R for Statistics

Prof. Ashwini Mathur

Descriptive Statistics ..

Overview of Descriptive analysis …

Descriptive statistics are used to **summarize data** in a way that provides insight into the information contained in the data.

For Example:

Examining the **mean or median** of numeric data or the **frequency of observations** for nominal data. **Plots** can be created that show the data and indicating summary statistics.

Descriptive statistics for different types of data

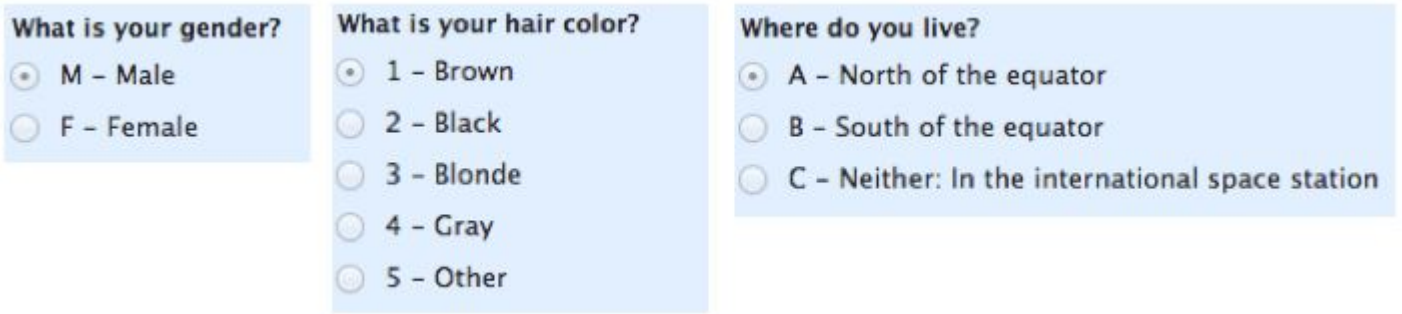
Choosing which summary statistics are appropriate depend on the type of **variable** being examined.

Different statistics should be used for **interval/ratio, ordinal, and nominal data**.

These statistical terms are also called as **Measurement scales**.

Nominal Data

Nominal scales are used for labeling variables, without any **quantitative** value. “Nominal” scales could simply be called “**labels**.”

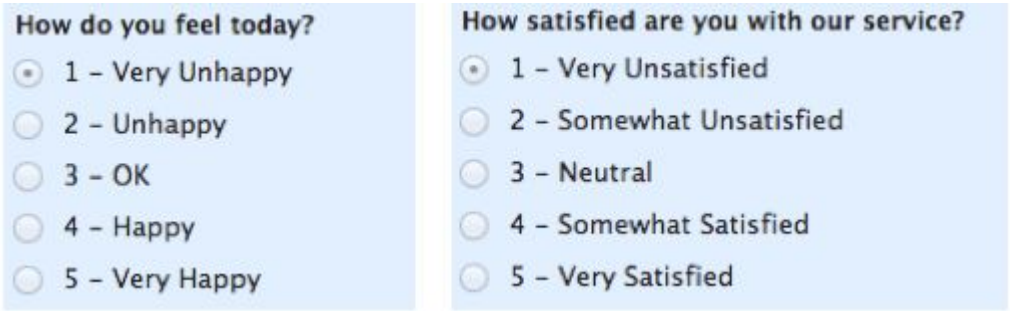


Ordinal Data

With ordinal scales, the *order* of the values is what's important and significant, but the differences between each one is not really known.

For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.

Ordinal scales are typically **measures of non-numeric concepts** like satisfaction, happiness, discomfort, etc.



Ordinal Data

Interval Data

Interval scales are numeric scales in which we know both the order and the exact differences between the values.

The classic example of an interval scale is **Celsius temperature** because the difference between each value is the same.

For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

Interval

Ratio

Ratio scales are the mostly used when it comes to data measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an absolute zero–which allows for a wide range of both **descriptive and inferential statistics** to be applied

Ratio scales provide a wealth of possibilities when it comes to statistical analysis.

These variables can be meaningfully added, subtracted, multiplied, divided (ratios).

Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Summary

Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔

Describing or examining data :

Location* is also called *central tendency. It is a measure of the values of the data.

For example, are the values close to 10 or 100 or 1000?

Measures of location include mean and median, as well as somewhat more exotic statistics like M-estimators or Winsorized means.

Cont..

Variation* is also called *dispersion. It is a measure of how far the data points lie from one another. Common statistics include standard deviation and coefficient of variation. For data that aren't normally-distributed, percentiles or the interquartile range might be used.

***Shape* refers to the distribution of values**. The best tools to evaluate the shape of data are histograms and related plots. Statistics include skewness and kurtosis, though they are less useful than visual inspection. We can describe data shape as normally-distributed, log-normal, uniform, skewed, bi-modal, and others.

Descriptive statistics for interval/ratio data

For this example, imagine that Ren and Stimpy have each held eight workshops to educating the public about water conservation at home.

They are interested in how many people showed up to the workshops.

Because the data formatted in a data frame, we can use the convention *Data$Attendees* to access the variable *Attendees*

within the data frame *Data*.

```
Input = ("
Instructor  Location  Attendees
Ren       North      7
Ren       North  22
Ren       North   6
Ren       North  15
Ren       South  12
Ren       South  13
Ren       South  14
Ren       South  16
Stimpy    North  18
Stimpy      North      17
Stimpy      North      15
Stimpy     North        9
Stimpy     South       15
Stimpy     South       11
Stimpy     South       19
Stimpy     South       23
")
```

R Syntax :

```
Data = read.table(textConnection(Input),header=TRUE)

Data            ### Will output data frame called Data

str(Data)       ###  but shows the structure of the data
frame

summary(Data)   ###   but summarizes variables in the data
frame
```

Functions *sum* and *length*

The sum of a variable can be found with the *sum* function, and the number of observations can be found with the *length* function.

```
sum(Data$Attendees)
```

```
232
```

```
length(Data$Attendees)
```

```
16
```

Statistics of location for interval/ratio data

Mean

The mean is the arithmetic average, and is a common statistic used with interval/ratio data. It is simply the sum of the values divided by the number of values. The *mean* function in R will return the mean.

```
sum(Data$Attendees) / length(Data$Attendees)

  14.5

mean(Data$Attendees)

  14.5
```

Important :

Caution should be used when **reporting mean values with skewed data**, as the **mean may not be representative of the center of the data**.

For example, imagine a town with 10 families, nine of whom have an income of less than $50, 000 per year, but with one family with an income of $2,000,000 per year. The mean income for families in the town would be $233,000, but **this may not be a reasonable way to summarize the income of the town**.

```
Income = c(49000, 44000, 25000, 18000, 32000, 47000, 37000, 45000, 36000,
2000000)

mean(Income)

   233300
```

Median

The **median is defined as the value below which are 50% of the observations**. To find this value manually, you would order the observations, and separate the lowest 50% from the highest 50%. For data sets with an odd number of observations, the median is the middle value. For data sets with an even number of observations, the median falls **half-way** between the two middle values.

The median is a robust statistic in that it is not affected by adding extreme values. For example, if we changed Stimpy's last *Attendees* value from 23 to 1000, it would not affect the median.

```
median(Data$Attendees)
```

```
15
```

```
###   Note that in this case the mean and median are close in value
###   to one another.  The mean and median will be more different
###   the more the data are skewed.
```

The **median is appropriate** for either **skewed or unskewed data**. The median income for the town discussed above is $40,500. Half the families in the town have an income above this amount, and half have an income below this amount.

```
Income = c(49000, 44000, 25000, 18000, 32000, 47000, 37000, 45000, 36000,
2000000)

median(Income)

   40500
```

Note that medians are sometimes reported as the “average person” or “typical family”. Saying, “The average American family earned $54,000 last year” means that the median income for families was $54,000. The “average family” is that one with the median income.

Mode

The **mode is a summary statistic** that is used **rarely in practice**, but is normally included in any discussion of mean and medians. When there are discrete values for a variable, the **mode is simply the value which occurs most frequently**. **For example**, in the Statistics Learning Center video in the *Required Readings* below, Dr. Nic gives an example of counting the number of pairs of shoes each student owns. The most common answer was 10, and therefore 10 is the mode for that data set.

For our Ren and Stimpy example, the value 15 occurs three times and so is the mode.

The *Mode* function can be found in the package *DescTools*.

```
library(DescTools)

Mode(Data$Attendees)

  15
```

DOUBT ???