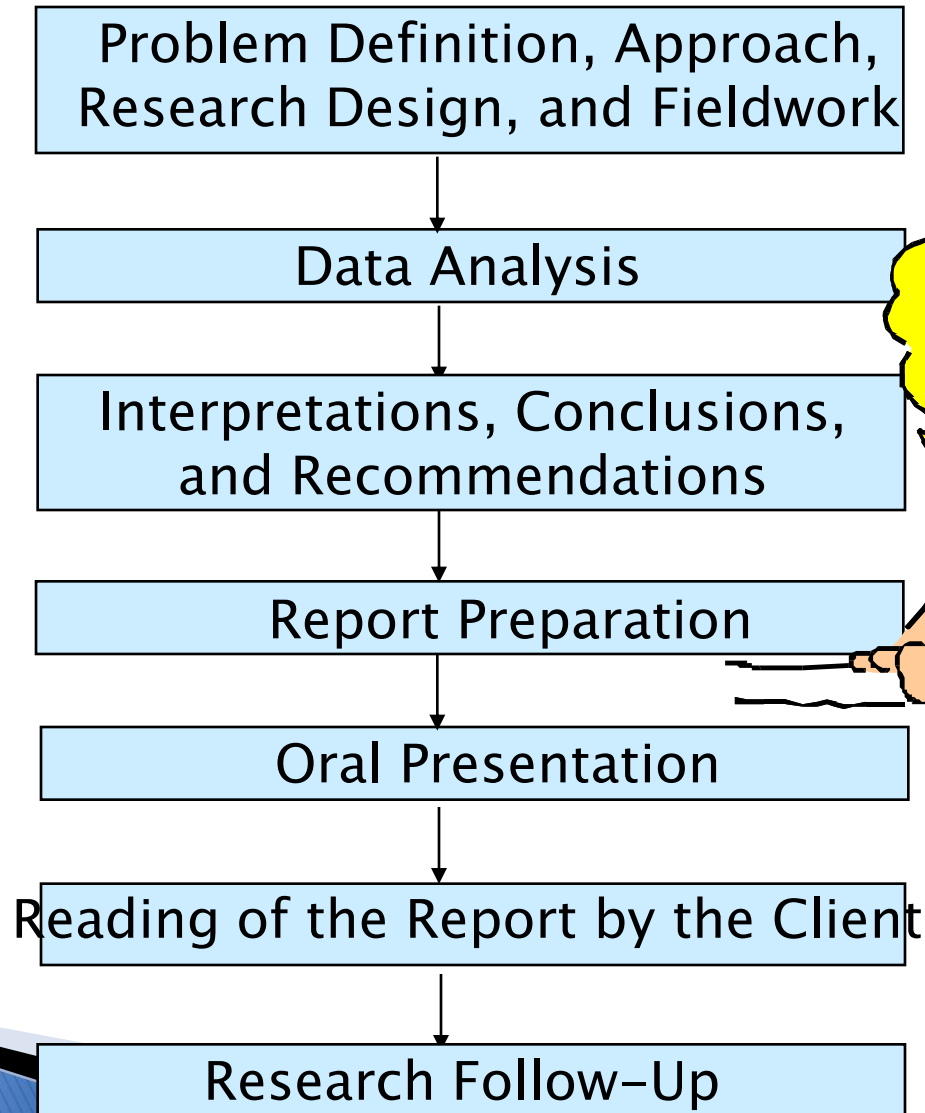
3. **Cramer's V Statistic** – To be used when number of rows are not equal to number of columns in a contingency table.

$$V = \sqrt{\frac{\chi^2}{n(f-1)}}$$

Minimum value of V equals zero when chi-square is equal to zero. The maximum value of chi-square equals $n(f-1)$ and in that case the maximum value of V equals 1.

The Report Preparation and Presentation Process

Fig. 22.1



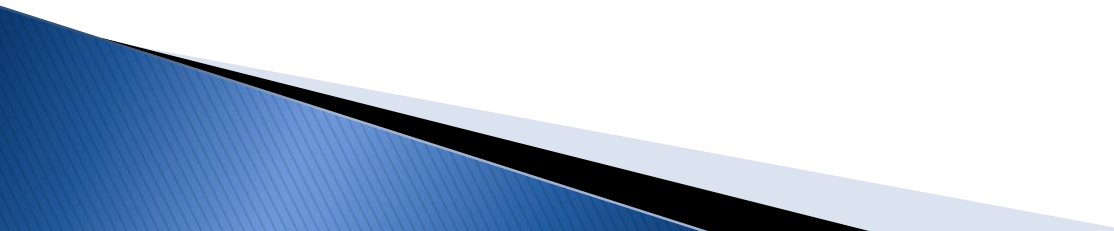


Report Format

- I. Title page
- II. Letter of transmittal
- III. Letter of authorization
- IV. Table of contents
- V. List of tables
- VI. List of graphs
- VII. List of appendices
- VIII. List of exhibits
- IX. Executive summary
 - a. Major findings
 - b. Conclusions
 - c. Recommendations



Report Format

- X. Problem definition
 - a. Background to the problem
 - b. Statement of the problem
 - XI. Approach to the problem
 - XII. Research design
 - a. Type of research design
 - b. Information needs
 - c. Data collection from secondary sources
 - d. Data collection from primary sources
 - e. Scaling techniques
 - f. Questionnaire development and pretesting
 - g. Sampling techniques
 - h. Fieldwork
- 

Report structure

