R – Programming 4th and 5th chapter Topics

Sampling Distribution

- Sampling distribution is the probability distribution of a statistic (such as a sample mean, proportion) that is obtained from repeatedly taking samples from a population.
- It shows how the statistic varies from sample to sample and helps us understand the spread and behaviour of the estimate around the true population parameter.

Distribution for a Sample Mean:

- The Distribution for a Sample Mean describes how the average value (mean) calculated from different random samples of the same size, taken from a population, varies.
- When we collect data from a population, we often take samples because it's more practical than measuring every single individual.
- The sample mean (average of the data in a sample) helps us estimate the true population mean**.**
- However, different samples can give slightly different means. Understanding how these sample means behave helps us make accurate predictions about the population.

Key Concepts:

1. Sample Mean ($\bar{X}$):
    - This is the average value of a sample, denoted as $\bar{X}$.
    - If you repeatedly take samples of the same size from a population and calculate their means, you will have many sample mean.
2. Distribution of the Sample Means:
    - The collection of all these sample means forms the sampling distribution of the sample mean.
    - This distribution shows how much the sample means vary from one sample to another.
    - Mean of the Sampling Distribution ($\mu_{\bar{X}}$): The mean of all the sample means is the estimation of the true population mean (μ)**.**

Distribution for a Sample Proportion

- The sampling distribution for a sample proportion refers to the distribution of the proportion of successes\failures that you calculate from multiple random samples taken from the population.

Key Concepts:

1. Sample Proportion ($\hat{p}$)

- For each sample you collect, the sample proportion is calculated as:

$$\hat{p} = x/n$$

- Where:
    - x is the number of successes (e.g., people who prefer the coffee brand).
    - n is the sample size.

2. Mean of the Sampling Distribution:

- The mean of the sampling distribution of the sample proportion ($\hat{p}$) is estimation of the true population proportion (π).

Standard Error (SE)

Standard Error quantifies how much a sample statistic (e.g., sample mean, sample proportions) is expected to vary from the true population parameter (e.g., population mean). It is the standard deviation of the sampling distribution.

- Factors that affect SE:
    - Sample Size (n): As the sample size increases, the SE decreases, meaning the estimate becomes more stable and accurate.
    - Population Variability (σ): Greater variability in the population results in a higher SE, indicating more potential for variation in estimates.

Confidence Interval

- A Confidence Interval is a range between a lower limit (l) and an upper limit (u) used to estimate a population parameter (like the mean) based on data from a sample.
- For example, if a survey estimates the average height of adults as between 165 cm and 175 cm with a 95% confidence level, it means we are 95% confident the true average height is within this range.

1. Confidence Level:
    - The level of confidence indicates how sure we are that the interval contains the true parameter.
    - Common confidence levels are 90%, 95%, and 99%.
2. Formula for a Confidence Interval:
    - For symmetrically distributed sample statistics, like means and proportions, a general formula is
    - **Confidence intervals = statistic ± critical value × standard error**
        - Statistic: The sample statistic (e.g., sample mean).
        - Critical Value: Based on the desired confidence level.
        - Standard Error (SE): Measures how much the sample statistic is expected to vary. For the mean, it is calculated as: **$s/\sqrt{n}$**

An Interval for a Mean

- To construct a confidence interval (CI) for a sample mean, the following steps and considerations are involved:

1. Sampling Distribution:
    - If the sample size is $n \geq 30$ (large enough), the Central Limit Theorem (CLT) suggests that the sampling distribution of the sample mean will be approximately normal.
    - If the true population standard deviation (σx) is known, use the normal (z-) distribution.
    - If σx is unknown, use the sample standard deviation (s) and apply the t-distribution
2. Calculate the Standard Error (SE):
    - The standard error of the sample mean is calculated as: SE= **$s/\sqrt{n}$**
    - Where s is the sample standard deviation, and n is the sample size.
3. Find the Critical Value:
    - The critical value depends on the chosen confidence level (e.g., 1.96 for a 95% CI using the z-distribution or a value from the t-table for smaller samples).
4. Construct the Confidence Interval:

- The formula for the confidence interval is **Confidence intervals = statistic ± critical value × standard error**

Interval for proportions

- To create a confidence interval (CI) for a sample proportion, follow these steps:

1. Sample Proportion (p^):
   - This is the estimated proportion based on your sample data (e.g., the proportion of people who prefer a certain product).
2. Standard Error (SE) Calculation:
   - Calculate the standard error (SE) for the sample proportion using the formula: **√ p^(1 – p^)/n.**
3. Critical Value:
   - The critical value is determined based on the chosen confidence level (e.g., 1.96 for a 95% confidence interval using the normal distribution).
4. Constructing the Confidence Interval:
   - Use the formula: CI=p^±(Critical Value)×SE

Hypothesis testing

- Hypothesis testing is a statistical method used to make decisions about a population parameter based on sample data.
- It helps determine if there is enough evidence to support or reject a specific claim about a population.

Components of Hypothesis.

1. Hypotheses
   - Hypothesis testing is a statistical method that uses data to evaluate two competing hypotheses.
   - The two hypotheses are the
     - Null hypothesis (H0) :- The null hypothesis is often (but not always) defined as an equality, =, to a null value
     - alternative hypothesis (Ha): - the alternative hypothesis, is often defined in terms of an inequality to the null value.
       - When HA is defined in terms of a less-than statement, with < , it is one – sided; this is also called as lower- tailed test.
       - When HA is defined in terms of a greater-than statement, with >, it is one-sided; this is also called an upper-tailed test.
       - When HA is merely defined in terms of a different-to statement, with it is two-sided; this is also called a two-tailed test.

2. Test Statistic
   - Once the hypotheses are formed, sample data are collected, and statistics are calculated according to the parameters detailed in the hypotheses. The test statistic is the statistic that's compared to the appropriate standardized sampling distribution to yield the p-value.
3. P-value
   - The p-value is the probability value that's used to quantify the amount of evidence, if any, against the null hypothesis. The exact nature of calculating a p-value is dictated by the type of statistics being tested and the nature of HA.
4. Significance level

- For every hypothesis test, a significance level, denoted α, is assumed.
- This is used to qualify the result of the test.
- The significance level defines a cutoff point, at which you decide whether there is sufficient evidence to view H0 as incorrect and favor HA instead.
- If the p-value is greater than or equal to α, then you conclude there is insufficient evidence against the null hypothesis, and therefore you retain H0 when compared to HA
- If the p-value is less than α, then the result of the test is statistically significant. This implies there is sufficient evidence against the null hypothesis, and therefore you reject H0 in favor of HA.
- Common or conventional values of α are α = 0.1, α = 0.05, and α = 0.01.

Linear Regression

- Linear Regression is a statistical method used to understand and model the relationship between one or more independent variables (predictors) and a dependent variable (target). There are two main types:
  - **Simple Linear Regression**: This involves one independent variable and one dependent variable.
  - **Multiple Linear Regression**: This involves two or more independent variables and one dependent variable.

Simple Linear Regression

- Simple Linear Regression is a method used to predict the value of one variable (called the dependent variable) based on the value of another variable (called the independent variable). It helps find a straight-line relationship between the two variables.

General terms in Simple Linear Regression

Defining the Model

- The goal here is to express the relationship between the independent variable (x) and the dependent variable (y) using the linear equation: **y=β0+β1x**
- Where,
  - y: The predicted value of the dependent variable.
  - β0: - The y-intercept of the line (value of y when x=0).
  - β1: - The slope of the line (how much y changes for a one-unit increase in x).
  - x: The value of the independent variable.

Estimating the Intercept (β0) and Slope (β1) Parameters:

- This step involves finding the values of β0 (intercept) and β1 (slope) that make the line "fit" the data as closely as possible.
- The Ordinary Least Squares (OLS) method is typically used for this purpose to minimize the error (difference) between the observed values and the values predicted by the line.

Formulas to Calculate β_1 and β_0:

- Slope (β_1):

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- Intercept (β_0):

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

- $\bar{x}$ and $\bar{y}$ are the mean (average) values of the independent and dependent variables.

An Example of a Linear Relationship

```
# Sample Data
experience <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)  # years of experience
salary <- c(30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000)  # salary in dollars
data <- data.frame(experience, salary)

# Displaying the data
print(data)
# Fit the Simple Linear Regression Model
model <- lm(salary ~ experience, data = data)

# Display the model summary
summary(model)
```

- Creating Sample Data: Let's create a dataset with salary and experience.
- Fitting the Simple Linear Regression Model: Use the lm() function in R to fit a linear model where salary is the dependent variable, and experience is the independent variable.
- Interpreting the Output: The summary() function provides key details about the model:

Example Plot

<u>Multiple Linear Regression</u>

- Multiple Linear Regression is an extension of Simple Linear Regression and is used when you want to model the relationship between a dependent variable and two or more independent variables.

Defining the Model

- The goal here is to express the relationship between the dependent variable (Y) and multiple independent variables (X1,X2,…,Xp) using a linear equation of the form:

  $$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$$

- Where,
  - Y: The predicted value of the dependent variable.

- β0: The intercept of the model, representing the predicted value of Y when all independent variables (X1,X2,…,Xp) are zero.
- β1,β2,…,βp: The coefficients (or parameters) of the model. Each βi represents the change in Y for a one-unit change in the corresponding Xi, holding all other variables constant.
- X1,X2,…,Xp: The independent variables (predictors or features) that influence the dependent variable Y.
- ϵ: The error term (or residual), which represents the difference between the predicted value and the actual observed value of Y. It captures the variability in Y that is not explained by the predictors

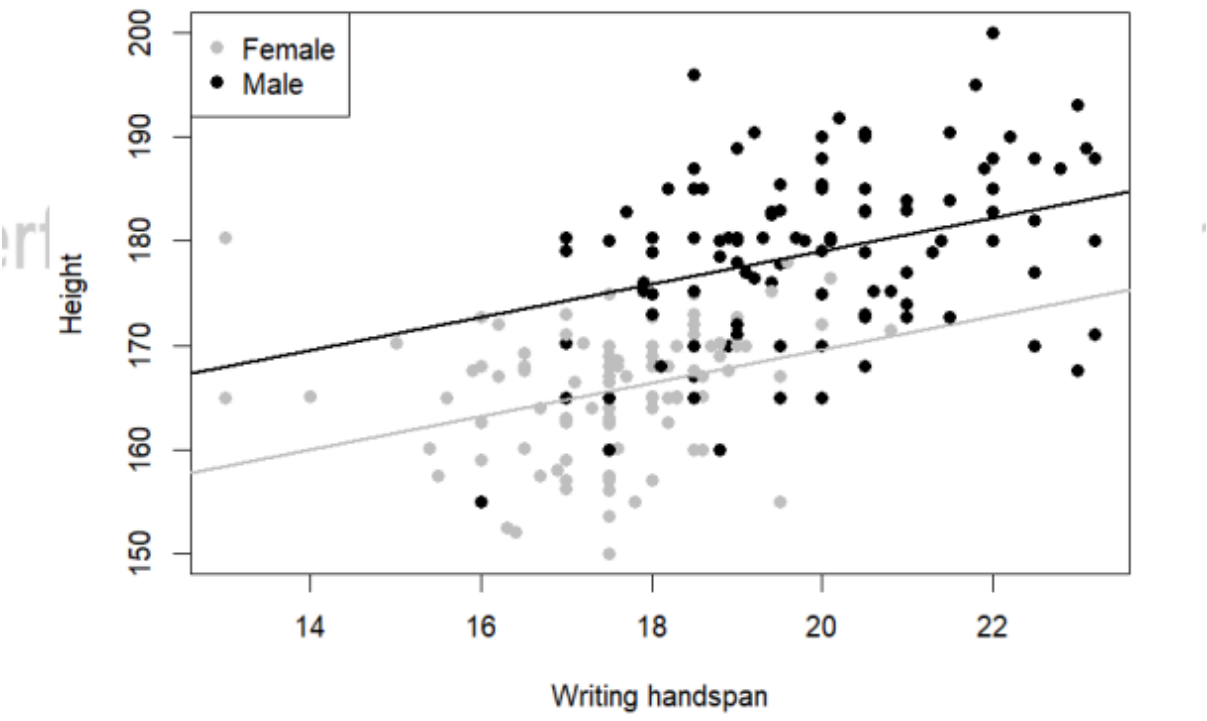
An example of Multiple Linear Regression

```
# Sample data
study_hours <- c(5, 3, 8, 2, 7, 4, 6, 1, 9, 10)  # Hours spent studying
attendance <- c(90, 70, 95, 60, 85, 75, 88, 50, 98, 100)  # Attendance rate (%)
final_score <- c(85, 65, 90, 60, 80, 70, 83, 55, 95, 98)  # Final exam score

# Combine into a data frame
student_data <- data.frame(study_hours, attendance, final_score)
print(student_data)  # View the data
# Fit the model
model <- lm(final_score ~ study_hours + attendance, data = student_data)

# Display the summary of the model
summary(model)
```

Example Plot



Linear Model selection and diagnostics

- ❖ Linear Model Selection is the process of choosing the best linear regression model that balances goodness-of-fit and model complexity.
- ❖ The objective is to create a model that explains the relationship between predictor variables and the response variable while avoiding overfitting or underfitting.

1. **Goodness-of-Fit** refers to how well the model captures the patterns in the data.
    - High goodness-of-fit means the model predicts the response variable accurately, while low goodness-of-fit indicates poor prediction ability.

2. **Model Complexity** involves the number of predictors and the model's structure. A model can be simple (few predictors) or complex (many predictors, polynomial terms, or interactions). A complex model may overfit, while a simple model might miss important relationships (underfitting).
3. **Principle of Parsimony:** The best model is the simplest one that explains the data well. Adding too many predictors can make the model unnecessarily complex, while removing important ones can make the model less accurate.
4. **Model selection algorithms are key to identifying the best model**:
    - o Forward Selection: Starts with no predictors and adds the most significant ones based on statistical criteria like p-values or AIC.
    - o Backward Elimination: Begins with all predictors and removes the least significant ones until the best subset is identified.
    - o Stepwise Selection: A combination of forward and backward selection, where predictors can be both added and removed to find the optimal model.

- ❖ Linear Model Diagnostics are methods used to assess the adequacy of a selected linear model.
- ❖ The goal is to ensure that the model assumptions hold and to identify potential issues that could undermine its validity.

1. **Residual Analysis**: Residuals (differences between observed and predicted values) are examined for randomness.
    - A good model should have residuals that are randomly scattered, indicating no unaccounted patterns.
    - Plots like **Residual vs. Fitted Values** and **Normal Q-Q Plots** help diagnose potential problems like non-linearity or non-normality of errors.
2. **Multicollinearity**: High correlation between predictors can destabilize the model. The Variance Inflation Factor (VIF) is used to check multicollinearity, where values above 10 indicate problematic predictors.
3. **Homoscedasticity**: The assumption that the residuals have constant variance is checked using the Residual vs. Fitted Plot. If the variance of residuals changes at different levels of the predicted values, it suggests heteroscedasticity, which can affect model validity.
4. **Influential Points:** Outliers or influential data points can disproportionately affect model coefficients. Cook's Distance is used to identify such points, and they may need to be removed if they heavily influence the model.
5. **Linearity:** A key assumption in linear regression is the linear relationship between predictors and the response. Scatterplots or polynomial transformations of predictors can be used to check this assumption.

In conclusion, model selection helps identify the best-fitting model, while model diagnostics ensure that the model assumptions are valid and the results are reliable. Combining both processes leads to robust and accurate linear regression models.