Introduction to Data Analysis

Kate Davis

February 12, 2015

Introduction to Data Analysis

Statistical Data Analysis is quantitative evaluation of **Numeric Data** and multiple data points with the same unit can be combined using basic arithmetic operations to form a new data point. Height (in mm), weight (in kg), temperature (degrees F), proportions (percent), and monetary values are examples of **continuous** numeric data points. **Discrete** numeric data are whole number data or count data, such as the number of sunspots per month, the number of apples in a bushel, or dice roll values. House numbers, credit scores, zipcodes and jersey numbers are examples of numbers that are not numeric data points, as none have units nor can these numbers be combined arithmetically to form another numeric data point.

Numeric Data points are numbers that represents value. Generally, each numeric data point has a unit of measure

Continuous Data has an infinite number of possible values within a given range, usually represented by real numbers, percentages or fractions

Discrete Data are data with a finite list of possible values within any given range, and are often integer or count data

Height in Whole Inches

Consider the numeric **Data Set** of **Height in Whole Inches** of our **Population**: students in MA3200 Section 2. Heights would be continuous data, but we have "discretized" this data by rounding to the nearest whole inch. The data, in the original order presented, is:

```
64 70 72 73 69 67 68 66 62 71 66 72 67 74
71 72 67 71 65 65 69 71 69 72 71 68 63 64
```

A **Data Set** is a collection of numeric data points. Each data point within a data set is called an observation, denoted x_i, where i is the number of the observation. For our data set, x_3 is 72. N denotes the number of observations.

A **Population** is any complete group or set of measure with at least one characteristic in common

This set of data has 28 data points. To better evaluate this data, lets sort it. We can begin to see patterns of multiple values, and can quickly see that the lowest or minimum value is 62 inches and the highest or maximum value is 74 inches. The **Range** is 12 inches.

The **Range** is difference between the maximum and minimum values of a data set

```
62 63 64 64 65 65 66 66 67 67 67 68 68 69
69 69 70 71 71 71 71 71 72 72 72 72 73 74
```

This data set has 13 discrete values for height, fewer than the range of 12 inches. There is a gap in observations between 62 inches and 74 inches, but all other height values in the range are represented.

To gain more knowledge about this dataset, we must describe the **distribution** of values across the measurement range, with a goal of using that information for predictions, estimations and other inferences about the population when a complete **census**

The **statistical distribution** can be estimated or inferred from a data

The Oxford English Dictionary defines **Distribution** as the *way in which something is shared out among a group or spread over an area*

A **Census** is a complete enumeration of every unit, everyone or everything in a population.

A **Statistical Distribution** assigns probabilities to the possible values of a data set

set, and is used to estimate the accuracy of these predictions, estimates, inferences.

Frequency Tables

We can create a **Frequency table** and **histogram** of the data set values. The height data is in whole inches, so we will start with using the integer height value as the class in integer order. A cumulative frequency column is added for additional calculation.

a **Frequency Table** is a summary of data point Frequency by class or interval

a **Histogram** is a chart that displays the distribution of a data set

Measures of Center

To understand more about the distribution of the height in inches of our studens, we first examine "centers" of the data: the mode, the median, and the mean.

The **mode** of this dataset is 71 inches with frequency 5 students. The mode is easily found from the frequency table. If there is one clear mode in a distribution, the dataset is said to be *unimodal*. A data set can have more than one mode, or be *multi-modal*.

The **Mode** of a data set is value that has the highest frequency

The **median** of this dataset is 69 inches with frequency of NA students. If the number n of data points is odd, this is a simple observation of $(n+1)/2$. If the number of data points is even, the arithmetic average of the nearest two data point values is the median.

The **Median** of a data set is the midpoint of the distribution, or the middle value of the data when sorted in ascending order. The median is the 50th percentile

The **mean**. The mean is the center that we will use to further examine the "spread" of the values.

The **Mean** of a data set refers to the arithmetic mean of the values, denoted $\bar{x}$

The mean we use is the arithmetic average, which is calculated by first adding the values of all the observations, then dividing by the number of observations.

$$\sum_{i=0}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_{27} + x_{28} + x_{29} + x_{30}$$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$= 64.00 + 70.00 + 72.00 + 73.00 + 69.00 + 67.00 + 68.00 + 66.00 + 62.00 + 71.00 + 66.00 + 72.00 + 67.00 + 74.00 + 71.00 + 72.00 + 67.00 + 71.00 + 65.00 + 65.00 + 69.00 + 71.00 + 69.00 + 72.00 + 71.00 + 68.00 + 63.00 + 64.00 = 1919.00$

In our data set, the mean height is 1919 divided by 28, or 68.53571 inches, which we round to 68.5

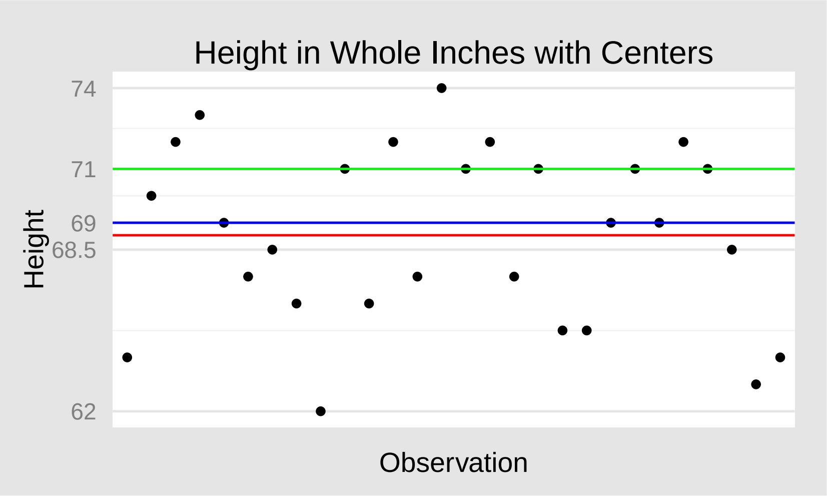


Figure 1: Heights in Observation order with Mode (Green), Median (Blue) and Mean (Red) lines

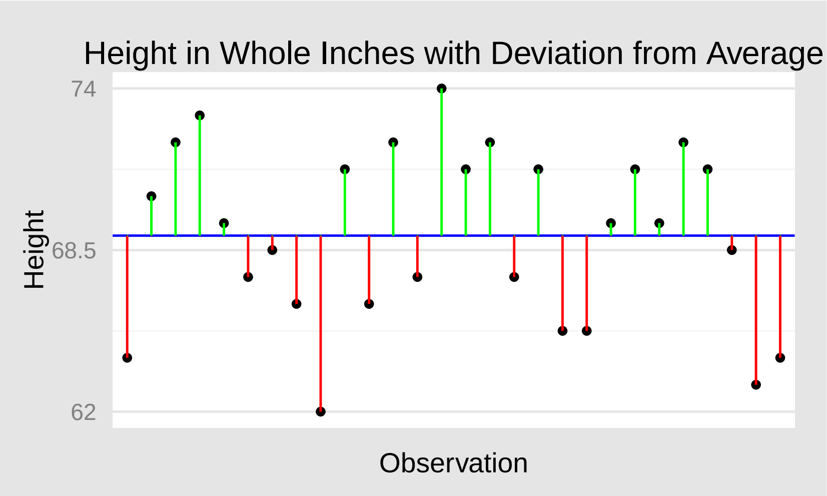
Measures of Spread

The "spread" of the distribution of a dataset can be quantified by range, first and third quartiles, variance and standard deviation.

We would like to measure the **Deviation** from the mean. The deviations from the mean are both positive and negative.

The **Deviation** is the amount by which a single measurement differs from a fixed value, such as the mean.

$$Dev_{\bar{x}} = (x_i - \bar{x})$$



The deviations are both positive and negative, and the sum of the deviations is zero, so this statistic alone is not suitable for further analysis. If we square the deviations, the sum is no longer zero; in fact, the sum of the squared deviations is the **Variance**. The variance of our dataset is 10.39158 square inches. To get back to our original unit of inches, we take the square root of the variance, 3.2, or **Standard Deviation**

The **Variance** is a measure of variability or spread $Var(X)$, often denoted by σ^2

The **Standard Deviation** of X, $StdDev(X)$, often denoted σ, is the standard measure of spread used in statistical analysis.

$$Dev_{\bar{x}}^2 = (x_i - \bar{x})^2$$

$$Var(X) \quad = \quad \frac{\sum_{i=1}^{N} Dev_{\bar{x}}^2}{N} \quad = \quad \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

$$StdDev(X) = \sqrt{Var(X)}$$

Table 1: Deviations

Observation	Height	$Dev_{\bar{x}}$	${Dev_{\bar{x}}}^2$
x_1	64	-4.5	20.6
x_2	70	1.5	2.1
x_3	72	3.5	12.0
x_4	73	4.5	19.9
x_5	69	0.5	0.2
x_6	67	-1.5	2.4
x_7	68	-0.5	0.3
x_8	66	-2.5	6.4
x_9	62	-6.5	42.7
x_{10}	71	2.5	6.1
x_{11}	66	-2.5	6.4
x_{12}	72	3.5	12.0
x_{13}	67	-1.5	2.4
x_{14}	74	5.5	29.9
x_{15}	71	2.5	6.1
x_{16}	72	3.5	12.0
x_{17}	67	-1.5	2.4
x_{18}	71	2.5	6.1
x_{19}	65	-3.5	12.5
x_{20}	65	-3.5	12.5
x_{21}	69	0.5	0.2
x_{22}	71	2.5	6.1
x_{23}	69	0.5	0.2
x_{24}	72	3.5	12.0
x_{25}	71	2.5	6.1
x_{26}	68	-0.5	0.3
x_{27}	63	-5.5	30.6
x_{28}	64	-4.5	20.6
Total	1919.0	0.0	291.0
Total/N	68.5	0.0	10.4
	Mean	Zero	Variance

The first and third quantiles can be found by examining either the sorted values or the frequency table, and taking the value of the observation at the first quarter and last quarter. For n observations, the first quartile is the value $(n+1)/4$th entry and the third quartile is the value at the $(n+1)*3/4$th entry, and similar to median, the second quartile, if the calculated entry value is not a whole number, the arithmetic mean of the nearest two observation values determines the mean.

In our height dataset of 28, the first quartile is the 7th observation, 66, and the third quartile is the 21th observation, 71.

```
62 63 64 64 65 65 66
66 67 67 67 68 68 69
69 69 70 71 71 71 71
71 72 72 72 72 73 74
```

Distribution Shapes: Histograms, Frequency Polygons, and Ogives

Once we have calculated the frequencies and centers of our datasets we can start to explore the shape and spread of the distribution of values with charts. All charts can be drawn from the frequency table data.

The **histogram** is a view of the overall pattern of the distribution. Histogram bars are evenly sized and each bar represents the same class levels of values, and is centered on the mean of the class. The height of the bar represents the number of observations in that class.

A **Histogram** is a visualization of a frequency table.

The mode can easily be seen on a histogram, and the median is the vertical line at which there is equal area to the left and to the right in the chart.

A histogram's shape can be symmetric, skewed right with more of the observations on the right or higher values, or skewed to the left with more of the observations on the left or lower values.

A frequency polygon simply displays the frequency for a class, and an ogive, or cumulative frequency polygon displays the cumulative frequency for a class.

The Empirical Cumulative Distribution shows the possible values of the variable, ordered with frequency, with the cumulative frequency. The proportion of cumulative frequency, often expressed as a percentage, is the number of observations that are less than or equal to the value.

The five number summary of a set of data are the 0th, 25th, 50th, 75th and 100the percentile.

The standard deviation, mean and quartiles are used to create a **boxplot**.

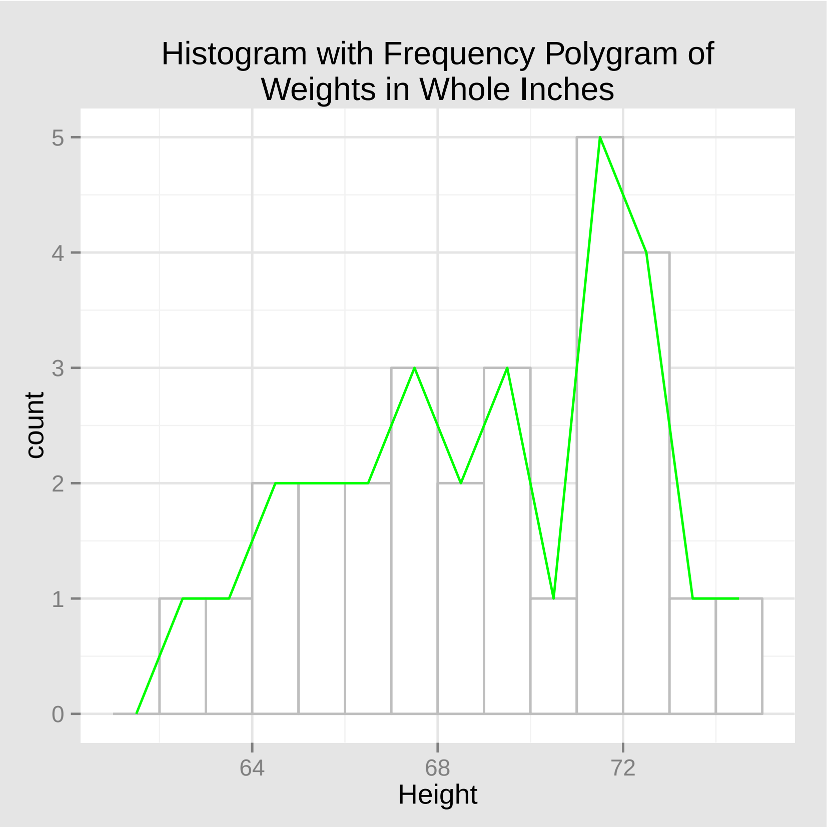


Figure 2: Histogram with Frequency Polygon. The Height data set is unimodal, skewed right, with out outlier on the left.

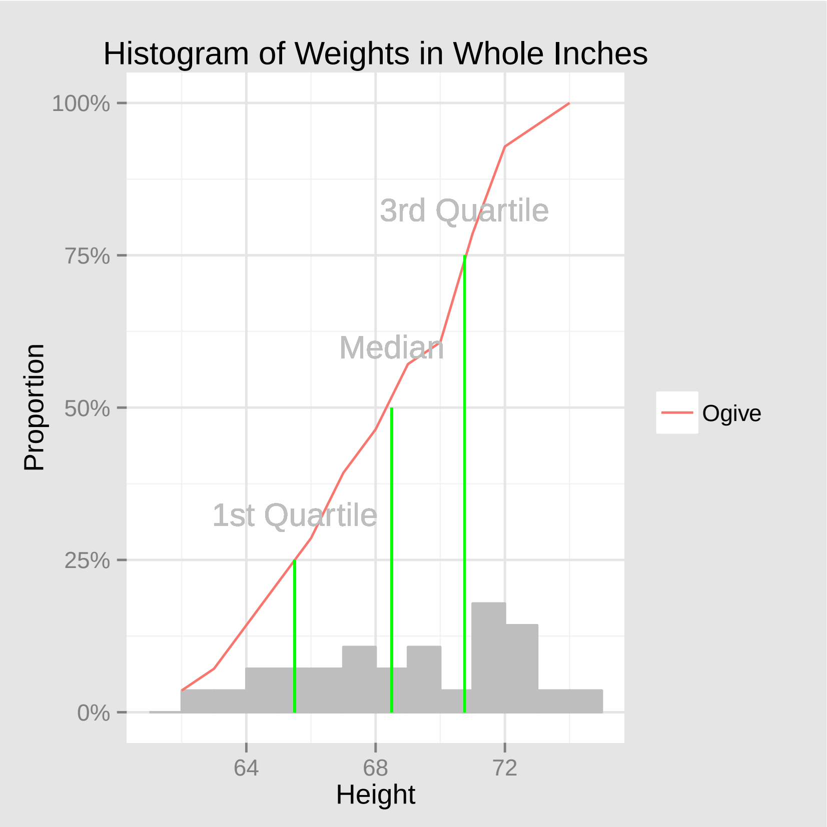


Figure 3: Histogram with Ogive (Cumulative Frequency Polygon).

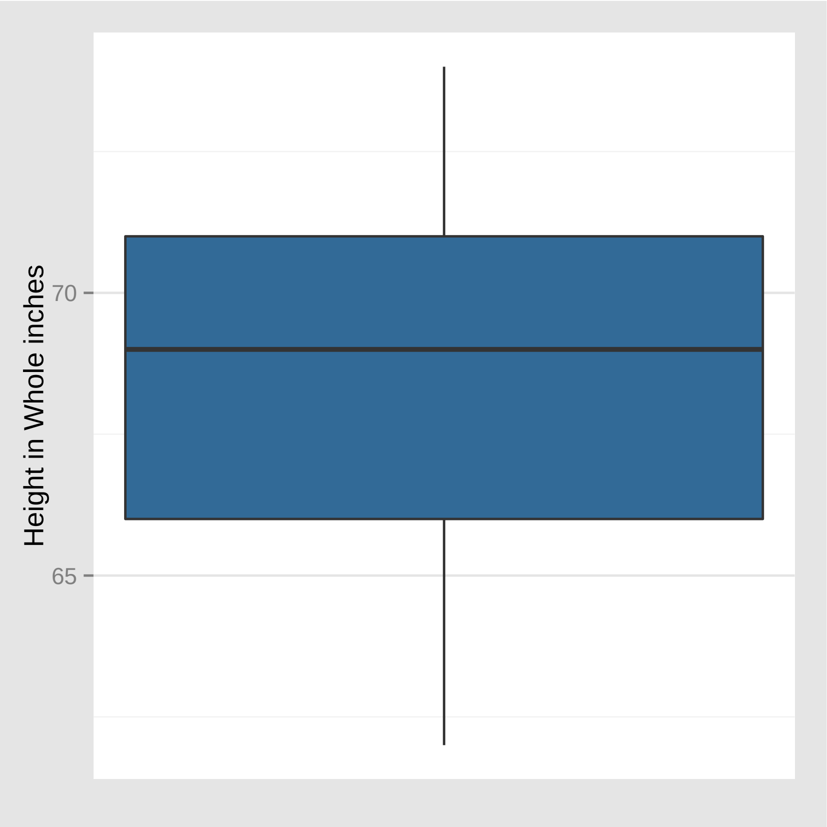


Figure 4: Boxplot