

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make informed decisions.

Biostatistics applies statistical methods to biological, health, and medical data to understand patterns and make predictions.

Key Uses of Biostatistics

- Hospital management and planning
Resource distribution
Tracking diseases and vaccination rates
Assessing risk factors
Diagnosing and choosing treatments
Understanding disease frequency
-

Types of Data

1. Qualitative (Categorical)

- *Nominal*: Categories without order (e.g., blood type)
Ordinal: Categories with order (e.g., disease severity: mild, moderate, severe)

2. Quantitative (Numerical)

- *Discrete*: Countable (e.g., number of surgeries)
Continuous: Measurable (e.g., weight, blood pressure)
Scales:
 - *Interval*: No true zero (e.g., temperature)
Ratio: Has a true zero (e.g., height)
-

Variables

- Qualitative Variable: Descriptive (e.g., gender)
Quantitative Variable: Numeric (e.g., age)
Discrete vs Continuous: Count vs measurable
-

Data Sources

- Primary: Collected firsthand for a specific purpose
Secondary: Pre-collected by others (e.g., medical records)
-

Data Presentation Tools

- Pie charts & bar charts: For qualitative data
Frequency tables, histograms, polygons, ogives: For quantitative data
Box plots, scatter plots, stem-and-leaf plots: For data distribution & relationships
-

Central Tendency Measures

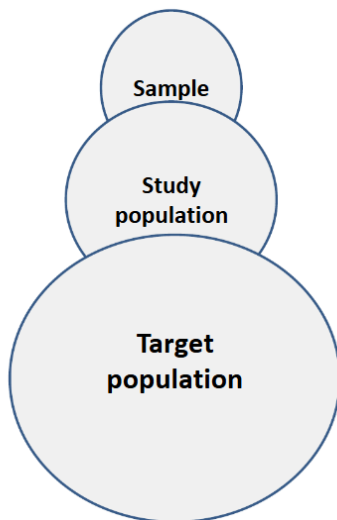
- Mean: Average
Median: Middle value
Mode: Most frequent value
-

Research Process Overview

Plan
Design study
Collect data
Process and analyze
Present results
Interpret and publish findings

Population & Sample

- Population: Entire group studied
Sample: Subset of the population
Generalizability: Apply results from sample → study population → target population



Summation Notation

- Σ (Sigma) is used to represent the sum of a series of numbers.
-

A. Qualitative (Categorical) Data

1. Nominal Scale: Categories with no order
Example: Marital status (married, not married, divorced)
Dichotomous: Only two categories (e.g., sex: male/female)
 2. Ordinal Scale: Categories with a logical order
Example:
Satisfaction: very happy, happy, unhappy, very unhappy
Malnutrition: mild, moderate, severe
Socio-economic status: high, middle, low
-

B. Quantitative (Numerical) Data

1. Interval Scale:
 - Equal spacing between values, no true zero
Example: Temperature (Celsius/Fahrenheit)
2. Ratio Scale:
 - Like interval, but has a true zero point
Example: Height, weight, blood pressure
3. Discrete Data:

- Countable whole numbers
Example: Number of surgeries in a month

4. Continuous Data:

- Can take any value within a range
Example: Time, weight

Variables

- Qualitative Variable: Describes categories
Quantitative Variable: Measures numeric value
 - Discrete: Countable (e.g., patients)
Continuous: Measurable (e.g., weight)

Sources of Data

1. Primary Data: Collected directly by researcher
Secondary Data: Collected by others, reused for new purposes

Data Collection Systems

1. Routine Systems:
Ongoing data collection (e.g., censuses, hospital records)
2. Non-Routine Systems (Ad Hoc):
 - Collected occasionally (e.g., surveys, surveillance)

Presenting Qualitative Data

- Pie Charts and Bar Charts (simple or grouped)
Relative Frequency Table:
Formula: Relative frequency = (frequency ÷ total)

Pie Chart Steps:

1. Create a frequency table
Calculate % for each category
Convert % to degrees: $\% \times 360$
Draw and divide the circle
-

Displaying and summarizing data in biostatistics:



Bar Chart (Bar Graph)

- Categories are placed on the horizontal (X) axis.
Frequencies or relative frequencies go on the vertical (Y) axis.
Draw vertical bars of equal width—height represents frequency.
Used for qualitative (categorical) data.
-



Cross Tabulation (Crosstabs)

- Shows the relationship between two qualitative variables (e.g., Gender and Spectacle Use).
Presented in a table format with counts and percentages:
Example:

	Wearing Glasses	Not Wearing	Total
Boys	5 (33.3%)	10 (66.7%)	15
Girls	10 (40%)	15 (60%)	25
Total	15 (37.5%)	25 (62.5%)	40



Frequency Distribution Table (Quantitative Data)

- Lists data values or intervals and how often they occur (frequency).
Can include:

- Relative frequency: $\text{frequency} \div \text{total}$
Cumulative frequency: running total of frequencies
Cumulative relative frequency: running total of percentages

Example Table:

Age Interval	Frequency (f)	Cumulative Freq	Rel. Freq (%)	Cum. Rel. Freq (%)
18–20.9	6	6	5%	5%
21–23.9	24	30	20%	25%
...

Constructing Frequency Tables

1. Group data into intervals
Count frequencies
Calculate relative and cumulative frequencies

Formula:

- Relative Frequency = $\text{Frequency} \div \text{Total}$
Class Width (C) = $(\text{Max} - \text{Min}) \div \text{No. of Classes}$
Sturges' Rule: $K = 1 + 3.322 \log(n)$ (to find number of classes)

Histogram vs. Bar Chart

- Histogram: For quantitative data; no gaps between bars; each bar represents an interval.
Bar Chart: For qualitative data; bars have gaps; each bar is a category.

Other Graphs for Quantitative Data:

- Frequency Polygon: Connects midpoints of histogram bars.
Ogive: Graph of cumulative frequency.

Stem-and-leaf plot: Shows distribution while retaining raw data.

Box and Whisker plot: Visualizes quartiles and outliers.

Scatter Plot: Shows relationships between two numerical variables (used in regression).

Ogive (Cumulative Relative Frequency Graph)

- Shows how many values are less than or equal to a certain amount. It's a line graph that displays cumulative relative frequency. Useful to estimate percentiles (e.g., "80% have BMI less than 30").
-

Stem-and-Leaf Plot

- Splits data into "stems" (e.g., tens) and "leaves" (units). Keeps all original data visible. Makes it easy to see distribution shape, min, max, and gaps.

Example:

Stem | Leaf

6 | 4 8

7 | 1 2 5 8

8 | 0 1 2

9 | 1

Box and Whisker Plot (Boxplot)

- **Shows spread and center of the data.**
Includes:
Minimum, Q1, Median (Q2), Q3, and Maximum
Whiskers show the full data range
Good for spotting outliers and comparing multiple groups.
-

Scatter Plot

- Used to visualize relationships between two quantitative variables (e.g., height and weight). Each point represents one observation.
-

Summation Notation (Σ)

- Σ means "sum of..."

- Example:
 - $\Sigma X = X_1 + X_2 + \dots + X_n$
 $\Sigma X^2 = X_1^2 + X_2^2 + \dots + X_n^2$
 $\Sigma (X - c)^2$ = sum of squared differences from a constant (often used in variance)
-

Measures of Central Tendency

1. Mean:

- Sum of all values ÷ number of values
Sensitive to outliers
Example: Mean of 4, 5.4, 4.6, 6.0 = $(20) \div 4 = 5$

2. Median:

- Middle value in ordered data
Not affected by outliers
If even number of values, take average of middle two.

3. Mode:

- Most frequent value(s)
Can be unimodal, bimodal, or no mode
Example: Mode of 3, 13, 13, 20 = 13
-

Trimmed Mean

- Removes a percentage of highest and lowest values (often 20%) to reduce impact of outliers.
Example: After trimming smallest and largest values, compute the mean of the rest.
-

Grouped Data: Mean, Median, Mode

- Mean (Grouped):
 - Use class midpoints:
Mean = $\Sigma (f_i \times m_i) \div \Sigma f_i$

- Median (Grouped):
 - Use class boundaries and cumulative frequency to estimate median.
 - Mode (Grouped):
 - Modal class = class with highest frequency.
 - Use:

$$\text{Mode} = Lo + (D1 / (D1 + D2)) \times \text{class width}$$
-

Measures of Dispersion

Range = Max – Min

Variance (S^2) = average squared deviation from mean

Standard Deviation (S) = square root of variance

Standard Error (SE) = S / \sqrt{n}

Quartiles & Interquartile Range (IQR) = $Q3 - Q1$

Coefficient of Variation (CV) = $(S / \text{Mean}) \times 100\%$

1. Range

- **Definition:** Difference between the largest and smallest values in a dataset.
Formula:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$
Example: For data = 4, 5, 6, 7, 14 \rightarrow Range = $14 - 4 = 10$
Note: Simple but not very reliable since it uses only two values.
-

2. Variation (SS) – Sum of Squares

- Measures how much the data values deviate from the mean, squared.
Formula:

$$SS = \sum (X_i - \text{Mean})^2$$
 Used as a base for calculating variance and standard deviation.
-

3. Variance (S^2)

- **Definition:** Average of the squared deviations from the mean.
Formula (Sample):

$$\text{Variance } (S^2) = SS / (n - 1)$$

Use $(n - 1)$ instead of n to avoid underestimating spread in small samples.
Example: If $SS = 20$, and $n = 4 \rightarrow S^2 = 20 / (4-1) = 6.67$

4. Standard Deviation (SD or S)

- Definition: Square root of variance. Brings units back to original scale.

Formula:

$$SD = \sqrt{(\text{Variance})}$$

Tells how much data varies around the mean.

Example: For variance 6.67 $\rightarrow SD = \sqrt{6.67} \approx 2.58$

5. Standard Error of the Mean (SEM or SE)

- Definition: Measures how precisely the sample mean estimates the population mean.

Formula:

$$SE = SD / \sqrt{n}$$

Smaller SE \rightarrow More reliable the sample mean.

Used in confidence intervals and hypothesis testing.

6. Quartiles and Interquartile Range (IQR)

- Q1: 25% of data below
Q2 (Median): 50% of data below
Q3: 75% of data below
IQR = $Q3 - Q1$
Used to describe spread and detect outliers
-

7. Outliers

- Definition: Data points far from others.
Rule of thumb:
A value is an outlier if it's:
Below $Q1 - 1.5 \times IQR$
Above $Q3 + 1.5 \times IQR$
Can distort mean and variance. Use boxplots or formulas to detect them.
-

8. Coefficient of Variation (CV)

- **Definition: A relative measure of variability.**
 - **Formula:**
$$CV = (SD / \text{Mean}) \times 100\%$$
 - **Useful for comparing variability across datasets with different units or scales.**
 - **Lower CV → More consistency**
-

✓ Example Summary (Data: 10, 12, 14, 16)

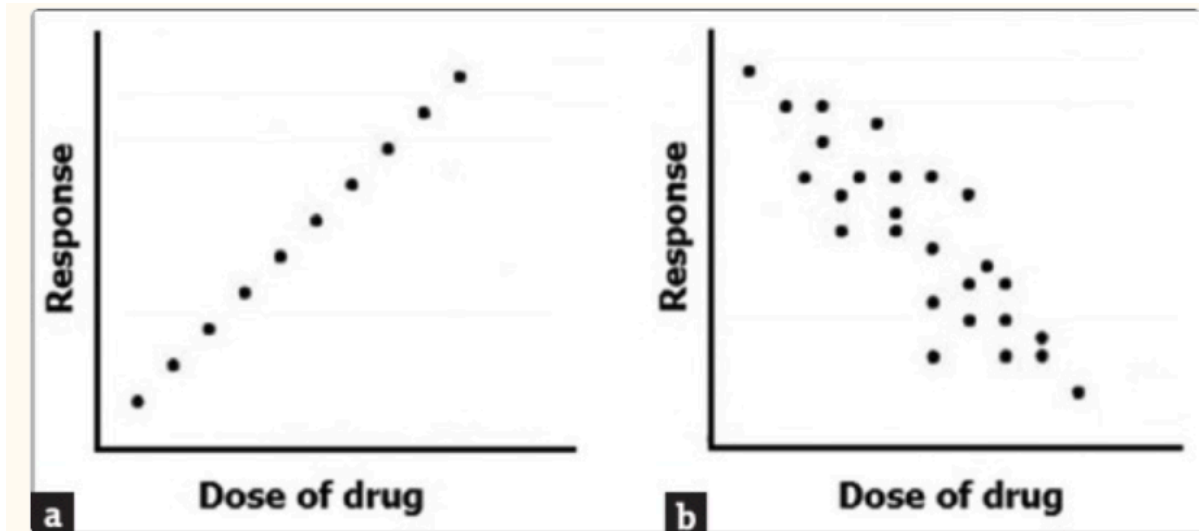
- **Mean = 13**
 - **Max = 16, Min = 10**
 - **Range = 6**
 - **SS = 20**
 - **Variance = 6.67**
 - **SD = 2.58**
 - **SE = $2.58 / \sqrt{4} = 1.29$**
 - **Quartiles: Q1 = 11, Median = 13, Q3 = 15**
 - **IQR = $15 - 11 = 4$**
 - **CV = $(2.58 / 13) \times 100 = 19.85\%$**
-

The Scatter Plot

To explore the relationship between two variables, we first use a scatter plot, where one variable is plotted against the other. Usually, the independent variable goes on the X-axis and the dependent variable on the Y-axis. In some cases (e.g., where it's unclear which is dependent), this choice doesn't matter.

- **Direct linear relationship: As X increases, Y increases (points form an upward line).**

- Inverse linear relationship: As X increases, Y decreases (points form a downward line).
- Curvilinear relationship: The relationship changes direction and forms a curve (e.g., drug dose vs. effect).



Correlation Coefficient

To measure the strength and direction of a linear relationship, we use the correlation coefficient, usually Pearson's r :

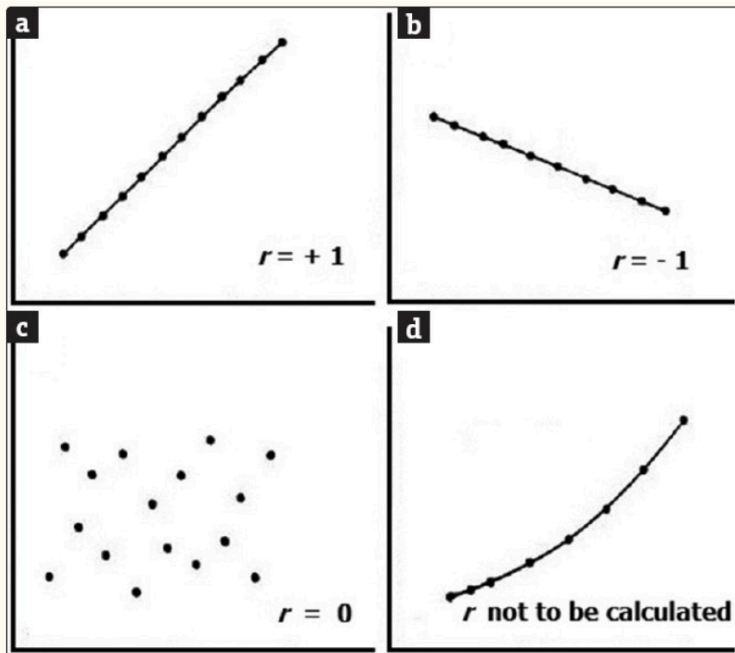
- Ranges from -1 to $+1$
- $+1$ = perfect positive correlation
- -1 = perfect negative correlation
- 0 = no linear relationship

A scatter plot helps confirm that the relationship is linear before calculating r .

Interpretation (general guide):

- > 0.7 : Strong correlation
- $0.5 - 0.7$: Moderate to good
- $0.3 - 0.5$: Fair

- < 0.3 : Weak



Nonparametric Correlation

If variables aren't normally distributed or are ordinal:

- Use Spearman's rho (ρ): ranks data before calculating correlation.
- Use Kendall's tau (τ) for small samples or when there are many tied ranks.

Testing and Confidence Interval

We can test if the observed correlation is real or due to chance using a t-test, assuming the null hypothesis that the population correlation is zero.

A confidence interval (CI) gives a range for the true correlation. This uses Fisher's z transformation to calculate.

Coefficient of Determination (r^2)

r^2 tells us how much of the variability in one variable is explained by the other.

- Example: $r = 0.7 \rightarrow r^2 = 0.49 \rightarrow 49\%$ of the variation in Y is due to X.

r^2 is used in regression to judge the model's goodness of fit.

Special Types of Correlation

- Point biserial (rPB): One variable is binary (e.g., yes/no), the other continuous.
 - Biserial (rB): Used when a continuous variable is artificially split into two groups.
 - Phi coefficient (ϕ): Correlation between two binary variables. Common in psychological or educational research.
-

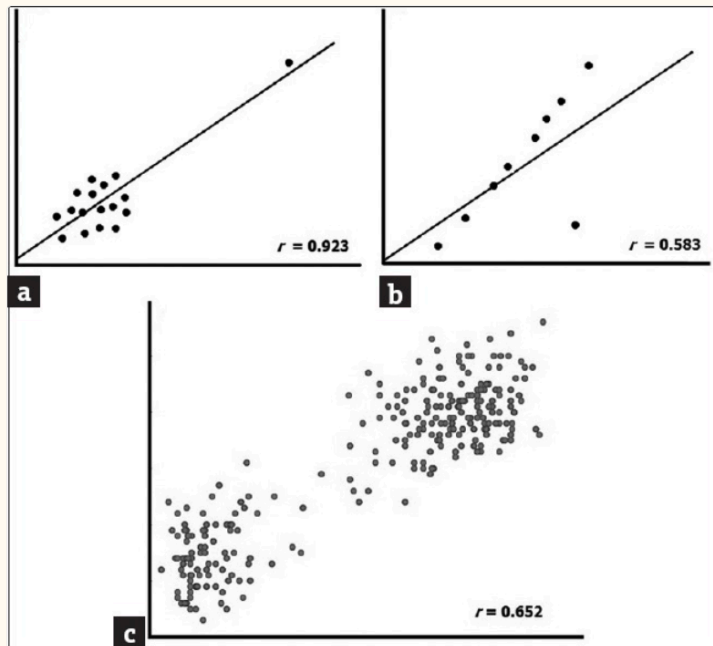
Simple Linear Regression

If two variables are correlated, we can use regression to predict one from the other using a straight-line equation:

$$y = a + bx$$

- a: Y-intercept
- b: Slope (how much Y changes for each 1 unit change in X)

The best-fit line is calculated using the least squares method, which minimizes the sum of the squares of the differences between observed and predicted values (residuals).



Pitfalls in Correlation and Regression Analysis

Assumptions and Limitations

Correlation and regression assume:

- A linear relationship between variables.
Independent observations (no repeated measures on the same subject).
For correlation, both variables should be random. For regression, only the dependent variable (y) needs to be random.

If these conditions aren't met, the results can be misleading.

Why Check a Scatter Plot First?

Scatter plots help avoid false conclusions:

- Nonlinear relationships can go undetected. A curved pattern might exist even if the correlation coefficient is close to zero.
- Outliers can distort results:
 - A univariate outlier (extreme x or y value) can falsely inflate correlation.
 - A bivariate outlier (normal x and y values, but odd combination) can hide a strong relationship.

- Clusters of data can falsely suggest a relationship, especially if the sample isn't random.
-

Dangers of Extrapolation

Using a regression equation beyond the range of observed data is risky. For example:

If $\text{Weight} = a + b \times \text{Height}$, plugging in a height of 0 gives a weight of a —which doesn't make sense.

This happens because the equation was created using real (non-zero) height values. Always stay within the observed data range and avoid applying equations to different populations (e.g., using adult data for children).

Correlation Is Not Causation

Just because two variables are correlated doesn't mean one causes the other. There may be:

- A third variable influences both.
A coincidence (random association).
Or various complex relationships like indirect or bidirectional causation.

Example: Early studies showed HRT use was linked to lower heart disease in women. Later trials found HRT actually increased heart risk. The initial correlation was due to higher socioeconomic status among HRT users—not the drug itself.

Key Point: Correlation may hint at a relationship but doesn't prove causality. Proving cause-effect requires deeper research, including controlled trials.

Correlation ≠ Agreement

Correlation tells you about a relationship, not whether two methods give the same result.

Two tests may be strongly correlated but not in close agreement. For example, if one consistently gives higher results than the other, the correlation may still be high—even though the agreement is poor.

Difference Between Correlation and Regression

SI	Correlation	Regression
No		

1	Shows how strongly and in what direction two variables are related.	Describes the exact relationship between variables using a mathematical equation.
2	A relative measure of association.	An absolute measure that shows the actual change.
3	Not affected by changes in origin or scale.	Unaffected by origin, but affected by scale (units matter).
4	Has no units.	Has units based on the variables.
5	The result is always a value between -1 and +1.	The relationship is shown as an equation like:
<ul style="list-style-type: none"> • $Y = a + bX$ • $Y = a + bX + cX^2$ (if nonlinear) Cannot be used for prediction. Can be used to predict the dependent variable from the independent one. 7 Zero correlation is possible (e.g., weight of wife vs. husband's income). There is no concept of "zero regression"—a line can always be fitted. 		

Example: Each of these terms refers to water quality parameters that help assess the health of an aquatic environment, especially for fish growth and survival. Here's what each indicates:

Key Water Quality Parameters and Their Meaning

Parameter	What It Indicates	Impact on Fish Growth
pH	Acidity or alkalinity of water (scale: 0–14). Ideal: ~6.5–8.5.	Outside this range, it can stress or kill fish.
Salinity	Amount of dissolved salts in water. Measured in ppt (parts per thousand).	Different fish thrive in different salinity (freshwater vs marine species).
DO (Dissolved Oxygen)	Amount of oxygen available in water (mg/L).	Essential for fish respiration. Low DO (<3 mg/L) can be deadly.

EC (Electrical Conductivity)	Ability of water to conduct electricity, indicating total ion concentration.	High EC often means pollution or high salinity; affects fish metabolism.
TDS (Total Dissolved Solids)	Total amount of inorganic salts and small amounts of organic matter.	High TDS can reduce oxygen levels and affect fish health.
Temperature	Warmth of the water (°C).	Affects metabolism, oxygen solubility, and breeding. Most fish have an optimal temperature range.
Alkalinity	Water's ability to neutralize acids (buffering capacity).	Stable alkalinity prevents sudden pH changes, which are stressful to fish.
Turbidity	Cloudiness of water caused by suspended solids.	High turbidity blocks light, affects photosynthesis, clogs gills, and reduces growth.
Fish Growth	The biological response to environmental conditions.	Influenced by all above parameters; optimal water quality supports faster, healthier growth.

Installing R

Here's a simplified version of the text:

Installing R and R Studio is key for anyone interested in data analysis, statistical computing, and graphical representation. This guide will walk you through the process of setting up both R and RStudio on your system, whether you're a beginner or an experienced analyst.

Why Use R Studio?

R Studio is a preferred tool because it provides:

1. **User-Friendly IDE:** It makes coding easier with features like syntax highlighting, code completion, and debugging tools.
2. **Project Management:** Organize multiple scripts, data files, and plots in one workspace.
3. **Data Visualization:** Create graphs and charts with ease using tools like ggplot2.
4. **Package Management:** Easily install and manage R packages.
5. **Markdown Support:** Combine code, visuals, and text in reports.
6. **Collaboration:** Share your work and use version control with Git.

How to Download and Install R and R Studio

For Windows:

1. **Install R:** Download R from [r-project.org](https://cran.r-project.org/).
2. <https://cran.r-project.org/bin/windows/base/>
3. **Install R Studio:** After installing R, download R Studio from their official website. <https://posit.co/download/rstudio-desktop/>
4. **Test Installation:** Open R Studio, type `print('Hello world!')` in the console to check if it works.

For Ubuntu:

1. **Install R:**
 - Open the terminal and run:
`sudo apt-get update`

```
sudo apt -y install r-base
```

- Check R version with `R --version`.

2. Install R Studio:

- Install gdebi package:

```
sudo apt-get install gdebi-core
```
- Download the latest R Studio `.deb` package.
- Install it using:

```
sudo gdebi rstudio-1.x.xxxx-amd64.deb
```
- Open R Studio by typing `rstudio` in the terminal.

Test by typing `print('Hello world!')` in R Studio to ensure it's working.

1. Basic Arithmetic Operations

- Addition:
`5 + 3`
- Subtraction:
`5 - 3`
- Multiplication:
`5 * 3`
- Division:
`5 / 3`
- Exponentiation:
`5^3`
- Modulo (remainder):
`5 %% 3`
- Integer Division:
`5 %/% 3`

2. Creating Variables

- Assign a value to a variable:

```
x <- 10
```

or

```
x = 10
```

- View the value of a variable:

```
x
```

3. Basic Functions

- Calculate the square root:

```
sqrt(16)
```

- Find the maximum value:

```
max(1, 2, 3, 4)
```

- Find the minimum value:

```
min(1, 2, 3, 4)
```

- Sum of numbers:

```
sum(1, 2, 3, 4)
```

- Mean of numbers:

```
mean(c(1, 2, 3, 4))
```

4. Creating Vectors

- Create a numeric vector:

```
v <- c(1, 2, 3, 4)
```

- Create a character vector:

```
v2 <- c("apple", "banana", "cherry")
```

- Access an element in a vector:

```
v[2] (this will return the second element in the vector)
```

5. Basic Data Structures

- Create a matrix:

```
m <- matrix(1:6, nrow = 2, ncol = 3)
```

- Create a data frame:

```
df <- data.frame(Name = c("Alice", "Bob"), Age = c(25, 30))
```

6. Basic Plotting

- Simple plot:

```
plot(1:10)
```

- Plotting a line graph:

```
plot(1:10, type = "l")
```

- Scatter plot:

```
plot(x, y) where x and y are numeric vectors.
```

7. Conditional Statements

if statement:

```
R
CopyEdit
if (x > 5) {

  print("x is greater than 5")

}
```

-

8. Loops

for loop:

```
R
CopyEdit
for (i in 1:5) {

  print(i)

}
```

-

while loop:

```
R
CopyEdit
i <- 1

while (i <= 5) {

  print(i)

  i <- i + 1

}
```

-

9. Install and Load Packages

- Install a package:
`install.packages("ggplot2")`
- Load a package:
`library(ggplot2)`

10. Getting Help

- Get help on a function:
`?mean`
- List all installed packages:
`installed.packages()`

These basic commands should help you start exploring and working with R. Let me know if you'd like more details on any specific command!

4o mini

You said:

clear the screen

ChatGPT said:

To clear the screen in R, you can use the following command:

R

CopyEdit

```
# For R Console
```

```
cat("\014")
```

This will clear the console screen.

If you're using RStudio, you can also use the keyboard shortcut:

- **Windows/Linux:** `Ctrl + L`
- **Mac:** `Command + L`

Sample data:

Structure of the Data:

- **Environments:** Bathinda, Hisar
- **Treatments:** Three treatments per environment (1, 2, 3)
- **Replications:** 12 replications for each treatment
- **Values:** Numeric data for each treatment and replication

```
# Example dataset creation
```

```
set.seed(123) # For reproducibility
```

Null Hypothesis: There is no significant difference between treatments

```
# Creating the data frame
```

```
d13 <- data.frame(
```

```
  env = rep(c("Bathinda", "Hisar"), each = 9), # Two environments
```

```
  treat = rep(c(1, 2, 3), times = 6),          # Three treatments
```

```
  replication = rep(1:3, times = 6),          # Three replications
```

```
  syield = c(
```

```
    1794, 1134, 718, # Bathinda, treat 1, 2, 3
```



```

3286, 2518, 757, # Hisar, treat 1, 2, 3
1500, 1200, 850, # Bathinda, treat 1, 2, 3
3100, 2200, 650, # Hisar, treat 1, 2, 3
1700, 1400, 900, # Bathinda, treat 1, 2, 3
3000, 2100, 800 # Hisar, treat 1, 2, 3
)
)

# Check the structure of the data
str(d13)

# Convert 'env', 'treat', and 'replication' to factors
d13$env <- factor(d13$env)
d13$treat <- factor(d13$treat)
d13$replication <- factor(d13$replication)

# Perform three-way ANOVA
aov_model <- aov(syield ~ env * treat * replication, data = d13)
summary(aov_model)

# Perform Tukey HSD for the 'treat' factor
tukey_result <- TukeyHSD(aov_model, "treat")
summary(tukey_result)
print(tukey_result)
tukey_result$treat

```

The results of the Tukey HSD test you have indicate pairwise comparisons between the treatment levels (labeled 1, 2, and 3) for the factor `treat`. Here's how to interpret these results:

Tukey HSD Results Summary:

You have three treatment groups:

- **Comparison 2 vs. 1:** The mean difference is **-638.0**.
 - **Confidence Interval:** The lower bound is **-1601.344**, and the upper bound is **325.34365**.
 - **p-value:** The adjusted p-value is **0.2219**.
 - **Conclusion:** Since the p-value is greater than 0.05, we fail to reject the null hypothesis. There is no significant difference between treatment 2 and treatment 1.
- **Comparison 3 vs. 1:** The mean difference is **-1617.5**.
 - **Confidence Interval:** The lower bound is **-2580.844**, and the upper bound is **-654.15635**.
 - **p-value:** The adjusted p-value is **0.0020**.
 - **Conclusion:** Since the p-value is less than 0.05, we reject the null hypothesis. This indicates a **significant difference** between treatment 3 and treatment 1.
- **Comparison 3 vs. 2:** The mean difference is **-979.5**.
 - **Confidence Interval:** The lower bound is **-1942.844**, and the upper bound is **-16.15635**.
 - **p-value:** The adjusted p-value is **0.0462**.
 - **Conclusion:** Since the p-value is less than 0.05, we reject the null hypothesis. This indicates a **significant difference** between treatment 3 and treatment 2.

Final Conclusion:

- There is no significant difference between **treatment 1 and treatment 2** ($p = 0.2219$).
- There is a significant difference between **treatment 1 and treatment 3** ($p = 0.0020$).
- There is a significant difference between **treatment 2 and treatment 3** ($p = 0.0462$).

These conclusions help you understand which treatment levels differ significantly in terms of `syield` (the yield measure). Let me know if you'd like more explanation or further steps!

```
=====
```

```
# Example dataset creation
```

```
set.seed(123) # For reproducibility
```

```
# Creating the data frame
```

```
d13 <- data.frame(
```

```
  env = rep(c("Bathinda", "Hisar"), each = 9), # Two environments
```

```
  treat = rep(c(1, 2, 3), times = 6),          # Three treatments
```

```
  replication = rep(1:3, times = 6),          # Three replications
```

```
  syield = c(
```

```
    1794, 1134, 718, # Bathinda, treat 1, 2, 3
```

```
    3286, 2518, 757, # Hisar, treat 1, 2, 3
```

```
    1500, 1200, 850, # Bathinda, treat 1, 2, 3
```

```
    3100, 2200, 650, # Hisar, treat 1, 2, 3
```

```
    1700, 1400, 900, # Bathinda, treat 1, 2, 3
```

```
    3000, 2100, 800 # Hisar, treat 1, 2, 3
```

```
  )
```

```
)
```

```
# Check the structure of the data
```

```
str(d13)
```

```
# Convert 'env', 'treat', and 'replication' to factors
```

```
d13$env <- factor(d13$env)
```

```
d13$treat <- factor(d13$treat)
```

```
d13$replication <- factor(d13$replication)
```

```
# Perform three-way ANOVA
```

```
aov_model <- aov(syield ~ env * treat * replication, data = d13)
```

```
summary(aov_model)
```

```
# Perform Tukey HSD for the 'treat' factor
```

```
tukey_result <- TukeyHSD(aov_model, "treat")
```

```
summary(tukey_result)
```

```
print(tukey_result)
```

```
tukey_result$treat
```

```
install.packages("lme4", dependencies = TRUE, type = "binary")
```

```
# Fit a Mixed Effects Model using lme4
```

```
library(lme4)
```

```
# Fit a fixed-effects model without any random effects
```

```
lm3_fixed = lm(syield ~ treat, data = d13)
```

```
# Check the summary of the fixed-effects model
```

```
summary(lm3_fixed)
```

```
# Perform Type I ANOVA for the mixed effects model
```

```
anova(lm3_fixed)
```

```
install.packages("car", dependencies = TRUE, type = "binary")
```

```
# Perform Type III ANOVA for the mixed effects model
```

```
library(car)
```

```
Anova(lm3_fixed, type = "III")
```

```
# Display summary of the mixed effects model
```

```
summary(lm3_fixed)
```

```
# Perform post-hoc analysis using lsmeans (emmeans is a modern equivalent)
```

```
library(lsmeans) # Or 'emmeans' if lsmeans is outdated
```

```
lsmeans(lm3_fixed, "treat")
```

```
library(emmeans)
```

```
emmeans(lm3_fixed, "treat")
```

```
=====
```

Output:

```
> # Example dataset creation
```

```
> set.seed(123) # For reproducibility
```

```
>
```

```
> # Creating the data frame
```

```
> d13 <- data.frame(
```

```
+ env = rep(c("Bathinda", "Hisar"), each = 9), # Two environments
```

```
+ treat = rep(c(1, 2, 3), times = 6), # Three treatments
```

```
+ replication = rep(1:3, times = 6), # Three replications
```

```
+ syield = c(
```

```
+ 1794, 1134, 718, # Bathinda, treat 1, 2, 3
```

```
+ 3286, 2518, 757, # Hisar, treat 1, 2, 3
```

```
+ 1500, 1200, 850, # Bathinda, treat 1, 2, 3
```

```
+ 3100, 2200, 650, # Hisar, treat 1, 2, 3
```

```
+ 1700, 1400, 900, # Bathinda, treat 1, 2, 3
```

```

+ 3000, 2100, 800 # Hisar, treat 1, 2, 3
+ )
+ )
>

> # Check the structure of the data

> str(d13)

'data.frame': 18 obs. of 4 variables:
 $ env      : chr  "Bathinda" "Bathinda" "Bathinda" "Bathinda" ...
 $ treat    : num  1 2 3 1 2 3 1 2 3 1 ...
 $ replication: int  1 2 3 1 2 3 1 2 3 1 ...
 $ syield   : num  1794 1134 718 3286 2518 ...

>

> # Convert 'env', 'treat', and 'replication' to factors

> d13$env <- factor(d13$env)

> d13$treat <- factor(d13$treat)

> d13$replication <- factor(d13$replication)

>

> # Perform three-way ANOVA

> aov_model <- aov(syield ~ env * treat * replication, data = d13)

> summary(aov_model)

          Df Sum Sq Mean Sq F value Pr(>F)
env         1  243369   243369    0.622 0.4455
treat        2 7965541 3982771   10.182 0.0026 **
env:treat     2  124652    62326    0.159 0.8545
Residuals   12 4693942   391162
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

> # Perform Tukey HSD for the 'treat' factor

```

```

> tukey_result <- TukeyHSD(aov_model, "treat")

> summary(tukey_result)

      Length Class  Mode
treat 12      -none- numeric

> print(tukey_result)

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = syield ~ env * treat * replication, data = d13)

$treat

      diff      lwr      upr      p adj
2-1  -638.0 -1601.344  325.34365 0.2219005
3-1 -1617.5 -2580.844 -654.15635 0.0020087
3-2  -979.5 -1942.844  -16.15635 0.0462118

> tukey_result$treat

      diff      lwr      upr      p adj
2-1  -638.0 -1601.344  325.34365 0.221900486
3-1 -1617.5 -2580.844 -654.15635 0.002008745
3-2  -979.5 -1942.844  -16.15635 0.046211790

>

> install.packages("lme4", dependencies = TRUE, type = "binary")

Error in install.packages : Updating loaded packages

> # Fit a Mixed Effects Model using lme4

> library(lme4)

>

> # Fit a fixed-effects model without any random effects

> lm3_fixed = lm(syield ~ treat, data = d13)

```

```
>
> # Check the summary of the fixed-effects model
> summary(lm3_fixed)
```

Call:

```
lm(formula = syield ~ treat, data = d13)
```

Residuals:

Min	1Q	Median	3Q	Max
-896.67	-508.67	-0.67	416.33	889.33

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2396.7	237.2	10.106	4.35e-08 ***
treat2	-638.0	335.4	-1.902	0.076517 .
treat3	-1617.5	335.4	-4.823	0.000224 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 580.9 on 15 degrees of freedom

Multiple R-squared: 0.6114, Adjusted R-squared: 0.5596

F-statistic: 11.8 on 2 and 15 DF, p-value: 0.0008336

```
>
```

```
>
```

```
> # Perform Type I ANOVA for the mixed effects model
```

```
> anova(lm3_fixed)
```

Analysis of Variance Table

Response: syield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treat	2	7965541	3982771	11.802	0.0008336 ***
Residuals	15	5061963	337464		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

```
> install.packages("car", dependencies = TRUE, type = "binary")
```

Error in install.packages : Updating loaded packages

```
> # Perform Type III ANOVA for the mixed effects model
```

```
> library(car)
```

```
> Anova(lm3_fixed, type = "III")
```

Anova Table (Type III tests)

Response: syield

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	34464067	1	102.127	4.35e-08 ***
treat	7965541	2	11.802	0.0008336 ***
Residuals	5061963	15		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>

```
> # Display summary of the mixed effects model
```

```
> summary(lm3_fixed)
```

Call:

```
lm(formula = syield ~ treat, data = dl3)
```

Residuals:

Min	1Q	Median	3Q	Max
-896.67	-508.67	-0.67	416.33	889.33

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2396.7	237.2	10.106	4.35e-08 ***
treat2	-638.0	335.4	-1.902	0.076517 .
treat3	-1617.5	335.4	-4.823	0.000224 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 580.9 on 15 degrees of freedom

Multiple R-squared: 0.6114, Adjusted R-squared: 0.5596

F-statistic: 11.8 on 2 and 15 DF, p-value: 0.0008336

>

> # Perform post-hoc analysis using lsmeans (emmeans is a modern equivalent)

> library(lsmeans) # Or 'emmeans' if lsmeans is outdated

The 'lsmeans' package is now basically a front end for 'emmeans'.

Users are encouraged to switch the rest of the way.

See help('transition') for more information, including how to

convert old 'lsmeans' objects and scripts to work with 'emmeans'.

> lsmeans(lm3_fixed, "treat")

treat	lsmean	SE	df	lower.CL	upper.CL
1	2397	237	15	1891	2902
2	1759	237	15	1253	2264
3	779	237	15	274	1285

Confidence level used: 0.95

Let's break down the analysis you performed, step by step:

1. Dataset Creation

- You created a dataset `d13` containing 18 observations across four variables:
 - `env`: Environment (Bathinda, Hisar)
 - `treat`: Treatment (3 levels)
 - `replication`: Replication (3 levels)
 - `syield`: Response variable (presumably the yield of some crop or measurement)

2. Structure of the Data

The dataset contains 18 rows with the following columns:

- `env`: The environment where the treatments are applied (2 levels: "Bathinda" and "Hisar").
- `treat`: The treatment applied (3 levels: 1, 2, 3).
- `replication`: The replication number for each combination (3 levels: 1, 2, 3).
- `syield`: The observed yield for each combination of `env`, `treat`, and `replication`.

3. Three-Way ANOVA

You performed a three-way ANOVA to assess the effects of:

- `env`, `treat`, and `replication` on `syield`.

The summary of the ANOVA gives the following key results:

- `env`: Not significant (p-value = 0.4455). The environment does not significantly affect the yield.
- `treat`: Significant (p-value = 0.0026). The treatment has a significant effect on the yield.

- **env:treat:** Not significant (p-value = 0.8545). There is no significant interaction between environment and treatment.
- **Residuals:** The residual sum of squares provides the remaining variability after accounting for the effects of **env**, **treat**, and **replication**.

4. Tukey HSD Test for **treat** Factor

The Tukey HSD test was conducted to compare the means between different levels of the **treat** factor:

- **treat 2 vs 1:** Not significant (p-value = 0.2219).
- **treat 3 vs 1:** Significant (p-value = 0.0020). Treatment 3 differs significantly from treatment 1.
- **treat 3 vs 2:** Significant (p-value = 0.0462). Treatment 3 differs significantly from treatment 2.

This shows that Treatment 3 significantly reduces **syield** compared to both Treatment 1 and Treatment 2.

5. Linear Model (**lm3_fixed**)

You fitted a fixed-effects linear model using **lm()** with **treat** as the independent variable and **syield** as the dependent variable.

The summary of the fixed-effects model provides:

- **Intercept:** Estimate of 2396.7 with a p-value of less than 0.001, which is significant.
- **treat2:** Estimate of -638, which is not statistically significant at the 0.05 level (p-value = 0.0765).
- **treat3:** Estimate of -1617.5, which is statistically significant (p-value = 0.00022).

The model explains about **61.14%** of the variance in **syield** (R-squared = 0.6114).

6. Type I and Type III ANOVA for Mixed Effects Model

- **Type I ANOVA:** This is for the fixed-effects model. It shows that **treat** is highly significant (p-value = 0.00083).

- **Type III ANOVA:** This tests the same fixed-effects model, confirming that `treat` remains significant (p-value = 0.00083).

7. Post-Hoc Analysis using `lsmeans` (or `emmeans`)

You performed post-hoc analysis using `lsmeans` to estimate the marginal means for the `treat` factor:

- **treat 1:** Estimated mean of 2397.
- **treat 2:** Estimated mean of 1759.
- **treat 3:** Estimated mean of 779.

The confidence intervals for these means confirm that `treat 3` is significantly lower than the other treatments.

Key Conclusions:

- **Treatment Effects:** Treatment 3 has the most significant negative impact on `syield`, reducing it compared to Treatment 1 and 2.
- **Interaction Effects:** There is no significant interaction between the environment and treatment (`env:treat`), meaning that the effect of treatment on yield is consistent across environments.
- **Statistical Significance:** Treatment differences are significant, with Treatment 3 showing the largest reduction in yield.