Week 2 - Data and Graphical Summaries

Created	August 9, 2023 8:29 PM
Date	Empty
Status	Empty
Add a property	

Add a comment…

Which graphical summaries to choose in a nutshell

1 Qual	2 Qual	1 Quant	2 Quant	1 Quant + 1 Qual
Simple Bar Plot	Double Bar Plot	Histogram Box Plot	Scatter Plot	Comparative Box Plot

1) Working with qualitative data

1.1) Data and Initial Data Analysis

Data: Data refers to information about the subjects being studied, which can come in various formats such as MRI scans, spreadsheet data, or surveys.

Initial Data Analysis (IDA): An IDA is the first general step in analyzing data. During this phase, investigators explore data to gain insights and ensure that later statistical analyses can be carried out effectively. The purpose of IDA is to minimize errors and inaccuracies in the results.

The process of IDA typically involves *(and this is usually the order of its phases)***:**

1. Checking the integrity and quality of the data
2. Exploring the structures of the data to understand what it is about
3. Data wrangling — including scraping, cleaning, tidying, reshaping, splitting, and combining data

4. Summarizing the data using graphical and numerical representations
5. Creating an IDA report if necessary.

1.2) Variables

Variable: A variable is a measurement or description of some data

p variables is said to have dimension p *(which is not intuitive btw, dimension is conventionally defined as (row * column))*

By convention, you should make variables as columns in order to make your data tidy.

Qualitative variable: A variable of which value is a number of categories that data can be classified into. A qualitative variable may be:

- ordinal (ordered, e.g. a grades variable may have ordered categories such as “A”, “B”, “C”, “D”, “E”, “F”)
- nominal (unordered, e.g. a cat species variable may have unordered categories such as “Birman”, “Ragdoll”, “Scottish fold”)

1.3) Graphical summaries for qualitative data

The aim of a graphical summary should always be to **highlight the best features of the data.**

As a general rule - it is advisable to avoid choosing pie charts because they’re often not very informative, plus, it starts to be difficult to comprehend when there are many variables.

Bar plots are usually the most effective graphical summary of qualitative data because it allows you to set one of the axes (usually the x-axis) as categories.

A simple bar plot is often sufficient for summarizing one qualitative variable. However, if you wish to summarize more than one variable, you may instead choose a:

- double bar plot
- stacked bar plot
- or side-by-side bar plot

we keep the number of qualitative variables presented in bar plots highest at 2 or 3 comprehensible.

1.3.1) Simple bar plot in R

- **with base R** (deprecated! use ggplot!)

```
barplot(table(Name_of_Dataset$Variable))
```

- **with ggplot**

```
ggplot(Name_of_Dataset, aes(Variable)) + geom_bar()
```

1.3.2) Double bar plot in R

```
ggplot(Name_of_Dataset, aes(Variable_1, fill = Variable_2)) + geom_bar()
```

1.3.3) Side-by-side bar plot in R

```
ggplot(
  Name_of_Dataset, aes(Variable_1, fill = Variable_2) + geom_bar(position = 'dodge')
)
```

2. Working with quantitative data

Quantitative variable: A variable of which value can be counted or measured. A quantitative variable may be:

- discrete (countable, e.g. number of cats, number of people)
- or continuous (measurable, e.g. height, weight)

Three plots that are most effective in representing quantitative data include:

- Histogram — useful for comparing the percentage of data in each class interval
- Box plot — useful for comparing multiple quantitative datasets, displaying the median, quartiles, and identifying outliers.
- Scatter plot — useful for exploring relationship, patterns, trends, and clusters between two quantitative variables.

2.1) Histogram

- The entire area of a histogram is 100%

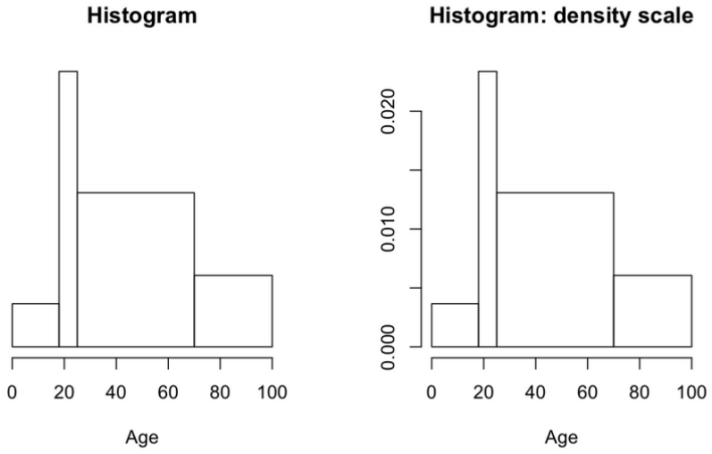
- The horizontal scale is divided into blocks which represent class intervals
- The area of each block represents the percentage of subjects in that class interval
- The height of each block represents crowding

IT IS ADVISED THAT: You use 10-15 bins at most, because having too many can over-condense a histogram!

2.1.1) Use density scale! vertical scales are useless!

In histograms, vertical scales do not mean anything. Instead, we use a density scale which is defined as follows:

$$\text{Height of each block} = \frac{\%\text{ in the block}}{\text{length of the class interval}}$$



The density scale normalizes the vertical axis of the histogram so that the area under each bar represents the proportion of data points within that bin.

2.1.2) Endpoint convention for points are in between two class intervals

A point X that falls on the border of two class intervals should be classified into the class of which endpoint is inclusive of X.

For example: if a histogram has a convention that the left endpoint is inclusive and the right is exclusive, and if 20 falls in between classes $a = [10, 20)$ and $b = [20, 30)$, then 20 would classify into the class interval b.

2.1.3) Constructing a histogram by hand

To construct a histogram by hand, you:

1. Create the distribution table of your quantitative data, like the example below.

Class intervals	Number of subjects in the interval	%	Height of block

[0,18)	29	6.6	0.004
[18,25)	72	16.4	0.023
[25,70)	259	58.9	0.013
[70,100)	80	18.2	0.006
Totals	440	100	

2. Draw the horizontal axis and blocks

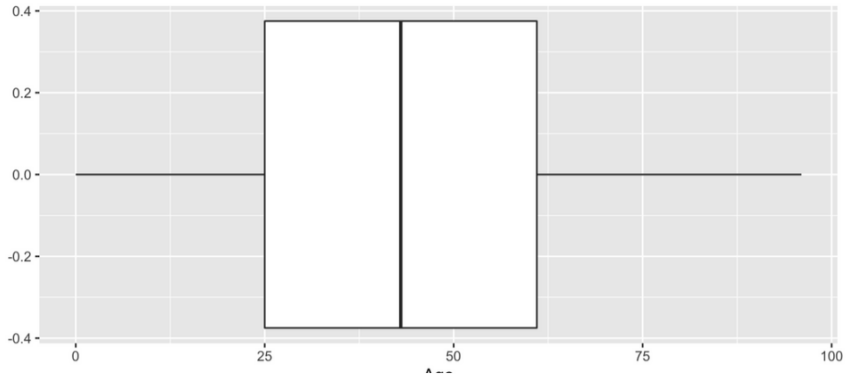
2.2) Box plot

The boxplot plots the median 50% of the data set and identifies any outliers.

2.2.1) Simple box plot in R

A simple box plot is effective for comparing multiple **quantitative** data sets.

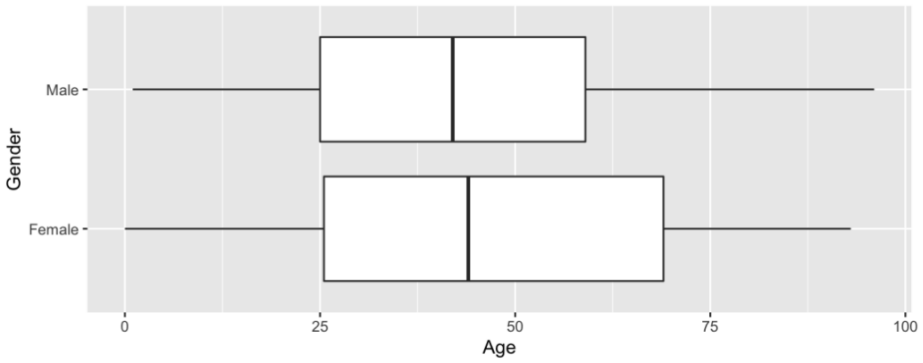
```
ggplot(Name_of_dataset, aes(Variable)) + geom_boxplot()
```



2.2.2) Comparative box plot in R

A comparative box plot is effective for **splitting up a quantitative variable by a qualitative variable.**

```
ggplot(Name_of_dataset, aes(Quant_var, Qual_var)) + geom_boxplot()
```



2.3) Scatter plot

A scatter plot is ideal for evaluating the relationship between **two quantitative variables.**

```
ggplot(Name_of_dataset, aes(y=Variable_1,x=Variable_2)) + geom_point()
```