STAT 3011: Discussion 015

Week 2

Talha Hamza
University of Minnesota
College of Science & Engineering

Fall 2025

Creating Objects in R

Key Concepts:

- **Objects**: Store data for reuse.
- **Syntax**: Use <- or = to assign values.
- **Preferred Assignment**: <- is preferred because it follows R's convention and avoids confusion with = used in function calls.
- **Console Execution**: Type the object's name and press `Enter`.

Review: Creating and Viewing Objects

Example Code:

```
# Assigning values to objects
x <- 10
y <- 5
z <- c(1, 2, 3) # this is a vector

# Displaying object values
z
```

Environment Pane:

- Objects you create appear here.
- Useful for tracking your variables.

RStudio Layout

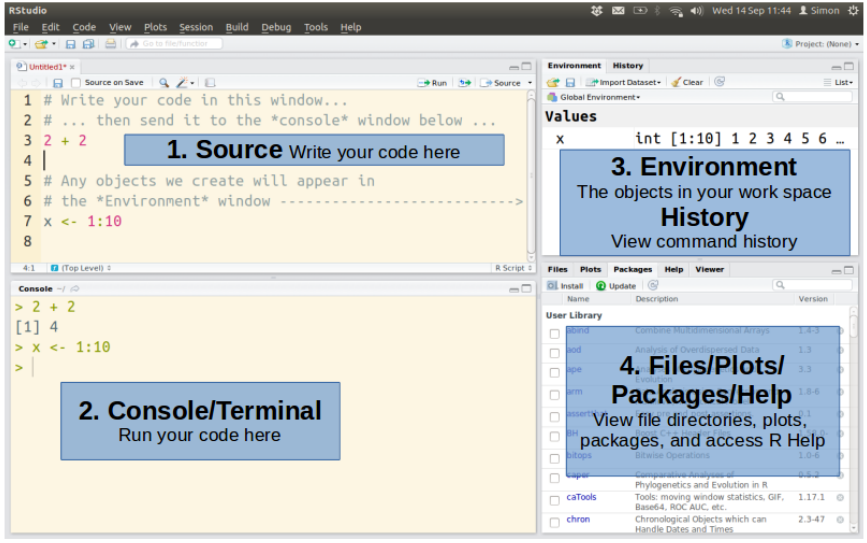

Image credit: `https://www.simonqueenborough.info/R/basic/first-look`

Key Statistical Concepts

Variable Types:

- **Categorical**: Qualitative data representing categories
- **Quantitative**: Numerical data that can be measured
  1. **Discrete**: Countable values (e.g., number of people)
  2. **Continuous**: Measurable values (e.g., weight, time)

Explanatory Variable: Variable thought to influence the response.
Response Variable: Outcome measured in a study.

Distribution Shapes:

- **Symmetric**: Balanced on both sides (e.g., normal distribution)
- **Skewed**: Asymmetric (left or right skewed)

Research Study Concepts

Study Design:

- **Population**: The entire group of interest.
- **Sample**: Subset of the population actually studied.
- **Subjects**: Individual participants in the study

Statistical Measures:

- **Parameter**: A number that summarizes a population.
- **Statistic**: A number that summarizes a sample.

Study Types:

- **Observational**: Researcher observes without intervention
- **Experimental**: Researcher assigns treatments

Exercise 3: Distribution Shapes

For each variable, identify the likely shape of the distribution:

- Salary of employees of a large company
- Heights of adult males

Exercise 3: Distribution Shapes

For each variable, identify the likely shape of the distribution:

- Salary of employees of a large company $\rightarrow$ Right-skewed distribution
- Heights of adult males

Exercise 3: Distribution Shapes

For each variable, identify the likely shape of the distribution:

- Salary of employees of a large company $\rightarrow$ Right-skewed distribution
- Heights of adult males $\rightarrow$ Normal, symmetric distribution

Exercise 3: Solution

Distribution Shapes:

- **Salary of employees**: Right-skewed distribution
  - Most employees cluster at lower salaries
  - Few executives with very high salaries pull mean upward
- **Heights of adult males**: Normal, symmetric distribution
  - Most heights cluster around the average
  - Few extremely short or tall individuals
  - Roughly symmetric on both sides of the mean

Exercise 4: Variable Types

Identify each variable as categorical or quantitative. If quantitative, continuous or discrete:

- Time to commute to work (drive, bus, walk, bike, etc.)
- The number of people in line at a box office
- The weight of a dog
- Shoe size (6, 6.5, 7, etc.)

Remember:

- Can you perform arithmetic operations on the values?
- Are the numbers used as labels or as measurable quantities?

Exercise 4: Variable Types

Identify each variable as categorical or quantitative. If quantitative, continuous or discrete:

- Time to commute to work (drive, bus, walk, bike, etc.) $\rightarrow$ Quantitative, continuous
- The number of people in line at a box office
- The weight of a dog
- Shoe size (6, 6.5, 7, etc.)

Remember:

- Can you perform arithmetic operations on the values?
- Are the numbers used as labels or as measurable quantities?

Exercise 4: Variable Types

Identify each variable as categorical or quantitative. If quantitative, continuous or discrete:

- Time to commute to work (drive, bus, walk, bike, etc.) $\rightarrow$ Quantitative, continuous
- The number of people in line at a box office $\rightarrow$ Quantitative, discrete
- The weight of a dog
- Shoe size (6, 6.5, 7, etc.)

Remember:

- Can you perform arithmetic operations on the values?
- Are the numbers used as labels or as measurable quantities?

Exercise 4: Variable Types

Identify each variable as categorical or quantitative. If quantitative, continuous or discrete:

- Time to commute to work (drive, bus, walk, bike, etc.) → Quantitative, continuous
- The number of people in line at a box office → Quantitative, discrete
- The weight of a dog → Quantitative, continuous
- Shoe size (6, 6.5, 7, etc.)

Remember:

- Can you perform arithmetic operations on the values?
- Are the numbers used as labels or as measurable quantities?

Exercise 4: Variable Types

Identify each variable as categorical or quantitative. If quantitative, continuous or discrete:

- Time to commute to work (drive, bus, walk, bike, etc.) $\rightarrow$ Quantitative, continuous
- The number of people in line at a box office $\rightarrow$ Quantitative, discrete
- The weight of a dog $\rightarrow$ Quantitative, continuous
- Shoe size (6, 6.5, 7, etc.) $\rightarrow$ Categorical, but ordered
  - Numbers represent categories, not measurable quantities
  - You wouldn't average shoe sizes meaningfully

Remember:

- Can you perform arithmetic operations on the values?
- Are the numbers used as labels or as measurable quantities?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest
- Sample
- Subjects
- Parameter
- Statistic
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest $\rightarrow$US college students
- Sample
- Subjects
- Parameter
- Statistic
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest $\rightarrow$US college students
- Sample $\rightarrow$500 UMN college students
- Subjects
- Parameter
- Statistic
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest $\rightarrow$US college students
- Sample $\rightarrow$500 UMN college students
- Subjects $\rightarrow$Individual college students
- Parameter
- Statistic
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest $\rightarrow$US college students
- Sample $\rightarrow$500 UMN college students
- Subjects $\rightarrow$Individual college students
- Parameter $\rightarrow$Average money US college students spend on decorating
- Statistic
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest $\rightarrow$US college students
- Sample $\rightarrow$500 UMN college students
- Subjects $\rightarrow$Individual college students
- Parameter $\rightarrow$Average money US college students spend on decorating
- Statistic $\rightarrow$Average money these 500 UMN students spend
- Is this a good sample? Why or why not?

Exercise 5: Sampling Concepts

Target whats to know how much money, on average, US college students spend on decorating their dorm rooms to determine if they should carry more merchandise. They reach out to students in an introductory statistics class to conduct a survey of 500 college students at the University of Minnesota

Identify:

- Population of interest →US college students
- Sample →500 UMN college students
- Subjects →Individual college students
- Parameter →Average money US college students spend on decorating
- Statistic →Average money these 500 UMN students spend
- Is this a good sample? Why or why not? →Not ideal, concentrated on one demographic region. Sample should be more random and representative of the national population

Exercise 6: Study Design

Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After a year, each group is given an achievement test to compare proficiency.

Identify:

- Explanatory variable
- Response variable
- Type of study (observational or experimental)

Exercise 6: Study Design

Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After a year, each group is given an achievement test to compare proficiency.

Identify:

- Explanatory variable $\rightarrow$Teaching method (traditional vs. reform)
- Response variable
- Type of study (observational or experimental)

Exercise 6: Study Design

Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After a year, each group is given an achievement test to compare proficiency.

Identify:

- Explanatory variable $\rightarrow$Teaching method (traditional vs. reform)
- Response variable $\rightarrow$Score on achievement test
- Type of study (observational or experimental)

Exercise 6: Study Design

Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After a year, each group is given an achievement test to compare proficiency.

Identify:

- Explanatory variable $\rightarrow$Teaching method (traditional vs. reform)
- Response variable $\rightarrow$Score on achievement test
- Type of study (observational or experimental) $\rightarrow$Experimental study

Lab 1: R Cheat Sheet

Basic Vector Operations:

- `vec0 <- c(1, 121, 100)` Create vector with specific values
- `vec1 = 3:100` Create sequence vector
- `length(vec1)` Get number of elements
- `mean(vec1)` Calculate mean
- `var(vec1)` Calculate variance

Vectorized Operations:

- `vec2 <- vec0*2` Multiply all elements by 2
- `vec3 <- vec0 + 10` Add 10 to all elements
- `vec4 <- vec0^2` Square all elements
- `vec23 <- vec2*vec3` Element-wise multiplication

Lab 1: R Cheat Sheet Conitnued

Data Import & Manipulation:

- `hur0 <- read.csv("path/to/file.csv", header = TRUE)` Import CSV file
- `hur0 <- read.csv(file.choose(), header = TRUE)` Interactive file selection
- `dim(hur0), nrow(hur0), ncol(hur0)` Dataset dimensions
- `names(hur0)` View column names
- `head(hur0), tail(hur0)` View first 6/last 6 rows

Subsetting Data:

- `hur3 <- hur0[,3:5]` Select columns 3-5
- `hur <- hur3[-c(193:235),]` Remove rows 193-235
- `hur$MaxWind` Access MaxWind column within hur dataframe

Visualization:

- `hist(hur$MaxWind)` Create histogram
- `?hist` Get function help

Questions?