Summarizing data

John Karuitha

Tuesday, October 17, 2023

Summarizing data

Usually, it is hard to get your mind around a large spreadsheet full of numeric and categorical data. One of the things to do once you get a dataset is to compute summary statistics to get a feel of the data. There are two important terms here;

- Statistic- We came across this term earlier. We explained it as values computed from data. Strictly speaking these are values computed from sample data.
- Parameter- These are values similar to the statistics but that refer to the entire population.

Summarizing data

- Usually, we do not have access to the entire population and hence we infer/estimate most parameters from sample data.
- We are at liberty to define the population. For instance, in a survey about the quality of food at the students canteen, the population is all the students and staff of Karatina University who use that canteen.
- What is the population of interest for pollsters trying to estimate the winner of a presidential election in a country like Kenya? What about for your COUNTY in reference to the governor?; How about the constituency for an MP? See the way the population varies depending on the context.

Sampling techniques [PROJECT]

- How can we draw a sample from a population?
- This will form part of the project work that you write on.

Summarizing data - the sampling frame

- As noted it is not possible in many cases to get hold of the entire population.
- At other times, the population is not well defined.
- The sampling frame is the list from which units are drawn for the sample. The 'list' may be an actual listing of units, as in a phone book from which phone numbers will be sampled, or some other description of the population, such as a map from which areas will be sampled.

Summarizing data - the sampling frame

- Frame error results when the sampling frame is not an accurate and complete representation of the population of interest [].
- The sampling frame is sometimes the same as the population. In cases where the population is poorly defined, the sampling frame may differ from the population, leading to Frame error.

Summarizing data- Sampling Frame

- In a country level survey of likely winners of presidential elections, the sampling frame is the list of all registered voters in Kenya.
- Is the list of all registered voters equal to the population (that will vote)? Maybe not- some are dead, have relocated to other countries, etc.
- Also, there are some youths who just attained the voting age who have not registered as voters. Towards 2022, they may register and hence the current voters register is not capturing the entire population.

Summarizing data- Sampling Frame

- Still the register of voters is a good representation of the population. That list is the sampling frame.
- Think of the population as an ideal situation that captures all the units of interest. The sampling frame is a list of all units of interest.
- If the sampling frame equals the population, you are lucky. If it does not, at least you have something to work with.

Measures of center

- We examine three measures of center
- The Mean
- The median
- The mode
- The quartiles

The mean

- The common mean is the arithmetic mean, the sum of all numeric entries divided by their counts.
- However, there are other types of mean, notably the geometric mean and the harmonic mean.

project work

- Define the three types of means, giving examples and the merits/demerits of each. When would you use each of these means?
- Also, keep the project as it forms part of your course notes.
- As agreed, the project will be written in R using R-Markdown and Latex and NOT Ms Word and other word processing software.
- I will introduce R Markdown in the course of the R lab sessions.

Example

- The following dataset represents the age of students in a primary school.
- 14 years, 12 years, 13.5 years, 16 years, 10 years, 15.4 years.
- Is the age represented above a discrete or a continuous numeric variable?
- Create a vector in R using the c function for the age of students. Name the vector age.
- Calculate the median of the age of students given previously. Do the calculation manually.
- Calculate the median of the age of students given previously. Do the calculation using the median function in R.

The median

- The median represents the center of the dataset, where 50% of the data is below and the other 50% above.
- You have to arrange the data in ascending or descending order to compute the median.
- Calculate the median of the age of students given previously. Do the calculation manually.
- Calculate the median of the age of students given previously. Do the calculation using the median function in R.
- The median is a robust average that is not duly influenced by extreme values.

Solutions; The arithmetic mean

- Note that age is a continous numeric variable; Check discrete vs continous.
- The mean is the sum of the age of students divided by the number of students.
- `mean = (14 + 12 + 13.5 + 16 + 10 + 15.4) / 6 = 13.48333 years`

```
## create a vector of age
age <- c(14, 12, 13.5, 16, 10, 15.4)

mean(age)
```

```
## [1] 13.48333
```

Solution; The median

- The median calculation starts with arranging the data in ascending or descending order.
- `10.0, 12.0, 13.5, 14.0, 15.4, 16.0`
- The middle figures are 13.5 and 14. We do not have one value as median. So we average these two.
- `(13.5+14) / 2 = 13.75 years`.
- Using R. Remember we already have a vector age saved in our workspace.

```
median(age)
```

```
## [1] 13.75
```

Median as a robust summary

- Unlike the mean, the median is not unduly affected by extreme values, the so-called outliers.
- Outliers are values that are not typical of the sample or population.
- In our age example, let us replace the last value 16 with 70 years.
- Clearly, a 70 years is odd in a group of pre-teens and teenagers. 70 years is an outlier in this case.
- Our new vector of age is `new_age = 10.0, 12.0, 13.5, 14.0, 15.4, 70`.

Median as a robust summary

- Let us compute the mean, we get 22.48333 in place of original 13.48333 years.
- Let us compute the median, we get 13.75 just like in the original.
- Note the way the arithmetic mean changes by a large amount while the median is not affected.
- When you have data that has extreme values, it is better to use the median in place of arithmetic mean.

The Mode

The mode is simply the most frequenty occurring element of a variable.

As an example, consider the following dataset.

x
1
2
2
4
3
2

In this case, 2 occurs 3 times. Hence 2 is the mode. In this case, the variable has only one mode. such a distribution is unimodal. It is possible for a variables to have two or more modes. A variable with two modes is bimodal. A distribution with two or more modes is multimodal. Hence a bimodal variable is also multimodal.

Quartiles

Quartiles are the three points in a data set that divide the data into 4 equal parts. The term quartile denotes `quarter` (1/4). Note that as the second quartile (2/4), the median is also a quartile. Hence we have 3 quartiles; 1st Quartile (1/4), second quartile (2/4) also called the median, and the third quartile (3/4). You must arrange the data in ascending or descending order to compute the quartiles including the median. Refer to our class discussion.