Chapter 1: Descriptive Statistics

Contents

1 Measure of Central Value

Central value refers to the location of the centre of the distribution of data and is usually defined by the average (arithmetic mean, or mean), median or mode.

1.1 The mean

The **mean** is defined as the arithmetic average of all the values. The average of a sample of n observations is the sum of the observations in the sample divided by the number of units on the samples

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Advantages:

- It is the most commonly used measure of location or central tendency for continuous variables.
- The arithmetic mean uses all observations in the data set. All observations are given equal weight.

Disadvantages:

- The mean is affected by extreme values that may not be representative of the sample (see example of the average wealth of patrons before and after Microsoft cofounder Bill Gates was added). The trimmed mean is computed after trimming a certain percentage of the smallest and largest values from the data. If they are values that skew the mean, they won't skew the trimmed mean.

Import the data set *temp_warm.txt* as `temperature`.

Compute the average `October` temperature

```
mean(temperature$October)
```

```
## [1] 11.00286
```

```
summary(temperature$October)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.50    8.65   10.20   11.00   13.30   19.40
```

Compute summary statistics for the `October` temperature by `Area`

```
library(dplyr)
by_area <- group_by(temperature, Area)
summarise(by_area, mean(October), n = n())
```

```
## # A tibble: 4 x 3
##   Area  `mean(October)`     n
##   <fct>           <dbl> <int>
## 1 East             7.95     8
## 2 North            7.41     8
## 3 South           16.4     10
## 4 West            10.9      9
```

1.2 The median

The **median** is the middle value of a group of an odd number of observations when the data is arranged in increasing or decreasing order. If the number of values is even, the median is the average of the two middle values.

Advantages:

- The median always exists and is unique.
- The median is not affected by extreme values.

Disadvantages:

- The values must be sorted in order of magnitude.
- The median uses only one (or two) observations.

Compute median `October` temperature

```
median(temperature$October)
```

```
## [1] 10.2
```

Compute median `October` temperature by `Area`

```
by_area <- group_by(temperature, Area)
summarise(by_area, mean(October), n = n(), median(October))
```

```
## # A tibble: 4 x 4
##   Area  `mean(October)`     n `median(October)`
##   <fct>           <dbl> <int>             <dbl>
## 1 East             7.95     8              8.05
## 2 North            7.41     8              7.6
## 3 South           16.4     10             17.4
## 4 West            10.9      9             11.1
```

1.3 Percentiles / Quartiles

A percentile is a generalization of the concept of the median. **Percentiles** are values that divide a distribution into two groups where the P^{th} percentile is larger than $P\%$ of the values.
Some specific percentiles have special names:

- First quartile: Q1 = the 25 percentile
- Median: Q2 = the 50 percentile

- Third quartile: Q3 = the 75 percentile

Compute quantiles of `October` temperature

```
q1 <- quantile(temperature$October, 0.25)
q1
```

```
##  25%
## 8.65
```

```
q3 <- quantile(temperature$October, 0.75)
q3
```

```
##  75%
## 13.3
```

```
quant <- quantile(temperature$October)
quant
```

```
##    0%   25%   50%   75%  100%
##  4.50  8.65 10.20 13.30 19.40
```

This means that 25% of the cities have `October` temperature less than `8.65`. This means that 75% of the cities have `October` temperature less than `13.3`.

```
summarise(group_by(temperature, Area), median(October), n = n(),
          quantile(October, 0.25), quantile(October, 0.75))
```

```
## # A tibble: 4 x 5
##   Area  `median(October)`     n `quantile(October, 0.25~ `quantile(October, 0.7~
##   <fct>             <dbl> <int>                    <dbl>                    <dbl>
## 1 East               8.05     8                     5.65                     9.73
## 2 North              7.6      8                     5.58                     9.02
## 3 South             17.4     10                    14.6                     18.2
## 4 West              11.1      9                     9.8                     11.5
```

1.4 The mode

The **mode** is the most frequent or most typical value. Usually we compute the mode for a categorical variable.

Advantages:

- Requires no calculations
- Represents the value that occurs most often

Disadvantages:

- The mode for continuous measurements is dependent on the grouping of the intervals. Without grouping, there may be many modes. The mode is not a very good summary of the dataset, as the mode can correspond to an extreme value.

Make a frequency table of the variable `Area`.

```
table <- table(temperature$Area)
table
```

```
##
##  East North South  West
##     8     8    10     9
```

Compute the mode.

```r
which.max(table)
```

```
## South
##     3
```

2 Measure of dispersion (or variability)

2.1 The range

The **range** is the easiest of all measures of dispersion to calculate. The range is the difference between the largest and the smallest value in the sample.

Range = Maximum value - Minimum value

Advantages:

- The range is easily understood and gives a quick estimate of dispersion.
- The range is easy to calculate when the sample size is small, ($n < 10$).

Disadvantages:

- The range is inefficient because it only uses the extreme values and ignores all other available data. The larger the sample size, the more inefficient the range becomes.

2.2 The variance

The **variance** is the mean square deviation of the observations from the mean.
To calculate the (sample) variance, we use the following equation:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} \tag{2}$$

The variation of a single datapoint is assessed using the value $(x_i - \overline{x})$.
A sum of these differences will give sense of the total variation. Just adding these values gives a value of 0, as terms cancel. Therefore the square is taken.
If there is a lot of spread in the data this sum will be relatively large. Of course, this sum can be large because n is large, not just because there is much spread. Therefore we divide the sum by a scale factor. We divide by $n-1$ to produce the sample variance.

Advantages:

- The variance is an efficient (unbiased) estimator.
- Variances can be added and averaged. It is of special value in the *Analysis of Variance* (ANOVA) to separate the components of variance. **Disadvantages**
- The calculation of the variance can be tedious without the aid of a calculator or computer.
- The variance is not linear and thus is not in the same units as the variable being evaluated.

Compute the variance of `October` temperature

```r
varT <- var(temperature$October)
varT
```

```
## [1] 18.69029
```

2.3 The standard deviation

The square root of the variance is known as the **standard deviation**. The symbol for the (sample) standard deviation is s.

$$s = \sqrt{s^2} \tag{3}$$

Advantages:

- The standard deviation is in the same dimension as the observed values
- The standard deviation is an efficient (unbiased) estimator.

Disadvantages:

- The calculations can be tedious without the aid of a good calculator.

Compute the standard deviation of `October` temperature

```
sdT <- sd(temperature$October)
sdT
```

```
## [1] 4.323226
```

2.4 Interquartile range

The difference between the 25^{th} and the 75^{th} quartiles is the **interquartile range**.

Compute the interquartile range of `October` temperature

```
IQRT <- IQR(temperature$October)
IQRT
```

```
## [1] 4.65
```