# Statistics for the Sciences

## Variables and Distributions

Xuemao Zhang
East Stroudsburg University

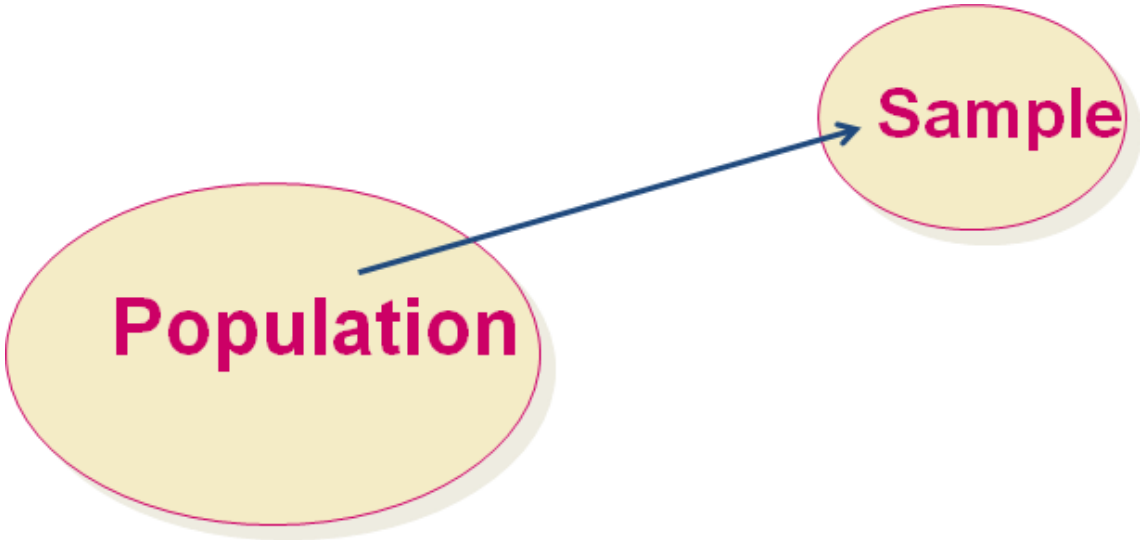
January 18, 2025

# Outline

- Sample and Population
- Types of data
  - Classification of variables and data
- Descriptive statistics and inferential statistics
  - Parameter and statistics
  - Descriptive statistics
  - Inferential statistics
- Distributions of univariate data

# Sample and Population

- An investigation will typically focus on a well-defined collection of subjects constituting a population of interest.




- Population : The complete collection of all subjects that are being considered.
- Sample: Subcollection of subjects selected from a population.

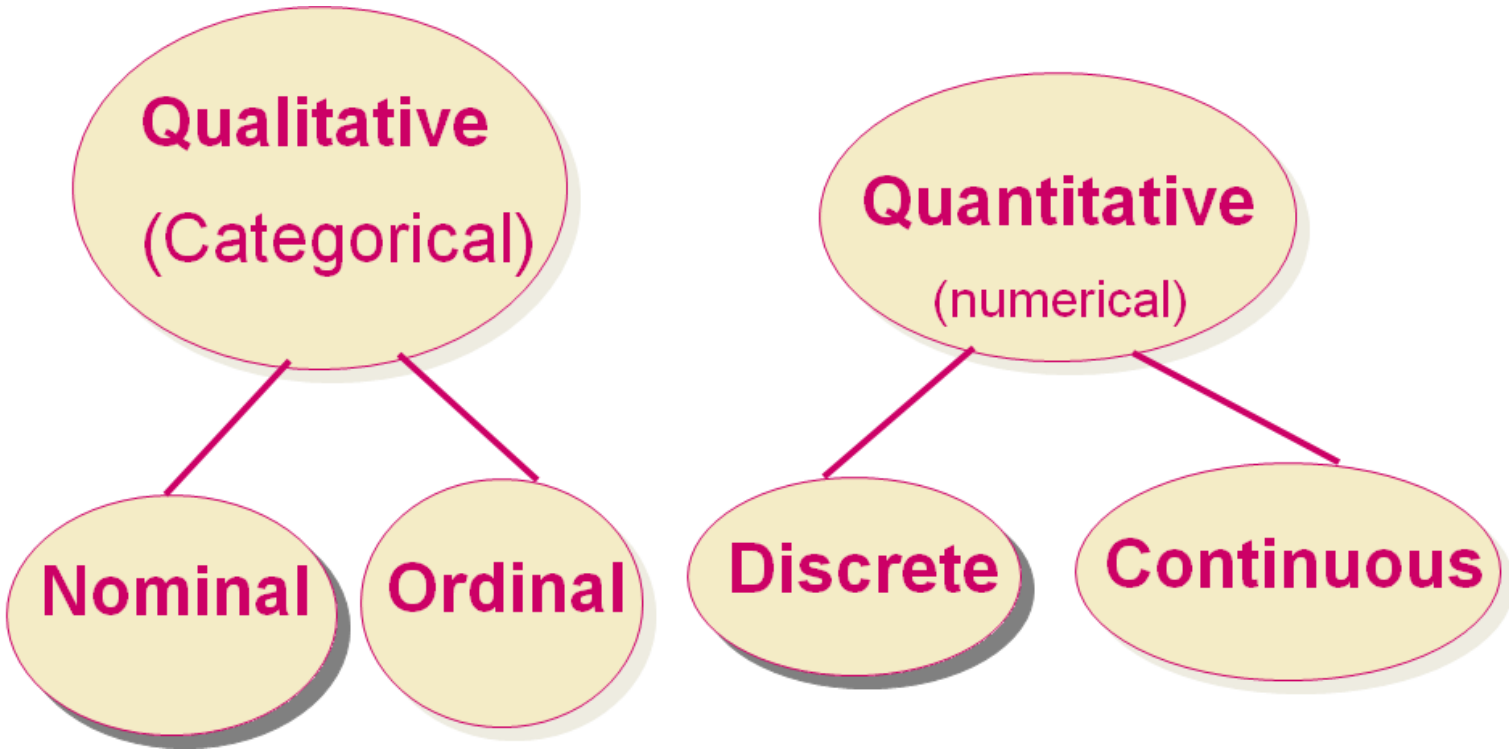
# Data

- A data set is a collection of measurements of one variable or several variables for some individuals or subjects.
- Often, a data set is a file, in which each column (or field) corresponds to an variable(or attribute) and each row corresponds to measurements of all variables for each subject. This type of data sets is called record-based data.
  - Example: Elephant seal foraging

```
##    male departwt distance FFAduration durationto durationfrom
## 1   Pop       NA      534          31         18           11
## 2   Alt      973      755          89          9            8
## 3   Pro      977     1210          77         12           18
## 4   Hal     1121       NA          NA         NA           NA
## 5   Blu       NA     1297          76         19           25
## 6   Dua      996     1487          68         18           23
## 7   Rov     1100     2073          69         29           25
## 8   Ric     1068     2181          46         21           42
## 9   Ori     1097       NA          NA         NA           NA
## 10  Jer     1199       NA          NA         NA           NA
```

# Variables

- A **variable** (or attribute) is a property or characteristic that can vary from one subject to another or frome one time to another.
  - A **data set** is obtained by measuring variables.
- Classification of variables by the type of measurements



# Categorical Variables

- Categorical variables take category or label/name values, and place an individual into one of several groups.
  - They cannot be used for computations.
- Categorical variables can be further classified using levels of measurement by looking at what is being measured.
  - **Nominal**, when there is no natural ordering among the categories.
    - Common examples would be like gender, eye color.
  - **Ordinal**, when there is a natural order among the categories, such as, ranking scales or letter grades.
    - Differences between data values either cannot be determined or are meaningless.
    - Examples: biological family and genus

# Categorical Variables

- Sengi example (Kaufman et al. (2013)) from Chapter 5:

```
##             family        genus   species bodymass brainmass        relat
## 1  Solenodontidae    Solenodon paradoxus    672.0      4723 laurasiather
## 2      Tenrecidae       Tenrec ecaudatus    852.0      2588     afrother
## 3      Tenrecidae      Setifer   setosus    237.0      1516     afrother
## 4      Tenrecidae Hemicentetes  semispin    116.0       839     afrother
## 5      Tenrecidae     Echinops  telfairi     87.5       623     afrother
## 6      Tenrecidae  Oryzorictes talpoides     44.2       580     afrother
## 7      Tenrecidae    Microgale    cowani     15.2       420     afrother
## 8      Tenrecidae    Limnogale  mergulus     92.0      1150     afrother
## 9      Tenrecidae    Microgale   dobsoni     31.9       557     afrother
## 10     Tenrecidae    Microgale  talazaci     48.2       766     afrother
##              relation2
## 1  other insectivore
## 2  other insectivore
## 3  other insectivore
## 4  other insectivore
## 5  other insectivore
## 6  other insectivore
## 7  other insectivore
## 8  other insectivore
```

# Categorical Variables

- All unique `family` in the data:

```
## [1] "Solenodontidae"  "Tenrecidae"      "Chrysochloridae" "Erinac
## [5] "Soricidae"       "Talpidae"        "Macroscelididae"
```

- Frequency table

```
##            family Freq
## 1 Chrysochloridae    2
## 2     Erinaceidae    4
## 3 Macroscelididae    5
## 4  Solenodontidae    1
## 5       Soricidae   27
## 6        Talpidae    5
## 7      Tenrecidae   12
```

# Numerical Variables

- Numerical variables take numerical values, and represent some kind of measurement.
- Numerical variables are often further classified by the number of values:
  - **Discrete**, when the variable takes on a finite or countably infinite number of values.
    - ⋆ Most often these variables indeed represent some kind of **count**.
  - **Continuous**, when the variable takes infinitely many values corresponding to the points on a real line interval
    - ⋆ Units should be provided.
    - ⋆ Our precision in measuring these variables is often limited by our instruments.
    - ⋆ Common examples would be like height (inches) and weight.

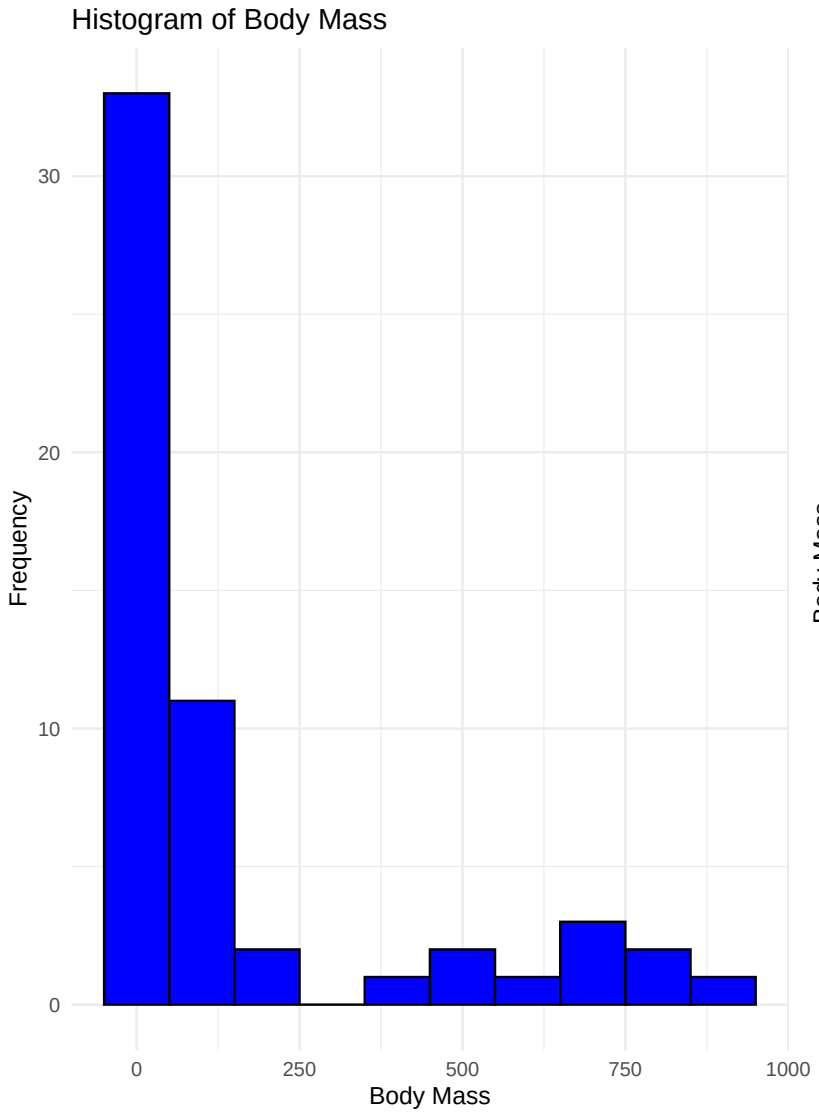
# Numerical Variables

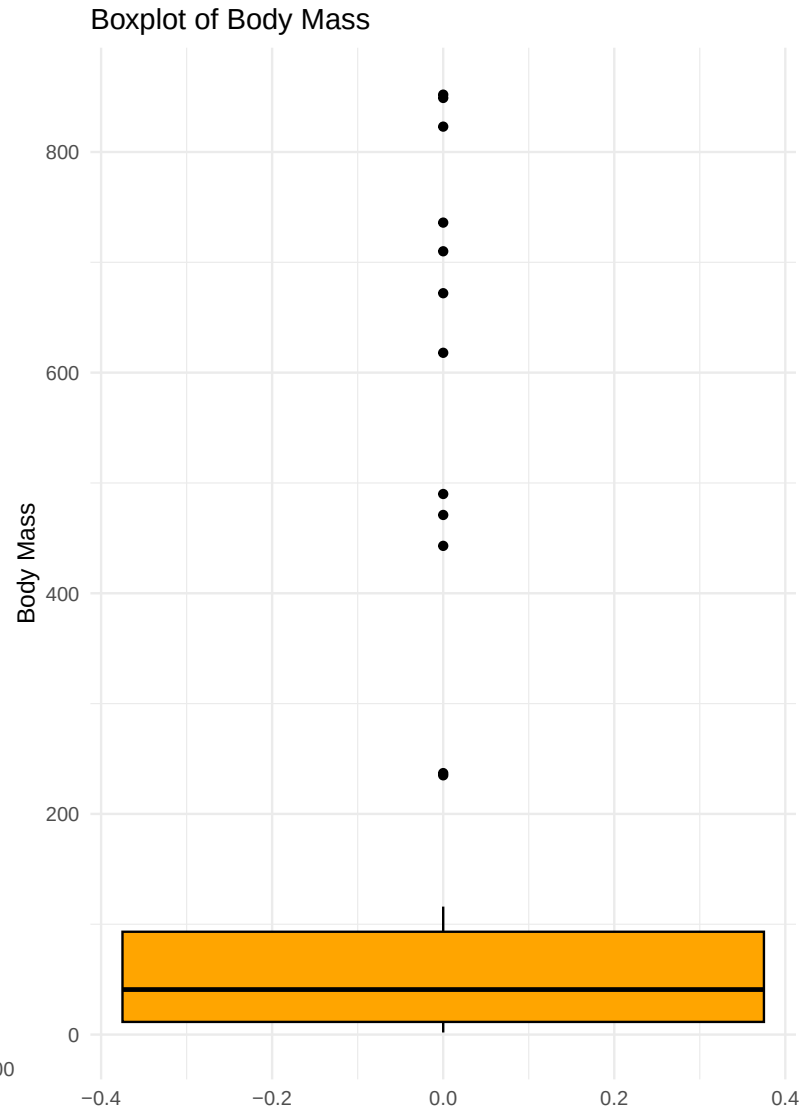
- Example: consider variable `bodymass` in the data set `kaufman`

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.90   11.47   40.80  154.09   93.20  852.00
```

# Numerical Variables

- Example: consider variable `bodymass` in the data set `kaufman`



# Univariate and Multivariate data

- A **univariate** data set consists of observations on a single variable.
  - For example, the following sample of lifetimes (hours) of brand D batteries put to a certain use is a numerical univariate data set:
    5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5
- We have **bivariate** data when observations are made on each of two variables.
- **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

# Parameter and Statistic

- **Parameter** is a numerical summary describing some variable of a population.
  - For example, population mean $\mu$, population proportion $p$
- **Statistic** is a numerical summary describing some variable of a sample
  - A statistic is an estimator of some parameter in a population.
  - For example, sample mean $\bar{x}$, sample proportion $\hat{p}$

# Descriptive statistics

- When desired information is available for all subjects in the population, we have what is called a **census**.
  - Census is costly and time-consuming. For example, the United States Bureau of the Census is conducting the U.S. Census every ten years.
- DESCRIPTIVE STATISTICS: Procedures used to summarize and describe a set of measurements.
  - Only Descriptive statistics is needed to analyze census data
  - Inferential statistics is needed for sample data

# Inferential statistics

- INFERENTIAL STATISTICS: Procedures used to draw conclusions or inferences about the population (parameter or distribution) from information contained in a **random sample** selected from the population.
  - When we cannot enumerate the whole population, we use both Descriptive Statistics and Inferential Statistics.
  - **Probability theory** is necessary for us to make statistical inferences.

# Distribution of a variable

- What does the word **distribution** mean?

# Distribution of a variable

- Distribution is the description of data values and frequencies of a (uni-variate)data set.
  - Location: where are the values for numerical data or what are the categories for categorical data
  - How often: frequency or relative frequency of data values or categories.

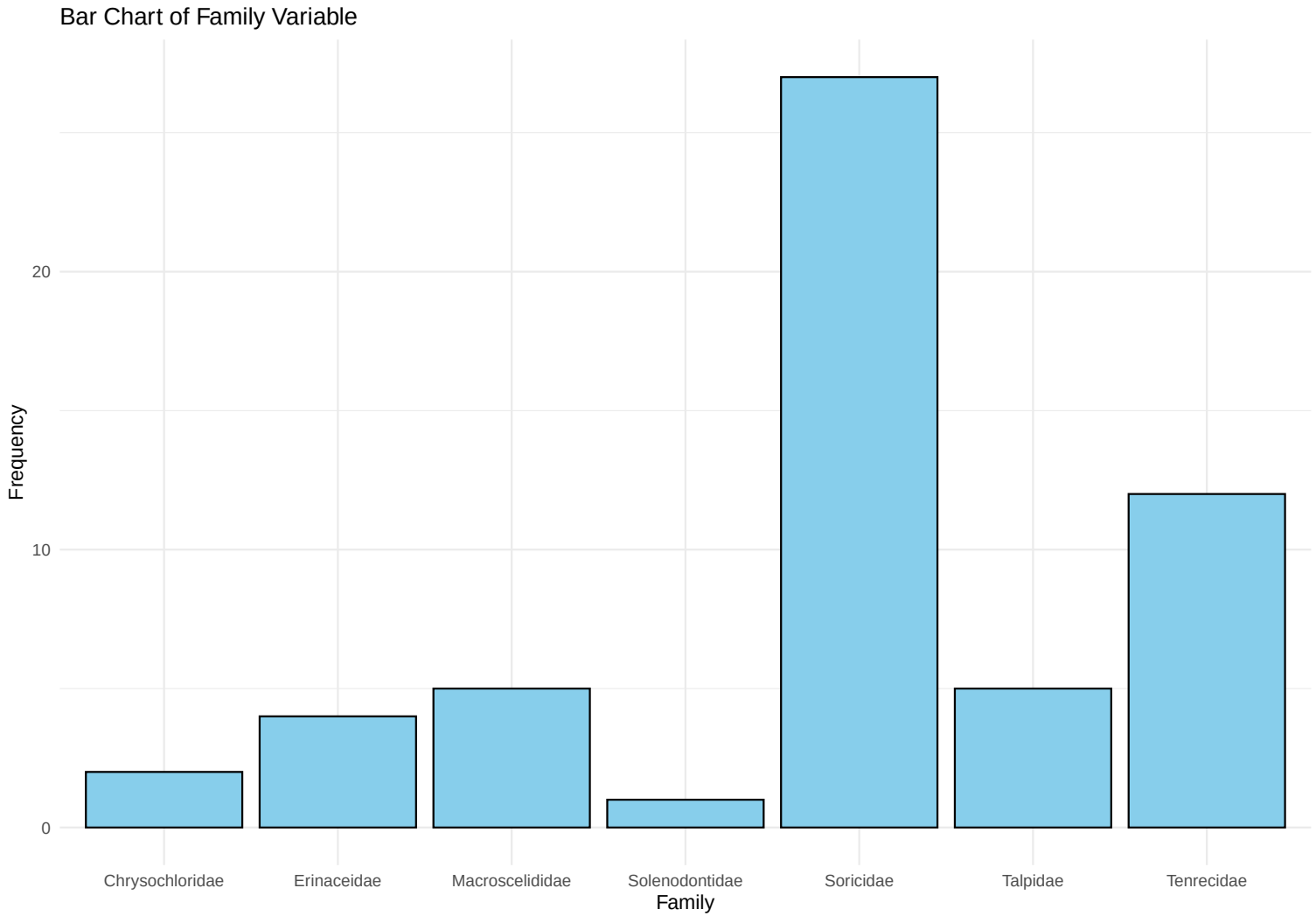
# Distribution of a catetorical variable

- For a categorical variable, we use frequency table or relative frequency table

```
##              family Freq Relative_Frequency
## 1 Chrysochloridae    2              0.0357
## 2     Erinaceidae    4              0.0714
## 3 Macroscelididae    5              0.0893
## 4  Solenodontidae    1              0.0179
## 5       Soricidae   27              0.4821
## 6        Talpidae    5              0.0893
## 7      Tenrecidae   12              0.2143
```

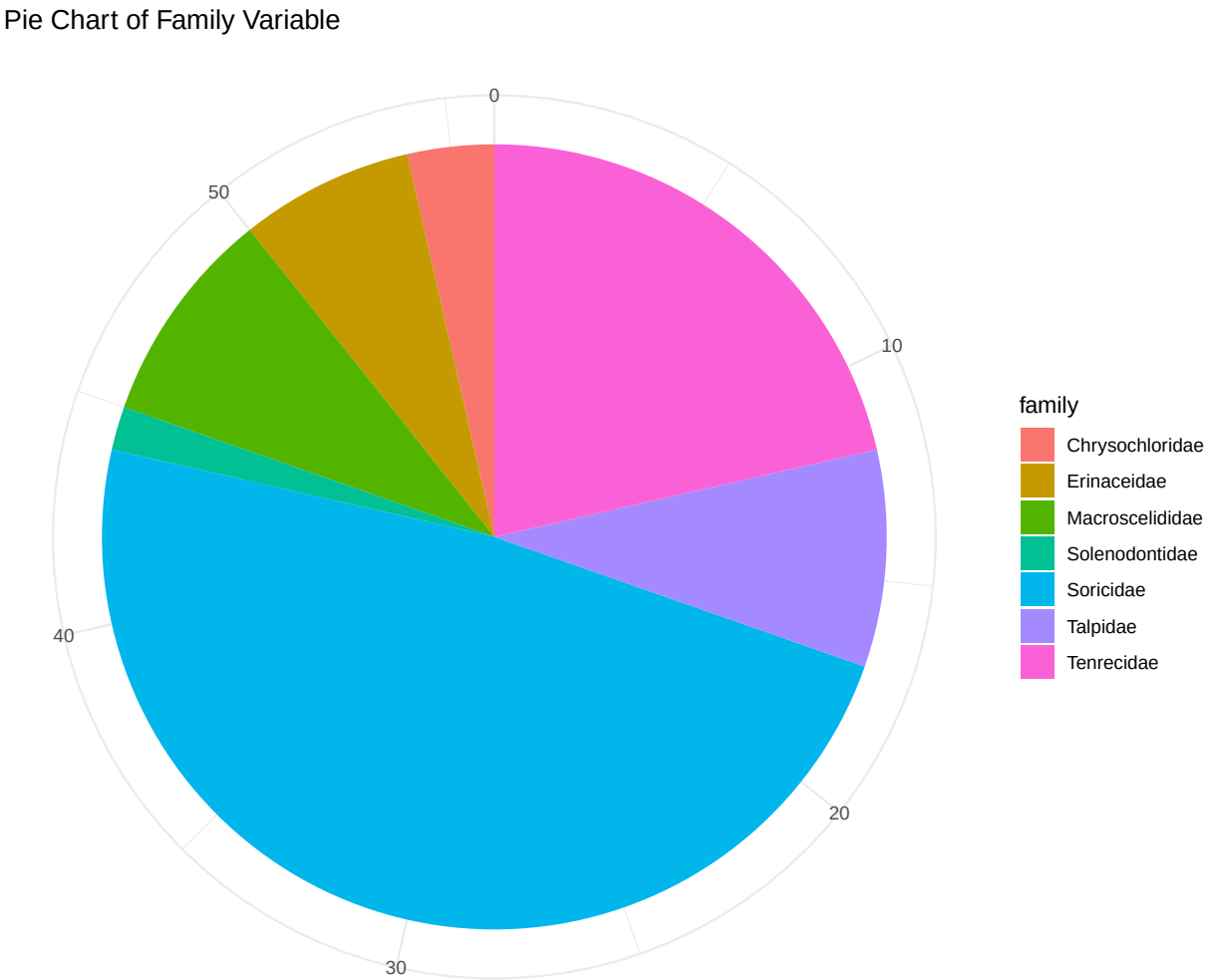
# Distribution of a catetorical variable

- Data visualization of a categorical variable using bar chart or pie chart



# Distribution of a catetorical variable

- Data visualization of a categorical variable using bar chart or pie chart



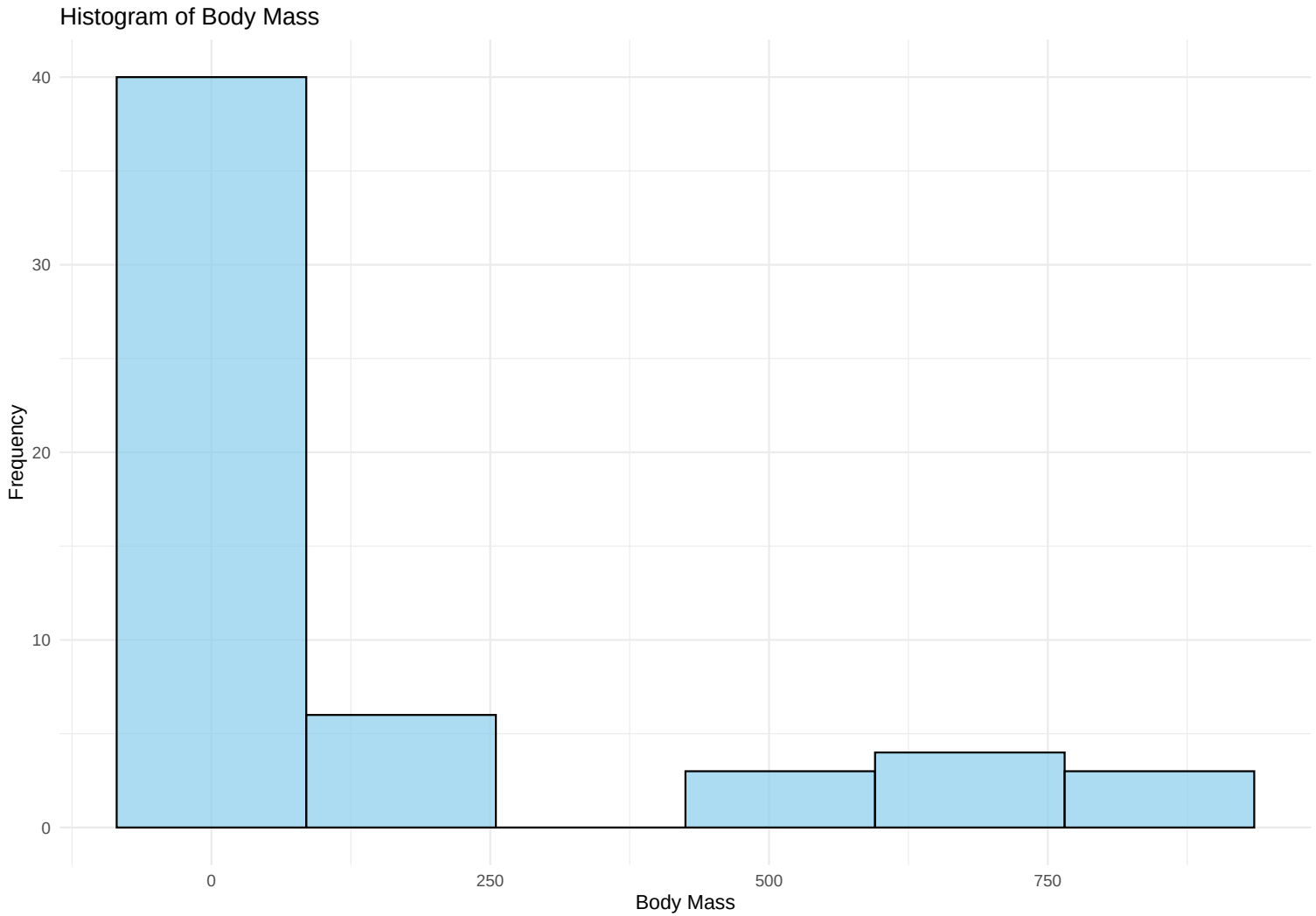
# Distribution of a continuous variable

- Frequency and relative frequency table is constructed in a way similar to the analysis of categorical variable. For example,

```
##   bodymass_intervals Freq Relative_Frequency
## 1          [1.9,144]   44              0.7857
## 2          (144,285]    2              0.0357
## 3          (285,427]    0              0.0000
## 4          (427,569]    3              0.0536
## 5          (569,710]    3              0.0536
## 6          (710,852]    4              0.0714
```

# Distribution of a continuous variable

- Then it results in a graph called **histogram**



# Characterizing a Distribution - Center

- **Mean** Denote the numerical variable by $x$
- $\bar{x} = \sum_{i=1}^{n} x_i/n$
- Location *parameter*: for example $\mu = E(X)$

```
## [1] "Mean of body mass = 154.086"
```

- **Median**

```
## [1] "Median of body mass = 40.8"
```

- **Mode** is the most frequently occurring value in a data set.
  - It is defined as a hump in a histogram

```
## [1] "Mode of body mass is 10.2"
```

# Characterizing a Distribution - Spread

- **Range**: max-min

```
## [1] "Range of body mass is 852 - 1.9 = 850.1"
```

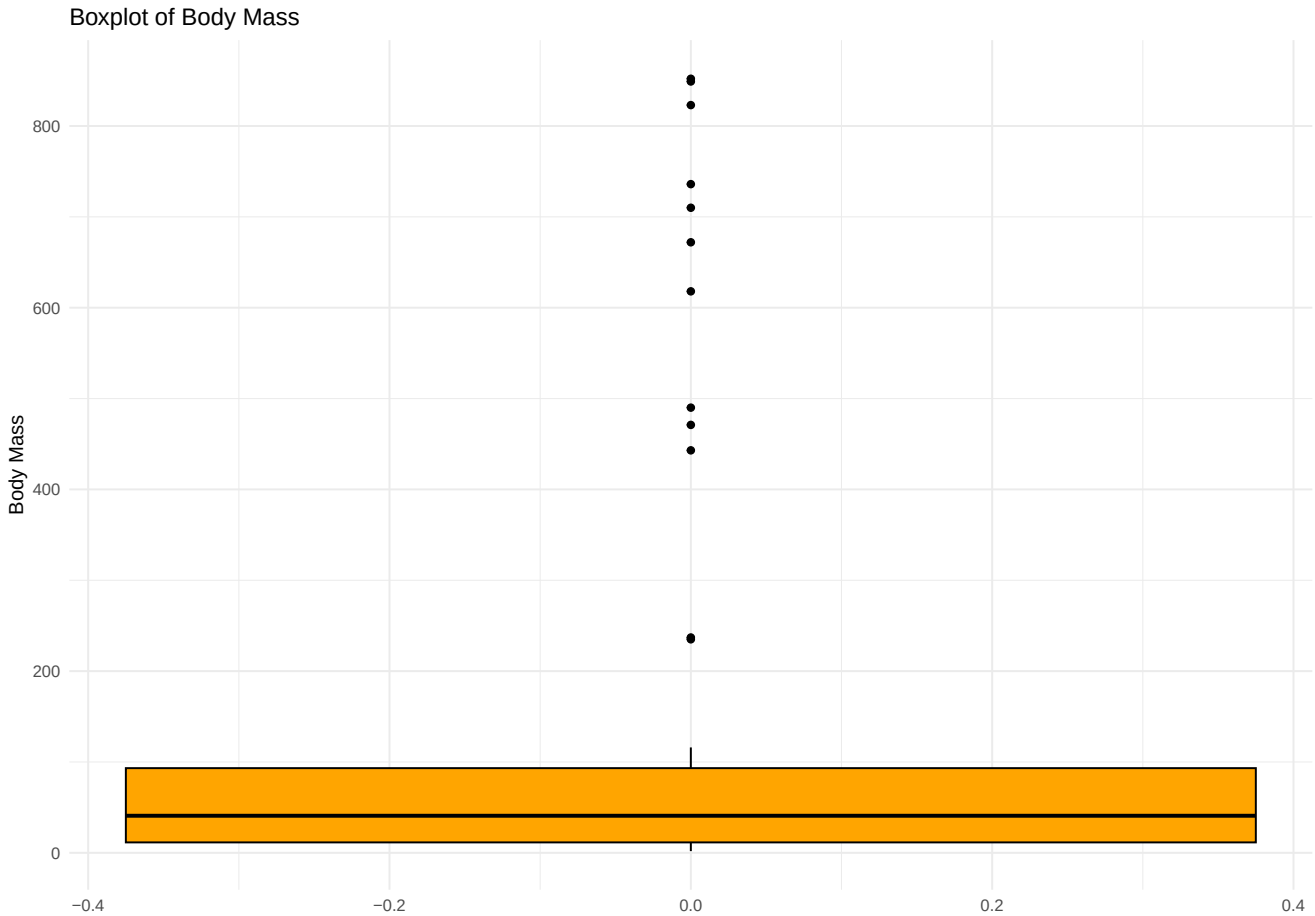
# Characterizing a Distribution - Spread

- **Percentiles** give the value below which a given percentage of the data values occur.
  - The 50th percentile is the median.
  - There are various algorithms calculating percentiles
- The three quartiles

```
##     25%     50%     75%
## 11.475 40.800 93.200
```

# Characterizing a Distribution - Spread

- IQR (Inter-Quartile Range) $= Q_3 - Q_1$
- Box-plot components
  - $Q_1$, $Q_2$(median) and $Q_3$
  - Outliers (determined by lower-fence and upper-fence)



# Characterizing a Distribution - Spread

- Sample **variance**
  - $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$
- Sample **standard deviation**
  - $s = \sqrt{s^2}$

```
## [1] "Variance of body mass = 64231.318"
```

```
## [1] "S.t.d. of body mass = 253.439"
```

- The **standard error** is the estimate of the standard deviation of a statistic when the statistics is considered as a random variable.
- For normally distributed data, the standard error (SE) of the sample mean $\bar{x}$ is $SE(\bar{x}) = \frac{s}{\sqrt{n}}$.

# License