Data Analysis using R

Dr. Hakeem–Ur–Rehman
IQTM–PU

Outline:

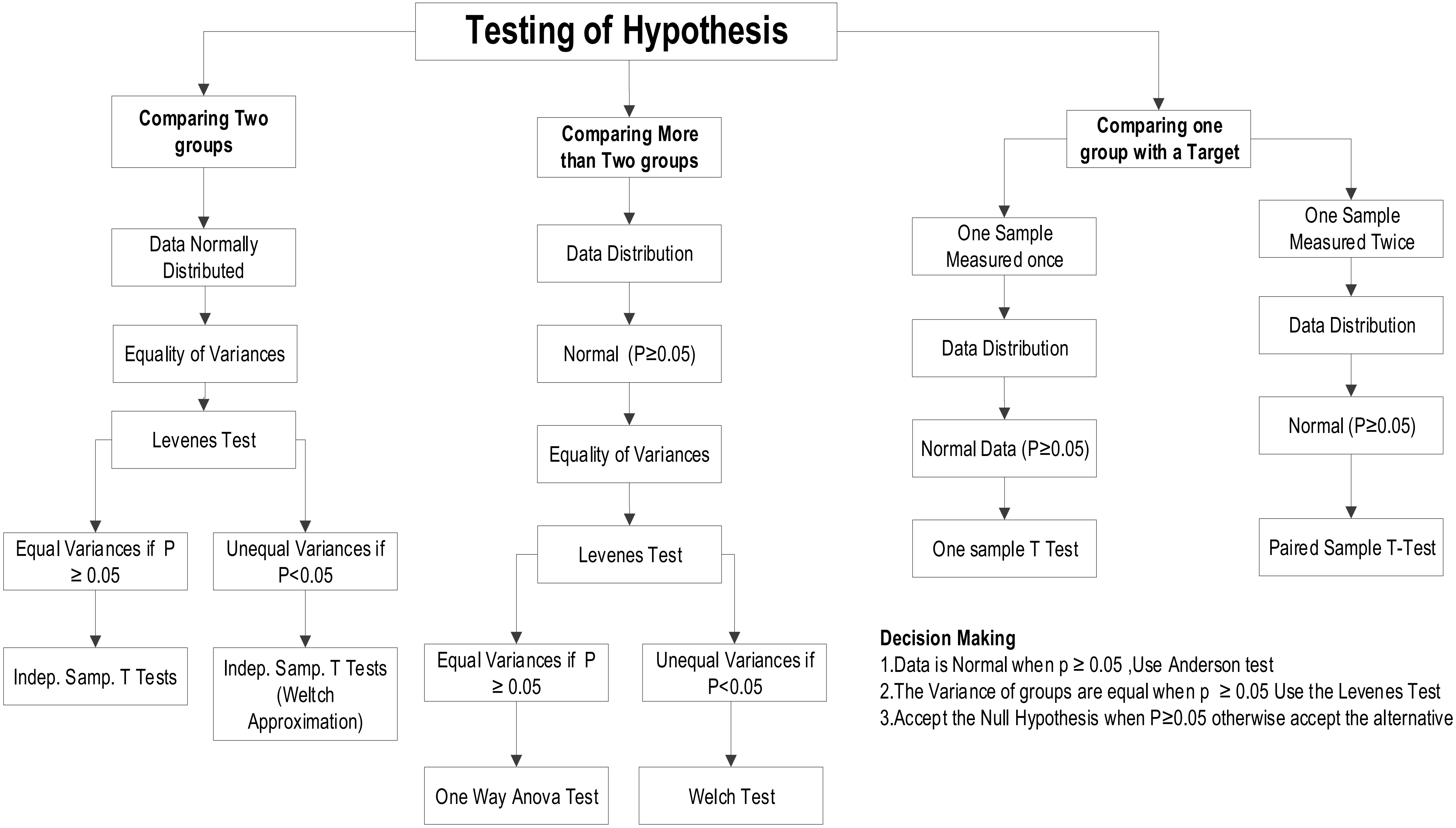
- Parametric Testing of Hypothesis
  - One-Sample t-test
  - Two samples Independent t-test
  - Paired t-test
  - One-Way ANOVA

- Correlation

- Simple Linear Regression

- Multiple Linear Regression

PARAMETRIC TESTS



Decision Making

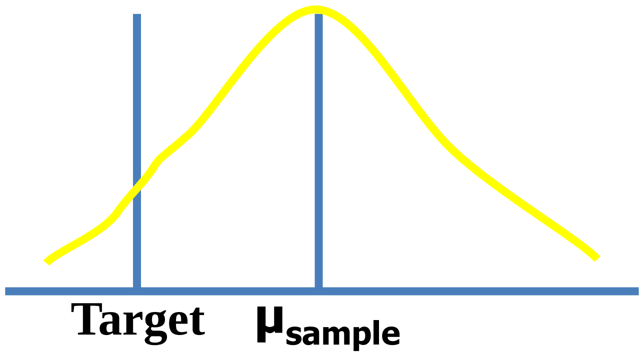
1.Data is Normal when $p \geq 0.05$,Use Anderson test

2.The Variance of groups are equal when $p \geq 0.05$ Use the Levenes Test

3.Accept the Null Hypothesis when $P \geq 0.05$ otherwise accept the alternative hypothesis

Testing of Hypotheses for Single Sample

- What are we testing?



Comparing a Single Mean

to a Specified Value

Comparing a Single Mean to a Specified Value

- Tests on the Mean of a Normal Distribution, Population Variance or S.D. Known
- Tests on the Mean of a Normal Distribution, Population Variance or S.D. Unknown ($n \geq 30$)
- Tests on the Mean of a Normal Distribution, Population Variance or S.D. Unknown ($n \leq 30$)

One Sample t-Test Using Minitab

Requirements:

1. Data must be **normally** distributed
2. σ is **unknown**

The shelf life of a carbonated beverage is of interest. Ten bottles are randomly selected and tested, and the following results are obtained:

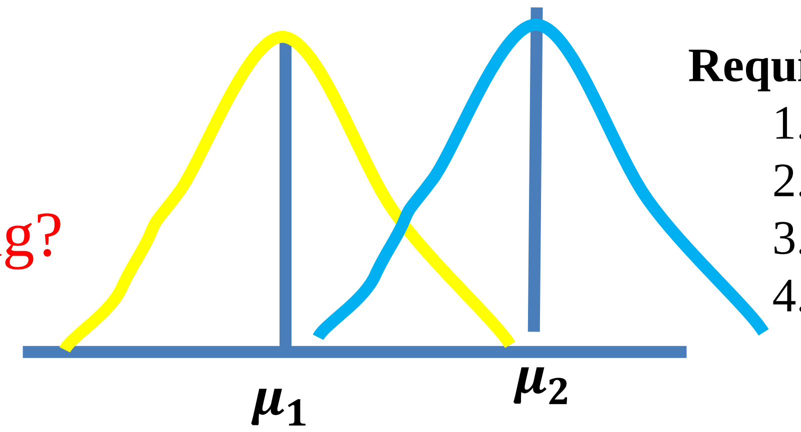
Days	
108	138
124	163
124	159
106	134
115	139

a. We would like to demonstrate that the mean shelf life exceeds 120 days. Set up appropriate hypotheses for investigating this claim.

b. Test these hypotheses using $\alpha = 0.01$. What are your conclusions?

Testing of Hypotheses for Two Independent Samples

What are we testing?



Requirements:

1. Two **normally** distributed but **independent** populations
2. σ is **unknown**
3. Variances of the two groups are **equal**
4. Sample observations should be **random**

Inferences About the Differences in Means:

1. Tests on the differences of Means of a Normal Distribution, Population Variances or S.D. are Known
2. Tests on the differences of Means of a Normal Distribution, Population Variances or S.D. are Unknown (n_1& $n_2 \geq$ 30).
3. Tests on the differences of Means of a Normal Distribution, Population Variances or S.D. are Unknown (n_1& $n_2 \leq$ 30). **Assume variances are equal**
4. Tests on the differences of Means of a Normal Distribution, Population Variances or S.D. are Unknown (n_1& $n_2 \leq$ 30). **Assume variances are unequal**

Comparing Two Population Variances:

1. Tests on the equality of two normal populations variance (F-test)
2. Tests on the equality of **two or more non-normal populations variance** (Levene's test)

Tests equality of Means & Variances Using R:

2-Sample (Independent) t Test: Example

Hospital comparison data:

A healthcare consultant wants to compare the patient satisfaction ratings of two hospitals. The consultant collects ratings from 20 patients for each of the hospitals.

Worksheet column	Description
Rating	The rating for the hospital: 1 to 100, with 100 being the best score
Hospital	The hospital that was rated: A or B

Open the sample data: HospitalComparison

Requirements:

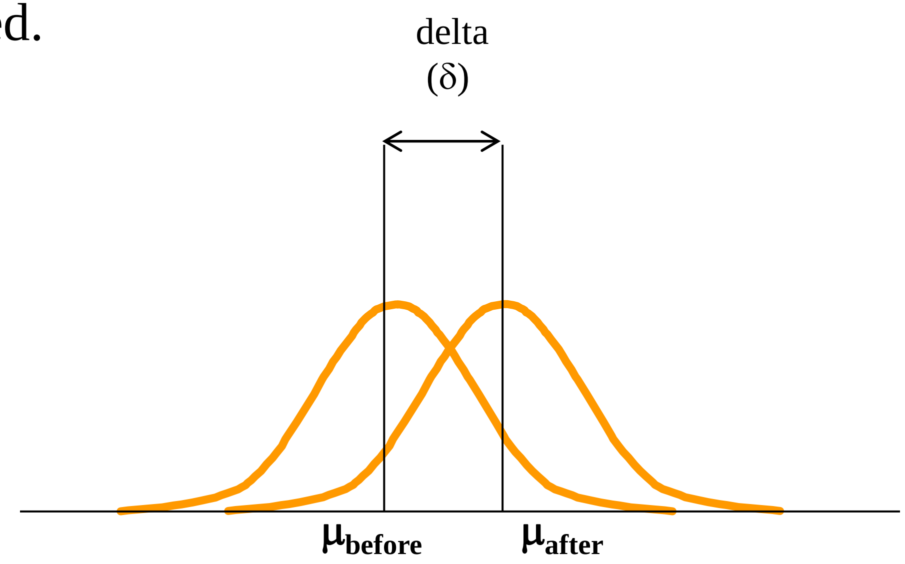
1. Two **normally** distributed but **independent** populations
2. σ is **unknown**
3. Variances of the two groups are **equal**
4. Sample observations should be **random**

Testing of Hypotheses for Two Paired Samples

- A Paired t-test is used to compare the Means of two measurements from the same samples generally used as a before and after test.

- <u>This is appropriate for testing the difference between two Means when the data are paired and the paired differences follow a Normal Distribution.</u>

- Use the Paired 't' command to compute a confidence interval and perform a Hypothesis Test of the difference between population Means when observations are paired.
  - H_o: $\mu_\delta = \mu_o$
  - H_a: $\mu_\delta \neq \mu_o$



- Where μ_δ is the population Mean of the differences and μ_0 is the hypothesized Mean of the differences, typically zero.

TEST OF MEANS (t-tests): PAIRED T-TEST

Resting heart rate data

A physiologist wants to determine whether a particular running program has an effect on resting heart rate. The heart rates of 20 randomly selected people were measured. The people were then put on the running program and measured again one year later. Thus, the before and after measurements for each person are a pair of observations.

Worksheet column	Description
Before	The resting heart rate of the person before the running program
After	The resting heart rate of the person after the running program
Difference	The difference between the person's resting heart rate before and after the running program

Open the sample data: RestingHeartRate

Requirements:

1. Data must be **normally** distributed

The Analysis of Variance (ANOVA): One-Way

- Tests the Equality of 2 or More Population Means
- Variables
  - One Nominal Scaled Independent Variable (Factor)
    - 2 or More Treatment Levels or Classifications
  - One Interval or Ratio Scaled Dependent Variable (Response)

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- All population means are equal
- No Treatment Effect

H_1: Not All μ_j are Equal (j = 1, 2, 3, 4)
- At Least 1 Pop. Mean is Different
- Treatment Effect

Assumptions:

1. **Normality**: Each group is Normally Distributed
2. **Homogeneity of Variance**: Variances of each group are equal

Parametric Vs Nonparametric Tests

Nonparametric test	parametric test
Sign test / Wilcoxon Signed Rank test	1-sample t-test
Sign Test / Wilcoxon Signed Rank test	Paired t-test
Mann-Whitney U test / Wilcoxon Sum Rank test	2-sample t-test
Kruskal-Wallis test	One-way ANOVA
Mood's Median test	One-way ANOVA
Friedman test	Two-way ANOVA

What is Regression?

Method of determining the statistical relationship between a *response (or output) and one or more predictor (or input) variables.*

$$Y = f\,(X_1, X_2, \ldots . X_n)$$

Where ‘Y’ is the <u>RESPONSE</u> and X_1 to X_n are the <u>PREDICTORS</u>

Types of Regression

<u>Simple Linear Regression…</u>

Is when the dependent variable is linearly proportional to just <u>ONE</u> independent variable.

<u>Multiple Regression…</u>

May be viewed as an extension of *simple regression analysis (where only one predictor is involved) to the* situation where there is more than ONE predictor to be considered.

QUESTIONS