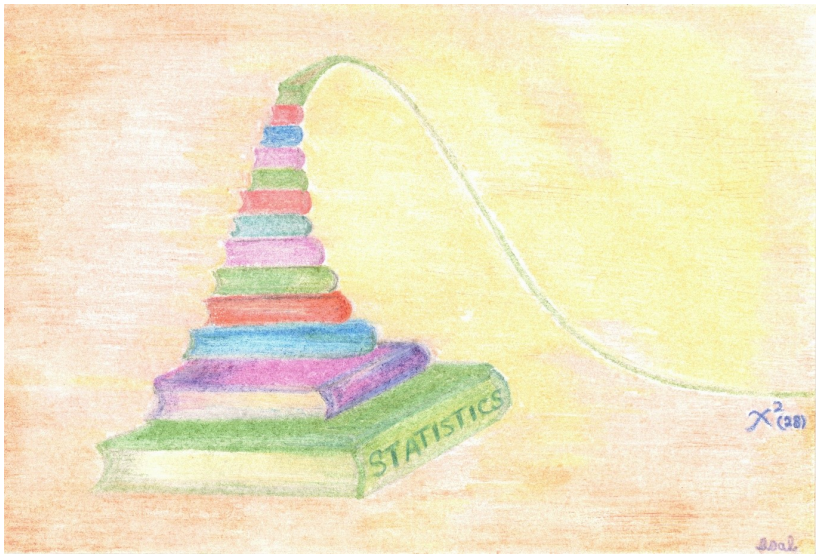
Data Analysis Activities

Asal Aslemand and Jaiditya Dev
University of Toronto, Mississauga

Contents

Introduction



This book contains a variety of statistical activities using real-life data sets. It is intended for developing statistical reasoning at an introductory to an intermediate level by explaining patterns in the data, interpreting and communicating results within the context of data, and making data-based arguments and inference. Each chapter corresponds to one statistical activity with a set of learning objectives.

Click on a link below to access a specific statistical activity.

Note: Solution to each statistical activity is shared at the instructor's discretion.

If you would like to learn how to set up R on your local computer to work on these activities, please visit the How to Set Up R for Data Analysis Activities page.

Chapter 1

Exploring Categorical Data

1.1 Activity Objective

The objective of this module is to equip you with the following:

- Use R/RStudio on a desktop/laptop to simulate confidence intervals for population parameters of interest.
- Construct plots and frequency tables for describing categorical data.
- Assess association between variables.

1.2 Context of Data

Survey of Consumer Finances (SCF, 2013) took a random sample of 6,015 adult Canadians and collected information on their level of education and whether or not they own a home. The variables "Education Level" and "Home Ownership" are measured as follows:

1. **Education Level**
    - No High School Diploma
    - High School Diploma or GED
    - Some College
    - College Degree
2. **Home Ownership**
    - Yes (Owns House/Condo/Ranch/Farm/Mobile Home/etc.)
    - No (Otherwise)

1.3 R Setup

To set up and use R for this activity, follow the general instructions provided here:
How to Access and Use R for Data Analysis Activities.

Once you've set up your project and loaded the provided files, proceed with the activity-specific instructions below.

1.4 Activity Instructions

1. Open the R script Home_SCF2013.R along with the data set Home_SCF2013.csv and run it line by line in RStudio.
2. Examine the data using the `str()` function to understand its structure and variable names.
3. Generate and interpret bar plots to visualize the relationship between education level and home ownership.
4. Modify the bar plot code (lines 124 to 143) to include your last name in the title:

```r
library(ggplot2)

# Exercise.
# We will construct a side-by-side (clustered) bar chart of the data
# bar.plot is a name where we want to save the plot and its features
# ggplot function will make a canvas,
# and will make the plot ready using the data set and its variables of interest
bar.plot = ggplot(Home2, aes(x = Education_Level, fill = Home_Ownership))
# We will add the bars to the plot of the data
# As well, we will add the legends and position it to the right-hand side
bar.plot = bar.plot + geom_bar(position = "dodge")
# We will add a label to the x-axis,
# We will differentiate the bars by filling in the levels of the response variable
# We will add a title and a subtitle to the plot
# And, we will centre the position of both the title and the subtitle
# Modify line 140 with your last-name in the subtitle
bar.plot = bar.plot + labs(xlab = "Education Level", fill = "Home Ownership",
                           title = "Bar Plot of Home Ownership and Education Level",
                           subtitle = "Constructed by You")
bar.plot = bar.plot + theme(plot.title=element_text(hjust=0.5),
                            plot.subtitle = element_text(hjust=0.5))
bar.plot
```

5. Save the modified bar plot as a `.jpeg` or `.png` file.

1.5 Related Questions

1. Suppose we are interested in investigating the relationship between **home ownership** and **education level**:
   - Identify the response and explanatory variable.
   - Identify the role of each variable, their type, and their scale of measurement.
2. Find the marginal proportions of the response variable, convert them to percentages, and interpret.
3. Find the conditional proportions of the response variable for each category of the explanatory variable. Convert them to percentages and interpret.
4. Compute the difference between the proportion of homeownership for the "No High School Diploma" group and the "College Degree" group. Interpret this difference in context.

5. Compute the ratio between the proportion of homeownership for the "No High School Diploma" group and the "College Degree" group. Interpret this ratio in context.

6. Using data-based arguments (i.e., comparing conditional proportions), describe the relationship between the variables **"home ownership"** and **"education level"**.

1.6 Submission Instructions

- Save your bar plot as a `.doc`, `.pdf`, `.jpeg`, or `.png` file.
- Provide your file to your instructor for feedback or assessment.

Chapter 2

Exploring Quantitative Data

2.1 Context of Data

The **Organisation of Economic Cooperation and Development (OECD)** gathers various information regarding OECD countries and its partners to promote policies that aim to improve the economic and social well-being of people around the world. From the **Better Life Index (BLI, 2021)**, a program conducted by OECD, you will analyze a quantitative variable named **Social Network Support**. This variable reflects the percentage of males and females aged 15 years and over in 41 OECD countries who perceive their social network as having relatives or friends that they can count on to help them in times of need and trouble.

2.2 R Setup

To set up and use R for this activity, follow the general instructions provided here:
How to Access and Use R for Data Analysis Activities.

Once you've set up your project and loaded the provided files, proceed with the activity-specific instructions below.

2.3 Activity Instructions

1. Open the R script `Social.R` and run it line by line in RStudio.
2. Begin by reading/importing the data file `Social.csv` into R.
3. Examine the data structure and variables using `str()` and summary statistics functions.
4. Create side-by-side boxplots to compare distributions of **Social Network Support** percentages for males and females.
5. Modify the plot title (line 86 in the R script) to include your last name. For example:

```
box.plot <- box.plot + labs(title = "Boxplot of Data Constructed by You")
```

6. Save the boxplot as a `.jpeg` or `.png` file.

2.4 Related Questions

After completing the activity, answer the following questions using your results:

1. Refer to your statistical analysis of percentages of perceived social network support for males and females. Use side-by-side boxplots and summary statistics to compare the distributions.

    - Compare the shapes, centers, and spreads of both distributions.
    - Identify potential outliers using the 1.5 IQR rule. For any outlier(s), determine how many standard deviations they are away from the overall mean of the distribution.

2. Use the boxplot and summary statistics for the differences between females' and males' percentages of perceived social network support (in each country) to describe what this plot reveals that the side-by-side boxplots do not.

    - Discuss the shape, center, and spread of the differences.
    - Identify suspect outliers using the 1.5 IQR rule and calculate their deviation from the overall mean.
    - Explain why the boxplot of differences is more insightful for understanding differences between males and females across OECD countries.

2.5 Submission Instructions

- Save your boxplot of differences in percentages of perceived social support network as a `.doc`, `.pdf`, `.jpeg`, or `.png` file.
- Share your file with your instructor or teaching assistant for feedback or assessment.

Chapter 3

Sampling Distributions Related to the Normal Population

3.1 Activity Objective

The objective of this module is to equip you with the following:

- Use R/RStudio on a desktop/laptop to simulate data from a Normal model.
- Describe the sampling distribution of sample means.
- Recognize sampling distributions related to the normal population.

3.2 Context of Data

At-term newborns in Canada vary in weight according to an approximate **Normal model**, with the following parameters: - Mean: 3500 grams - Standard Deviation: 500 grams

The objective of this activity is to explore the sampling distributions of sample means and understand how they relate to the normal distribution.

3.3 R Setup

To set up and use R for this activity, follow the general instructions provided here:
How to Access and Use R for Data Analysis Activities.

Once you've set up your project and loaded the provided files, proceed with the activity-specific instructions below.

3.4 Activity Instructions

1. Open the R script `Newborn_Weights.R` and start running the provided code.
2. Modify the R code as follows:
   - **Line 7**: Insert a unique seed number.

```
# Enter your seed number in line 8.
set.seed( )
```

- **Line 39**: Add your last name to the title of the plot.

```
# Make a change to line 40. Include your last name in the main (title) argument.
hist(Xbar,
   main = "Histogram of Sample Means \n Conducted by You",
   xlab = "Xbar: Sample Means for Total Weights (in grams) of New Borns",
   col = "lavender")
```

3. Simulate 100 samples of size $n = 30$ from the population distribution of newborn weights. Store the means in a variable named `Xbar`.
4. Construct a frequency histogram of the sampling distribution of sample means. Obtain summary statistics and interpret:
   - Shape
   - Center
   - Spread
5. Experiment with other sample sizes ($n = 2, 5, 15, 30, 100, 300$) using different replication numbers ($n = 1000$):
   - Plot the sampling distributions of sample means for these sample sizes.
   - Identify which sample size gives the smallest **Standard Error**.

3.5 Related Questions

After completing the activity, answer the following questions:

1. **Sampling Distribution Analysis**:
   - Using the boxplot and Normal QQ plot, describe the sampling distribution of your sampled data. Justify your answer with both visualizations.
2. **IQR Extrapolation**:
   - Refer to the summary statistics. Write a sentence that extrapolates the **IQR** for this sample into a statement about the population. Ensure your wording is precise.
3. **Sampling Distribution of Means**:
   - For $n = 30$, describe the sampling distribution of sample means (shape, center, and spread). Use summary statistics and histogram to support your analysis.
4. **Effect of Sample Size**:
   - Compare sampling distributions for different sample sizes ($n = 2, 5, 15, 30, 100, 300$).
   - Identify which sample size gives the smallest **Standard Error** and explain why.

3.6 Submission Instructions

- Save the histogram of the sampling distribution of sample means for newborn weights ($n = 30$) as a `.doc`, `.pdf`, `.jpeg`, or `.png` file.
- Provide your file to your instructor or teaching assistant for feedback or assessment.

Chapter 4

Applications of Limit Theorem

4.1 Activity Objective

The objective of this module is to equip you with the following: - Use R/RStudio on a desktop/laptop to construct sampling distribution of the sample mean. - Explore the concept of central limit theorem - Apply the Normal Approx. to the Binomial

4.2 Context of Data

The General Social Survey (GSS, 2013) reports the total number of groups, organizations, and associations that individuals, persons 12 years of age and older in Canada, participated in during the past 12 months. A researcher took a random sample of 100 persons 12 years of age and older from this population. She asked the randomly selected persons to indicate the total number of groups, organizations, and associations that they participated in during the past 12 months. She stored her data in a CSV file and named the variable of interest "group".

4.3 Activity Instructions

Complete the following steps and work on answering the related questions:

1. Download the following two files to your computer:
   - Data File: Volunteering.csv
   - R Script: Volunteering.R
2. Follow the instructions provided on the **How to Access and Use R for Data Analysis Activities** page to set up your R environment:
   How to Access and Use R for Data Analysis Activities.
3. Once you've set up your project folder and uploaded the files, proceed with the steps below:
   - Open the R script `Volunteering.R` and start running the provided codes.
4. There are two required R code modifications:
   - **Lines 26 to 30** require you to input a seed number into line 30. Put it into line 30 and run the line of code in R.

```
# R code modification #1: Obtain sampling Distribution of Sample Means
# Pick a seed number so that each time you run your sampling,
# you will obtain the same result.
```

```
# Enter your seed number in line 30.
set.seed( ) # Enter your seed number here
```

- **Lines 40 to 45** in the R script describe a small exercise to make a change to line 43. You will insert your last name in the title of the plot. Make the change to the line mentioned and upload your plot using this assignment page.

```
# R code modification #2: Plot the Sampling Distribution of Sample Means
# for Total Number of Group Participation
# Make a change to line 43. Include your last name in the main (title) argument.
  hist(Xbar,
     main = "Histogram of Sample Means \n Conducted by You",
     xlab = "Xbar: Sample Means for Total Number of Group Participation",
     col = "lavender")
```

4.4 Related Questions

Based on your R analysis (using your R outputs), answer the following related questions:

1. Refer to the plots of data (boxplot and Normal QQ plot). Does it appear that this data could have come from a normal distribution? Use **both** plots to justify your answer.
2. Refer to the summary statistics of the data. Write a sentence that extrapolates the **IQR** for this sample into a statement about the population – be careful with your wording.
3. Use the sample mean and sample standard deviation of the total number of group participation as the parameter values for the population mean and population standard deviation. If the population distribution of the total number of group participation was normal, 20% of all cases in the specified age range participated in at least how many group organizations? Find x.
4. For this question, treat the "group" data as a population data ($N = 100$). Note the mean as the population mean and SD as the population SD. Draw 60 samples of size $n = 50$ each from this supposed population distribution to approximate the sampling distribution of sample means total number of group participation. Store the result of your experiment in a variable named `Xbar`. Construct a frequency histogram and obtain summary statistics for the sampling distribution of sample means total number of group participation. Describe the shape, center, and spread of the sampling distribution of sample means total number of group participation for 50 persons in the specified age group. Name the theorem that you use to answer this question.

4.5 Submission Instructions

Using this assignment page, upload your histogram for the sampling distribution of sample means total number of group participation. The accepted file formats for submission are `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 5

Estimation with Confidence Intervals

5.1 Activity Objective

The objective of this module is to equip you with the following:

- Use R/RStudio on a desktop/laptop to simulate confidence intervals for population parameters of interest.
- Construct a confidence interval for one mean (known sigma) from a Normal Population.
- Construct a confidence interval for one mean (unknown sigma, small sample) from a Normal Population.
- Construct a confidence interval for one proportion.
- Construct a confidence interval for one variance.

5.2 Context of Data

Suppose **STA123 course grades** are Normally distributed. Instead of collecting data from students, we will simulate (generate) some data for our example. To simulate data, we need to know the population mean and standard deviation of STA123 course grades. Here we assume the mean STA123 course grades is **70** and the standard deviation is **5**.

Using the sampled data, we obtain a **95% CI for population mean STA123 course grades**. Next, we carry out an experiment for CI interpretation. We know that a CI changes each time with a study. If we repeat the same study again and again, (1-α)% of the time the obtained confidence intervals would cover the true population parameter value. This can be shown through a simulation study or experiment. Using the STA123 course grades example, we can conduct an experiment using the following steps:

1. Generate a set of STA123 course grades data with 150 students from the population.
2. Calculate the observed sample mean of STA123 course grades and the standard error of $\bar{X}$.
3. Calculate the confidence interval.
4. Check whether the confidence interval contains the population parameter value.
5. Repeat steps (1)-(4) **10,000 times** and count the total number of times that the confidence intervals contain the population value.
6. For a **95% CI**, one would expect about **9500 times** the CIs contain the population value.

Next, we draw a plot of the first 100 simulated confidence intervals and indicate those which do not contain the true population mean value of 70.

5.3 Activity Instructions

1. Download the following file to your computer:
   - R Script: Interval-Estimates-Simulations.R
2. Follow the instructions provided on the **How to Access and Use R for Data Analysis Activities** page to set up your R environment:
   How to Access and Use R for Data Analysis Activities.
3. Once you've set up your project folder and uploaded the file, proceed with the steps below:
   - Open the R script `Interval-Estimates-Simulations.R` and start running the provided code.
4. There are two required R code modifications:
   - **Line 15**: You need to input a seed number. Choose your seed number and put it into line 15 and run the line of code in R.
   - **Line 96**: You need to insert your last name in the title of the plot. Make the change to this line and upload your plot using this assignment page.

5.4 Related Questions

Based on your R analysis (using your R outputs), answer the following related questions:

1. Refer to the plots of your sampled data (histogram, boxplot, and Normal QQ plot). Describe the distribution of sampled data.
2. Use the results of your sampled data to construct a **95% CI** for μ.
3. Briefly describe the result of your simulation. What do you observe?

5.5 Submission Instructions

Using this assignment page, upload your plot of the **95% confidence intervals** using your simulated data. The accepted file formats for submission are `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 6

Introduction to Hypothesis Testing and Concepts

6.1 Activity Objective

The objective of this module is to equip you with the following:

- Use R/RStudio on a desktop/laptop to carry out hypothesis testing and to estimate error probability via simulation.
- Employ hypothesis testing for:
  - **One Proportion**
  - **One Mean**
    - Sigma known, Normal population:
      - Large sample
      - Small sample
    - Sigma unknown, estimate sigma (replace with sample standard deviation):
      - Large sample, CLT applies
      - Small sample, assess (or know) whether the random sample comes from a Normal population.
  - **One Variance**
    - Independent, (small or large) random sample.

6.2 Context of Data

Suppose **STA123 course grades** are Normally distributed with a mean of **70** and a standard deviation of **5**.

Estimate the error probability in a t-test of $H_0 : \mu = 70$ versus $H_a : \mu > 70$, when the underlying population is Normal with $\mu = 70$ and $\sigma = 5$. Take a random sample of $n = 25$ and test at $\alpha = 0.05$.

Using the STA123 course grades example, we can conduct an experiment using the following steps: 1. Generate a set of **10,000** STA123 course grades data with 25 students from the population. 2. For each of the generated random samples, conduct a t-test. 3. Store the observed t-test statistics and the p-values in a matrix. 4. Count the number of p-values that are ≤ 0.05 (α-level) to estimate the probability of error.

6.3 Activity Instructions

1. Download the following file to your computer:
   - R Script: Hypothesis Testing-Error Probability-Simulation.R
2. Follow the instructions provided on the **How to Access and Use R for Data Analysis Activities** page to set up your R environment:
   How to Access and Use R for Data Analysis Activities.
3. Once you've set up your project folder and uploaded the file, proceed with the steps below:
   - Open the R script `Hypothesis Testing-Error Probability-Simulation.R` and start running the provided code.
4. There are two required R code modifications:
   - **Line 10**: You need to input a seed number. You are provided with a unique seed number that you can retrieve from in your Grade page (see Seed Number). Find your seed number and put it into line 10 and run the line of code in R.

```
# R code modification #1:
# Enter your seed number inside the function set.seed( )
# set.seed() function in R will create reproducible results.
set.seed()
```

   - **Line 75**: You need to insert your last name in the title of the plot. Make the change to this line and upload your plot using this assignment page.

```
# R code modification #2:
# Make a change to the main (title) of the plot by adding your last name to it.
library(ggplot2)
hist.plot = ggplot(data = HT.Array)
hist.plot = hist.plot + geom_histogram(aes(x = t.Stat,
                                           y = after_stat(count / sum(count)),
                                           fill = P.value <= 0.05),
                                       binwidth = 0.5,
                                       color = 'black')
hist.plot = hist.plot + scale_fill_manual(values=c("lightgrey", "darkorange"),
                                          name="P.values <= 0.05",
                                          labels=c("No", "Yes"))
hist.plot = hist.plot + labs(x = "t Statistics",
                             y = "Proportion")
hist.plot = hist.plot + ggtitle('Simulated t Statistics Conducted by You')
hist.plot + theme_bw() + theme(plot.title=element_text(hjust=0.5))
```

6.4 Related Questions

Based on your R analysis (using your R outputs), answer the following related questions:

1. Briefly describe the result of your simulation. What do you observe?

6.5 Submission Instructions

Using this assignment page, upload your histogram of t-statistics based on your simulated data. The accepted file formats for submission are `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 7

Errors in Tests, Statistical Power and Sample Size

7.1 Activity Objective

The objective of this module is to equip you with the following:

- Obtain statistical power.
- Use statistical power to find sample sizes.
- Use R/RStudio on a desktop/laptop to explore the relationship between statistical power and sample sizes for different mean values.

7.2 Context of Data

Suppose it is claimed that the mean STA123 course grade is 70. The population SD is 4 (suppose sigma is known). Suppose that the population distribution of STA123 course grades is Normal. Suppose we test $H_0 : \mu = 70$ versus $H_a : \mu < 70$. Suppose the true μ value is something else (i.e., 68). So H_0 (the null hypothesis) is wrong. How likely is that to happen? This is the power of the test:

- In R, the function is: `power.t.test`.
- The argument delta in this function is the difference between null and true mean.
- We also need a sample size. Let's consider various sample sizes and their effect on power.
- We will examine the relationship between power and sample sizes for different mean values (possible true μ values).
- We plot the power curves.

7.3 Activity Instructions

1. Use Chrome or Firefox to download the following file to your computer:
   - R Script: Statistical Power.R
2. Follow the **How to Access and Use R for Data Analysis Activities**. to prepare your environment.
3. Upload the downloaded R script into the appropriate folder in your project directory.
4. Open the R script and start running the provided code.
5. **Required R code modification**:

- In **line 44**, you need to insert your last name in the title of the plot. Make the change to this line and upload your plot using this assignment page.

```
# Plot the power curve for different true mean values as n sample size increases
power.curve = ggplot(my_data, aes(x = n, y = power, colour = mean)) +
  geom_point() + geom_line() +
  geom_hline(yintercept = 1, linetype="dashed") +
  xlab("sample size") +
  # Make a change to the main (title) of the plot by adding your last name to it
  ggtitle('Relationship Between Power and Sample Sizes for Different Mean Value
  Constructed by You') +
  theme(plot.title=element_text(hjust=0.5))
power.curve
```

7.4 Related Questions

Based on your R analysis (using your R outputs), answer the following related question:

- Refer to the plot of power curves. What do you observe?

7.5 Submission Instructions

Using this assignment page, upload your plot of power curves. The accepted file formats for submission are `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 8

Comparing Groups

8.1 Activity Objective

The objective of this module is to equip you with the following:

- Comparing groups.
- Use R/RStudio in a desktop/laptop environment to test relationships between a quantitative variable and a categorical variable.

8.2 Context of Data

Every year, the US releases to the public a large data set containing information on births recorded in the country. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of 1,000 cases from the data set released in 2014.

We would like to know: is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who do not smoke? We will use data from this sample to try to answer this question.

8.3 Activity Instructions

Complete the following steps and work on answering the related question:

1. **Download the R script**:
   - Download the following R script to your computer:
     - babies.R
2. **Set up your R environment**:
   - Follow the instructions provided on the page mentioned below to set up your R environment: **How to Access and Use R for Data Analysis Activities**.
3. **Upload the R script**:
   - Copy the downloaded R script into the `Module 9` folder inside your project directory.
4. **Run the R script**:
   - Open the R script in RStudio and start running the provided codes.
5. **Modify the R code**:

- In **line 33**, you need to insert your last name in the title of the plot. Make the change to this line and save your updated plot.

```
# Construct a side-by-side boxplots of weights of babies by mothers' smoking habit.
# Give an appropriate title and x-y labels (modify the code below).
box.plot <- ggplot(babies, aes(x = habit, y = weight, fill = habit))
box.plot <- box.plot + geom_boxplot()
box.plot <- box.plot + scale_fill_manual(values = c("lightgrey", "darkorange"),
                                         name = "Habit")
box.plot <- box.plot + labs(x = "Mothers' Smoking Habit",
                            y = "Babies' Weights")
# Center and add a title to the side-by-side boxplots.
box.plot <- box.plot + ggtitle("Boxplots Constructed by You") # LINE 33
box.plot <- box.plot + theme_update(plot.title = element_text(hjust = 0.5))
box.plot
```

8.4 Related Questions

Based on your R analysis (using your R outputs), answer the following related questions:

1. Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.
2. State the null and alternative hypotheses, the value for the observed test statistic, the reference distribution, the p-value, and the conclusion using plain English.

8.5 Submission Instructions

Using this assignment page, upload your side-by-side boxplots of the data. The accepted file formats for submission are: `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 9

Analysis of Categorical Data

9.1 Activity Objective

The objective of this module is to equip you with the following:

- Conduct Goodness of Fit Test for Analysis of One-way Table
- Conduct Chi-Square Test of Independence for Analysis of Two-way Table (Association Between Two Categorical Variables)
- Measure and Detect Pattern of Association
- Use R to analyze categorical data via RStudio (local setup instructions can be found on the **How to Access and Use R for Data Analysis Activities** page).

9.2 Context of Data

The Canadian Community Health Survey (CCHS, 2018) is a cross-sectional survey that collects information related to health status for the Canadian population. The CCHS data is collected from persons aged 12 and over living in the ten provinces and three territories. Excluded from the sampling frame are individuals living on Reserves, institutional residents (health institutions, prisons, etc.), full-time members of the Canadian Forces, and youth aged 12 to 17 living in foster homes. In this activity, you will investigate the relationship between perception of mental health among youth and their gender.

9.3 Activity Instructions

Follow these steps to complete the activity and answer the related questions:

1. **Download the required files**:
   - Data File: CCHS.csv
   - R Script: CCHS.R
2. **Set up your R environment**: Follow the instructions provided on the page mentioned below to set up your R environment:
   How to Access and Use R for Data Analysis Activities.
3. Open the R script `CCHS.R` from your project folder in RStudio.
4. Run each line of code in the R script, starting with reading/importing the `CCHS.csv` file.

5. **Modify the R code**:
   - Line 307 in the script requires you to insert your last name into the bar plot title.

```
# Bar Plots of Age and Positive Mental Health Perception by Gender
bar.plot <- ggplot(data = crosstabl, aes(x = Age, y = Percent,
                                         fill = Positive.Mental.Health))
bar.plot <- bar.plot + geom_bar(stat = "identity", position = "dodge")
bar.plot <- bar.plot + scale_fill_manual(values = c("orange", "grey"),
                                         name = "Positive Mental Health Perception")
bar.plot <- bar.plot + scale_y_continuous(breaks = seq(0, 100, 10))
bar.plot <- bar.plot + labs(y = "Percentages within Age",
                            title = "Conditional Distribution of Positive Perception
                            of Mental Health on Age by Gender",
                            subtitle = "Constructed by You") # LINE 307
bar.plot = bar.plot + facet_wrap(~ Gender, scales = "free_x")
bar.plot <- bar.plot + theme_bw()
bar.plot <- bar.plot + theme(plot.title=element_text(hjust=0.5),
                             plot.subtitle = element_text(hjust=0.5))
bar.plot
```

6. Save your R outputs, including the statistical results, to a Word document or R Markdown file (optional).

9.4 Related Questions

Based on your data analysis (using your R outputs), answer the following related questions:

9.4.1 Question 1: Gender and Mental Health Perception

1. Find the conditional distribution of having a positive perception of mental health for males and for females.
2. Find and interpret the estimated ratio of the conditional proportion of having a positive perception of mental health for males and for females.
3. Test whether there is a significant association between having a positive perception of mental health for males and for females. Include:
   - Null and alternative hypotheses
   - Observed test statistic
   - Reference distribution
   - P-value
   - Conclusion (plain English).
4. What is the absolute value of the adjusted standardized residuals? Interpret this in the study's context.
5. Use R to find the 95% confidence interval for the difference between population proportions of males and females who perceived having positive mental health. Interpret this interval in context.
6. Find the estimated odds of having a positive perception of mental health for males and the estimated odds for females. Describe the strength of association using the estimated odds ratio.

9.4.2 Question 2: Age and Mental Health Perception

1. Find the conditional distribution of having a positive perception of mental health for different age groups.
2. Test whether there is a significant association between having a positive perception of mental health and age group. Include:
   - Null and alternative hypotheses

- Observed test statistic
- Reference distribution
- P-value
- Conclusion (plain English).

3. What is the absolute value of the adjusted standardized residuals? Interpret this in the study's context.

9.5 Submission Instructions

Using this assignment page, upload your conditional distribution of perception of mental health on different ages by gender. The accepted file formats for submission are `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 10

Correlation and Introduction to Simple Linear Regression Model

10.1 Activity Objective

The objective of this module is to equip you with the following:

- Describe the nature of the relationship between two quantitative variables using a scatterplot of the data.
- Interpret the correlation coefficient estimate within the context of the data.
- Fit a simple linear regression model to data.
- Describe what the regression line indicates.
- Interpret the estimated regression coefficients within the context of the data.
- Check the necessary assumptions about the random errors for the regression model.
- Describe the proportion of variation in the response variable accounted for by the regression model.
- Test whether there is a significant relationship between the variables in the data.
- Analyze relationships between quantitative variables using R (local setup instructions can be found on the **How to Access and Use R for Data Analysis Activities** page).

10.2 Context of Data

The Organisation of Economic Cooperation and Development (OECD) gathers various information regarding OECD countries and its partners to promote policies that improve the economic and social well-being of people around the world. The "Better Life Index" (BLI, 2017), a program conducted by OECD, includes information about Educational Attainment. This variable refers to the percentage of adults aged 25 to 64 holding at least an upper secondary degree over the same age as defined by ISCED Classification (International Standard Classification of Education).

You will analyze the relationship between males' and females' percentage of educational attainment.

10.3 Activity Instructions

1. **Download the required files**:
   - Data file: EduAttain-OECD2017.csv
   - R Script: EduAttain.R
2. **Set up your R environment**: Follow the instructions provided on the **How to Access and Use R for Data Analysis Activities** page to set up your R environment.
3. **Run the R script**:
   - Open the `EduAttain.R` script in your R project folder.
   - Execute each line of code sequentially.
4. **Copy and save results**:
   - Copy each result/R output that appears in the R console to a Word document or similar. Optionally, save the results in an R Markdown file.
5. **Modify the R script**:
   - In **line 15**, make a change to include your last name on the scatterplot of the data. Bring the result of your R output to your tutorial.

```r
# Obtain the scatterplot of the data
# Load the library ggplot
library(ggplot2)
Plot = ggplot(EduAttain, aes(x=Females_Edu, y=Males_Edu))
Plot = Plot + geom_point(shape=19, color="blue")
Plot = Plot + xlab("Percent Females Educational Attaintment")
Plot = Plot + ylab("Percent Males Educational Attaintment")
Plot = Plot + ggtitle("Scatterplot of Males vs Females Educational Attainment",
                      subtitle = "Constructed by You")
Plot = Plot + theme(plot.title=element_text(hjust=0.5),
                    plot.subtitle = element_text(hjust=0.5))
```

10.4 Related Questions

1. Refer to the scatterplot of the data. Describe the nature of the relationship between the percentage of males' and females' educational attainment in OECD countries.
2. What is the estimated correlation coefficient? Interpret this number in this context.
3. What does the regression line tell us?
4. Give the least-squares prediction equation (the equation of the regression line).
5. Interpret the estimated slope in the regression equation in this context.
6. Interpret the estimated y-intercept in the regression equation in this context.
7. What is the predicted value for the percentage of males' educational attainment when females' educational attainment is 80%?
8. For the United Kingdom, the percentage of educational attainment for males is 81%, and for females is 80%. What is the residual for this observation?
9. What is the residual standard error of the regression model?
10. What is the estimated standard deviation of males' educational attainment y for any fixed value of females' educational attainment x?
11. Confirm that this data has no regression outlier.
12. State and check the necessary assumptions about the random errors in the regression model.
13. What percentage of variation in males' educational attainment is accounted for by the regression model?

14. Test whether there is a significant relationship between males' and females' percentages of educational attainment. In your answer, be sure to state the null and alternative hypotheses, the value for the observed test statistic, the reference distribution, the p-value, and, if possible, the directional conclusion using plain English.

10.5 Submission Instructions

Using this assignment page, upload your scatterplot of the data with a histogram of each variable in the data added to its margins. The accepted file formats for submission are: `.doc`, `.pdf`, `.jpeg`, or `.png` files.

Chapter 11

How to Set Up R for Data Analysis Activities

Follow these instructions to access and use R for working with the provided CSV and R script files:

1. **Download the Required Files**

   Each activity comes with two files:

   - A CSV data file.
   - An R script.

   You can download these files from the respective activity pages in this resource.

---

2. **Set Up Your Project Directory**
   - On your computer, create a folder named `Stats Activities`.
   - Inside this folder, create a sub-folder for the activity you are working on (e.g., `Activity 1`).

---

3. **Start RStudio and Create a New Project**
   - Open RStudio.
   - Go to `File > New Project > Existing Directory`.
   - Browse to the folder you created for the activity (e.g., `Stats Activities/Activity 1`).
   - Click `Create Project`.

---

4. **Upload the Files to RStudio**
   - Copy the downloaded CSV and R script files into the activity folder.
   - Use the `Files` pane in RStudio (bottom-right) to ensure they appear in your project directory.

---

5. **Run the R Script**
   - Open the R script file in RStudio by clicking on it in the `Files` pane.

- Highlight each line of code in the script and click `Run` (top-right of the editor pane) or press `Ctrl+Enter` (Windows) / `Cmd+Enter` (Mac).

---

6. **Save and Export Outputs**
   - Save the bar plot or other visualizations by clicking the `Export` button in the `Plots` pane or using the `ggsave()` function in R.
   - Save your R console outputs into a text or Word document for reference.

---

7. **Submit Your Work**
   - Share your results with your instructor or teaching assistant for assessment.
   - Accepted file formats include `.doc`, `.pdf`, `.jpeg`, or `.png`.