Institute for Computational Systems Biology
University of Hamburg

Lecture: Foundations of Systems Biology, WS 2025-26

Group:
Student names:

Exercise sheet 03 - Biostatistics I

1. "Know your Data" - Load, inspect and familiarize with a given data set (**3 points**)

Download data "phenoData.csv" from a study on human myocardial disease from here:
https://github.com/mpmorley/MAGNet?tab=readme-ov-file

File can also be found in the moodle course.

For more information on the patient cohort, feel free to visit https://www.med.upenn.edu/magnet/

a) Use a (non)-coding tool of your choice, e.g., Excel, R (studio) or python, to load and inspect the data. What is the dimension of the given data table? (**0.5 points**)

b) Types of data: Inspect the data in each column with respect to data types. What are quantitative data, i.e., discrete or continuous data? What are qualitative (categorial) data, i.e., binary, nominal or ordinal data? Fill in **Table 1** or provide a **screenshoot** of the **filled table**. (**1 points**)

c) Missingness: Determine the missingness, i.e., absolute and relative numbers of missing values "NA", of the dataset: (1) entire data set, (2) number of rows with at least one NA value, (3) number of NA values per columns (fill **Table 1** or **provide screenshoot of table**). (**0.5 points**)

d) Regarding the data in column "Diabetes" and "RIN". What would be your suggestion to deal with either of the missing values? (**0.5 points**)

e) Data filtering: Remove all rows that contain NAs. What is the dimension of the data frame after removal of NAs? **From now on: continue working with the NA-filtered data table.** (**0.5 points**)

Table 1: Overview of variables in the given data table including unique variable names, data type, and missingness (excluding columns: *sample_name, Library.Pool*, *minexpr*, *disease_race*).

Column header	**Abbreviations**	**Name unique entries for categorical data**	**Data type**	**'NA' (abs)**	**'NA' (rel)**

tissue_source	*NF - non-failing*	*{'Cardiectomy'}*
		{'NF'}
			Categorical, nominal	*0*	*0*

etiology	DCM - Dilated Cardiomyopathy HCM - hypertrophic cardiomyopathy NF - non-failing PPCM - Peripartum Cardiomyopathy				
gender	-				
race	AA - afro-american				
age	-	*Numbers, integers*	*Quantitative, discrete*	*0*	*0*
weight	-				
height	-				
hw	Heart weight				
lv_mass	Left ventricle				
afib	Atrial fibrillation				
VTVF	Pulseless ventricular tachycardia (VT) and Ventricular fibrillation (VF)				
Diabetes	-				
Hypertension	-				
LVEF	Left ventricle ejection fraction				
RIN	RNA integrity value				
TIN.median	Median transcript integrity number				

2. Describe your data - Use the NA-filtered data table (Task 1e). (**2 points**)

a) Determine min, max, median, mean, mode, variance, standard deviation, first and third quartile, IQR for columns *age*, *weight*, *height* and *RIN* (fill Table 2 or provide a screenshot of the filled table). (**0.5 points**)

Table 2: Descriptive statistical metrics of age, weight, height, and RIN.

	age	**weight**	**height**	**RIN**
min	*21*			
max	*76*			
mean	*53*			
median	*54*			
mode	*56*			
variance	*107.05*			
std	*10.35*			
Q(0.25)	*49*			
Q(0.75)	*59.25*			
IQR	*10.25*			

b) Plot histograms of age and height, describe and compare the histograms using the statistical metrics in Table 2, e.g. what is the range, variability, etc. of the data? Are there any unexpected instances and/or outliers in the data? (**1 points**)

c) The integrity of RNA is crucial for gene expression studies. A RIN value of 1 indicates the presence of very small RNA pieces, i.e., most of the RNA is highly degraded. Usually a RIN between 7 and 10 denotes "good" integrity. How would you describe the RNA integrity of the present study using the descriptive values in Table 2? (**0.5 points**)

3. Quantitative values: boxplots - Use the NA-filtered data table (Task 1e). (**2 points**)

a) Create a figure with four boxplots on the same axes using the data for age, weight, height and TIN.median. Describe the datasets accordingly ("five-number summary", Table 3 or screenshot) and include explanations on the boxes, whiskers and outliers. (**1.5 points**)

Table 3: "Five-number summary" of boxplots.

	age	**weight**	**height**	**TIN.median**
Lower whisker	*34*			
Upper whisker	*67*			
Median	*54*			
Q(0.25)	*49*			
Q(0.75)	*59.25*			
Number of Outliers	*7*			

b) Is the figure from 3a), with the four data sets, an example of a good or bad data visualization? Explain why. (**0.5 points**)

4. Categorical values: frequency tables, pie charts, stacked bar charts - <u>Use the NA-filtered data table (Task 1e)</u>. (**1.5 points**)

a) Create a stacked bar chart for the data of VTFV, Diabetes and Hypertension, with “VTFV”, “Diabetes” and “Hypertension” on the x-axis and absolute frequency of “Yes” and “No” as stacked bar on the y-axis. (**0.5 points**)

b) For gender, race and etiology, create both frequency tables as well as pie charts. (**1 points**)

Table 4: Frequency table for gender, race and etiology.

Gender	
Male	
Female	
Race	
AA	
Caucasian	
Etiology	
DCM	
NF	
HCM	

5. Study and data interpretation (**1.5 points**)

a) Is the study data representative of a population? (**0.5 points**)

b) What other measurements of the study cohort would you like to acquire to draw insightful conclusions from the data with respect to elucidating myocardial disease? (**0.5 points**)

c) What could be biases in the data? (**0.5 points**)