Research Methodology: Basic Statistics in R

Yesoda Bhargava

January 11, 2021

What is R?

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced.
- R can be extended (easily) via packages.

R Vs. Python

- Python is often praised for being a general-purpose language with an easy-to-understand syntax, R's functionality was developed with statisticians in mind, thereby giving it field-specific advantages such as great features for data visualization.
- R is more functional, Python is more object-oriented.
- R has more data analysis functionality built-in, Python relies on packages.
- R has more statistical support in general.
- Python can be used when data analysis is a part of a larger web-based solution.

Data Used for R Practical

- Download **data.csv** from the link here.
- This data is collated from the Centre for Disease Control and Prevention, USA.
- The following variables can be found in the data:
  - birthsex
  - age
  - genhlth
  - cholesterol
  - blood_pressure
  - diabetes
  - diabage
  - exercise
  - bmi

Type of Variable

- Generally, variable definition helps in understanding the type of variable.
- Before looking into each variable type, it is better to look at their definitions, if data is taken from third party.
- Classify these variables as Quantitative Continuous, Quantitative Categorical, Qualitative Categorical, Nominal, Ordinal.

Variable Sex Definition

Label: Are you male or female?
Section Name: Sex at Birth
Module Number: 28
Question Number: 1
Column: 620
Type of Variable: Num
SAS Variable Name: BIRTHSEX
Question Prologue:
Question: What was your sex at birth? Was it male or female?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Male	30,527	46.95	47.80
2	Female	34,227	52.64	51.66
7	Don't know/Not Sure	99	0.15	0.29
9	Refused	165	0.25	0.25
BLANK	Not asked or Missing	353,250	.	.

Variable Age Definition

Label: Imputed age in six groups
Section Name: Calculated Variables
Module Number: 8
Question Number: 15
Column: 1986
Type of Variable: Num
SAS Variable Name: _AGE_G
Question Prologue:
Question: Six-level imputed age category

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 24 Notes: 18 <= _IMPAGE <= 24	25,104	6.00	12.22
2	Age 25 to 34 Notes: 25 <= _IMPAGE <= 34	43,903	10.50	17.42
3	Age 35 to 44 Notes: 35 <= _IMPAGE <= 44	49,470	11.83	16.31
4	Age 45 to 54 Notes: 45 <= _IMPAGE <= 54	61,072	14.60	16.10
5	Age 55 to 64 Notes: 55 <= _IMPAGE <= 64	84,018	20.09	16.52
6	Age 65 or older Notes: _IMPAGE => 65	154,701	36.99	21.43

Variable GENHLTH Definition

Label: General Health
Section Name: Health Status
Core Section Number: 1
Question Number: 1
Column: 101
Type of Variable: Num
SAS Variable Name: GENHLTH
Question Prologue:
Question: Would you say that in general your health is:

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Excellent	65,946	15.77	17.35
2	Very good	135,814	32.47	31.39
3	Good	133,631	31.95	32.27
4	Fair	59,725	14.28	14.09
5	Poor	22,105	5.29	4.68
7	Don't know/Not Sure	741	0.18	0.15
9	Refused	280	0.07	0.07
BLANK	Not asked or Missing	26	.	.

Variable Cholesterol Definition

Label: Ever Told Blood Cholesterol High
Section Name: Cholesterol Awareness
Core Section Number: 5
Question Number: 2
Column: 115
Type of Variable: Num
SAS Variable Name: TOLDHI2
Question Prologue:
Question: Have you ever been told by a doctor, nurse or other health professional that your blood cholesterol is high?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	144,170	36.61	30.86
2	No—Go to Section 06.01 CVDINFR4	245,065	62.23	68.01
7	Don't know/Not Sure—Go to Section 06.01 CVDINFR4	4,198	1.07	1.04
9	Refused—Go to Section 06.01 CVDINFR4	392	0.10	0.09
BLANK	Not asked or Missing Notes: Section 05.01, CHOLCHK2, is coded 1, 9, or Missing	24,443	.	.

Variable Blood Pressure Definition

Label: Ever Told Blood Pressure High
Section Name: Hypertension Awareness
Core Section Number: 4
Question Number: 1
Column: 112
Type of Variable: Num
SAS Variable Name: BPHIGH4
Question Prologue:
Question: Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure? (If ´Yes´ and respondent is female, ask ´Was this only when you were pregnant?´.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	169,634	40.56	32.52
2	Yes, but female told only during pregnancy—Go to Section 05.01 CHOLCHK2	3,072	0.73	0.90
3	No—Go to Section 05.01 CHOLCHK2	239,873	57.35	65.32
4	Told borderline high or pre-hypertensive—Go to Section 05.01 CHOLCHK2	4,117	0.98	0.90
7	Don´t know/Not Sure—Go to Section 05.01 CHOLCHK2	1,064	0.25	0.25
9	Refused—Go to Section 05.01 CHOLCHK2	504	0.12	0.11
BLANK	Not asked or Missing	4	.	.

Variable Diabetes Definition

Label: (Ever told) you had diabetes
Section Name: Chronic Health Conditions
Core Section Number: 6
Question Number: 11
Column: 127
Type of Variable: Num
SAS Variable Name: DIABETE4
Question Prologue:
Question: (Ever told) (you had) diabetes? (If ´Yes´ and respondent is female, ask ´Was this only when you were pregnant?´. If Respondent says pre-diabetes or borderline diabetes, use response code 4.)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	57,401	13.72	11.08
2	Yes, but female told only during pregnancy—Go to Section 07.01 HAVARTH4	3,782	0.90	1.03
3	No—Go to Section 07.01 HAVARTH4	347,091	82.98	85.57
4	No, pre-diabetes or borderline diabetes—Go to Section 07.01 HAVARTH4	9,149	2.19	2.10
7	Don't know/Not Sure—Go to Section 07.01 HAVARTH4	605	0.14	0.17
9	Refused—Go to Section 07.01 HAVARTH4	231	0.06	0.05
BLANK	Not asked or Missing	9	.	.

Variable Diabetes Age Definition

Label: Age When Told Had Diabetes
Section Name: Chronic Health Conditions
Core Section Number: 6
Question Number: 12
Column: 128-129
Type of Variable: Num
SAS Variable Name: DIABAGE3
Question Prologue:
Question: How old were you when you were told you had diabetes?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 97	Age in years [97 = 97 and older] Notes: _ _ Code age in years, 97 = 97 or older	53,213	92.71	93.43
98	Don't know/Not sure	3,893	6.78	6.01
99	Refused	294	0.51	0.56
BLANK	Not asked or Missing Notes: Section 06.11, DIABETE4, is coded 2, 3, 4, 7, 9, or Missing	360,868	.	.

Variable Exercise Definition

Label: Exercise in Past 30 Days
Section Name: Exercise (Physical Activity)
Core Section Number: 11
Question Number: 1
Column: 223
Type of Variable: Num
SAS Variable Name: EXERANY2
Question Prologue:
Question: During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	288,516	72.60	73.51
2	No—Go to Section 11.08 STRENGTH	107,745	27.11	26.24
7	Don't know/Not Sure—Go to Section 11.08 STRENGTH	620	0.16	0.14
9	Refused—Go to Section 11.08 STRENGTH	548	0.14	0.11
BLANK	Not asked or Missing	20,839	.	.

Variable BMI Definition

Label: Computed body mass index
Section Name: Calculated Variables
Module Number: 8
Question Number: 19
Column: 1998-2001
Type of Variable: Num
SAS Variable Name: _BMI5
Question Prologue:
Question: Body Mass Index (BMI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 9999	1 or greater Notes: WTKG3/(HTM4*HTM4) (Has 2 implied decimal places)	382,065	100.00	100.00
BLANK	Don't know/Refused/Missing Notes: WTKG3 = 777 or 999 or HTM4 = 777 or 999	36,203	.	.

Table 1: Nature of variables.

Variable	Continuous	Quantitative Categorical	Qualitative Categorical	Nominal	Ordinal
Sex					
Age					
GENHLTH					
Cholesterol					
Blood Pressure					
Diabetes					
Diabage					
Exercise					
BMI					

Table 2: Nature of variables.

Variable	Continuous	Quantitative Categorical	Qualitative Categorical	Nominal	Ordinal
Sex	No	No	Yes	Yes	No
Age	No	Yes	No	No	No
GENHLTH	No	No	Yes	No	Yes
Cholesterol	No	No	Yes	Yes	No
Blood Pressure	No	No	Yes	Yes	No
Diabetes	No	No	Yes	Yes	No
Diabage	Yes	No	No	No	No
Exercise	No	No	Yes	Yes	No
BMI	Yes	No	No	No	No

Summarization of each variable : Univariate Analysis

- Since SEX is a categorical variable, we use tabulate feature.
- R Command: **table(data$sex)**
- For proportion **round(prop.table(table(data$sex))*100,2)**
- Similar commands for other categorical variables: nominal/ordinal/qualitative categorical/quantitative categorical.

```
> table(data$sex)

     1      2
189849 228419
> round(prop.table(table(data$sex))*100,2)

    1     2
45.39 54.61
```

- Sometimes in summary we also get to see the blank/missing/NA values.
- Example: Summarize *genhlth*.
- R Command: **table(data$genhlth)**

```
> table(data$genhlth)

    1     2     3     4     5     7     9
65946 135814 133631 59725 22105   741   280
```

- What is the meaning of each of these values? Check definition of variable *genhlth*
- What do you find?
- *Genhlth*=7 means Don't Know and *Genhlth*=9 means Refused.
- These data may need to be removed from the dataset.
- Command : **data_new=subset(data, data$genhlth<7)**

```
> table(data_new$genhlth)

    1     2     3     4     5
65946 135814 133631 59725 22105
```

- You just cleaned your variable genhlth.

Other Variable Summarization

- Similarly try for other variables.
- A good descriptive summary will also mention about the invalid entries.
- They will need to be removed before any analysis/modelling.
- In certain situations these values can be imputed also and there are methods for that.
- **Homework**: Summarize *age*, *cholesterol*, *blood_pressure*, *diabetes*,*diabage*, *exercise*.
- Let us summarize BMI.

BMI summarization

- BMI is a quantitative continuous variable after we look at its definition.

```
> summary(data_new$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  12.00   24.02   27.34   28.33   31.45   99.84   35985
```

- What do you observe?
- Why do you think there are so many NA values?
- We need to remove records without any information on BMI.
- Remember we are using **data_new** data set because this is the data without missing values for **genhlth**.
- This becomes our data for further analyses.
- In the end, we want data which contains values for all variables for all rows/respondents.
- Remove NA values: **data_new=data_new[!is.na(data_new$bmi),]**
- Histogram: **hist(data_new$bmi)**
- Subsetting data: **data_new=subset(data_new, data_new$bmi<=50)**

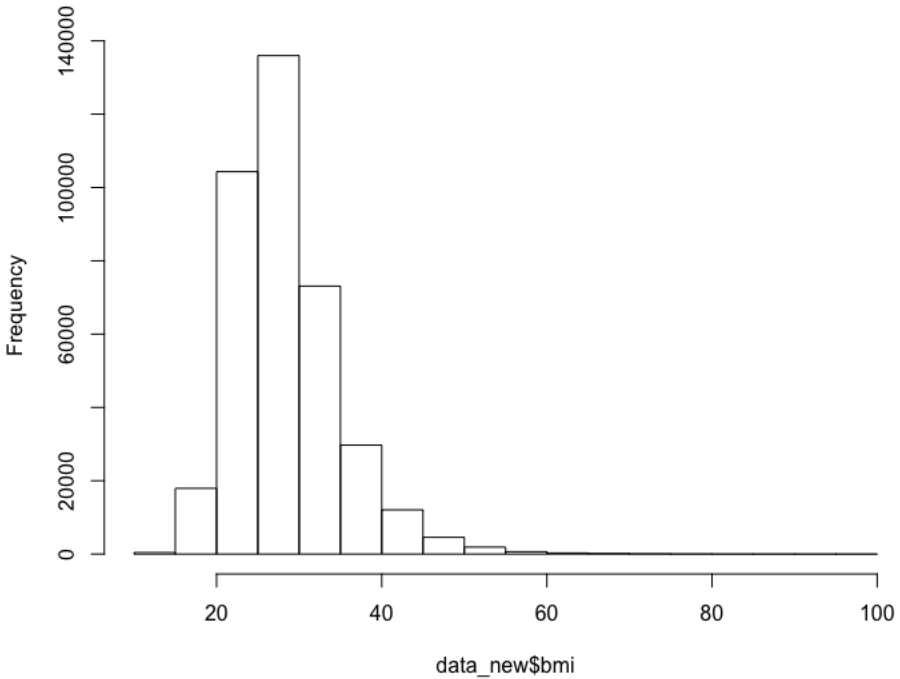

Figure 1: Histogram of BMI in the dataset.

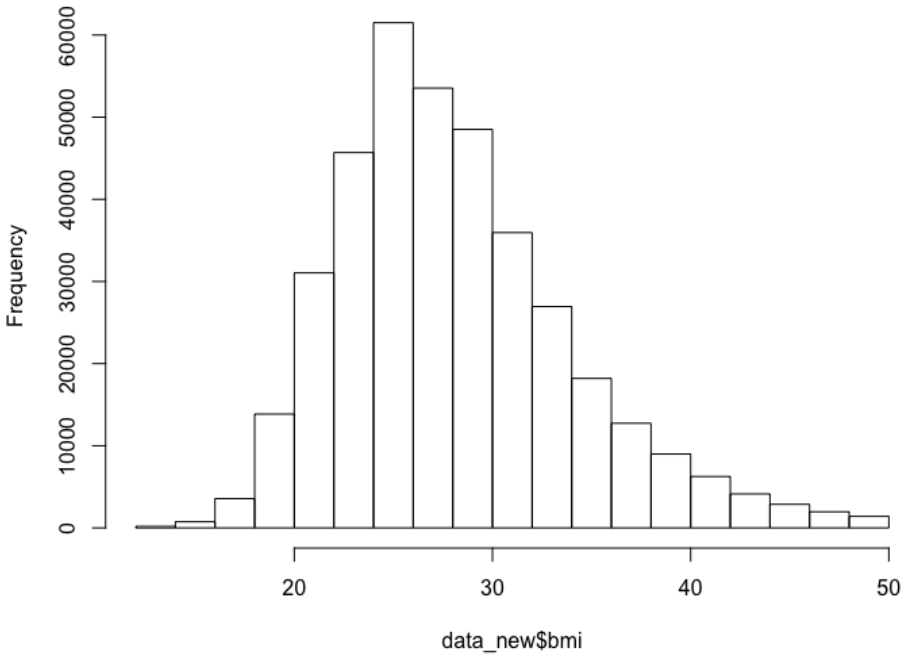

Figure 2: Histogram of BMI in the dataset after removing extreme values, BMI <50.

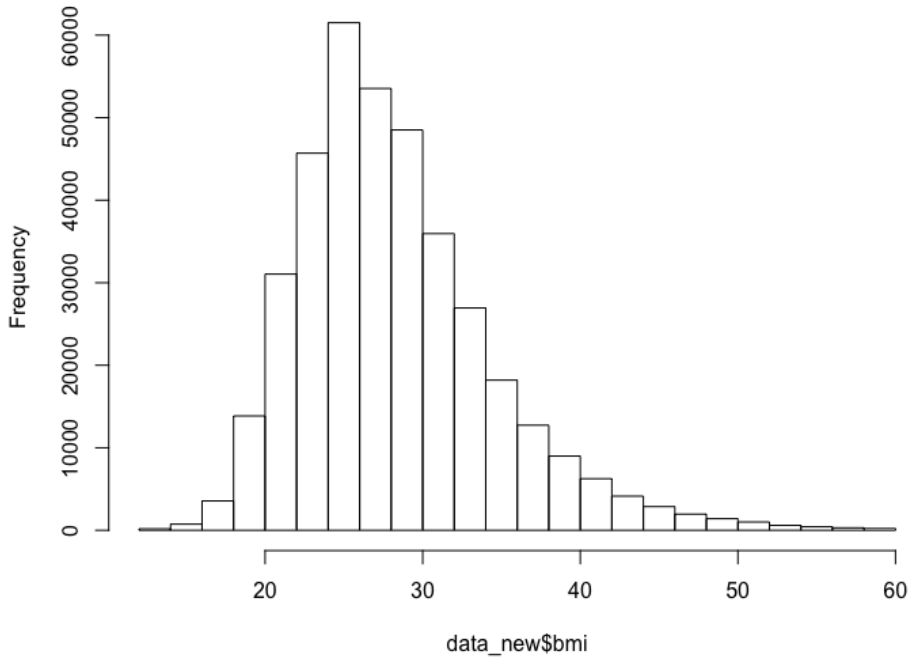

Figure 3: Histogram of BMI in the dataset after removing extreme values, BMI <60.

What is a Boxplot?

- Box-plot is a 5 number summary for a continuous quantitative variable.
- Alternative way to visualize the distribution of the variable in the dataset.

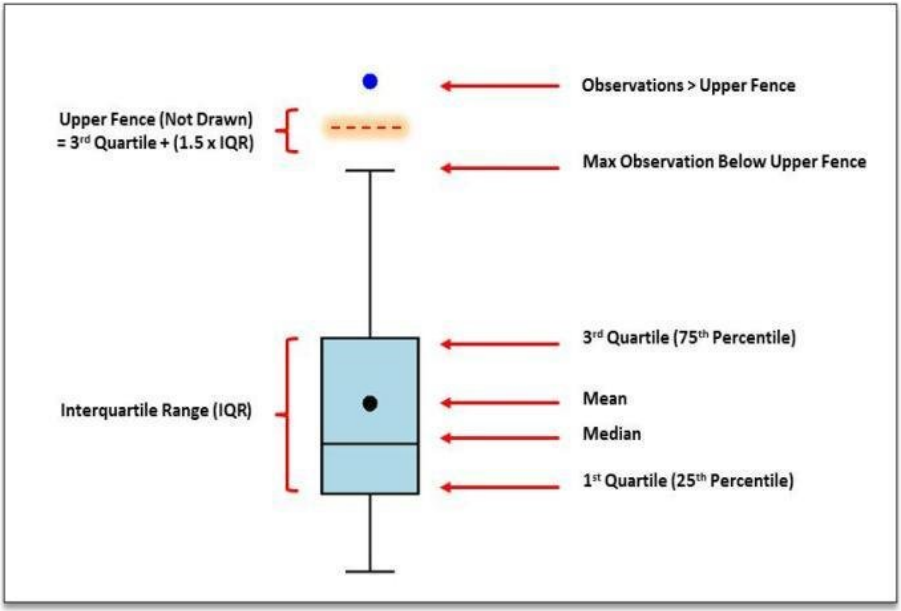

Figure 4: What is a Boxplot?

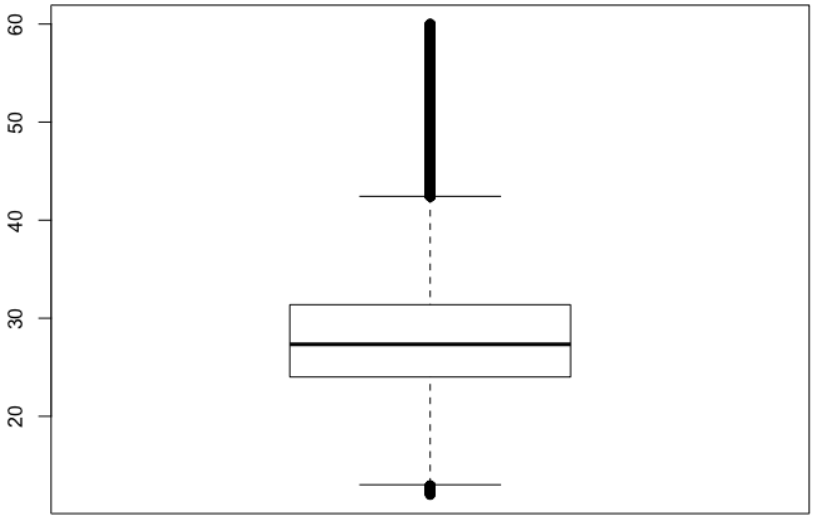

Figure 5: Boxplot for BMI in the dataset. Command: **boxplot(data_new$bmi)**

- Next lecture: Bivariate Analysis
- Analysis of two categorical variables, categorical and continuous variable.