DEPARTMENT OF COMMERCE AND FINANCIAL STUDIES
BHARATHIDASAN UNIVERSITY,
TIRUCHIRAPPALLI – 620024
MBA (Financial Management)

- **Course Code: FMCC10/21**
- **Course Name : Business Research Methods**
- **Unit – IV / Topic : Data Analysis through Statistical Software**
- **Course Teacher: Dr. K. Rajalakshmi**
- **Email ID: rajalakshmi7409@gmail.com**

Scheme of the presentation

- Testing of Hypotheses
- Analysis of Variance
- Non-Parametric Test
- Multiple Regression
- Correlation
- Factor Analysis
- Discriminant Analysis

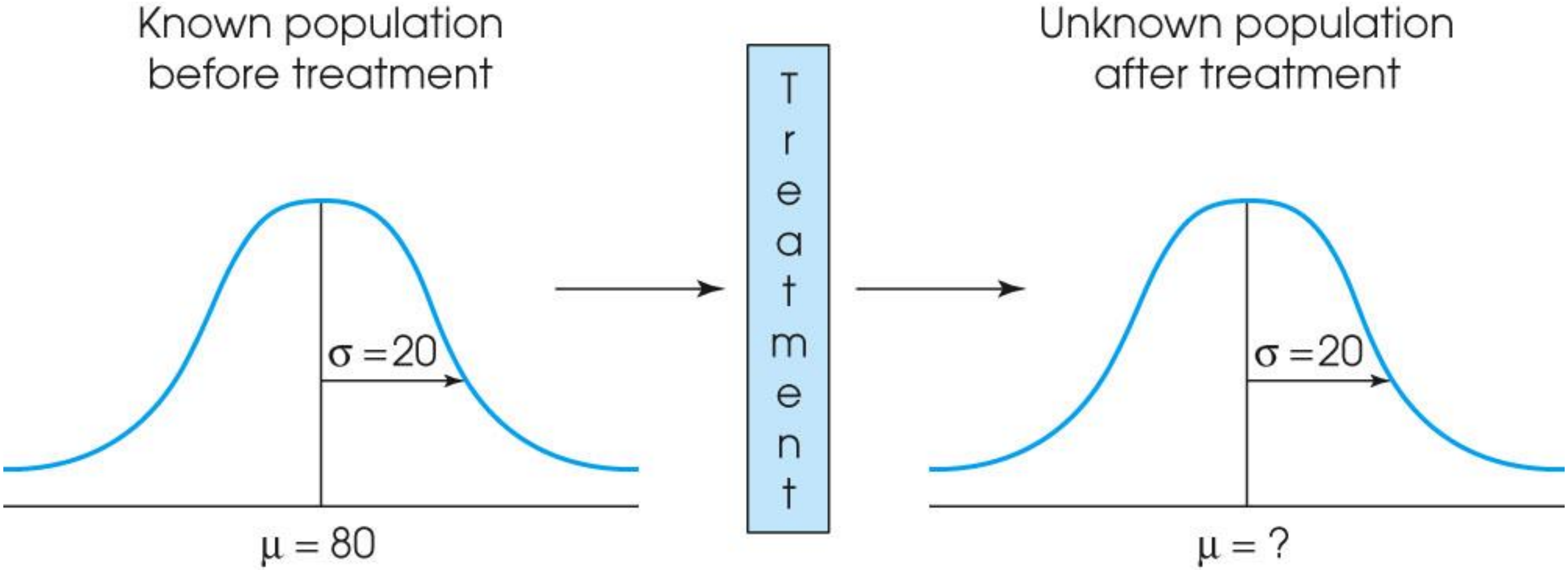
Hypothesis Testing

- The general goal of a hypothesis test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.
- Hypothesis testing is a technique to help determine whether a specific treatment has an effect on the individuals in a population.

Hypothesis Testing

The hypothesis test is used to evaluate the results from a research study in which

1. A sample is selected from the population.
2. The treatment is administered to the sample.
3. After treatment, the individuals in the sample are measured.


Known population
before treatment
σ = 20
μ = 80
Treatment
Unknown population
after treatment
σ = 20
μ = ?

Hypothesis Testing (cont.)

- If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect.
- However, it is also possible that the difference between the sample and the population is simply sampling error.

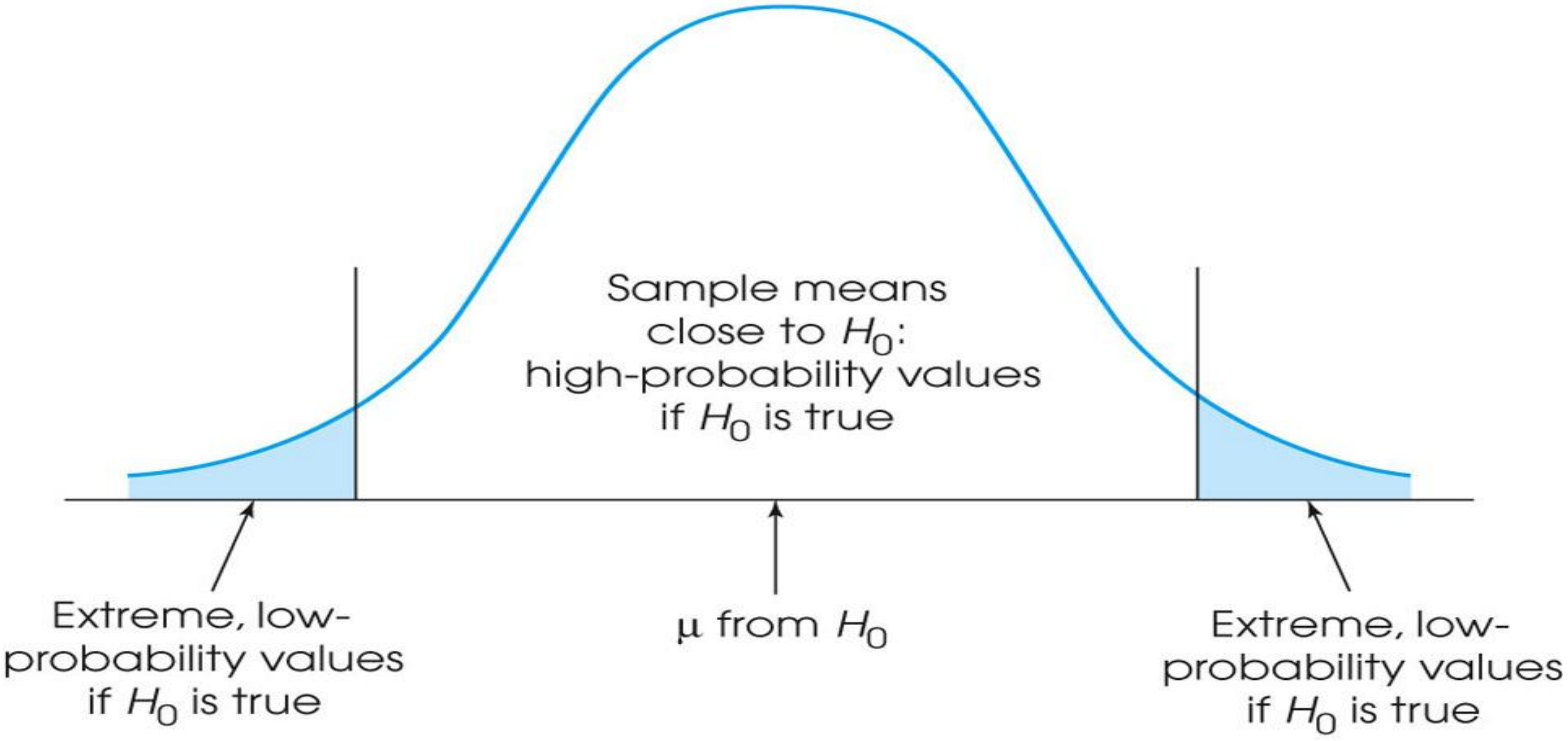
Hypothesis Testing (cont.)

- The purpose of the hypothesis test is to decide between two explanations:

  1. The difference between the sample and the population can be explained by sampling error (there does not appear to be a treatment effect)

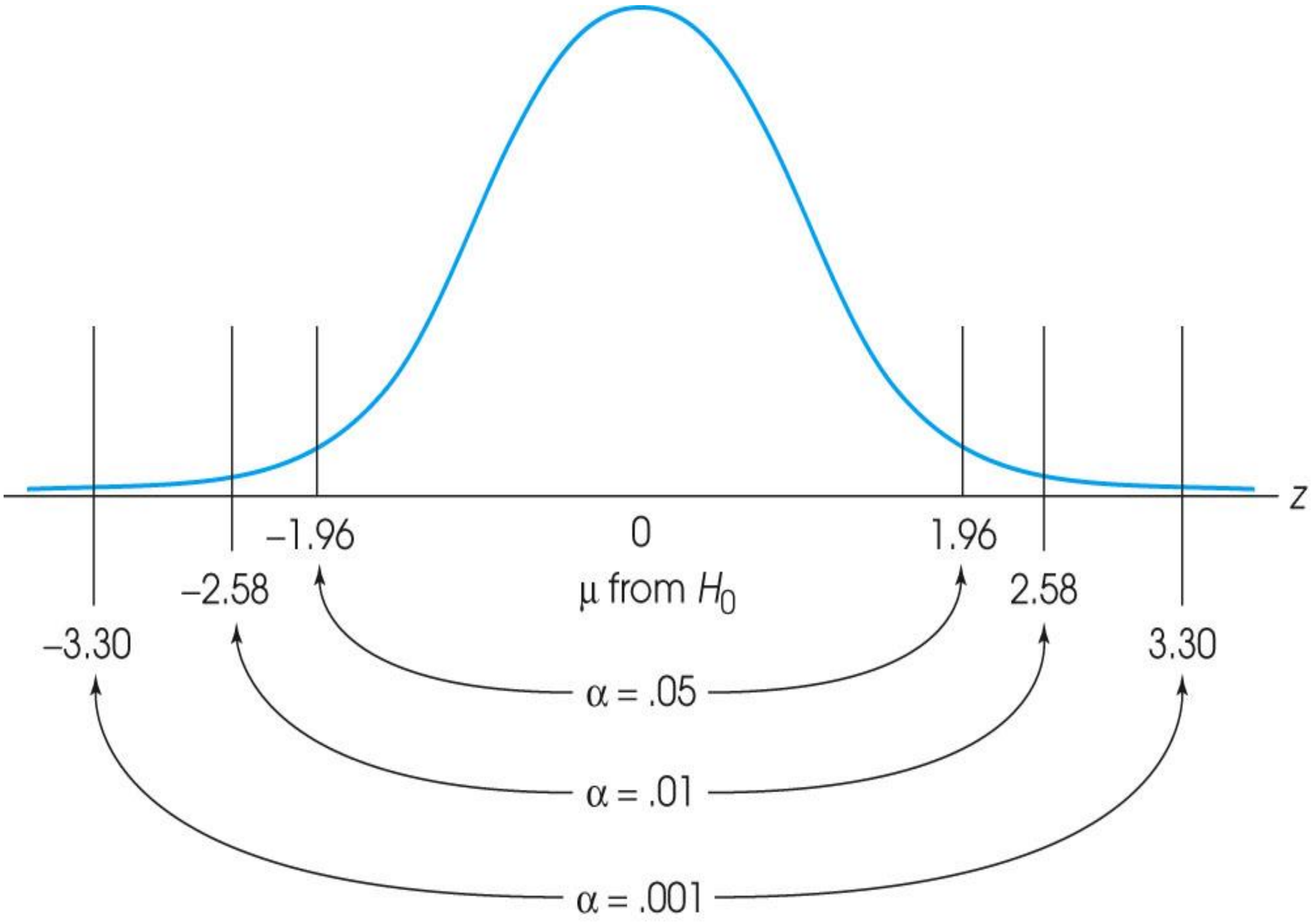
  2. The difference between the sample and the population is too large to be explained by sampling error (there does appear to be a treatment effect).

The distribution of sample means
if the null hypothesis is true
(all the possible outcomes)



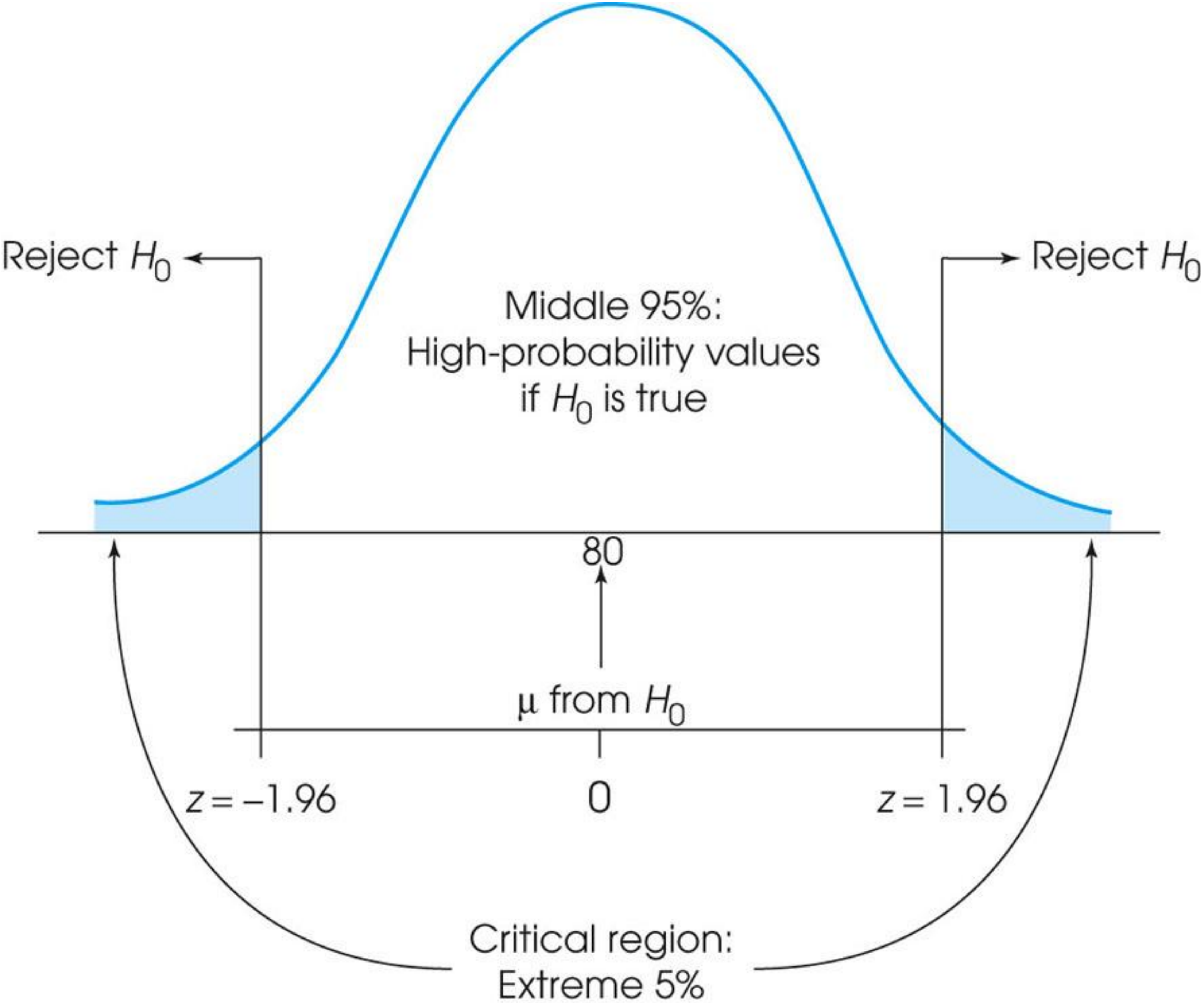
Null Hypothesis

State the hypotheses and select an α level. The **null hypothesis**, H**0**, always states that the treatment has no effect (no change, no difference). According to the null hypothesis, the population mean after treatment is the same is it was before treatment. The **α level** establishes a criterion, or "cut-off", for making a decision about the null hypothesis. The alpha level also determines the risk of a Type I error.


z
–3.30
–2.58
–1.96
0
μ from H_0
1.96
2.58
3.30
α = .05
α = .01
α = .001

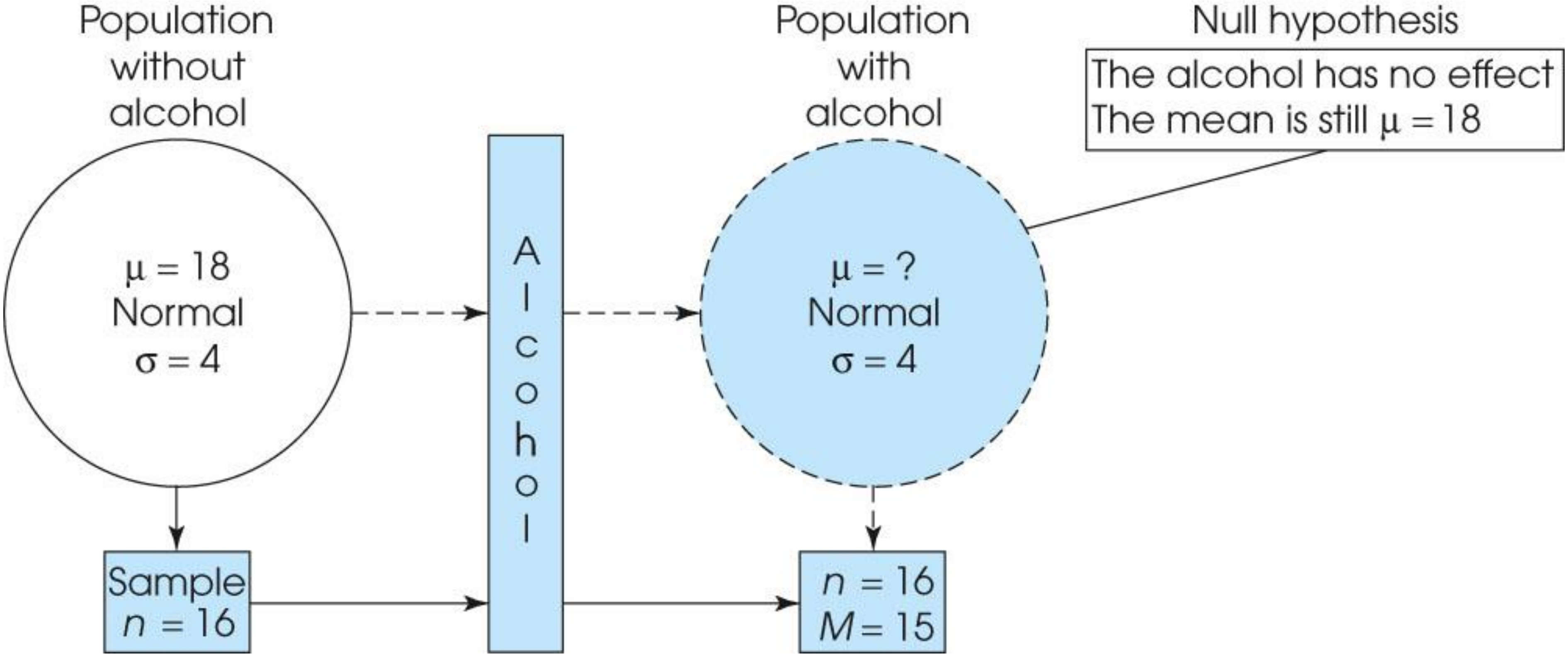
Critical Region

The **critical region** consists of outcomes that are very unlikely to occur if the null hypothesis is true. That is, the critical region is defined by sample means that are almost impossible to obtain if the treatment has no effect. The phrase “almost impossible” means that these samples have a probability (p) that is less than the alpha level.


Reject H_0
Reject H_0
Middle 95%:
High-probability values
if H_0 is true
80
μ from H_0
$z = -1.96$
0
$z = 1.96$
Critical region:
Extreme 5%

Test Statistic

The **test statistic** (in this chapter a z-score) forms a ratio comparing the obtained difference between the sample mean and the hypothesized population mean versus the amount of difference we would expect without any treatment effect (the standard error). A large value for the test statistic shows that the obtained mean difference is more than would be expected if there is no treatment effect. If it is large enough to be in the critical region, we conclude that the difference is **significant** or that the treatment has a significant effect. In this case we reject the null hypothesis. If the mean difference is relatively small, then the test statistic will have a low value. In this case, we conclude that the evidence from the sample is not sufficient, and the decision is fail to reject the null hypothesis.


Population
without
alcohol
μ = 18
Normal
σ = 4
Sample
n = 16
A
l
c
o
h
o
l
Population
with
alcohol
μ = ?
Normal
σ = 4
n = 16
M = 15
Null hypothesis
The alcohol has no effect
The mean is still μ = 18

Errors in Hypothesis Tests

- Just because the sample mean (following treatment) is different from the original population mean does not necessarily indicate that the treatment has caused a change.
- You should recall that there usually is some discrepancy between a sample mean and the population mean simply as a result of sampling error. Because the hypothesis test relies on sample data, and because sample data are not completely reliable, there is always the risk that misleading data will cause the hypothesis test to reach a wrong conclusion.
- Two types of error are possible.

Type I Errors

- **Type I error** occurs when the sample data appear to show a treatment effect when, in fact, there is none.
- In this case the researcher will reject the null hypothesis and falsely conclude that the treatment has an effect.
- Type I errors are caused by unusual, unrepresentative samples. Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though the treatment has no effect.
- The hypothesis test is structured so that Type I errors are very unlikely; specifically, the probability of a Type I error is equal to the alpha level.

Type II Errors

- **Type II error** occurs when the sample does not appear to have been affected by the treatment when, in fact, the treatment does have an effect.
- In this case, the researcher will fail to reject the null hypothesis and falsely conclude that the treatment does not have an effect.
- Type II errors are commonly the result of a very small treatment effect. Although the treatment does have an effect, it is not large enough to show up in the research study.

Directional Tests

- When a research study predicts a specific direction for the treatment effect (increase or decrease), it is possible to incorporate the directional prediction into the hypothesis test.
- The result is called a **directional test** or a **one-tailed test**. A directional test includes the directional prediction in the statement of the hypotheses and in the location of the critical region.

Directional Tests (cont.)

- For example, if the original population has a mean of μ = 80 and the treatment is predicted to increase the scores, then the null hypothesis would state that after treatment:

  H0: $\mu \leq 80$ (there is no increase)

- In this case, the entire critical region would be located in the right-hand tail of the distribution because large values for M would demonstrate that there is an increase and would tend to reject the null hypothesis.

Measuring Effect Size

- A hypothesis test evaluates the *statistical significance* of the results from a research study. The test determines whether or not it is likely that the obtained sample mean occurred without any contribution from a treatment effect.
- The hypothesis test is influenced not only by the size of the treatment effect but also by the size of the sample. Thus, even a very small effect can be significant if it is observed in a very large sample. significant effect does not necessarily mean a large effect, it is recommended that the hypothesis test be accompanied by a measure of the **effect size**.
- We use Cohen=s d as a standardized measure of effect size.
- Much like a z-score, **Cohen=s d** measures the size of the mean difference in terms of the standard deviation.

Power of a Hypothesis Test

The **power** of a hypothesis test is defined is the probability that the test will reject the null hypothesis when the treatment does have an effect. The power of a test depends on a variety of factors including the size of the treatment effect and the size of the sample.

Analysis of variance

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Cont.,

F-Statistic

The statistic which measures if the means of different samples are significantly different or not is called the F-Ratio. Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.

Assumptions of ANOVA

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

1. An ANOVA can only be conducted if there is no relationship between the subjects in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).

Cont.,

2.The different groups/levels must have equal sample sizes.

3.An ANOVA can only be conducted if the dependent variable is normally distributed, so that the middle scores are most frequent and extreme scores are least frequent.

4.Population variances must be equal (i.e. homoscedastic). Homogeneity of variance means that the deviation of scores (measured by the range or standard deviation for example) is similar between populations.

Cont.,

Types of ANOVA Tests

1. One-Way Anova
2. Two-Way Anova

One-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable. The independent variable divides cases into two or more mutually exclusive levels, categories, or groups. The one-way ANOVA test for differences in the means of the dependent variable is broken down by the levels of the independent variable.

Two-way ANOVA

A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable. The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

Understanding the ANOVA F-value

1.The test statistic for an ANOVA is denoted as F. The formula for ANOVA is F = variance caused by treatment/variance due to random chance.

2. ANOVA F value -there is a significant difference between the levels of the independent variable, when $p < .05$. So, a higher F value indicates that the treatment variables are significant.

Parametric Tests

1.Parametric tests are normally involve to data expressed in absolute numbers or values rather than ranks; an example is the Student’s t-test.

2.The results of a parametric test depends on the validity of the assumption.

3.Parametric tests are most powerful for testing the significance.

Non-Parametric Tests

1.Where we can not use the assumptions & conditions of parametric statistical procedures, in such situation we apply non-parametric tests.

2.It covers the data techniques that do not rely on data belonging to any particular distribution.

Cont.,

3.In this statistics is based on the ranks of observations and do not depend on any distribution of the population.

4.In non-parametric statistics, the techniques do not assume that the structure of a model is fixed.

5.It deals with small sample sizes, and, these are user friendly compared with parametric statistics and economical in time.

Nonparametric tests are used when either:

□ Sample is not normally distributed.

□ Sample size is small.

□ The variables are measured on nominal or ordinal scale.

There is at least one nonparametric equivalent for each parametric general type of test. Broadly, these tests fall into the following categories:

Cont.,

Test of differences between groups (independent samples)

□ Test of differences (dependent samples)

□ Test of relationships between variables.

The concepts and procedure to undertake Run Test, Chi-Square Test, Wilcoxon Signed Rank Test, Mann-Whitney Test and Kruskal-Wallis Test are discussed here under:

Run Test: Run Test is used to examine the randomness of data.

Chi-Square Test: Examine the association between two or more variables measured on categorical scales.

Cross tabulation: Analysis is used when trying to summarize the intersections of independent and dependent variables and to examine the relationship (if any) between those variables.

Cont.,

Mann- Whitney U Test: Generally, the t-test for independent samples is used, if two samples are compared over their mean value for some variable interest. Nonparametric alternatives for the test are the Wald-Wolfowitz Run test, the Mann Whitney U test, and the Kolmogorov-Smirnov two sample test.

Wilcoxon Signed Rank Test: Wilcoxon Signed Rank Test (also known as Wilcoxon Matched Pair Test) is the non-parametric version of dependent sample t-test or paired sample t-test. Sign test is the other nonparametric alternative to the paired sample t-test. If the variables of interest are dichotomous in nature (Male and Female or Yes and No) then McNemar’s Chi-Square test is used.

Cont.,

Wilcoxon Signed Rank Test: It is also a nonparametric version for one sample t-test. Wilcoxon Signed Rank Test compares the medians of the groups under two situations (paired samples) or it compares the median of the group with hypothesized median (one sample).

Kruskal –Wallis Test: It is used with multiple groups. It is the non-parametric version of one-Way ANOVA. Median test is another nonparametric alternative to one-Way ANOVA. Kruskal –Wallis Test compares medians of more than two independent groups.

Correlation

- Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables.
- A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:

I. Positive and Negative II. Simple, Partial and Multiple

III. Linear and non-linear

Positive, Negative and Zero Correlation:

Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive.

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative.

Zero Correlation: Actually it is not a type of correlation but still it is called as zero or no correlation.

Simple, Partial and Multiple Correlation:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation.

Partial Correlation: In case of partial correlation one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant.

Multiple Correlation: When three or more variables are studied, it is a case of multiple correlation.

Linear and Non-linear Correlation:

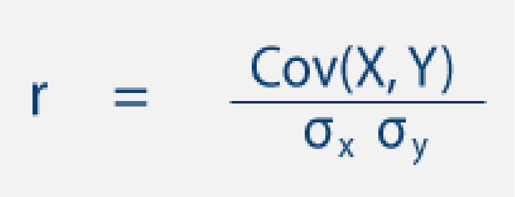
Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear.

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear.

Karl Pearson's Coefficient of Correlation:

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. This method is widely used in practice and the coefficient of correlation is denoted by the symbol "r".

$$r = \frac{\text{Covariance(x,y)}}{\text{S.D.(x) S.D.(y)}} \quad \text{(or)} \quad r = \frac{\text{Cov(X,Y)}}{\sigma_x \, \sigma_y}$$



$$r = \frac{\text{Cov(X, Y)}}{\sigma_x \, \sigma_y}$$

Where,

Cov(X,Y) – Covariance of(x, y)

$\sigma_x \, \sigma_y -$ Standard deviation of (x,y)

Regression

- A study of measuring the relationship between associated variables, where in one variable is dependent on another independent variable, called as Regression. It is developed by Sir Francis Galton in 1877 to measure the relationship of height between parents and their children.

- Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

Therefore, with the help of simple linear regression model we have the following two regression lines.

1. **Regression line of Y on X:**

This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable).

Regression line of Y on X : $Y - \dot{Y} = byx\,(X - \dot{X})$

(Or)

$$Y = a + bX$$

2. **Regression line of X on Y:**

This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable).

Regression line of X on Y : $X - \dot{X} = bxy\,(Y - \dot{Y})$

(Or)

$$X = a + bY$$

Factor Analysis

- Factor Analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) “factors.”
- The factors typically are viewed as broad concepts or ideas that may describe an observed phenomenon. For example, a basic desire of obtaining a certain social level might explain most consumption behavior. These unobserved factors are more interesting to the social scientist than the observed quantitative measurements.
- Factor analysis is generally an exploratory/descriptive method that requires many subjective judgments. It is a widely used tool and often controversial because the models, methods, and subjectivity are so flexible that debates about interpretations can occur.

- The method is similar to principal components although, as the textbook points out, factor analysis is more elaborate. In one sense, factor analysis is an inversion of principal components. In factor analysis we model the observed variables as linear functions of the “factors.” In principal components, we create new variables that are linear combinations of the observed variables.
- In both PCA and FA, the dimension of the data is reduced. Recall that in PCA, the interpretation of the principal components is often not very clean. A particular variable may, on occasion, contribute significantly to more than one of the components. Ideally we like each variable to contribute significantly to only one component. A technique called factor rotation is employed towards that goal. Examples of fields where factor analysis is involved include physiology, health, intelligence, sociology, and sometimes ecology among others.

Objectives of Factor Analysis

- ❖ Understand the terminology of factor analysis, including the interpretation of factor loadings, specific variances, and communalities;
- ❖ Understand how to apply both principal component and maximum likelihood methods for estimating the parameters of a factor model;
- ❖ Understand factor rotation, and interpret rotated factor loadings.

Discriminant Analysis (DA)

Discriminant Analysis (DA) is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

Discriminant Analysis undertakes the same task as multiple linear regressions by predicting an outcome. However, multiple linear regressions is limited to cases where the dependent variable on the Y axis is an interval variable so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values.

Objectives Discriminant Analysis

- Development of discriminant functions
- Examination of whether significant differences exist among the groups, in terms of the predictor variables.
- Determination of which predictor variables contribute to most of the intergroup differences.
- Evaluation of the accuracy of classification

Discriminant analysis linear equation

Discriminant analysis involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D= v1\ X1+ v2\ X2+ v3X3+ \ldots + viXi + a$$

Where,

D = discriminate function

v = the discriminant coefficient or weight for that variable

X = respondent’s score for that variable

a = a constant

i = the number of predictor variables