MATH1324 Introduction to Statistics Assignment 1

Last Updated 19.07.2019

Modeling Body Measurements

Overview

The goal of this assignment is simple. **You must determine if one of the body measurements fits a normal distribution.** To do this, you are going to use the Body Measurements Dataset (bdims.csv) which is located under Google Drive -> Data Repository.

This assignment is worth 10% and must be uploaded to the Assignment 1 Turnitin link by **1/09/2019**.

Body Measurements Dataset Description

Body girth measurements and skeletal diameter measurements, as well as age, weight, height, and gender, are given for 507 physically active individuals - 247 men and 260 women.

Data Source: Heinz G, Peterson LJ, Johnson RW, Kerk CJ. 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education 11(2).

Variables in the dataset: Nine skeletal measurements (diameter measurements) and twelve girth (or circumference) measurements, as well as age, weight, height, and gender, are available in this dataset. Variable names and short descriptions are given below:

- **bia.di:** Respondent's biacromial diameter in centimeters.
- **bii.di:** Respondent's biiliac diameter (pelvic breadth) in centimeters.
- **bit.di:** Respondent's bitrochanteric diameter in centimeters.
- **che.de:** Respondent's chest depth in centimeters, measured between spine and sternum at nipple level, mid-expiration.
- **che.di:** Respondent's chest diameter in centimeters, measured at nipple level, mid-expiration.
- **elb.di:** Respondent's elbow diameter in centimeters, measured as sum of two elbows.
- **wri.di:** Respondent's wrist diameter in centimeters, measured as sum of two wrists.
- **kne.di:** Respondent's knee diameter in centimeters, measured as sum of two knees.
- **ank.di:** Respondent's ankle diameter in centimeters, measured as sum of two ankles.
- **sho.gi:** Respondent's shoulder girth in centimeters, measured over deltoid muscles.
- **che.gi:** Respondent's chest girth in centimeters, measured at nipple line in males and just above breast tissue in females, mid-expiration.
- **wai.gi:** Respondent's waist girth in centimeters, measured at the narrowest part of torso below the rib cage as average of contracted and relaxed position.

- **nav.gi:** Respondent's navel (abdominal) girth in centimeters, measured at umbilicus and iliac crest using iliac crest as a landmark.
- **hip.gi:** Respondent's hip girth in centimeters, measured at at level of bitrochanteric diameter.
- **thi.gi:** Respondent's thigh girth in centimeters, measured below gluteal fold as the average of right and left girths.
- **bic.gi:** Respondent's bicep girth in centimeters, measured when flexed as the average of right and left girths.
- **for.gi:** Respondent's forearm girth in centimeters, measured when extended, palm up as the average of right and left girths.
- **kne.gi:** Respondent's knee diameter in centimeters, measured as the sum of two knees.
- **cal.gi:** Respondent's calf maximum girth in centimeters, measured as the average of right and left girths.
- **ank.gi:** Respondent's ankle minimum girth in centimeters, measured as the average of right and left girths.
- **wri.gi:** Respondent's wrist minimum girth in centimeters, measured as the average of right and left girths.
- **age:** Respondent's age in years.
- **wgt:** Respondent's weight in kilograms.
- **hgt:** Respondent's height in centimeters.
- **sex:** Respondent's gender, 1 if the respondent is male, 0 if female.

Assignment Instructions

1- You are required to select **ONLY ONE MEASUREMENT** from the dataset for this investigation. You must decide which measurement to deal with. You don't need to include all variables.

2- Since males and females tend to have different body dimensions, you are required to investigate the normality assumption of the selected variable **separately in men and women**. Let's say that you selected biacromial diameter measurement as a variable of interest. Then you should investigate if this measurement fits a normal distribution in men and in women separately. Keep in mind that there will be some cases in which men's distribution may fit a normal distribution where else female's distribution may not fit a normal distribution, or vice a versa.

3- You need to import this dataset into RStudio and tidy it up (e.g., you may need to define the variable sex as a factor and define labels for it) using R functions.

4- You need to give summary statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) for your variable of interest separately in men and in women using R functions.

5- Then you will use R to summarise the empirical distribution of body measurement separately in men and women and compare it to a normal distribution. You need to do this visually by plotting the histogram with normal distribution overlay.

6- You will end by discussing the extent to how your theoretical normal distribution fits the empirical data and make recommendations regarding the modeling of this body measurement.

Submission

Assignment 1 must be completed using the R Notebook template available here:

R Notebook Template - Assignment 1

Reports are limited to 6 pages maximum (this includes code). Information for using R Notebooks can be found here. The R Notebook template must be updated with your student ID details and your responses and code for the following sections. You must use the headings and chunks provided in the template. You can add more chunks to explain your approach if required.

Report Section Descriptions

The report will be in a reproducible R Notebook format with written sections, R code and output. The report will be composed of the following sections (see Template above).

1. **Problem Statement [Plain text]**: Write a clear and concise problem statement that guides your investigation. Explain which variable you choose and outline the approaches taken for normal distribution fitting.
2. **Load Packages [R Chunk]:** This section is not marked.
3. **Data [R Chunk]:** Import the body measurements data and prepare it for analysis. Show your code.
4. **Summary Statistics [R Chunk]:** Calculate descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) of the selected measurement grouped by sex.
5. **Distribution Fitting [Plain Text and R Chunk] :** Compare the empirical distribution of selected body measurement to a normal distribution separately in men and in women. You need to do this visually by plotting the histogram with normal distribution overlay and also using some tests. Show your code.
6. **Interpretation [Plain text]:** Going back to your problem statement, what insight has been gained from the investigation? Discuss the extent to how your theoretical normal distribution fits the empirical data.

The report must be uploaded as a **PDF** with your code chunks showing. The easiest way to achieve this is to **Preview** your notebook → **Open in Browser** (Chrome) → Right click on the report in Chrome → Click **Print** and Select the **Destination** Option to **Save as PDF**.

This assignment is worth 10% and must be uploaded to the Assignment 1 Turnitin link by **1/09/2019**.

Extensions will only be granted in accordance with the RMIT University Extension and Special Consideration Policy. No exceptions. Assignments submitted late will be penalised (see Course Information for further details).

Collaboration

You are permitted to discuss and collaborate on the assignment with your classmates. However, the write-up of the report must be an individual effort. Assignments will be submitted through Turnitin, so if you've copied from a classmate, it will be detected. It is your responsibility to ensure you do not copy or do not allow another classmate to copy your work. If plagiarism is detected, both the copier and the student copied from will be responsible. It is good practice to never share assignment files with other students. You should ensure you understand your responsibilities by reading the RMIT University website on academic integrity. Ignorance is no excuse.

Assignment 1 Marking Rubric

Criteria	Not acceptable (0)	Needs Improvement (3)	Excellent (5)
Problem Appearance of all plots you draw in the assignment, **(5%)**	No problem statement was provided	A problem statement was provided, but it was not clear or complete.	A clear and accurate problem statement was provided.
Data (20%)	The data import and management was insufficient or inappropriate.	There were some issues with the data import and management that needed improvement.	Data import and management was appropriate.
Summary Statistics (25%)	No summary statistic was provided OR the summary statistics were inaccurate.	Summary statistics were provided for the variable of interest but not separately given by sex OR some of the summary statistics were missing.	Summary statistics were accurate and given grouped by sex.
Fitting (25%)	The attempt to compare the empirical distribution to the theoretical distribution is inappropriate or missing.	The attempt to compare the empirical distribution to the theoretical distribution needed improvement. For example: - Investigation was not provided separately for men and women. - Empirical distribution was given but the theoretical distribution was missing. - Explanation and/or the approach	The attempt to compare the empirical distribution to the theoretical distribution was appropriate and well explained.

		taken needed improvement.	
Interpretation (25%)	The interpretation of the extent to how the normal distribution fits empirical data was missing.	The interpretation of the extent to how the normal distribution fits empirical data was provided but needed improvement. For example: -Interpretation was not provided separately for men and women. - Explanation and/or the approach taken needed improvement.	The interpretation of the extent to how the normal distribution fits empirical data was appropriate and well explained.