APPM 4570/5570

Unit 1: Exploratory Data Analysis (EDA)

(Ch 1.1, 1.3, 1.10 -- 1.13, 2.4.3, 2.5)

Learning Objectives

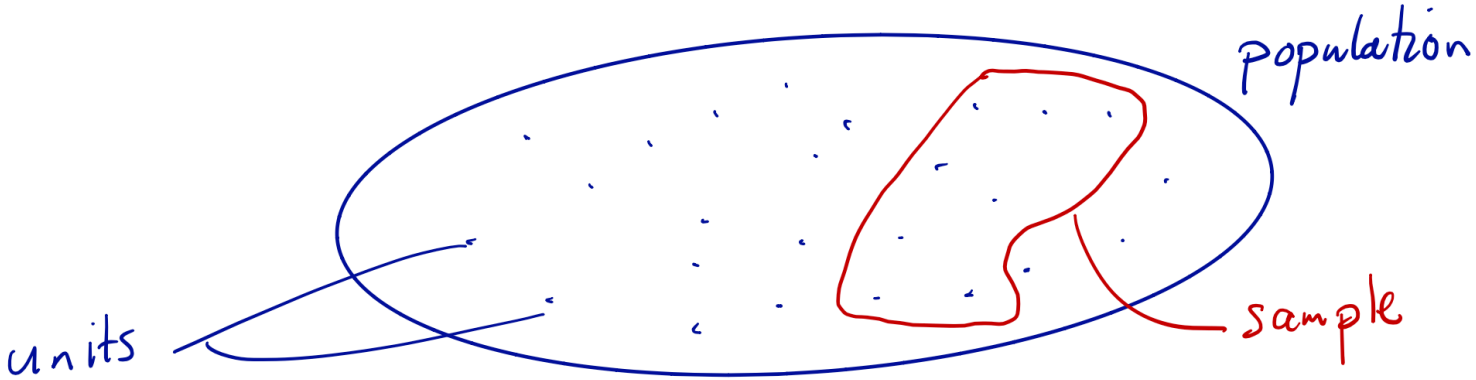
At the end of this unit, students should be able to:

1. Define a *population, sample, sample frame, variable of interest* and identify these concepts in particular examples.
2. Describe what is meant by "statistical inference".
3. Define, calculate, and interpret three measures measures of center.
4. Describe situations in which one measure might be better than another.
5. Define, calculate, and interpret three measures of variation.
6. Define, calculate, and interpret quantiles and percentiles.
7. Produce and interpret histograms; identify whether the distribution of a variable is unimodal or multimodal, based on a histogram.
8. Produce and interpret boxplots, and scatterplots.
9. Perform meaningful exploratory data analysis in R.

Populations and Samples

Statisticians hope to learn about some *characteristic/variable* of a *population*. But we often can't see the whole population; so, we investigate a *sample.*

- **Definition**: A *population* is a collection of units (units can be people, widgets, servings of food, kittens, songs, Tweets, etc.)
- **Definition**: A *sample* is a subset of the population.
- **Definition**: A *characteristic/variable* of interest (VoI) is something to be measured for each unit.



Populations and Samples

Example: An insurance company surveying damage in a particular town after a hurricane. In this case, the Population is..? Reasonable sample? VoI?

Population: Addresses in the town

Sample: Actual addresses observed (say *n* of them)

Variable of Interest: "Damaged" or "Not damaged"

Example: CU might want to study the average GPA of juniors who are engineering majors at CU. In this case, the Population is..? Reasonable sample? VoI?

Population: GPA in junior year of any engineering major ever

Sample: Current juniors (say *n* of them)

Variable of Interest: GPA above or below 3.2?

Populations and Samples

Statisticians learn about a characteristic in a population by studying a sample.

A major component of this course is to figure out how to make the jump from sample to population – **Statistical Inference.**

Before we learn about *inference*, we're first going to learn how to explore data. This is helpful for summarizing, recognizing patterns, etc.

Exploratory Data Analysis (Descriptive Statistics)

There are two main types of explorations: **numerical** and **graphical.**

Populations and Samples

Statisticians learn about a characteristic in a population by studying a sample.

A major component of this course is to figure out how to make the jump from sample to population – **Statistical Inference.**

Before we learn about *inference*, we're first going to learn how to explore data. This is helpful for summarizing, recognizing patterns, etc.

<u>Exploratory Data Analysis (Descriptive Statistics)</u>

There are two main types of explorations: **numerical** and **graphical.**

Numerical Summaries: Sample Statistics

- The calculation and interpretation of certain **summarizing numbers** can help us gain an understanding of the data.

- These sample numerical summaries are called *sample statistics*.

- A statistic is typically any value that is calculated from a sample.

Sample Statistics: Measures of Centrality

Summarizing the "**center**" of the sample data is a popular and important characteristic of a set of numbers. The goal here is to capture something like the "typical" unit with respect to the variable of interest.

<u>3 popular types of center:</u>

1. Mean
2. Median
3. Mode

The Sample Mean

Given a set of numbers $x_1, x_2, \ldots, x_n$, the most familiar measure of the center is the *mean* (arithmetic average).

Sample mean, $\bar{x}$, of observations $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Advantages? Easy to calculate, equally weights all info

Disadvantages? Not robust to outliers.

The Sample Median

Median: Middle value when observations are ordered smallest to largest, denoted by $\tilde{x}$.

To calculate: Order the *n* observations smallest to largest (repeated values included and find the middle one.

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

Population Mean, Median and Mode

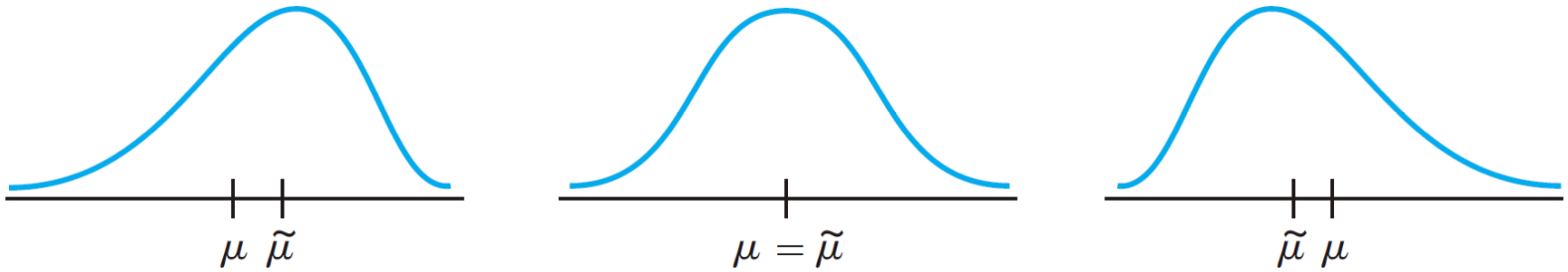
- The greek letter μ denotes the **population mean**, this is the arithmetic average of all the elements of the population which can (in theory) be calculated.

- We denote the **population median** by $\tilde{\mu}$.

- The **mode** of a data set is the **value that occurs most often** (if there is one) and the **population mode** is denoted by $\hat{\mu}$ and the **sample mode** is denoted as $\hat{x}$.

The Mean vs. the Median

The population mean and median will not generally be identical:



(a) Negative skew (b) Symmetric (c) Positive skew

Three different shapes for a population distribution

Which population characteristic is most important?

Example - Median

- Suppose the amount of active ingredient in a brand name drug and a generic drug (in grams) is sampled:

•Brand:	5.6	5.1	6.2	6.0	5.8	6.5	5.8	5.5
•Generic:	5.3	4.1	7.2	6.5	4.8	4.9	5.8	5.0

Example - Median

- Suppose the amount of active ingredient in a brand name drug and a generic drug (in grams) is sampled:

- Brand: 5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5
- Generic: 5.3 4.1 7.2 6.5 4.8 4.9 5.8 5.0

- Brand sample sorted: 5.1 5.5 5.6 5.8 5.8 6.0 6.2 6.5

- Generic sample sorted: 4.1 4.8 4.9 5.0 5.3 5.8 6.5 7.2

Example - Median

- Suppose the amount of active ingredient in a brand name drug and a generic drug (in grams) is sampled:

- Brand: 5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5
- Generic: 5.3 4.1 7.2 6.5 4.8 4.9 5.8 5.0

- Brand sample sorted: 5.1 5.5 5.6 5.8 5.8 6.0 6.2 6.5
- So we see that the median is: $\frac{5.8 + 5.8}{2} = 5.8 \text{ grams} = \tilde{x}$
- Generic sample sorted: 4.1 4.8 4.9 5.0 5.3 5.8 6.5 7.2
- Here we see the median is: $\frac{5.0 + 5.3}{2} = 5.15 \text{ grams} = \tilde{x}$

Quartiles and Percentiles - Quartiles

Recall that the **median** divides data set into **two** parts of equal size. But we could divide data into more than two halves.

• **Definition**: **Quartiles** divide the data sample as nearly as possible into quarters. There are 3 Quartiles: the 1st Quartile (q_1), the 2nd Quartile (q_2) and the 3rd Quartile (q_3).

•*Example*: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

Quartiles and Percentiles - Quartiles

Recall that the **median** divides data set into **two** parts of equal size. But we could divide data into more than two halves.

• **Definition**: **Quartiles** divide the data sample as nearly as possible into quarters. There are 3 Quartiles: the 1st Quartile (q_1), the 2nd Quartile (q_2) and the 3rd Quartile (q_3).

•*Example*: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

• The median is 6, i.e. the 2nd Quartile is $q_2 = 6$

Quartiles and Percentiles - Quartiles

Recall that the **median** divides data set into **two** parts of equal size. But we could divide data into more than two halves.

• **Definition**: **Quartiles** divide the data sample as nearly as possible into quarters. There are 3 Quartiles: the 1st Quartile (q_1), the 2nd Quartile (q_2) and the 3rd Quartile (q_3).

•*Example*: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.7,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$

Quartiles and Percentiles - Quartiles

Recall that the **median** divides data set into **two** parts of equal size. But we could divide data into more than two halves.

• **Definition**: **Quartiles** divide the data sample as nearly as possible into quarters. There are 3 Quartiles: the 1st Quartile (q_1), the 2nd Quartile (q_2) and the 3rd Quartile (q_3).

•*Example*: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.7,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$
- The upper half of the data is {6,7,8,9,10,17}, 3rd Quartile is $q_3 = 8.5$

•***Note***: The 1st Quartile is the median of the lower half of the data, the 2nd Quartile is the median of the data and the 3rd Quartile is the median of the upper half of the data.

Percentiles & Quantiles

The *first quartile* is also know as the 25th percentile, the *second quartile* is also known as the 50th percentile and the *third quartile* is also known as the 75th percentile.

Definition: The ***p*-th percentile** of a data sample, for a number *p* between 0 and 100, divides the sample so that as nearly as possible *p*% of the sample values are less than the *p*-th percentile and (100-*p*)% of the sample values are greater.

For example:

- If a sample value is in the 99th percentile than it is greater than 99% of the data points.
- The 50th percentile of the data is the median.

Percentiles & Quantiles

The *first quartile* is also know as the 25th percentile, the *second quartile* is also known as the 50th percentile and the *third quartile* is also known as the 75th percentile.

Definition: The ***p*-th percentile** of a data sample, for a number *p* between 0 and 100, divides the sample so that as nearly as possible *p*% of the sample values are less than the *p*-th percentile and (100-*p*)% of the sample values are greater.

For example:

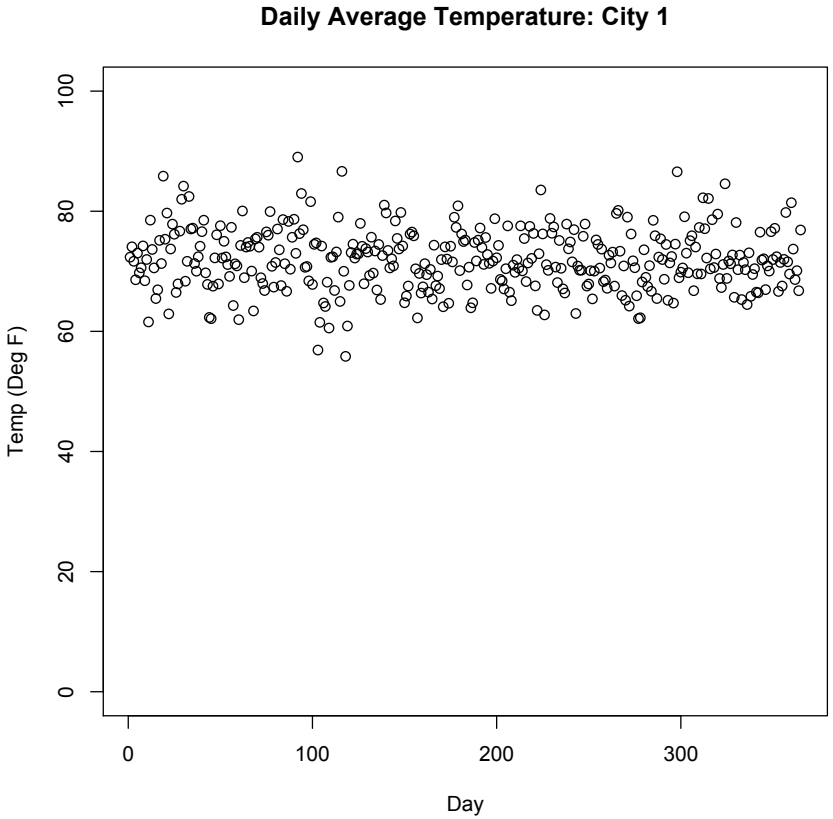
- If a sample value is in the 99th percentile than it is greater than 99% of the data points.
- The 50th percentile of the data is the median.

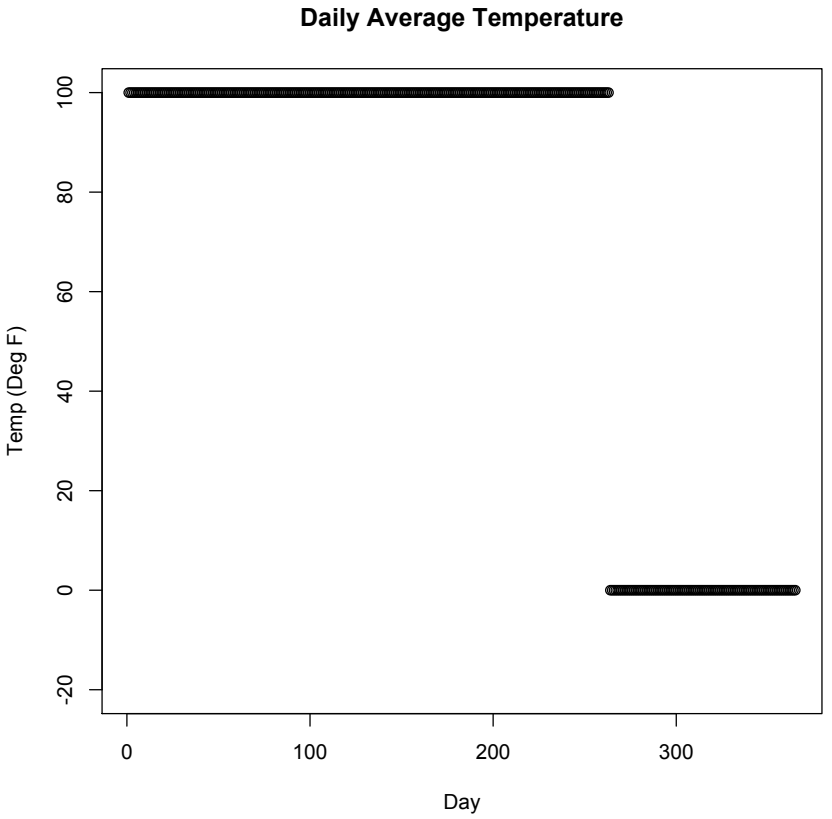
The ***p*-th percentile** is the same as the ***p*/100 quantile**. For example the **25th percentile** is the **0.25 quantile** (since 25%=0.25) which is also know as the the **1st Quartile**.

Variability

So far, we've learned techniques for visualizing our data and measures of center. What about the *spread* of the data?

Example: A tail of two cities.





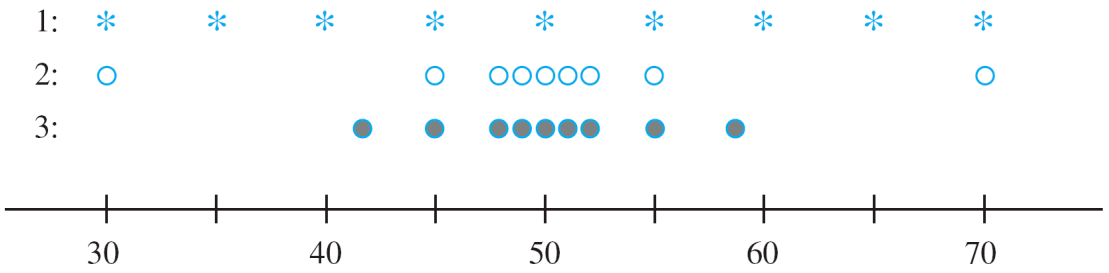
Variability

The simplest measure of variability is the **range**, i.e the difference between the largest and smallest sample value.

Variability

The simplest measure of variability is the **range**, i.e the difference between the largest and smallest sample value.

The value of the range for sample 1 is much larger than it is for sample 3, reflecting **more variability** in the first sample than in the third:

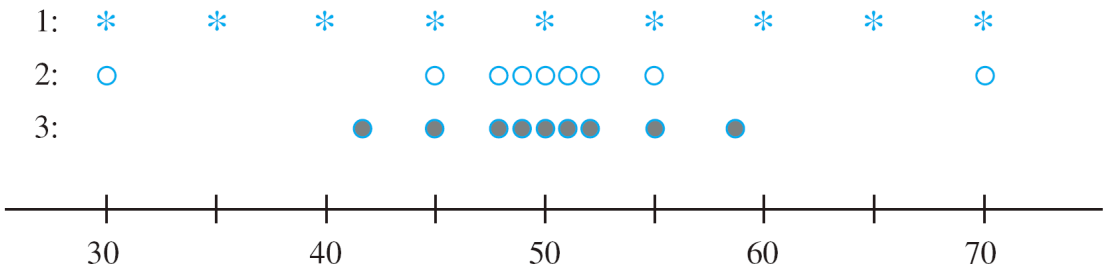


Samples with identical measures of center but different amounts of variability

Variability

The simplest measure of variability is the **range**, i.e the difference between the largest and smallest sample value.

The value of the range for sample 1 is much larger than it is for sample 3, reflecting **more variability** in the first sample than in the third:



Samples with identical measures of center but different amounts of variability

Figure above shows dotplots of three samples with the same mean and median, yet the extent of **spread about the center** is different for all three samples. The range ignores the *n-2* values that are not the extreme values.

Variability

A more robust measure of variation takes into account **deviations from the mean:**

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

Can we combine the deviations into a single quantity by finding the average deviation?

Variability

A more robust measure of variation takes into account **deviations from the mean:**

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

Can we combine the deviations into a single quantity by finding the average deviation? No! If we sum the deviations we have,

$$\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

Variability

A more robust measure of variation takes into account **deviations from the mean:**

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

Can we combine the deviations into a single quantity by finding the average deviation? No! If we sum the deviations we have,

$$\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

Consider instead the **squared deviations**:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2.$$

Variability

A more robust measure of variation takes into account **deviations from the mean:**

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

Can we combine the deviations into a single quantity by finding the average deviation? No! If we sum the deviations we have,

$$\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

Consider instead the **squared deviations**:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2.$$

Rather than use the average squared deviation, $\sum(x_i - \bar{x})^2/n$,

in samples, we divide the sum of squared deviations by $n - 1$ (this yields an "**unbiased**" estimator of the population variance) rather than by n (which gives a "**biased**" estimator of the population variance, more on this later).

Variability

Definition: The **sample variance,** denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

Variability

Definition: The **sample variance,** denoted by s^2, is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation,** denoted by *s,* is the (positive) square root of the variance:

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

Note that s^2 and s are both nonnegative. The unit for s is the same as the unit for each of the x_i.

Variability

www.fueleconomy.gov contains a wealth of information about fuel efficiency (mpg). Consider the following sample of $n = 11$ efficiencies for the 2009 Ford Focus equipped with an automatic transmission:

Car	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	27.3	−5.96	35.522
2	27.9	−5.36	28.730
3	32.9	−0.36	0.130
4	35.2	1.94	3.764
5	44.9	11.64	135.490
6	39.9	6.64	44.090
7	30.0	−3.26	10.628
8	29.7	−3.56	12.674
9	28.5	−4.76	22.658
10	32.0	−1.26	1.588
11	37.6	4.34	18.836
	$\sum x_i = 365.9$	$\sum (x_i - \bar{x}) = .04$	$\sum (x_i - \bar{x})^2 = 314.106$

Variability

The numerator of s^2 is $S_{xx} = 314.106$

Variability

The numerator of s^2 is S_{xx} = 314.106, from which we get

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314.106}{11-1} = 31.41 \Rightarrow s = 5.60$$

The size of a representative deviation from the sample mean 33.26 is roughly 5.6 mpg.

Variability

The numerator of s^2 is S_{xx} = 314.106, from which we get

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314.106}{11-1} = 31.41 \Rightarrow s = 5.60$$

The size of a representative deviation from the sample mean 33.26 is roughly 5.6 mpg.

Population Variance and Population St. Dev: We will use σ^2 to denote the population variance and σ to denote the population standard deviation.

Graphics: Histograms

A **histogram** is a graphical representation of the distribution of numerical data, i.e. the proportion of data points that fall into particular class intervals known as “bins”

Construct a histogram:

1. “Bin” the measured values of the VoI. (The bins are usually consecutive and non-overlapping.)

2. Frequency histogram: count how many values fall into each bin/interval (usually equal size) and draw a bar plot accordingly.

Graphics: Histograms

A **histogram** is a graphical representation of the distribution of numerical data, i.e. the proportion of data points that fall into particular class intervals known as "bins"

Construct a histogram:

1. "Bin" the measured values of the VoI. (The bins are usually consecutive and non-overlapping.)

2. Frequency histogram: count how many values fall into each bin/interval (usually equal size) and draw a bar plot accordingly.

3. Density histogram: select bin width (does not have to be equal) and determine **relative frequency** of each bin then height of each rectangle is height = rel. freq/width and the total area under the density histogram equals 1.

Example

Charity is a big business in the United States. The Web site `charitynavigator.com` gives information on roughly 5500 charitable organizations.

Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities.

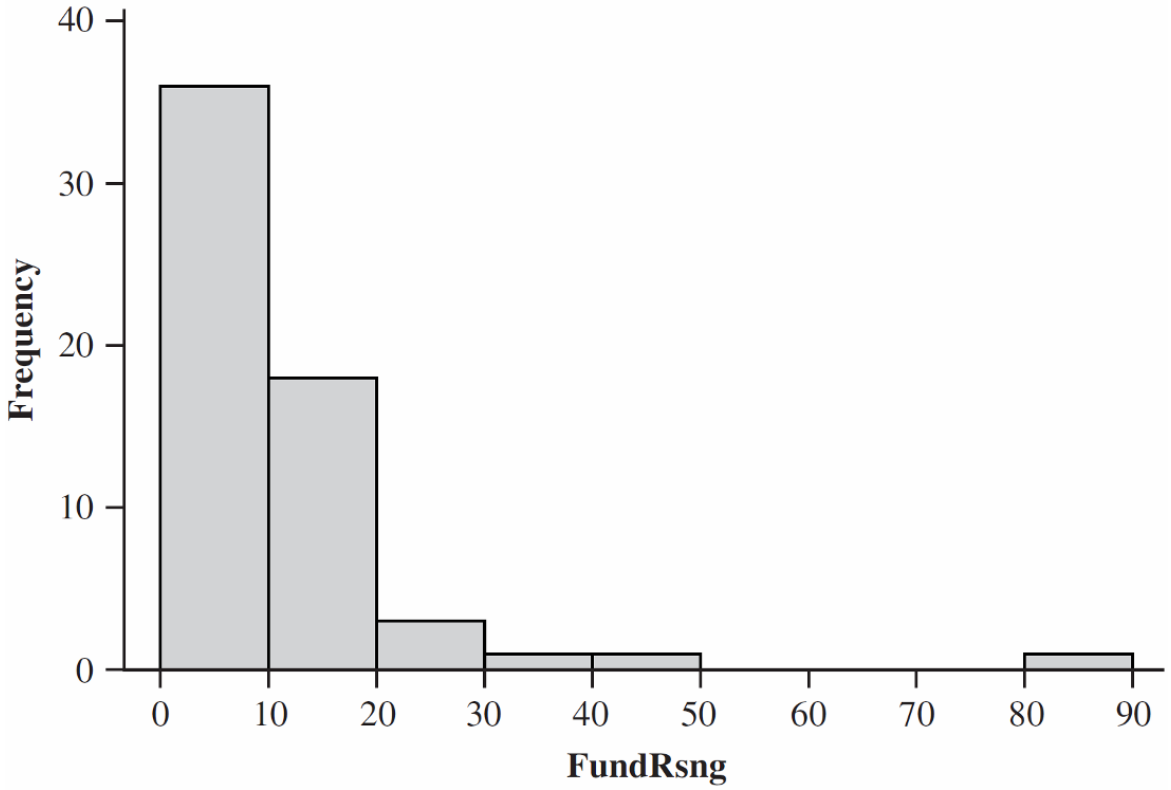
Example

Here are the data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Example

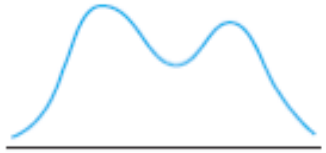
We can see from the histogram that a substantial majority of the charities in the sample spend less than 20% on fundraising:



Graphics: Histograms

Histograms come in a variety of shapes:

- **Unimodal** histogram: single peak
- **Bimodal** histogram: two different peaks
- **Multimodal** histogram: many different peaks



Bimodality: Can occur when the data set consists of observations on two quite different kinds of individuals or objects.

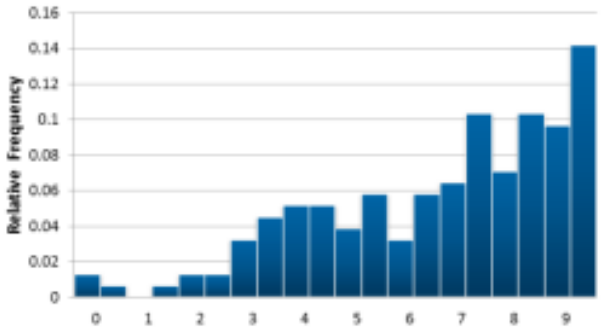
Multimodality occurs when it consists of many different kinds of observations.

A unimodal histogram is *positively skewed* if the right or upper tail is stretched out compared with the left or lower tail and *negatively skewed* if the stretching is to the right.

The Mean vs. the Median

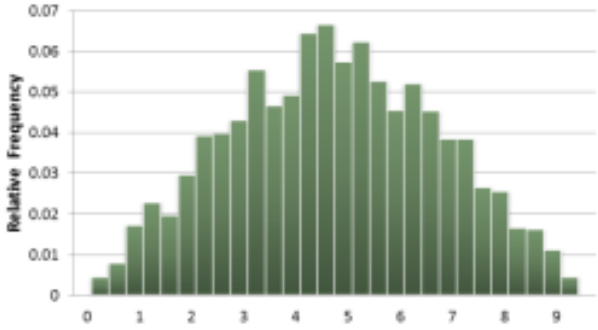
A histogram will be ***skewed to the left*** or ***negatively skewed*** if the median is **greater** than the mean, i.e.

If $\overline{x} < \tilde{x}$ then the histogram is left skewed.



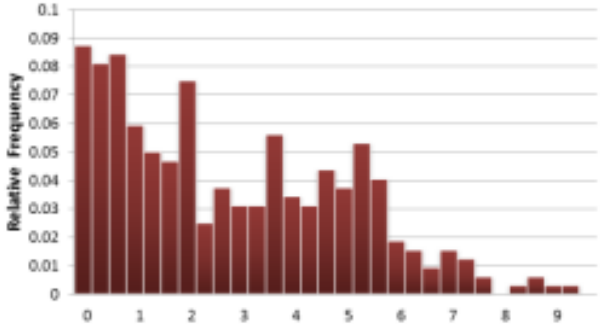
A histogram is ***symmetric*** if its mean is **equal** to it's median, i.e.

If $\overline{x} = \tilde{x}$ then the histogram is symmetric.



A histogram will be ***skewed to the right*** or ***positively skewed*** if the median is **less** than the mean, i.e.

If $\overline{x} > \tilde{x}$ then the histogram is right skewed.



Graphics: Boxplots

Definition: The **interquartile range**, or IQR, of a data set is the difference between the 3rd Quartile and the 1st Quartile, that is, the interquartile range is IQR = $q_3 - q_1$.

Note:

- About half of the data will be between the 1st Quartile and 3rd Quartile.
- The *interquartile range* is the *distance* required to span the middle half of the data.
- An **outlier** is a data point that is unusually large or small.
- Lower Threshold: Data point x_i is an *outlier* if $x_i < q_1 - 1.5 \bullet (\text{IQR})$
- Upper Threshold: Data point x_i is an *outlier* if $x_i > q_3 + 1.5 \bullet (\text{IQR})$

Graphics: Boxplots

Example: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.1,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$
- The upper half of the data is {6,7,8,9,10,17}, 3rd Quartile is $q_3 = 8.5$

Graphics: Boxplots

Example: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.1,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$
- The upper half of the data is {6,7,8,9,10,17}, 3rd Quartile is $q_3 = 8.5$

- The IQR here would be $q_3 - q_1 = 8.5 - 3.5 = 5$ units

Graphics: Boxplots

Example: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.1,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$
- The upper half of the data is {6,7,8,9,10,17}, 3rd Quartile is $q_3 = 8.5$

- The IQR here would be $q_3 - q_1 = 8.5 - 3.5 = 5$ units
- Here the *lower threshold for an outlier* is

$$q_1 - 1.5 \bullet (\text{IQR}) = 3.5 - 1.5(5) = 3.5 - 7.5 = -4$$

- And the *upper threshold for an outlier* is

$$q_3 + 1.5 \bullet (\text{IQR}) = 8.5 + 1.5(5) = 8.5 + 7.5 = 16$$

- So a data point x_i is an *outlier* if $x_i < -4$ or if $x_i > 16$

Graphics: Boxplots

Example: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }

- The median is 6, i.e. the 2nd Quartile is $q_2 = 6$
- The lower half of the data is {-1.1,2.2,3,4,5,6}, so 1st Quartile is $q_1 = 3.5$
- The upper half of the data is {6,7,8,9,10,17}, 3rd Quartile is $q_3 = 8.5$

- The IQR here would be $q_3 - q_1 = 8.5 - 3.5 = 5$ units
- Here the *lower threshold for an outlier* is

  $$q_1 - 1.5 \bullet (\text{IQR}) = 3.5 - 1.5(5) = 3.5 - 7.5 = -4$$

- And the *upper threshold for an outlier* is

  $$q_3 + 1.5 \bullet (\text{IQR}) = 8.5 + 1.5(5) = 8.5 + 7.5 = 16$$

- So a data point x_i is an *outlier* if $x_i < -4$ or if $x_i > 16$
- So we have one outlier, namely 17.

Graphics: Boxplots

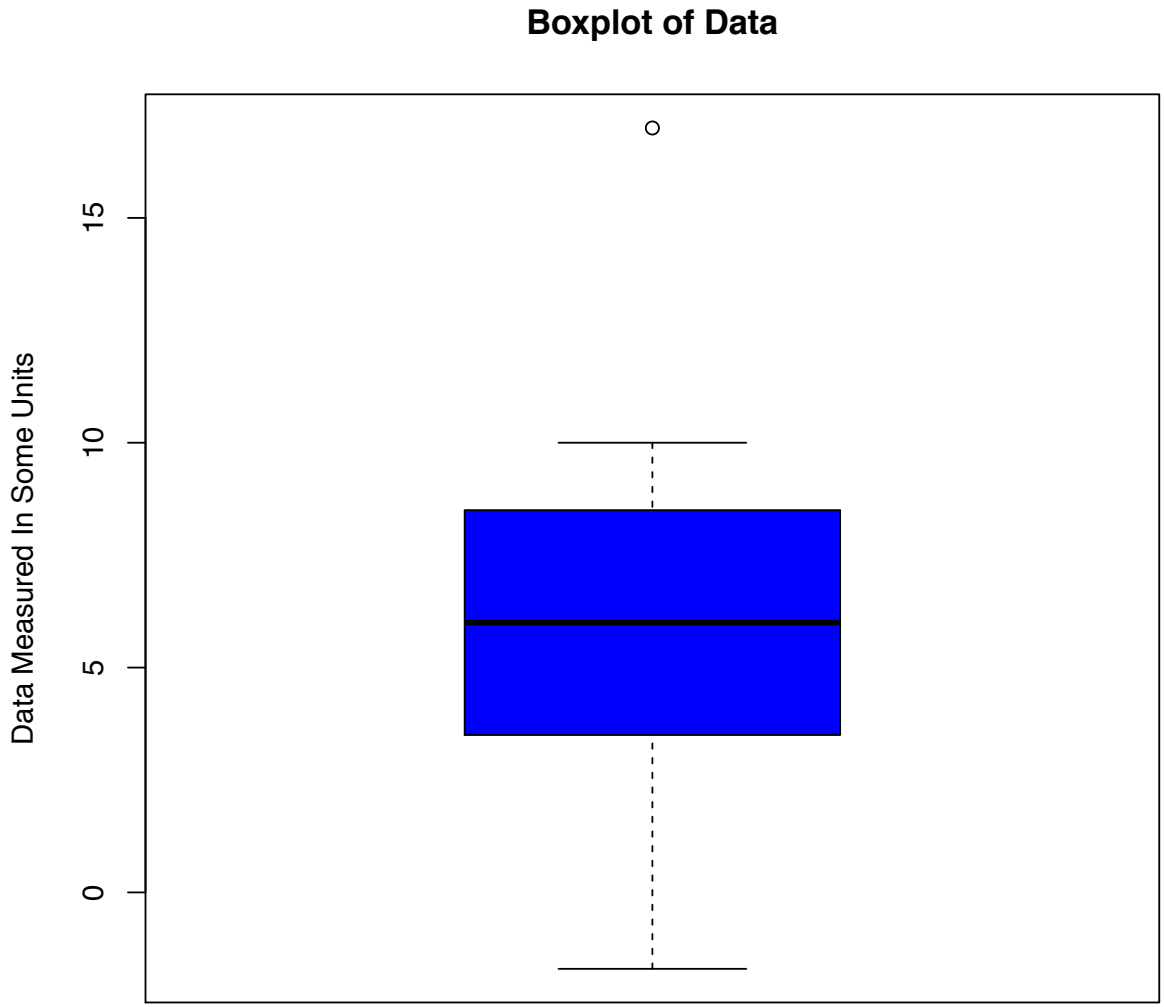
A **boxplot** is a diagram that uses the IQR to describe the spread of the data and identify any outliers.

A boxplot is a plot consisting of:

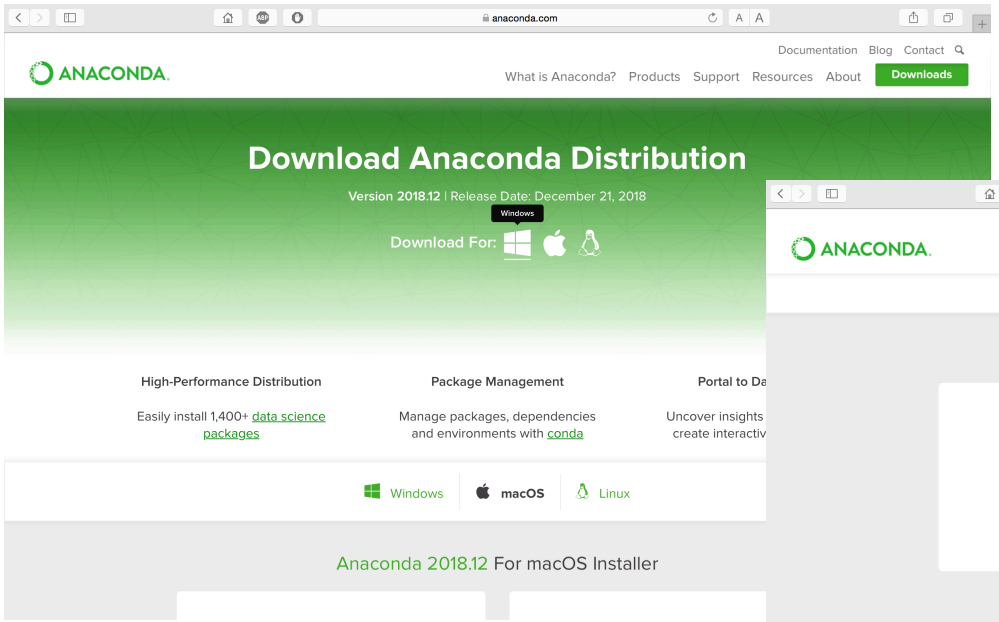
1. A **box** whose *bottom side* is the first quartile and whose *top side* is the third quartile.
2. A **horizontal line** drawn at the median
3. Extending from the top and bottom of the box are vertical lines (called "**whiskers**") that end at the most extreme data points that are not outliers.
4. The **outliers** are plotted individually.

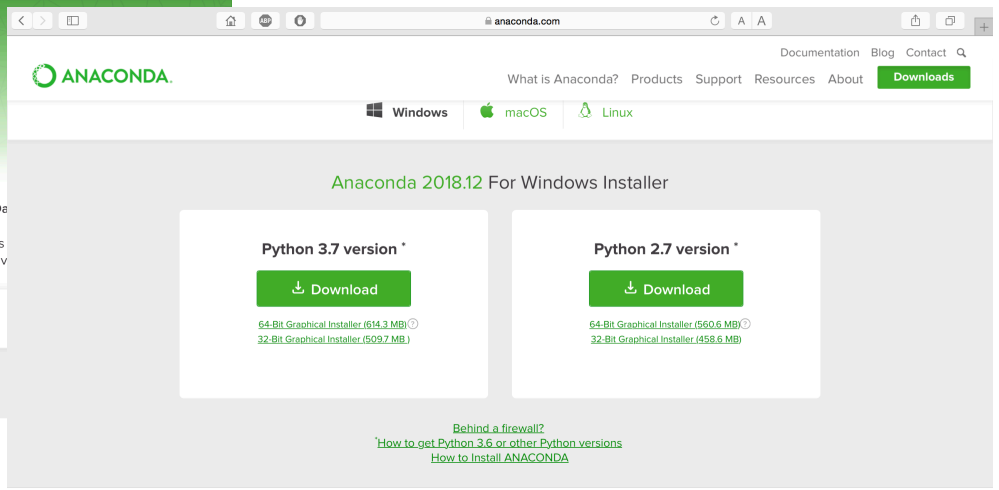
Graphics: Boxplots

Example: Consider the set { -1.7, 2.2, 3, 4, 5, 6, 7, 8, 9, 10, 17 }



Download anaconda and R: https://www.anaconda.com





R Essentials bundle

Rather than install each R language package individually, you can get the R Essentials bundle. It includes over 100 of the most popular scientific packages for the R programming language.

You can install the R Essentials bundle with this command:

```
conda install -c r r-essentials
```

https://docs.anaconda.com/anaconda/packages/r-language-pkg-docs/