Tool to Analyse Descriptive Statistics

Asst. Prof. Ashwini Mathur

Statistics of variation for interval/ratio data

Standard deviation

The standard deviation is a measure of variation which is commonly used with interval/ratio data. It's a measurement of **how close the observations in the data set are to the mean**.

Normally distributed data—68% of data points fall within the **mean** \pm **1 standard deviation**, 95% of data points fall within the **mean** \pm **2 standard deviations**, and 99.7% of data points fall within the **mean** \pm **3 standard deviations**.

Because the mean is often represented with the letter *mu*, and the standard deviation is represented with the letter *sigma*,

sd(Data\$Attendees)

4.830459

Standard deviation may not be appropriate for skewed data.

Standard error of the mean

Standard error of the mean is a **measure that estimates how close a** calculated mean is likely to be to the true mean of that population.

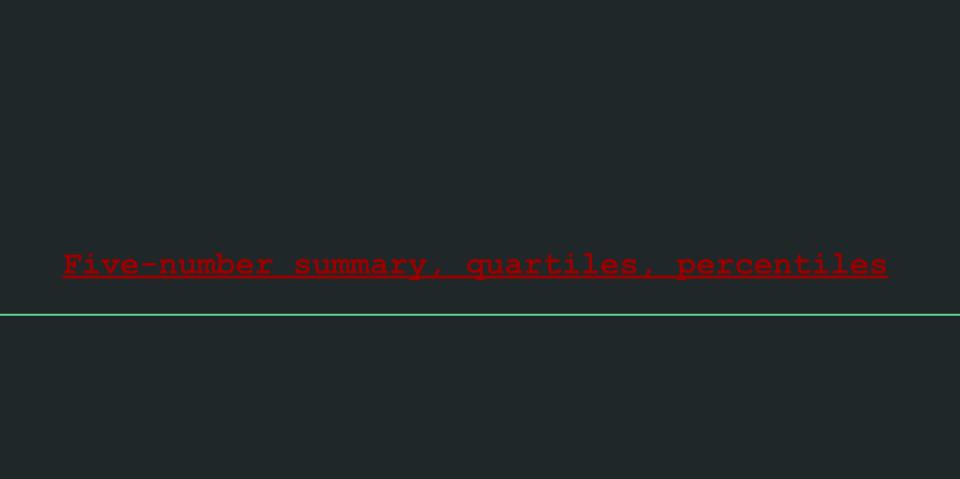
It is commonly used in tables or plots where multiple means are presented together.

For example, we might want to present the mean attendees for Ren with the standard error for that mean and the mean attendees for Stimpy with the standard error that mean.

The **standard error** is the standard deviation of a data set divided by the square root of the **number of observations**. It can also be found in the output for the *describe* function in the *psych* package, labelled *se*.

```
sd(Data$Attendees) /
  sqrt(length(Data$Attendees))
 1.207615
library (psych)
describe (Data$Attendees)
   vars n mean sd median trimmed mad min max range skew kurtosis
                                                          se
    ###
        se indicates the standard error of the mean
```

Standard error of the mean **may not be appropriate** for skewed data.



Five-number summary, quartiles, percentiles

The median is the same as the 50th percentile,

because 50% of values fall below this value. Other percentiles for a data set can be identified to provide more information.

Typically, the **o**th, **25**th, **50**th, **75**th, **and 100**th percentiles are reported. This is sometimes called the **five-number summary**.

These values can also be called the minimum, 1st quartile, 2nd quartile, 3rd quartile, and maximum.

The five-number summary is a useful measure of variation for skewed interval/ratio data or for ordinal data.

25% of values fall below the 1st quartile and 25% of values fall above the 3rd quartile. This leaves the middle 50% of values between the 1st and 3rd quartiles, giving a sense of the range of the middle half of the data.

This range is called the *interquartile range* (IQR).

Percentiles and quartiles are relatively robust, as **they aren't affected much by a few extreme values**. They are appropriate for both skewed and unskewed data.

summary (Data\$Attendees)

```
Min. 1st Qu. MedianMean 3rd Qu. Max.
6.00 11.75 15.00 14.50 17.25 23.00
```

The five-number summary and the mean

Optional technical note on calculating percentiles

It may have struck you as odd that the 3rd quartile for *Attendees* was reported as 17.25. After all, if you were to order the values of *Attendees*, the 75th percentile would fall between 17 and 18. But **why does R go with 17.25 and not 17.5**?

```
sort(Data$Attendees)
6  7  9 11 12 13 14 15 15 16 17 18 19 22 23
```

The answer is that there are several different methods to calculate percentiles, and they may give slightly different answers. For details on the calculations, see *?quantiles*.

Percentiles other than the 25th, 50th, and 75th can be calculated with the quantiles function. For example, to calculate the 95th percentile:

```
quantile(Data$Attendees, .95)
    95%
    22.25
```

For *Attendees*, the default type 7 calculation yields a 75th percentile value of 17.25, whereas the type 2 calculation simply splits the difference between 17 and 18 and yields 17.5. The type 1 calculation doesn't average the two values, and so just returns 17.

```
quantile(Data$Attendees, 0.75, type=7)
    75%
 17.25
quantile(Data$Attendees, 0.75, type=2)
   75%
 17.5
```