Additional Material - Data Cleaning

Contents

This document explains when we should use one of the three measures of central tendency (mean, median and mode) to replace missing values in our data.

1 Mean:

Mean is the most commonly used measure of central tendency..There are different types of mean, they are, arithmetic mean, weighted mean, geometric mean (GM) and harmonic mean (HM). If mentioned without an adjective (as mean), it generally refers to the arithmetic mean. This is to just find the average value(Manikandan (2011b)). This is used in most data replacement scenarios because its formula is simple and easy to understand. Let us consider the following data:

```
v_mean<-c(9,10,11,NA,NA,14,15,16,NA,18)
```

We can see that the data points are linearly distributed that is the difference between two consecutive data elements is uniform. In such a case, using mean to replace **NA** values is the best possible course of action. We use the mean() function to calculate the mean of our data:

```
v_mean[is.na(v_mean)]<-mean(v_mean,na.rm = TRUE)
v_mean
```

```
##  [1]  9.00000 10.00000 11.00000 13.28571 13.28571 14.00000 15.00000 16.00000
##  [9] 13.28571 18.00000
```

Using this command, we replaced all the **NA** values of the vector v with its mean. This mean was calculated after removing the **NA** values.

However, the mean of observations can't always be used to replace missing values. The following are the drawbacks(Manikandan (2011b)) of `mean`:

- The `mean` is highly sensitive to outliers.
- Time-series data can't be represented accurately by `mean`.

Let us now move on to the `median`, which can be used to compensate for some of the drawbacks of `mean`.

2 Median:

Median is the value which occupies the middle position when all the observations are arranged in an ascending/descending order. It divides the frequency distribution exactly into two halves. Fifty percent of observations in a distribution have scores at or below the median (Manikandan (2011a)). It is best used when the data is skewed that is, it has a lot of outliers. Also, data where mathematical average doesn't give an accurate picture can often be better described by the `median`. For example, consider the following data:

```
v_median<-c(7,36,8,24,NA,667,NA,3,NA,15)
```

We notice here that the data is skewed noticeably - we have single and double digit values interspersed with very large triple digit values. In such a case, calculating the `mean` of the data is going to be meaningless:

```
mean(v_median, na.rm=TRUE)
```

```
## [1] 108.5714
```

Does this value come close to most of the values of our data? No!In such a case, we should instead be looking to replace our data with the `median` value. We use median() function to calculate the median of our data:

```
v_median[is.na(v_median)]<-median(v_median,na.rm = TRUE)
v_median
```

```
##  [1]   7  36   8  24  15 667  15   3  15  15
```

Now our replaced values make a lot more sense.

The disadvantages(Manikandan (2011a)) of `median` are :

- It does not take into account the precise value of each observation and hence does not use all information available in the data.
- Unlike mean, median is not amenable to further mathematical calculation and hence is not used in many statistical tests.
- If we pool the observations of two groups, median of the pooled group cannot be expressed in terms of the individual medians of the pooled groups.

3 Mode:

Mode is defined as the value that occurs most frequently in the data(Manikandan (2011a)). It is useful in situations where either `mean` or `median` don't adequately represent the data. For example, consider the following dummy data representing the incomes of an agrarian economy:

```
v_mode=c(1000,1000,2500,NA,1000,NA,1000000,NA,4000000,
         1000,1000,2000,NA,2500,150000,2000,1000,2500,2500)
```

As we can see, the majority of the income is concentrated amongst the farmers, with a few richer middlemen and one millionaire leader. In such a case, both the `mean` and `median` are not great measures of data:

```
mean(v_mode,na.rm=TRUE)
```

```
## [1] 344666.7
```

Clearly the `mean` of the data doesn't even come close to representing all the values contained here. Let us try median now:

```
median(v_mode,na.rm=TRUE)
```

```
## [1] 2000
```

`Median` focuses too much on the middle values, but that is not where most of the economy is concentrated. Thus, in such a case, we will use `mode` to replace our **NA** values. Since R doesn't have a built-in `mode` function, we have to write a user-defined function for the same:

This function takes two inputs - vector x which we pass as input to find the mode and na.rm, which will instruct the function about whether we should remove **NA** values or not. Then, this function checks if our na.rm is set to TRUE, and if that is the case it will only take those entries which are not **NA**. Then, it stores all the unique values in a vector ux, and finally it returns the element that appears the most number of times in the original vector x using the which.max command.

```
Mode <- function(x,na.rm=FALSE) {
  if (na.rm){
    x = x[!is.na(x)]
  }
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

Now we will use this function to find the `mode` of our data.

```
v_mode[is.na(v_mode)]<-Mode(v_mode,na.rm=TRUE)
v_mode
```

```
##  [1]    1000    1000    2500    1000    1000    1000 1000000    1000 4000000
## [10]    1000    1000    2000    1000    2500  150000    2000    1000    2500
## [19]    2500
```

As we can see we have replaced our **NA** values with the`mode` of the data, which is 1000. This fits in perfectly with the scenario we had described.

The disadvantages(Manikandan (2011a)) of `mode` are:

- It is not used in statistical analysis as it is not algebraically defined.
- The fluctuation in the frequency of observation is more when the sample size is small.

Manikandan, S. 2011a. “Measures of Central Tendency: Median and Mode.” *Journal of Pharmacology and Pharmacotherapeutics* 2 (3): 214.

———. 2011b. “Measures of Central Tendency: The Mean.” *Journal of Pharmacology and Pharmacotherapeutics* 2 (2): 140.