Elementry Statistics

Raw Data

Raw data refers to the original, unprocessed data collected from various sources .It's the data in its most basic form, without any modifications or analysis. Think of it as the raw ingredients before they are cooked into a meal.

Variable in R

A variable in R is like a container that holds data.You can store different type of data in variables, such as numbers, text, or more complex data structures.For example:

```
x <- 10;x            # Here, x is a variable storing the number 10
```

```
## [1] 10
```

```
name <- "Alice";name # Here, name is a variable storing the text "Alice"
```

```
## [1] "Alice"
```

Numeric Variables

Numeric variables store numbers and can be either integers or real numbers (with decimals).They are used for mathematical calculations. For example:

```
age <- 25;age           # Integer
```

```
## [1] 25
```

```
height <- 175.5;height  # Real number
```

```
## [1] 175.5
```

Categorical Variables

Categorical variables represent categories or groups. They can be stored as factors or characters in R. Factors are useful for statistical modeling because they can have a fixed set of possible values (levels).For example:

```r
gender <- factor(c("male", "female", "female", "male"))
gender
```

```
## [1] male   female female male
## Levels: female male
```

String Variables

String variables (also known as character variables) store text data. They are used for any data that is not numeric, such as names or descriptions.For example:

```r
name <- "Alice"
name
```

```
## [1] "Alice"
```

```r
description <- "This is a sample text."
description
```

```
## [1] "This is a sample text."
```

Univariate and Bivaraite Data

Uni-variate data involves analyzing single variable.The main goal is to describe the data and find patterns within it.Here are some common methods for univariate analysis in R:

1. Summary Statistics: These include measures like mean, median, mode, variance, and standard deviation.

```r
data <- c(5, 10, 15, 20, 25)
summary(data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5      10      15      15      20      25
```

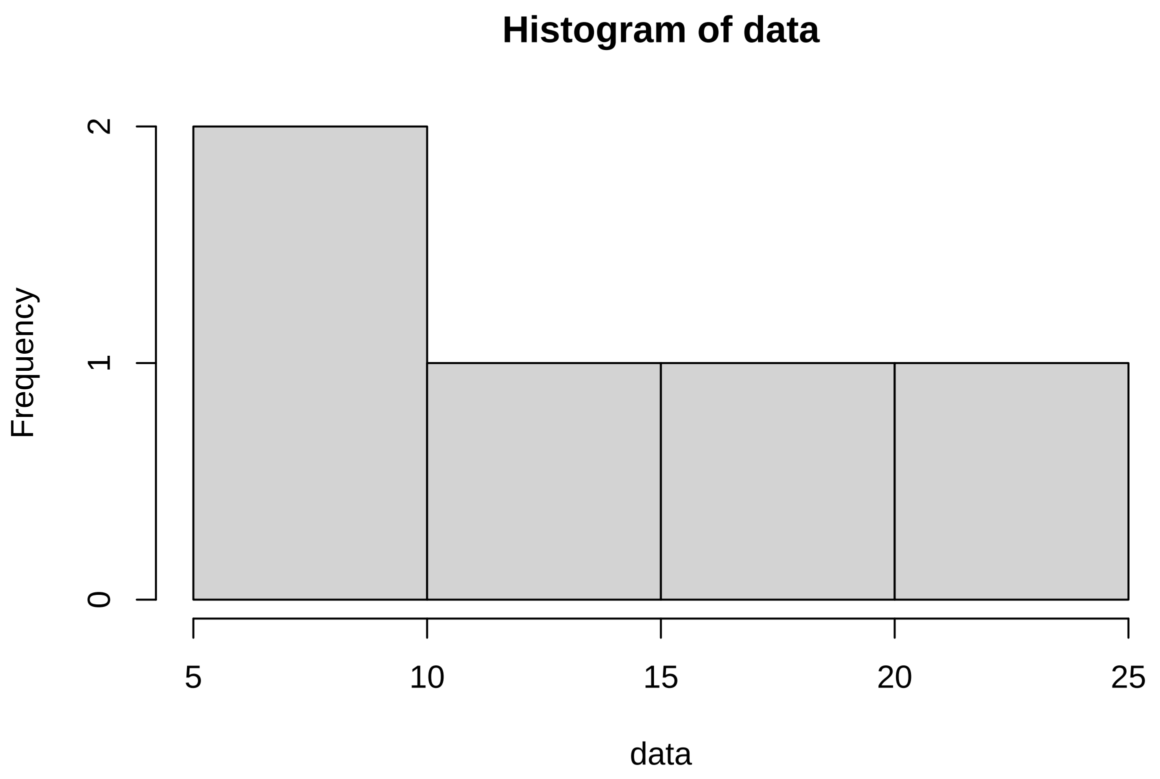
2. Frequency Tables: These show how often each value occurs.

```r
table(data)
```

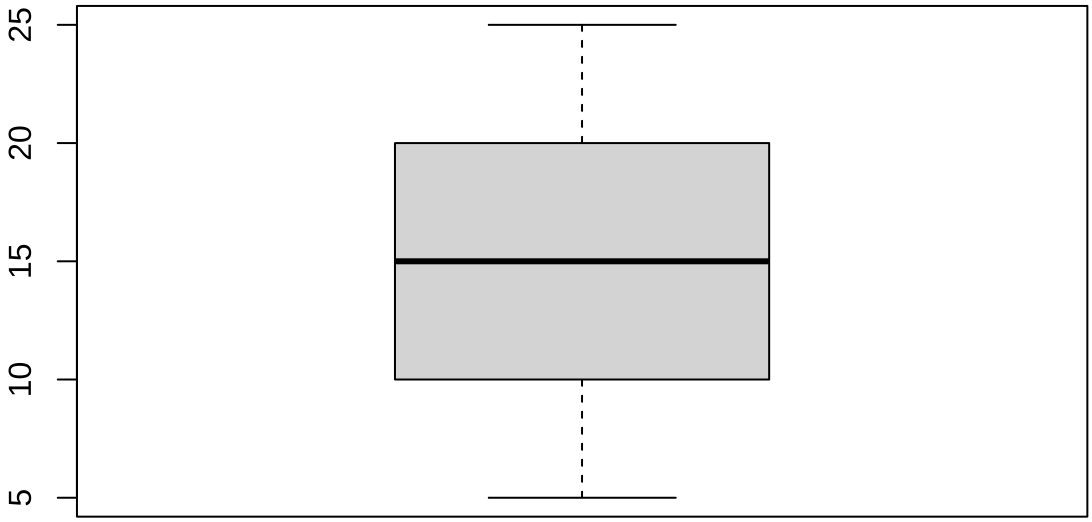
```
## data
##  5 10 15 20 25
##  1  1  1  1  1
```

3. Charts: Visualizations like histograms, box plots,and bar charts help to understand the distribution.

```r
hist(data)
```



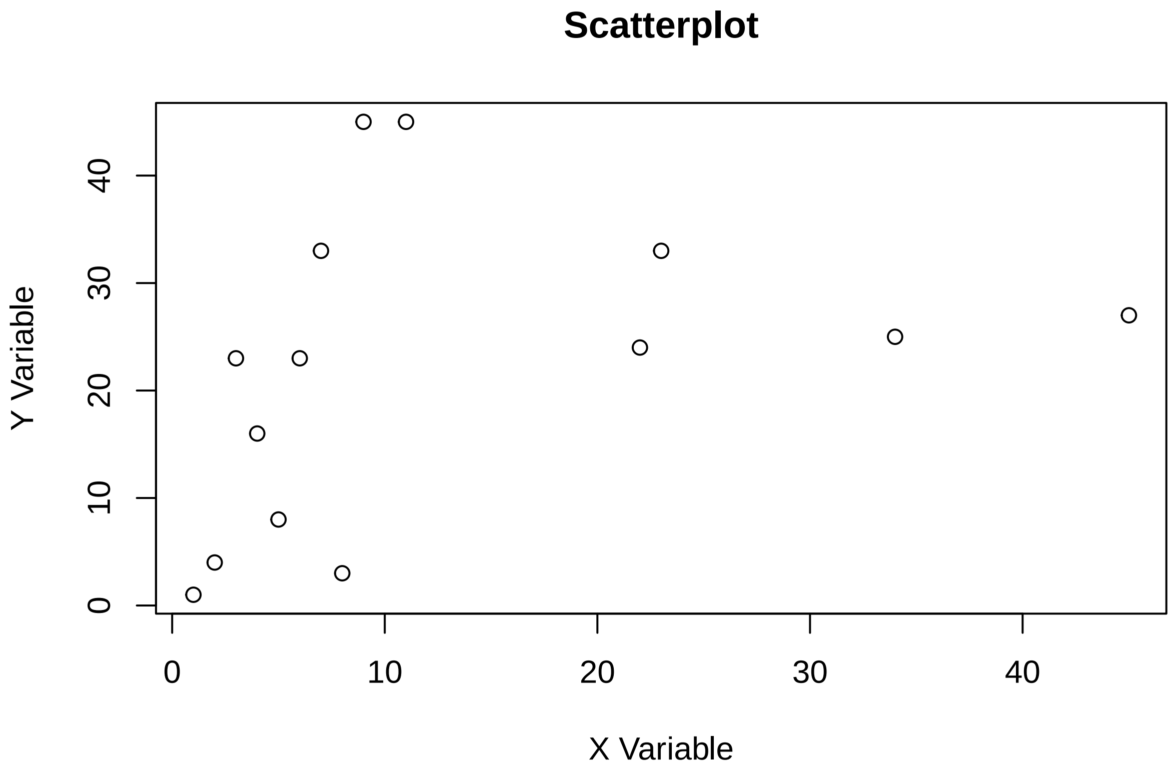
```r
boxplot(data)
```



Bivariate data involves analyzing the relationship between two variables. Here are some common methods for bivariate analysis in R:

1. **Scatter plots** These plots show the relationship between two continuous variables.

```
X<-c(1,2,4,6,8,9,7,5,3,11,23,22,34,45)
Y<-c(1,4,16,23,3,45,33,8,23,45,33,24,25,27)
plot(X, Y, main="Scatterplot", xlab="X Variable",
     ylab="Y Variable")
```



2. **Correlation Coefficients** measures the strength and direction of the relationship between two variables.

```
cor(X, Y)
```

```
## [1] 0.3387931
```

3. **Simple Linear Regression** models the relationship between two variables by fitting a linear equation.

```
model <- lm(Y ~ X)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.355 -10.182  -2.195   6.464  24.277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.4106     5.3486   3.255  0.00689 **
## X             0.3681     0.2951   1.247  0.23605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.11 on 12 degrees of freedom
## Multiple R-squared:  0.1148, Adjusted R-squared:  0.04101
## F-statistic: 1.556 on 1 and 12 DF,  p-value: 0.2361
```

Parameter

A parameter is a value that describes a characteristic of an entire population. Since it's often impractical to measure every member of a population, parameters are usually unknown and are estimated using statistics.Parameters are denoted by Greek letters.For example:

1. **Population mean( )**: The average of all values in the population.
2. **Population standard deviation( )**: The measure of the spread of values in the population.

Statistic

A statistic is a value that describes a characteristic of a sample, which is a subset of the population.Statistics are used to estimate parameters.They are denoted by Latin letters.For example:

1. **Sample mean ($\bar{x}$)** The average of all values in the sample.
2. **Sample standard deviation (s)**: The measure of the spread of values in the sample.

```r
# Sample data: heights of 30 students (in cm)
heights <- c(160, 162, 165, 170, 172, 168, 167, 169, 171, 173,
             160, 161, 164, 166, 170, 175, 176, 178, 180, 182,
             160, 162, 165, 170, 172, 168, 167, 169, 171, 173)

# Calculate sample mean (statistic)
sample_mean <- mean(heights)
sample_sd <- sd(heights)

# Descriptive Statistics
# Display results
sample_mean
```

```
## [1] 168.8667
```

```r
sample_sd
```

```
## [1] 5.864759
```

Mean

The mean is the average of a set of numbers. It's calculated by summing all the values and dividing by the number of values.

```r
data <- c(1, 2, 3, 4, 5)
mean(data)
```

```
## [1] 3
```

Median

The median is the middle value in a sorted list of numbers. If the list has an even number of values, the median is the average of the two middle numbers.

```r
median(data)
```

```
## [1] 3
```

Mode

The mode is the value that appears most frequently in a data set.R doesn't have a built-in function for mode, but you can create one.

```r
mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
mode(data)
```

```
## [1] 1
```

Percentage

Percentage represents a proportion out of 100. You can calculate it by dividing the part by the whole and multiplying by 100.

```r
part <- 25
whole <- 200
percentage <- (part / whole) * 100
percentage
```

```
## [1] 12.5
```

Count

Count is the number of observations in a data set.

```
length(data)
```

```
## [1] 5
```

Proportion

Proportion is the fraction of the total that a particular group represents.

```
group <- c(1, 1, 2, 2, 2, 3, 3)
prop.table(table(group))
```

```
## group
##         1         2         3 
## 0.2857143 0.4285714 0.2857143
```

Quantiles

Quantiles divide a data set into equal-sized intervals.The most common quantiles are quartiles (dividing data into four parts).

```
quantile(data)
```

```
##   0%  25%  50%  75% 100% 
##    1    2    3    4    5
```

Percentiles

Percentiles indicate the value below which a given percentage of observations fall.

```
quantile(data, probs = c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75% 
##   2   3   4
```

Five Summary Statistics

The five-number summary includes the minimum, first quartile (Q1),median, third quartile (Q3), and maximum.

```
summary(data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       1       2       3       3       4       5
```

Variance

Variance measures the spread of a set of numbers.It's the average of the squared differences from the mean.

```r
var(data)
```

```
## [1] 2.5
```

Standard Deviation

Standard deviation is the square root of the variance and provides a measure of the spread of values around the mean.

```r
sd(data)
```

```
## [1] 1.581139
```

Interquartile Range (IQR)

IQR is the range between the first quartile (Q1) and the third quartile (Q3). It measures the spread of the middle 50% of the data.

```r
IQR(data)
```

```
## [1] 2
```

Covariance

Covariance measures the relationship between two variables.A positive covariance indicates that the variables tend to move in the same direction,while a negative covariance indicates they move in opposite directions.

```r
x <- c(1, 2, 3, 4, 5)
y <- c(2, 4, 6, 8, 10)
cov(x, y)
```

```
## [1] 5
```

Correlation

Correlation measures the strength and direction of the relationship between two variables. It ranges from -1 to 1.

```r
cor(x, y)
```

```
## [1] 1
```

Outliers

Outliers are data points that are significantly different from the rest of the data. You can identify them using box plots or statistical tests.

```r
boxplot(data)
```