To Do List

Read Chapter 1
Do Problems 1-20

Assignment 1 is due Friday
September 23
Tutorial Test 1 is on Wednesday
September 28

Last Class: Numerical Summaries

Types of numerical measures:

1) **Measures of location (sample mean, median, and mode)**

2) **Measures of variability or dispersion (sample variance, sample standard deviation, range, and interquartile range (IQR))**

Today's Class: Ways of Summarizing Data Continued

1) Numerical Summaries (finish)

Measures of shape (sample skewness and sample kurtosis)

2) Graphical Summaries

Note: Numerical and graphical summaries will be useful in identifying a suitable probability model for the data.

Measures of Shape

Numerical measures which describe the shape of the data:

a) sample skewness

b) sample kurtosis

Sample Skewness and Kurtosis

In this course we most often look at the values of sample skewness and sample kurtosis for a given data set to see if the values are close to the values for Gaussian (Normal) data.

We do this as part of an investigation of whether it is reasonable to assume a Gaussian model for a given data set.

Sample Skewness

$$\text{sample skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}}$$

What are the units of sample skewness?

Sample Skewness

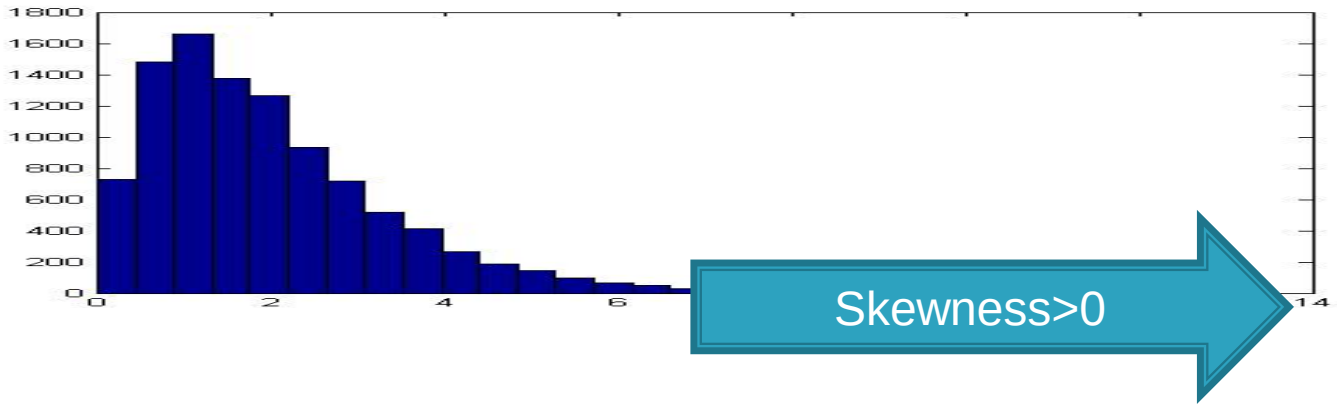
Sample skewness measures the asymmetry of the data.

Data that look symmetric (Gaussian or Uniform) have a sample skewness close to 0.

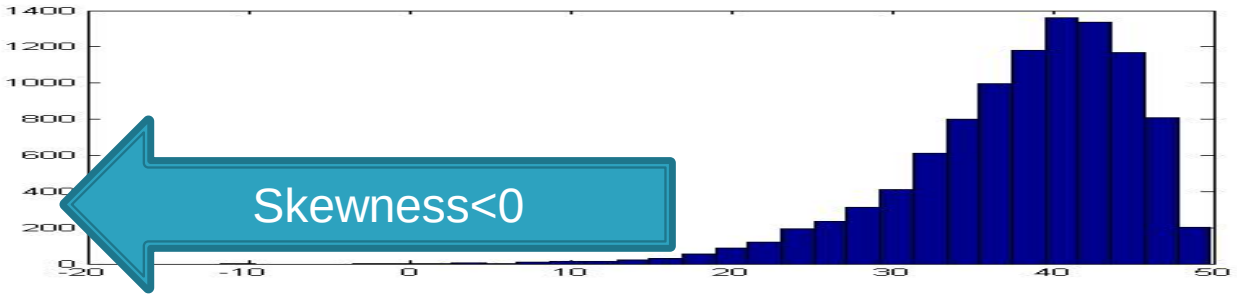
Data that have a long right tail have a sample skewness that is positive.

Data that have a long left tail have a sample skewness that is negative.

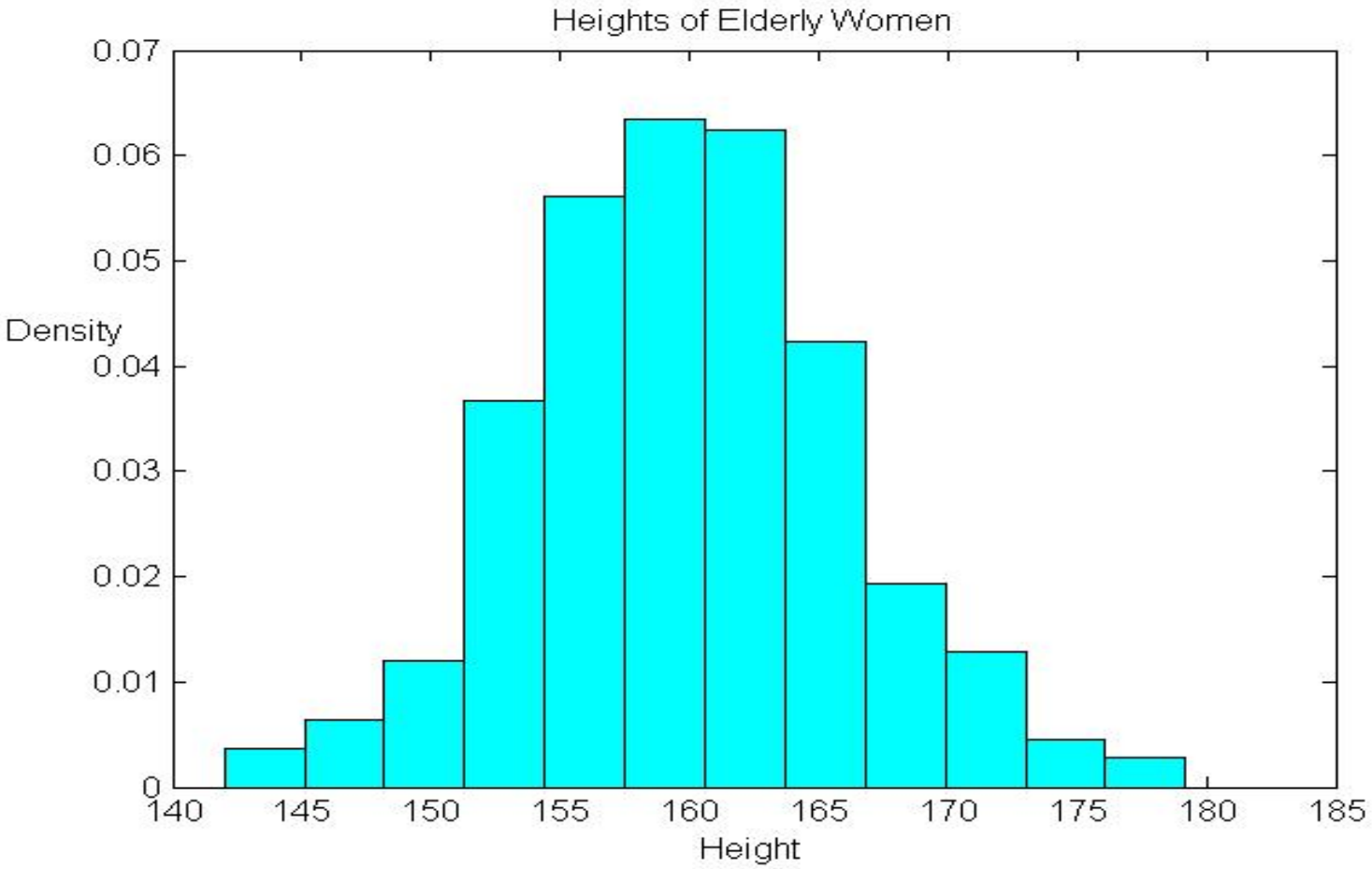
If the right tail is longer than the left tail then we say the data are “skewed to the right” or “positively skewed”.



If the left tail is longer than the right tail then we say the data are “skewed to the left” or “negatively skewed”.

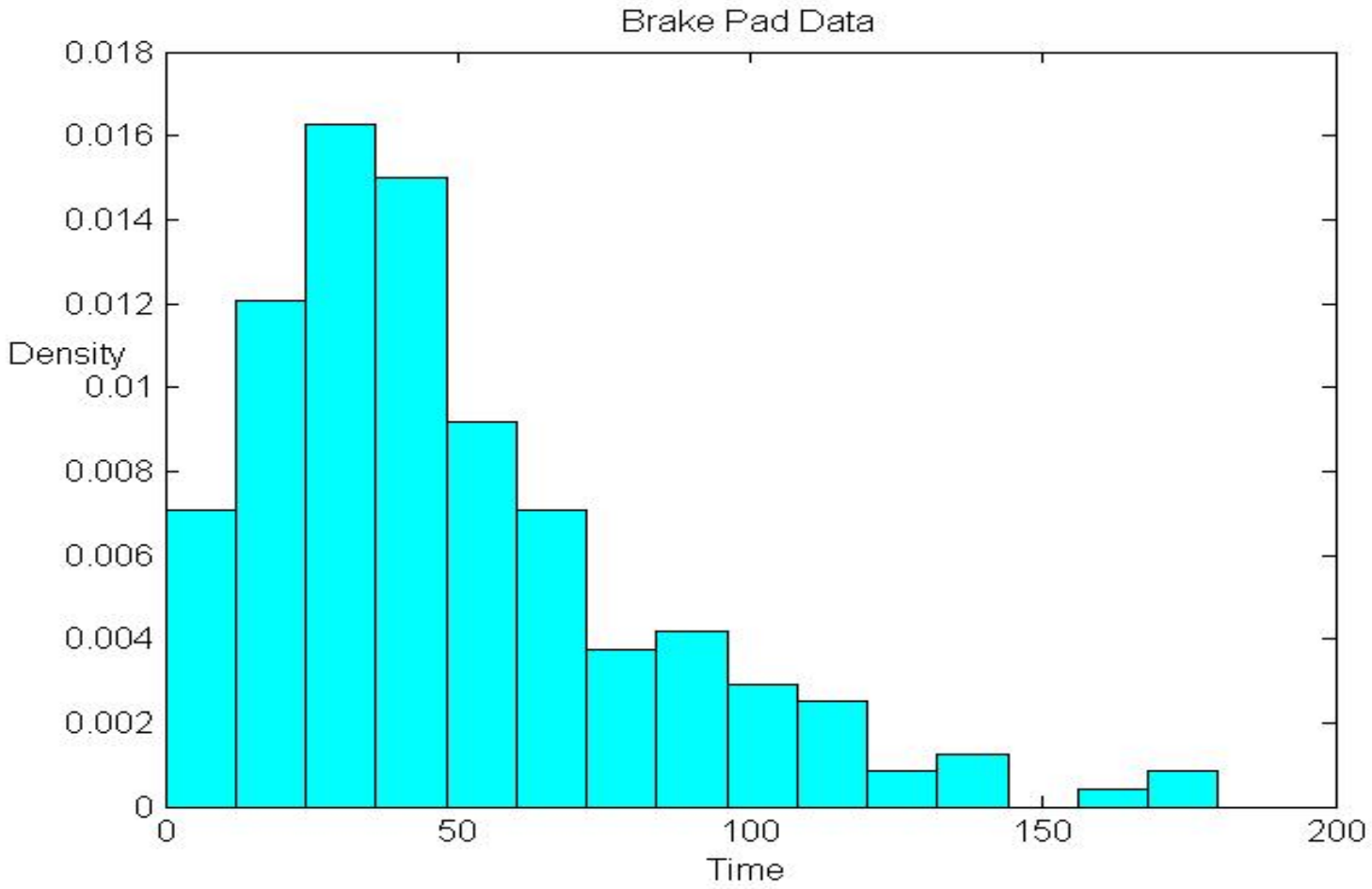


Skewness for Heights of Elderly Women Data Set



Skewness = 0.13

Lifetimes of Brake Pads in Thousands of Kilometers (Ex. 1.3.3)



Skewness = 1.28

Guidelines for Skewness

If data are generated from a Gaussian model then the skewness should be close to 0.

Values between -1 and +1 are considered to be “close to 0” for Gaussian data.

Sample Kurtosis

$$\text{sample kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right]^2}$$

What are the units of sample kurtosis?

Sample Kurtosis

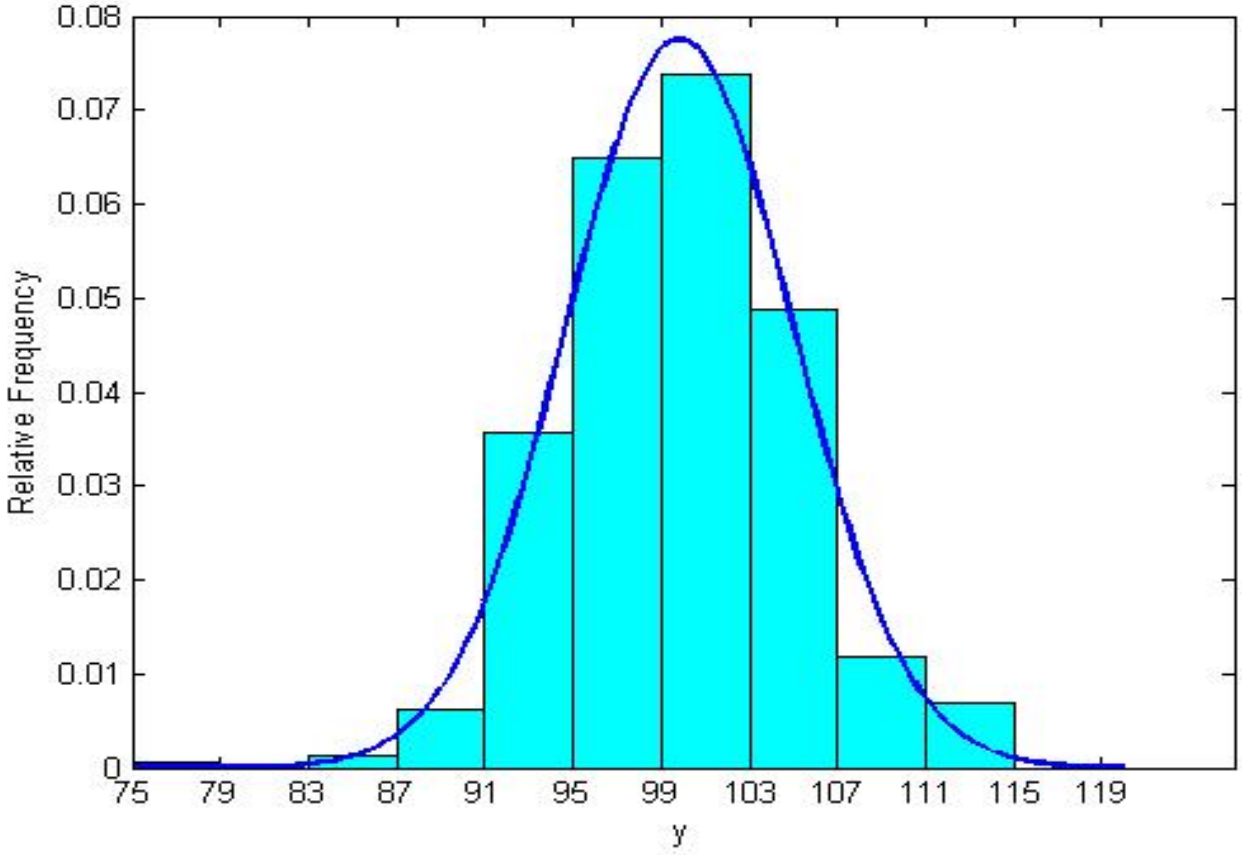
Sample kurtosis measures whether the data are concentrated in a central peak or in the tails.

Data that look Gaussian or bell-shaped have a sample kurtosis close to 3.

Data that are very peaked have a sample kurtosis larger than 3.

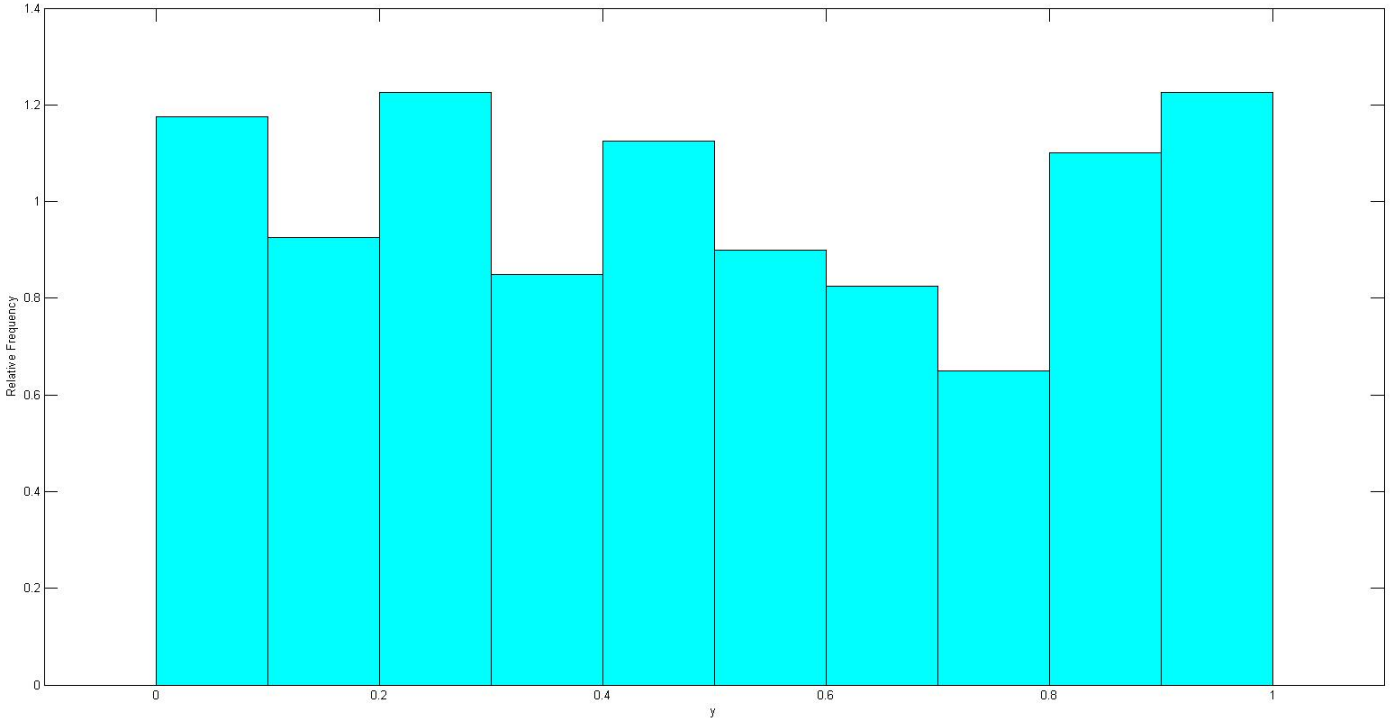
Data that look Uniform have a sample kurtosis close to 1.2.

Kurtosis: Example 1



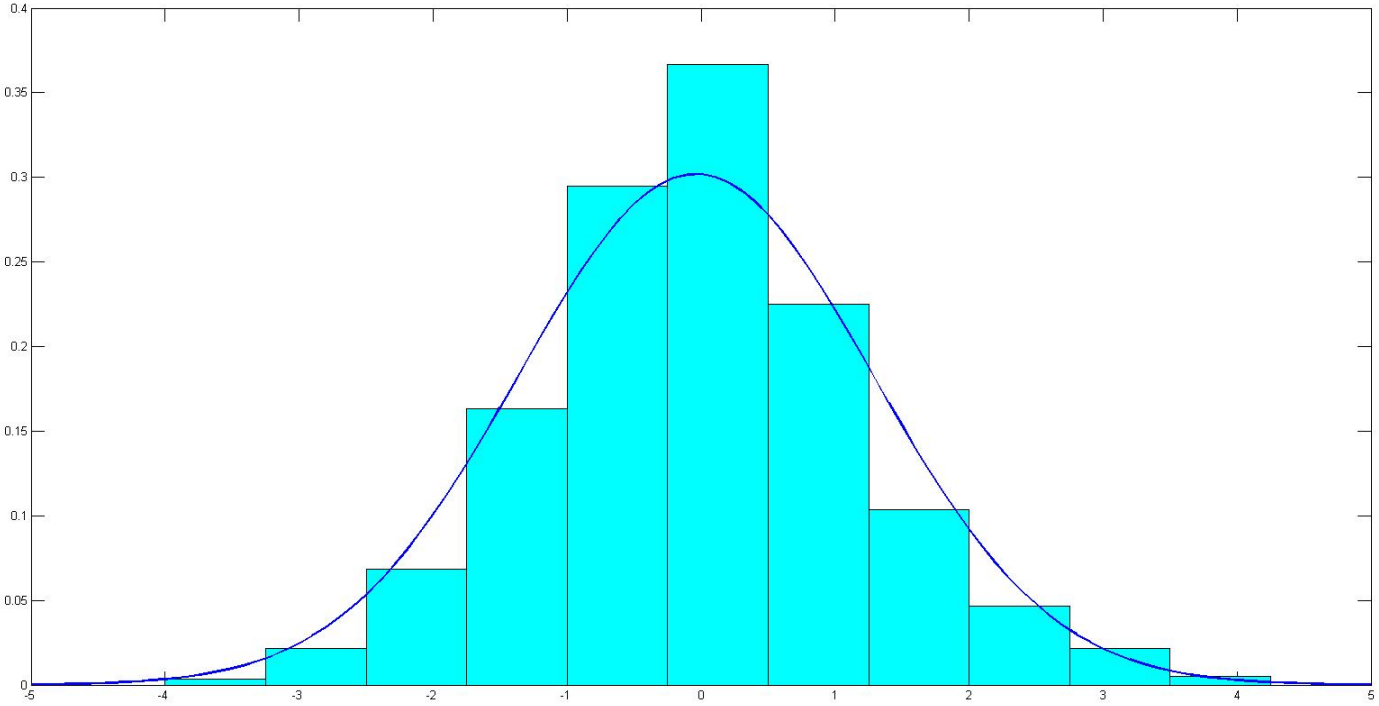
sample kurtosis = 3.61

Kurtosis: Example 2



sample kurtosis = 1.77

Kurtosis: Example 3



sample kurtosis = 6.95

Guidelines for Kurtosis

If data are generated from a Gaussian model then the kurtosis should be close to 3.

Values between 2 and 4 are considered to be “close to 3” for Gaussian data.

Summarizing Gaussian Data

Since the distribution of Gaussian data is reasonably symmetric about the sample mean then

$$\left[\bar{y} - 2s,\ \bar{y} + 2s\right]$$

is a useful summary of the data.

Recall that for Gaussian data approximately 95% of the data should lie in this interval.

How do we summarize non-Gaussian data?

Five-number Summary

Any one of the numerical summaries (mean, median, mode, sample variance, sample standard deviation, range, interquartile range, sample skewness, sample kurtosis) taken on its own does not really provide a good summary of the entire data set.

A useful summary which is more informative is the five-number summary:

$$y_{(1)},\ q(0.25),\ q(0.5),\ q(0.75),\ y_{(n)}$$

where $y_{(1)}=\min(y_1,\ldots,y_n)$, q(0.5) = median, & $y_{(n)}=\max(y_1,\ldots,y_n)$

Five Number Summary for Old Faithful Data

For example for the Old Faithful data the five number summary is:

$y_{(1)}$, q(0.25), q(0.5), q(0.75), $y_{(299)}$

= 43 59 76 83 108 (minutes)

These five numbers summarize the original 299 observations with respect to numerical information about the location and spread of the data.

Graphical Summaries

1) Histograms
2) Empirical Cumulative Distribution Function (e.c.d.f.)
3) Boxplots
4) Run Charts
5) Scatterplots
6) Bar Charts, Pie Charts

Graphical Summaries

The basics (common sense):

1) All graphs should be displayed at an appropriate size.

2) Graphics should have clear titles which are fairly self explanatory.

3) Axes should be labelled and units given where appropriate.

4) The choice of scales should be made with care.

5) Graphics should not be used without thought; there may well be better ways of displaying the information.

Histograms

The idea is that we want to create a graphical summary of the data that we could use to compare with a p.d.f. for a continuous random variable or a p.f. for a discrete random variable.

This is helpful in determining what probability model could be used to model the data (more on this later).

Histograms

Let the observed data be represented as $\{y_1, y_2, \ldots, y_n\}$.

Partition the range of *y* into k non-overlapping intervals $I_j = [a_j, a_{j-1})$ for j=1,2,…,k.

Let f_j = number of values from $\{y_1, y_2, \ldots, y_n\}$ that are in I_j. The f_j are called the observed frequencies.

Draw a rectangle above each of the intervals with height proportional to the observed frequency or relative frequency.

Histograms Cont'd

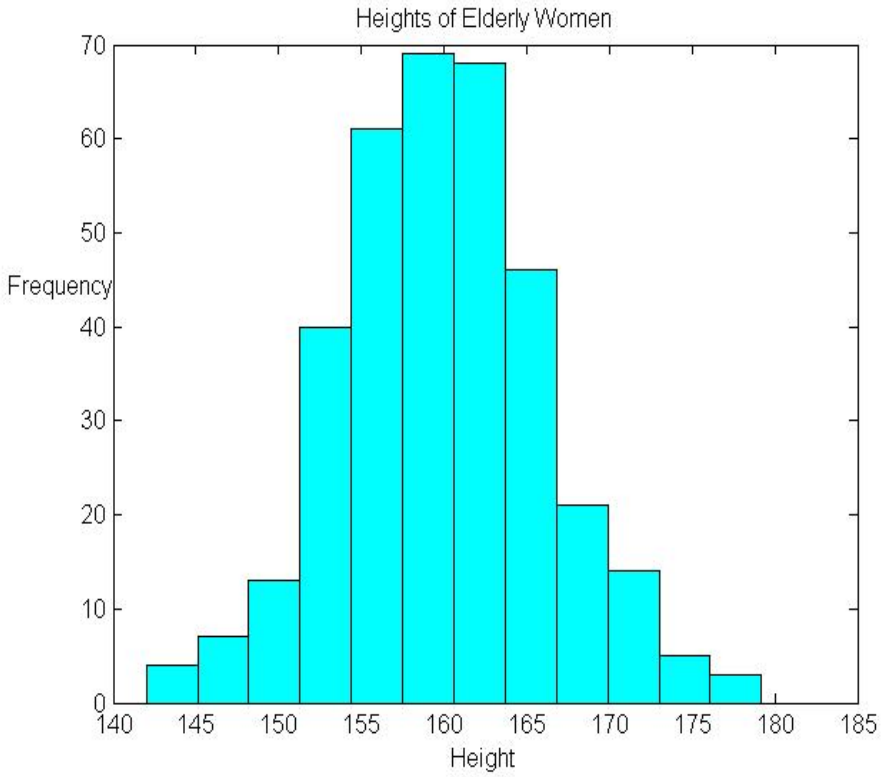
In a "standard" histogram the intervals are of equal width and the heights are equal to the frequencies or the relative frequencies.

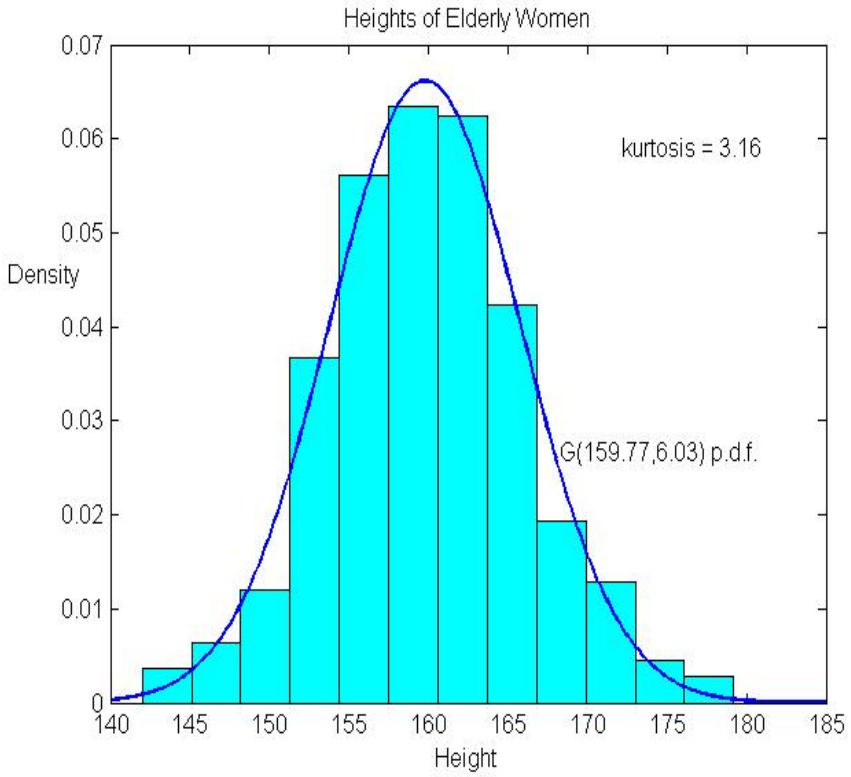
In a "relative frequency" histogram the intervals may not be of equal width. The height of the rectangle is chosen so that the area of the rectangle equals f_j *ln*, that is,

$$\text{height} = \frac{f_j / n}{\left(a_j - a_{j-1}\right)}$$

In this case the sum of the areas of the rectangles equals one. In comparing data to a p.d.f. only a relative frequency histogram can be used.

Histograms for the Heights of Elderly Women Data Set





Total area of rectangles = 1
Area under Gaussian p.d.f. = 1.

Empirical c.d.f.

The idea is that we want to create a graphical summary of the data that we could use to compare with a c.d.f. of a random variable.

This is also helpful in determining what probability model could be used to model the data (more on this later).

Example

- **Ten observations randomly generated from a Uniform(0,1) distribution:**

0.76 0.43 0.52 0.45 0.01 0.85 0.63 0.39 0.72 0.88

- **Order data from smallest to largest:**

0.01 0.39 0.43 0.45 0.52 0.63 0.72 0.76 0.85 0.88

- **Based on the data, what is the approximate probability of observing a value below a given value y?**

- **For example, if *y* = 0.5, then there are four values, 0.01 0.39 0.43 0.45, less than 0.5 so we would estimate the probability to be 4/10.**

Empirical c.d.f.

In general we estimate the probability of values below a given *y* using

$$\frac{\text{number of observations} \leq y}{n}$$

Definition of the empirical c.d.f.

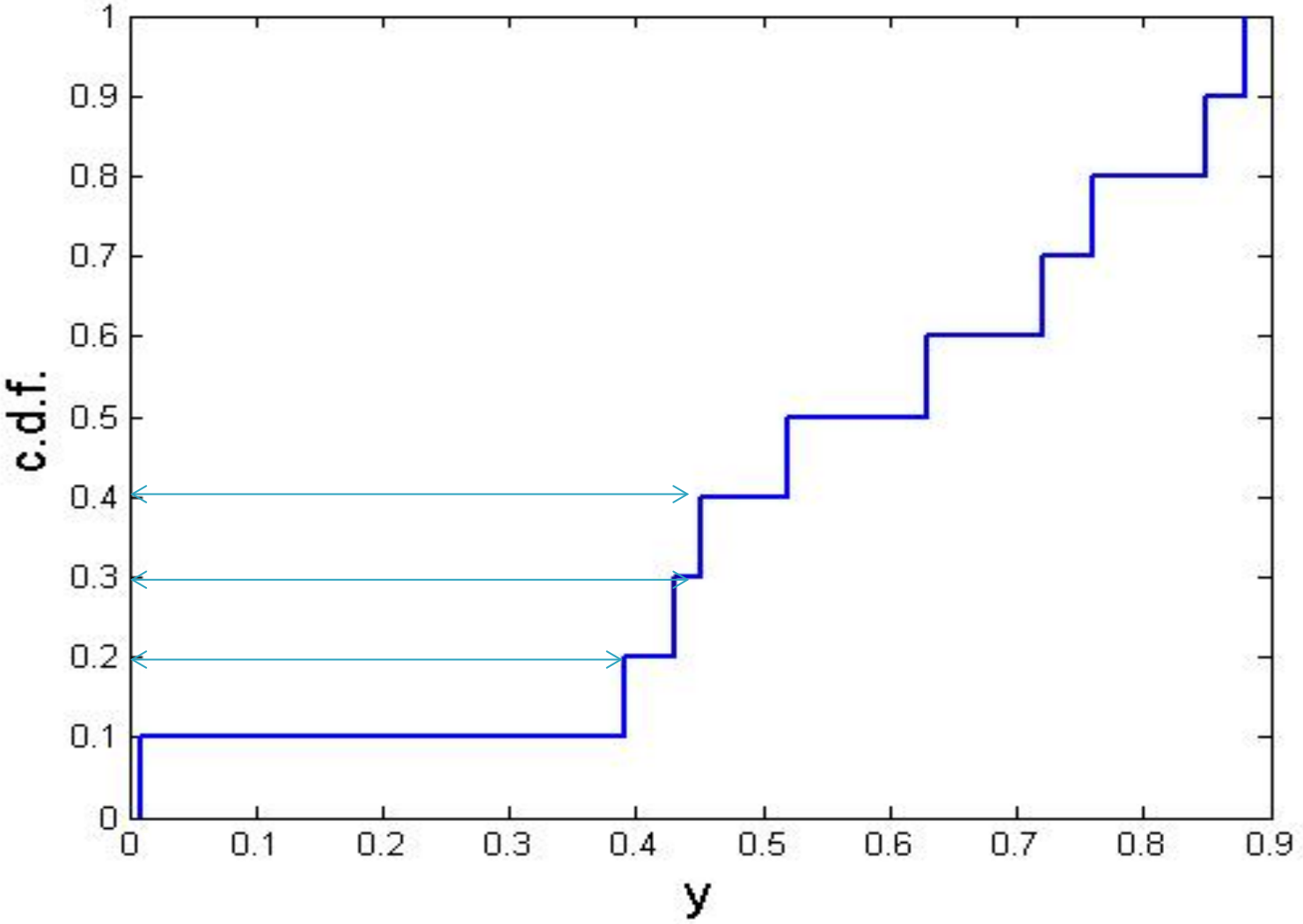
$$\hat{F}(y) = \frac{\text{number of values in } \{y_1, y_2, \ldots, y_n\} \text{ which are } \leq y}{n}$$

defined for all real values *y*.

Example:

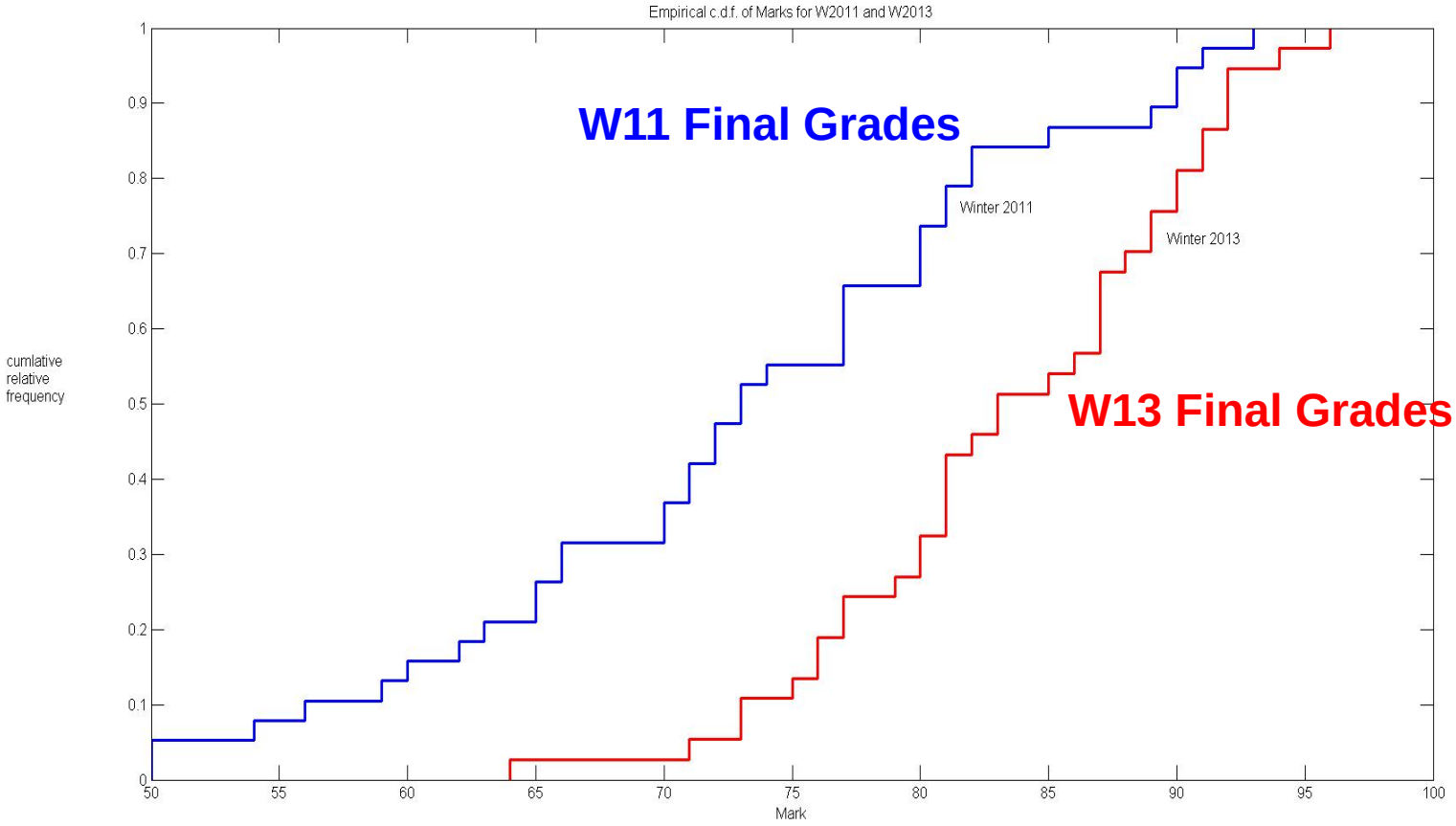
0.01 0.39 0.43 0.45 0.52 0.63 0.72 0.76 0.85 0.88

Example: 0.01 0.39 0.43 0.45 0.52
0.63 0.72 0.76 0.85 0.88



Empirical c.d.f.'s can also be used for comparing the variate values of two or more groups.

Empirical c.d.f. for Final Grades W11 and W13



STAT 231 Final Grade