UNIT - IV

Data Analysis & Interpretation

ANALYSIS and **INTERPRETATION** provide answers to the research questions postulated in the study.

ANALYSIS means the ordering, manipulating, and summarizing of data to obtain answers to research questions. Its purpose is to reduce data to intelligible and interpretable form so that the relations of research problems can be studied and tested.

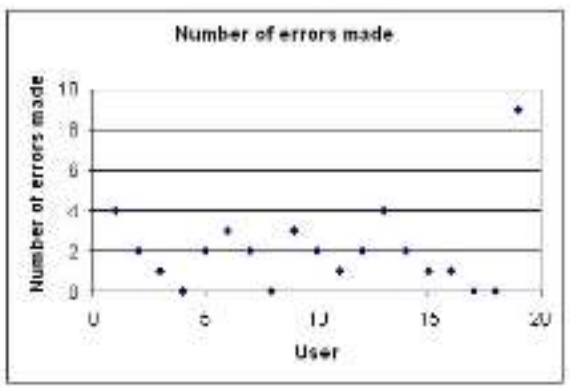
INTERPRETATION gives the results of analysis, makes inferences pertinent to the research relations studied, and draws conclusions about these relations.

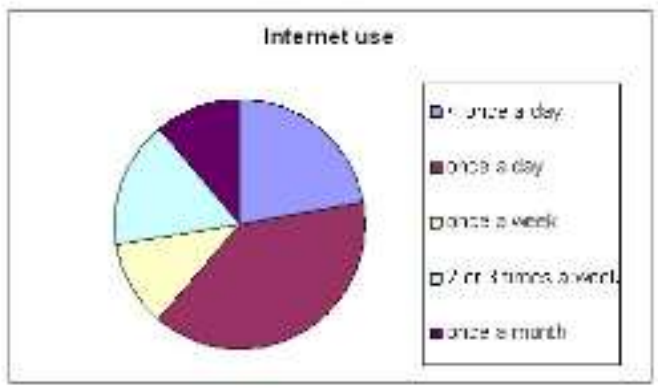
Quantitative and qualitative

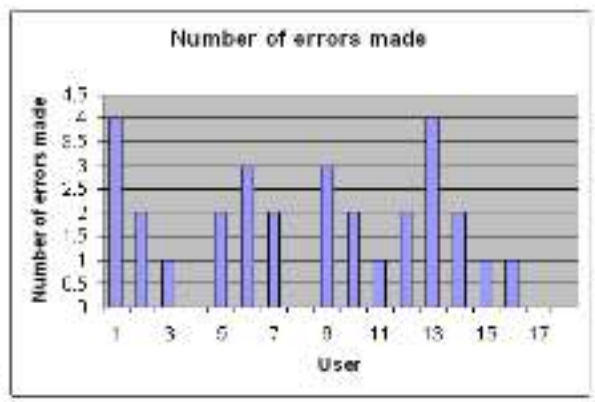
- Quantitative data – expressed as numbers
- Qualitative data – difficult to measure sensibly as numbers, e.g. count number of words to measure dissatisfaction
- Quantitative analysis – numerical methods to ascertain size, magnitude, amount
- Qualitative analysis – expresses the nature of elements and is represented as themes, patterns, stories
- Be careful how you manipulate data and numbers!

Simple quantitative analysis

- Averages
  - Mean: add up values and divide by number of data points
  - Median: middle value of data when ranked
  - Mode: figure that appears most often in the data
- Percentages
- Be careful not to mislead with numbers!
- Graphical representations give overview of data







Simple qualitative analysis

- Recurring patterns or themes
  - Emergent from data, dependent on observation framework if used
- Categorizing data
  - Categorization scheme may be emergent or pre-specified
- Looking for critical incidents
  - Helps to focus in on key events

Figure 8.8 Building the affinity diagram of Indian ATM usage

Source: Figure 1, A. DeAngeli, U. Athavamker, A. Joshi, L. Coventry and G.I. Johnson (2004) "Introducing ATMs in India: a contextual inquiry", *Interacting with Computers* 16(1), 29–44. Reproduced with permission.

STATISTICS is simply a tool in research. In fact, according to Leedy (1974:21), statistics is a language which, through its own special symbols and grammar, takes the numerical facts of life and translates them meaningfully.

Statistics thus gathers numerical data. The variations of the data gathered are abstracted based on group characteristics and combined to serve the purpose of description, analysis, interpretation, and possible generalization. According to McGuigan (1987), in research, this is known as the process of **concatenation** where the statements are "chained together" with other statements.

There are two kinds of statistics:

- **DESCRIPTIVE STATISTICS** – allows the researcher to describe the population or sample used in the study.

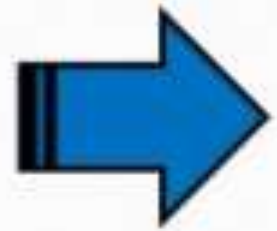
- **INFERENTIAL STATISTICS** – draws inferences from sample data and actually deals with answering the research questions postulated which are, in some cases, cause and effect relationship.

DESCRIPTIVE STATISTICS

❖ describes the characteristics of the population or the sample. To make them meaningful, they are grouped according to the following measures:

- ✓ Measures of Central Tendency or Averages
- ✓ Measures of Dispersion or Variability
- ✓ Measures of Noncentral Location
- ✓ Measures of Symmetry and/or Asymmetry
- ✓ Measures of Peakedness or Flatness

Measures of Central Tendency or Averages. These include the mean, mode and the median. The ***mean*** is a measure obtained by adding all the values in a population or sample and dividing by the number of values that are added (Daniel, 1991:19-20). The ***mode*** is simply the most frequent score of scores (Blalock, 1972:72). The ***median*** is the value above and below which one half of the observations fall (Norusis, (1984:B-63)



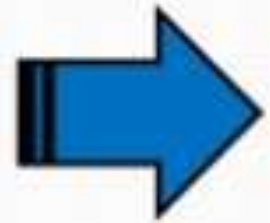
Measures of Dispersion or Variability. These include the variance, standard deviation, and the range. According to Daniel (1991:24), a measure of dispersion conveys information regarding the amount of variability present in a set of idea.

The **VARIANCE** is a measure of the dispersion of the set of scores. It tells us how much the scores are spread out. Thus, the variance is a measure of the spread of the scores; it describes the extent to which the scores differ from each other about their mean.

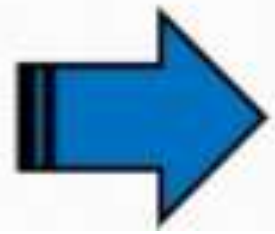
STANDARD DEVIATION thus refers to the deviation of scores from the mean. Where ordinal measures of 1-5 scales (low-high) are used, data most likely have standard deviations of less than one unless, the responses are extremes (i.e., all ones and all fives).

The **RANGE** is defined as the difference between the highest and lowest scores (Blalock, 1972:77)

Measures of Noncentral Location. These include the quantiles (i.e., percentiles, deciles, quartiles). The word *percent* means *"per hundred"*. Therefore, in using percentages, size is standardized by calculating the number of individuals who would be in a given category if the total number of cases were 100 and if the proportion in each category remains unchanged. Since proportions must add to unity, it is obvious that percentages will sum up to 100 unless the categories are not mutually exclusive or exhaustive (Blalock, 1972:33)



Measures of Symmetry and/or Asymmetry. These include the ***skewness*** of the frequency distribution. If a distribution is asymmetrical and the larger frequencies tend to be concentrated toward the low end of the variable and the smaller frequencies toward the high end, it is said to be ***positively skewed.*** If the opposite holds, the larger frequencies being concentrated toward the high end of the variable and the smaller frequencies toward the low end, the distribution is said to be ***negatively skewed*** (Ferguson and Takane, 1989:30).



Measures of Peakedness or Flatness of one distribution in relation to another is referred to as ***Kurtosis***. If one distribution is more peaked than another, it may be spoken of as more ***leptokurtic.*** If it is less peaked, it is said to be more ***platykurtic***.

Parametric Test

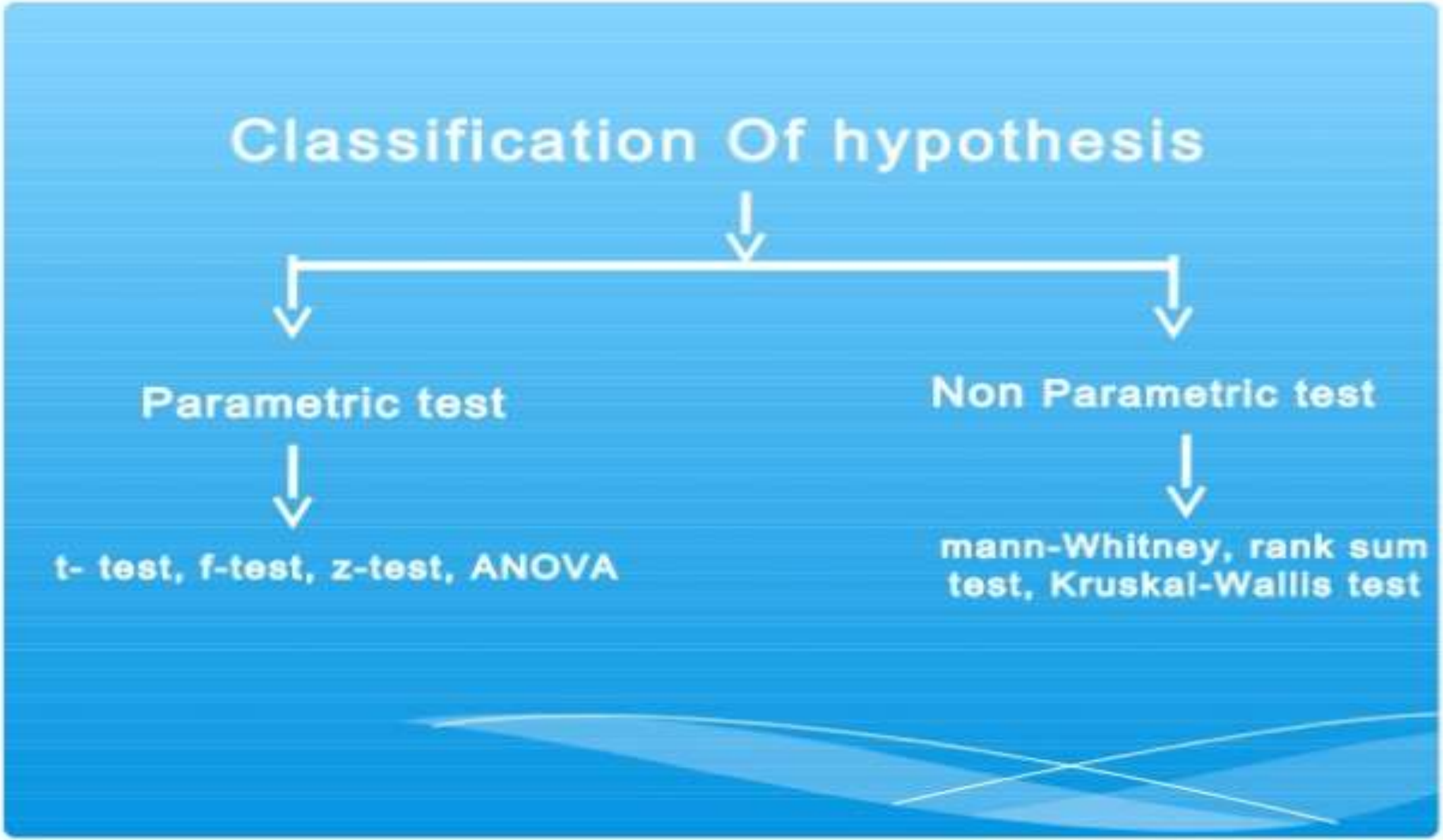
* If the information about the population is completely known by means of its parameters then statistical test is called parametric test

* Eg: t- test, f-test, z-test, ANOVA

Nonparametric test

* If there is no knowledge about the population or paramters, but still it is required to test the hypothesis of the population. Then it is called non-parametric test

* Eg: mann-Whitney, rank sum test, Kruskal-Wallis test

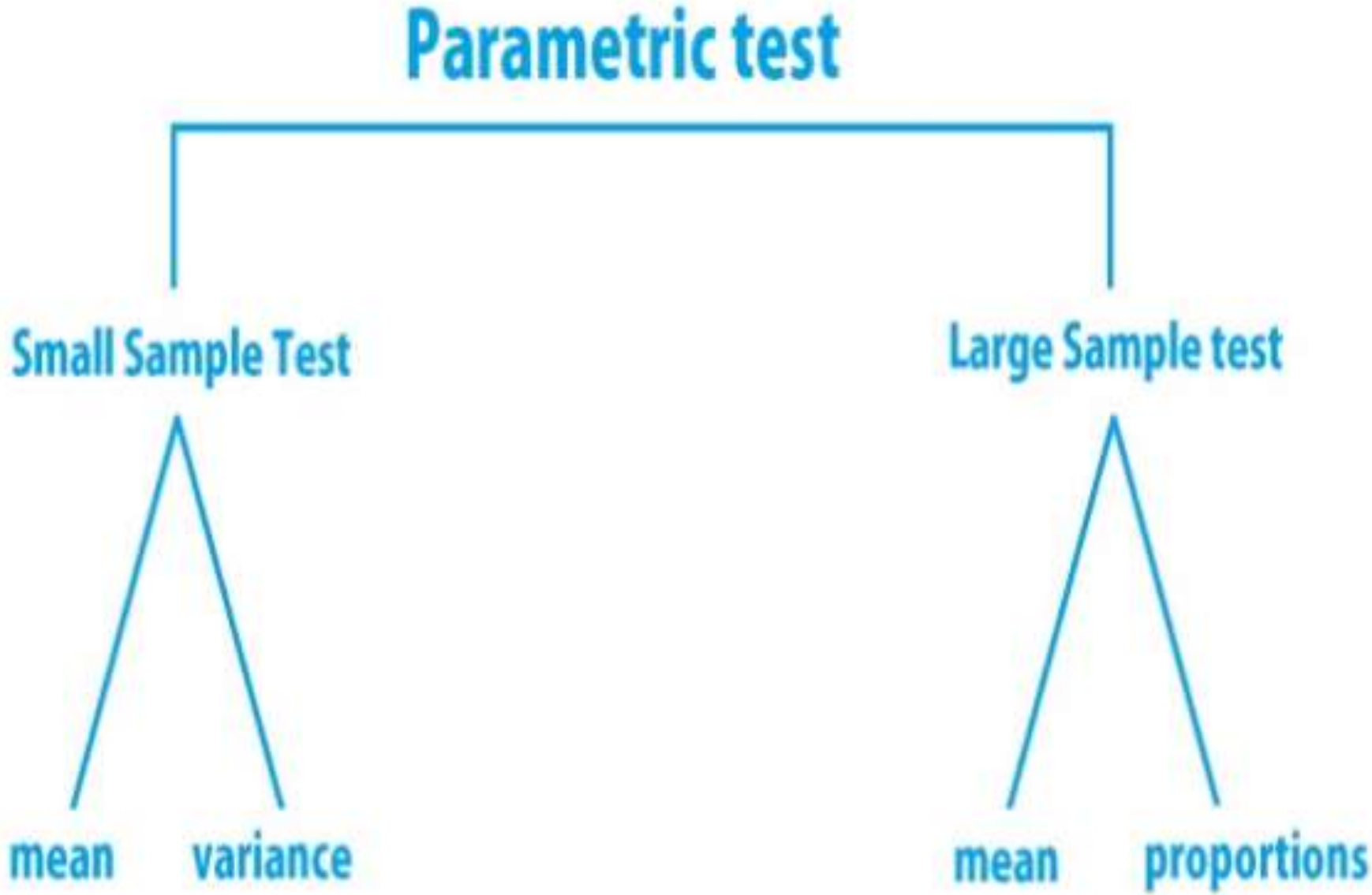

Classification Of hypothesis
Parametric test
t- test, f-test, z-test, ANOVA
Non Parametric test
mann-Whitney, rank sum test, Kruskal-Wallis test

Difference between parametric and Non parametric

Parametric	Non Parametric
Information about population is completely known	No information about the population is available
Specific assumptions are made regarding the population	No assumptions are made regarding the population
Null hypothesis is made on parameters of the population distribution	The null hypothesis is free from parameters

Difference between parametric and Non parametric

Parametric	Non Parametric
Test statistic is based on the distribution	Test statistic is arbritary
Parametric tests are applicable only for variable	It is applied both variable and artributes
No parametric test excist for Norminal scale data	Non parametric test do exist for norminal and ordinal scale data
Parametric test is powerful, if it exist	It is not so powerful like parametric test


Parametric test
Small Sample Test
Large Sample test
mean
variance
mean
proportions

NONPARAMETRIC TESTS

Two types of data are recognized in the application of statistical treatments, these are: ***parametric data and nonparametric data***.

Parametric data are measured data and parametric statistical tests assume that the data are normally, or nearly normally, distributed (Best and Kahn, 1998:338).

Nonparametric data are *distribution free* samples which implies that they are free, of independent of the population distribution (Ferguson and Takane, 1989:431).

The tests on these data do not rest on the more stringent assumption of normally distributed population (Best and Kahn, 1998:338).

APPROPRIATE STATISTICAL METHODS BASED ON THE RESEARCH PROBLEM AND LEVELS OF MEASUREMENT

❖ the statistical methods appropriate to any studies are always determined by the research problem and the measurement scale of the variables used in the study.

CHI - SQUARE

❖ the most commonly used nonparametric test. It is employed in instances where a comparison between observed and theoretical frequencies is present, or in testing the mathematical fit of a frequency curve to an observed frequency distribution.

T - TESTS

❖ provides the capability of computing student's t and probability levels for testing whether or not the difference between two samples means is significant (Nie, et al., 1975:267). This type of analysis is the comparison of two groups of subjects, with the group means as the basis for comparison.

Two types of T-tests may be performed:

❖***Independent Samples*** - cases are classified into 2 groups and a test of mean differences is performed for specified variables;

❖ ***Paired Samples*** – for paired observations arrange casewise, a test of treatment effects is performed.

CORRELATION ANALYSIS

❖ Correlation is used when one is interested to know the relationship between two or more paired variables. According to Blalock (1972:361), this is where interest is focused primarily on the exploratory task of finding out which variables are related to a given variable.

Nonparametric Methods

- Sign Test
- Wilcoxon Signed-Rank Test
- Mann-Whitney-Wilcoxon Test
- Kruskal-Wallis Test
- Rank Correlation

Nonparametric Methods

- Most of the statistical methods referred to as parametric require the use of interval- or ratio-scaled data.
- Nonparametric methods are often the only way to analyze nominal or ordinal data and draw statistical conclusions.
- Nonparametric methods require no assumptions about the population probability distributions.
- Nonparametric methods are often called distribution-free methods.

Nonparametric Methods

- In general, for a statistical method to be classified as nonparametric, it must satisfy at least one of the following conditions.
  - The method can be used with nominal data.
  - The method can be used with ordinal data.
  - The method can be used with interval or ratio data when no assumption can be made about the population probability distribution.

Sign Test

- A common application of the sign test involves using a sample of n potential customers to identify a preference for one of two brands of a product.
- The objective is to determine whether there is a difference in preference between the two items being compared.
- To record the preference data, we use a plus sign if the individual prefers one brand and a minus sign if the individual prefers the other brand.
- Because the data are recorded as plus and minus signs, this test is called the sign test.

Sign Test: Small-Sample Case

- The small-sample case for the sign test should be used whenever $n \leq 20$.
- The hypotheses are

$H_0 : p = .50$ No preference for one brand over the other exists.

$H_a : p \neq .50$ A preference for one brand over the other exists.

- The number of plus signs is our test statistic.
- Assuming H_0 is true, the sampling distribution for the test statistic is a binomial distribution with $p = .5$.
- H_0 is rejected if the p-value $\leq$ level of significance, α.

Sign Test: Large-Sample Case

- Using H_0: $p = .5$ and $n > 20$, the sampling distribution for the number of plus signs can be approximated by a normal distribution.
- When no preference is stated (H_0: $p = .5$), the sampling distribution will have:

$$\text{Mean: } \mu = .50n$$
$$\text{Standard Deviation: } \sigma = \sqrt{.25n}$$

- The test statistic is:

$$z = \frac{x - \mu}{\sigma}$$

(x is the number of plus signs)

- H_0 is rejected if the p-value $\leq$ level of significance, α.

Sign Test: Large-Sample Case

- Example: Ketchup Taste Test

As part of a market research study, a sample of 36 consumers were asked to taste two brands of ketchup and indicate a preference. Do the data shown on the next slide indicate a significant difference in the consumer preferences for the two brands?

Sign Test: Large-Sample Case

- Example: Ketchup Taste Test

18 preferred Brand A Ketchup (+ sign recorded)

12 preferred Brand B Ketchup (– sign recorded)

6 had no preference

The analysis will be based on a sample size of 18 + 12 = 30.

Sign Test: Large-Sample Case

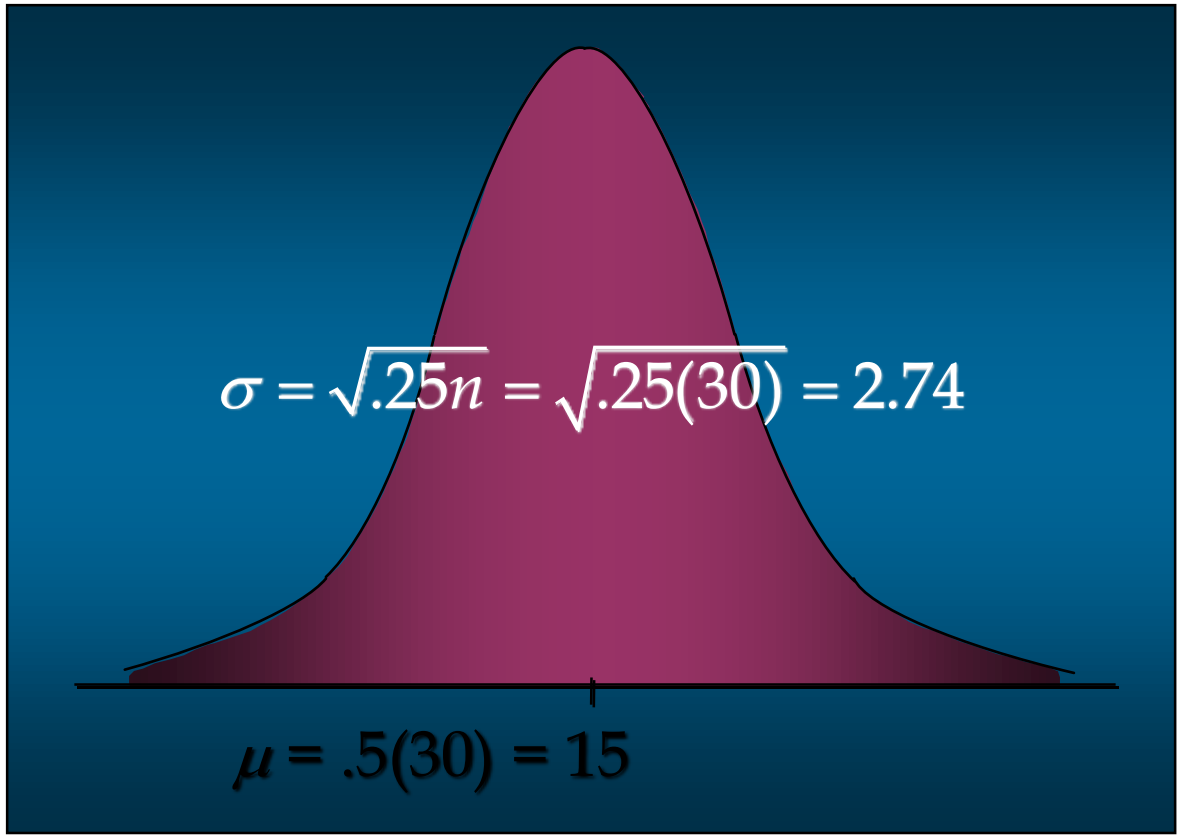
- Hypotheses
  - $H_0 : p = .50$

    No preference for one brand over the other exists
  - $H_a : p \neq .50$

    A preference for one brand over the other exists

Sign Test: Large-Sample Case

- Sampling Distribution for Number of Plus Signs



Sign Test: Large-Sample Case

- Rejection Rule

  Using .05 level of significance:

  Reject H_0 if p-value $\leq .05$

- Test Statistic

  $z = (x - \mu)/\sigma = (18 - 15)/2.74 = 3/2.74 = 1.10$

- p-Value

  p-Value $= 2(.5000 - .3643) = .2714$

Sign Test: Large-Sample Case

- Conclusion
  - Because the p-value > α, we cannot reject H_0. There is insufficient evidence in the sample to conclude that a difference in preference exists for the two brands of ketchup.

Rank Correlation

- The Pearson correlation coefficient, r, is a measure of the linear association between two variables for which interval or ratio data are available.
- The Spearman rank-correlation coefficient, r_s, is a measure of association between two variables when only ordinal data are available.
- Values of r_s can range from –1.0 to +1.0, where
  - values near 1.0 indicate a strong positive association between the rankings, and
  - values near -1.0 indicate a strong negative association between the rankings.

Rank Correlation

- Spearman Rank-Correlation Coefficient, r_s

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

where: n = number of items being ranked

x_i = rank of item i with respect to one variable

y_i = rank of item i with respect to a second variable

$d_i = x_i - y_i$

Test for Significant Rank Correlation

- We may want to use sample results to make an inference about the population rank correlation p_s.
- To do so, we must test the hypotheses:

$H_0 : p_s = 0$ (No rank correlation exists)

$H_a : p_s \neq 0$ (Rank correlation exists)

Rank Correlation

- Sampling Distribution of r_s when $p_s = 0$
  - Mean

$$\mu_{r_s} = 0$$

  - Standard Deviation

$$\sigma_{r_s} = \sqrt{\frac{1}{n-1}}$$

  - Distribution Form

Approximately normal, provided $n \geq 10$

Rank Correlation

- Example: Crennor Investors

Crennor Investors provides a portfolio management service for its clients. Two of Crennor's analysts ranked ten investments as shown on the next slide. Use rank correlation, with $\alpha = .10$, to comment on the agreement of the two analysts' rankings.

Rank Correlation

- Example: Crennor Investors

- Analysts' Rankings

Investment	A	B	C	D	E	F	G	H	I	J
Analyst #1	1	4	9	8	6	3	5	7	2	10
Analyst #2	1	5	6	2	9	7	3	10	4	8

- Hypotheses

$H_0 : p_s = 0$ (No rank correlation exists)

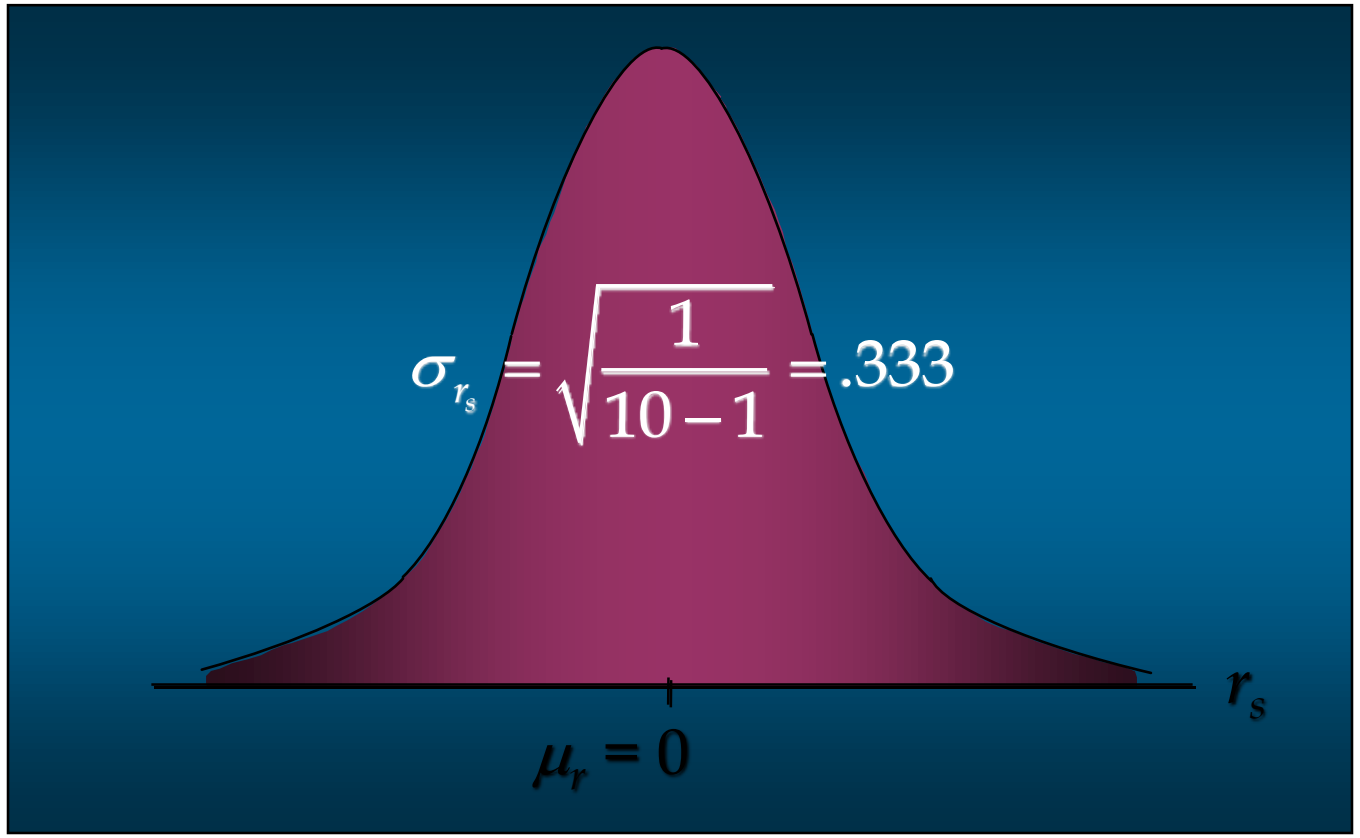
$H_a : p_s \neq 0$ (Rank correlation exists)

Rank Correlation

Investment	Analyst #1 Ranking	Analyst #2 Ranking	Differ.	(Differ.)^2
A	1	1	0	0
B	4	5	-1	1
C	9	6	3	9
D	8	2	6	36
E	6	9	-3	9
F	3	7	-4	16
G	5	3	2	4
H	7	10	-3	9
I	2	4	-2	4
J	10	8	2	4
				Sum = 92

Rank Correlation

- Sampling Distribution of r_s Assuming No Rank Correlation



Rank Correlation

- Rejection Rule

With .10 level of significance:

Reject H_0 if p-value $\leq$ .10

- Test Statistic

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(92)}{10(100-1)} = 0.4424$$

$$z = (r_s - \mu_r)/\sigma_r = (.4424 - 0)/.3333 = 1.33$$

- p-Value

$$p\text{-Value} = 2(.5000 - .4082) = .1836$$

Rank Correlation

- Conclusion
  - Do no reject H_0. The p-value > α. There is not a significant rank correlation. The two analysts are not showing agreement in their ranking of the risk associated with the different investments.

Tools to support data analysis

- Spreadsheet – simple to use, basic graphs
- Statistical packages, e.g. SPSS
- Qualitative data analysis tools
  - Categorization and theme-based analysis
  - Quantitative analysis of text-based data
- Nvivo and Atlas.ti support qualitative data analysis
- CAQDAS Networking Project, based at the University of Surrey (http://caqdas.soc.surrey.ac.uk/)