To Do List

Do Problems 1-20 in Chapter 1

Assignment 1 is due Friday September 23

Tutorial Test 1 is on Wednesday September 28 – see detailed instructions posted on Learn

Last Class: Numerical Measures for Summarizing Univariate Data

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))**

3) **Measures of shape (sample skewness and sample kurtosis)**

Today's Class: Finish 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

Boxplots

Another graphical way to summarize the data is to use a boxplot.

A boxplot also gives a graphical summary about the shape of the distribution.

(Compare the five number summary.)

How to construct a boxplot:

1) Draw a box with ends at q(0.25) and q(0.75) so the box height = IQR.

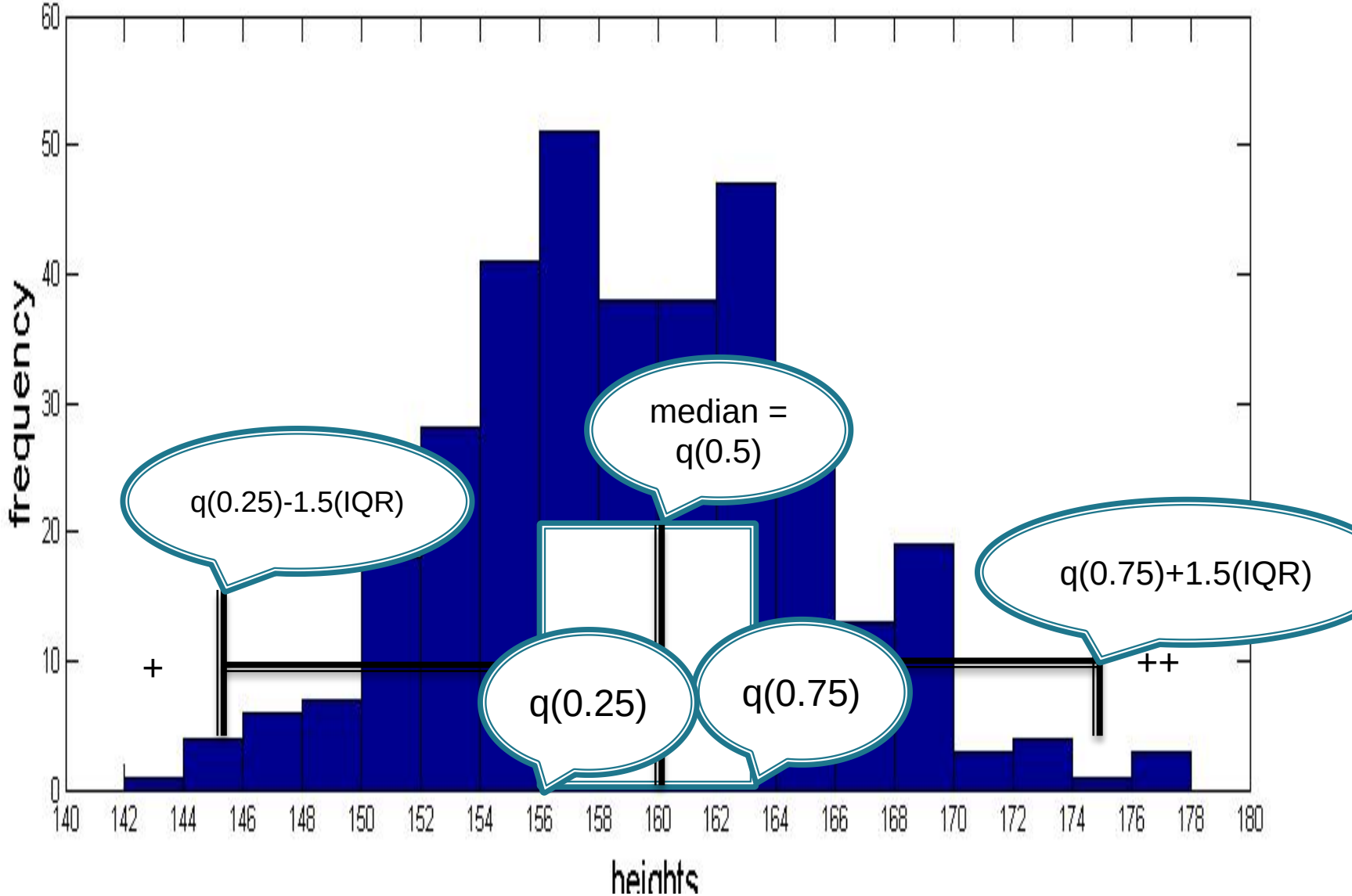
2) Draw a line in the box at q(0.5) = median.

3) Draw two lines or “whiskers” extending up and down from the box.

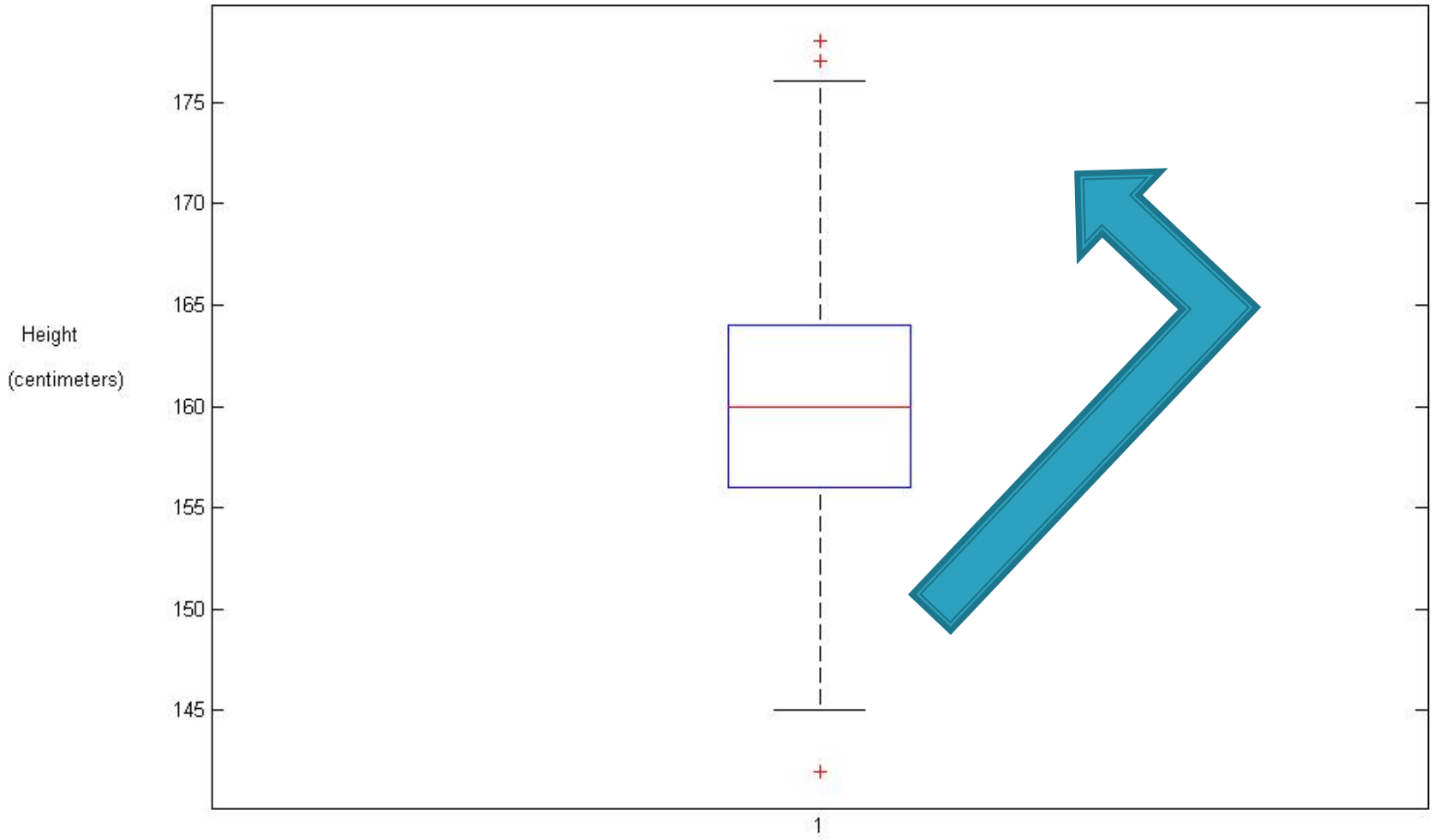
4) Draw short horizontal lines at the smallest observation that is larger than q(0.25) - 1.5*IQR and at the largest observation that is smaller than q(0.75) + 1.5*IQR.

5) Plot any additional points beyond these lines individually using a special symbol like “+” or “*”. These points are called “outliers”.

Boxplot for Heights of Elderly Women

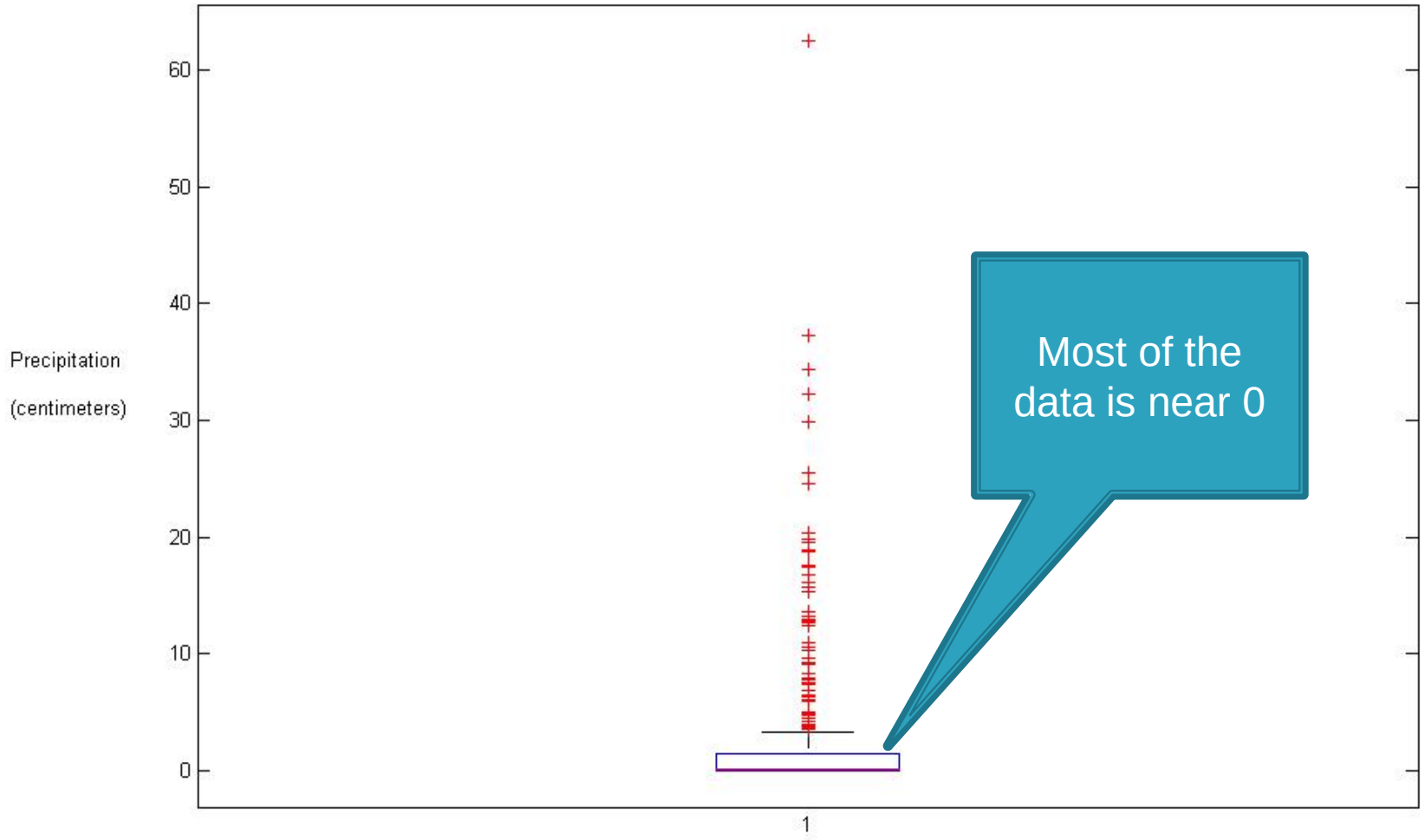


Boxplot for Heights of Elderly Women

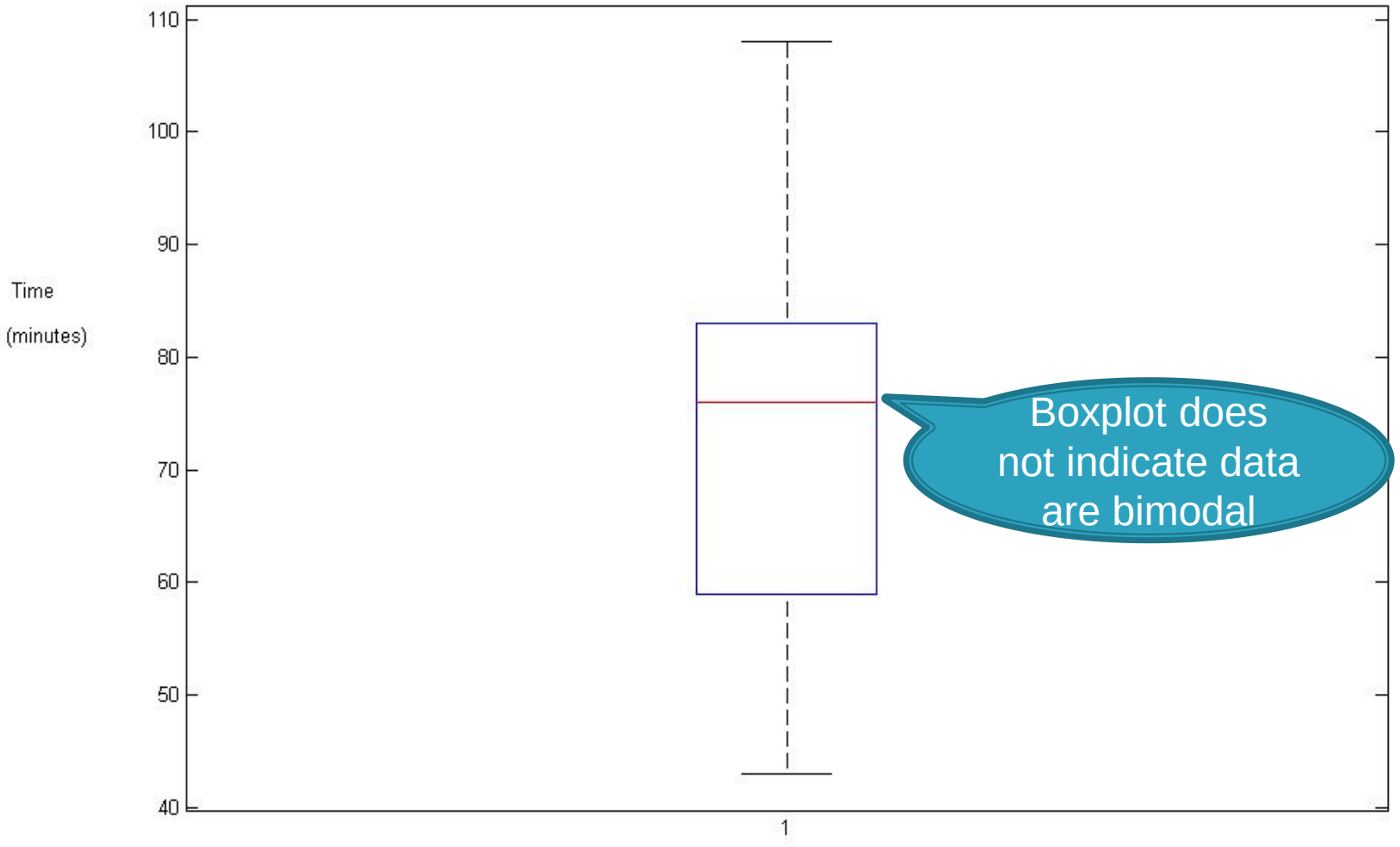


Rotate to get Boxplot

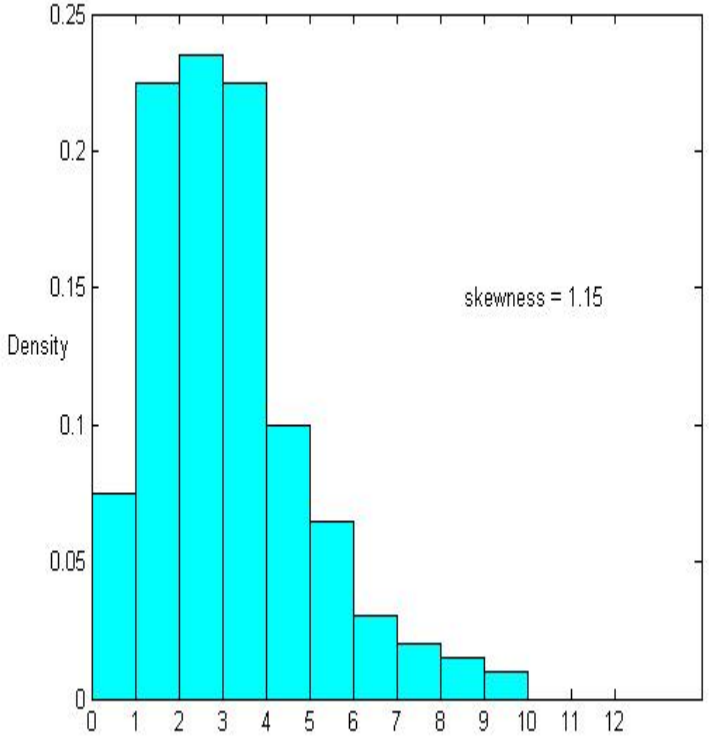
Boxplot of the Precipitation Data for 2010

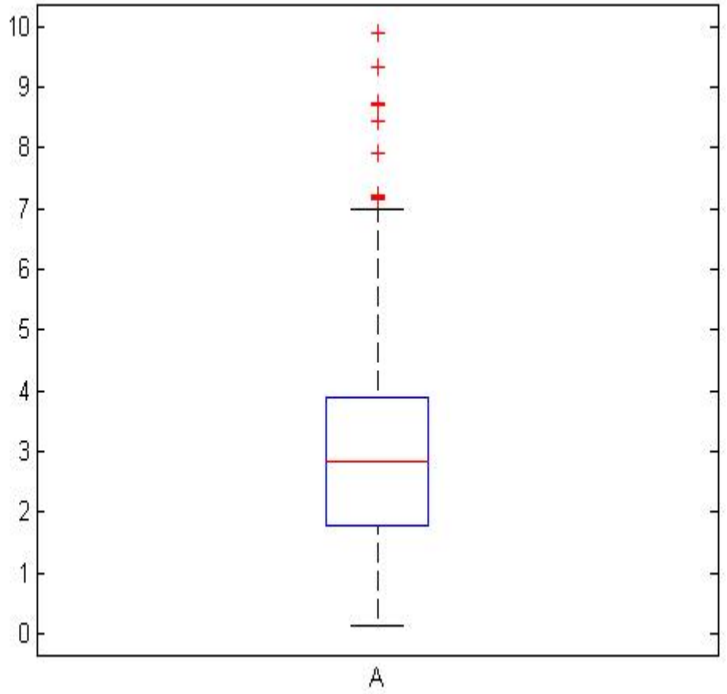


Boxplot for the Old Faithful Data

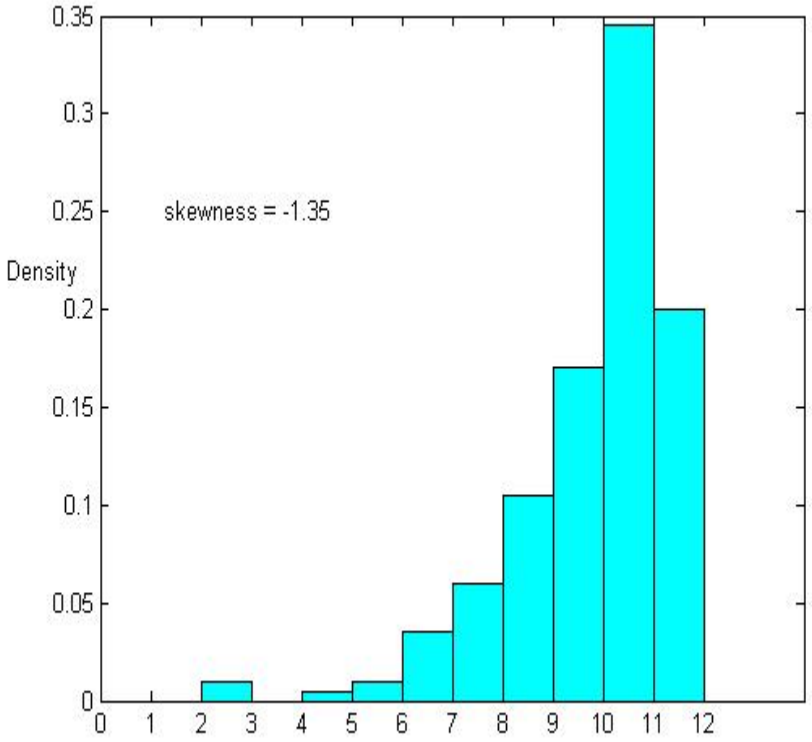


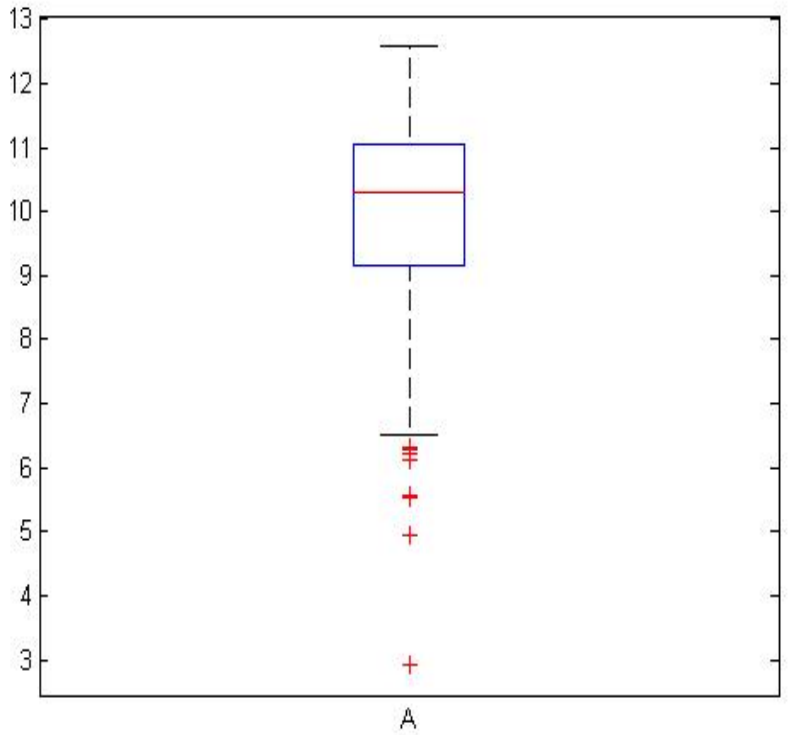
Positive Skewnes: Histogram and Boxplot



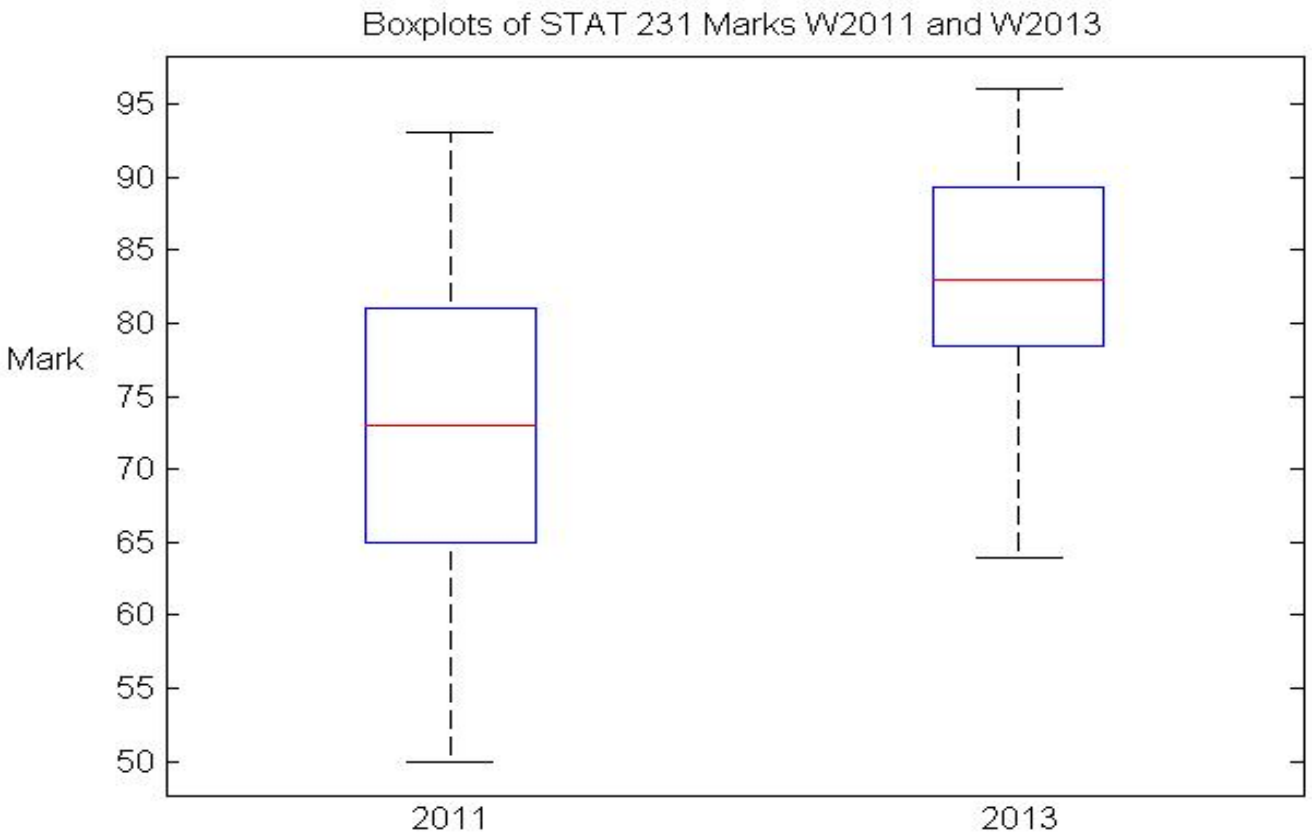


Negative Skewnes: Histogram and Boxplot





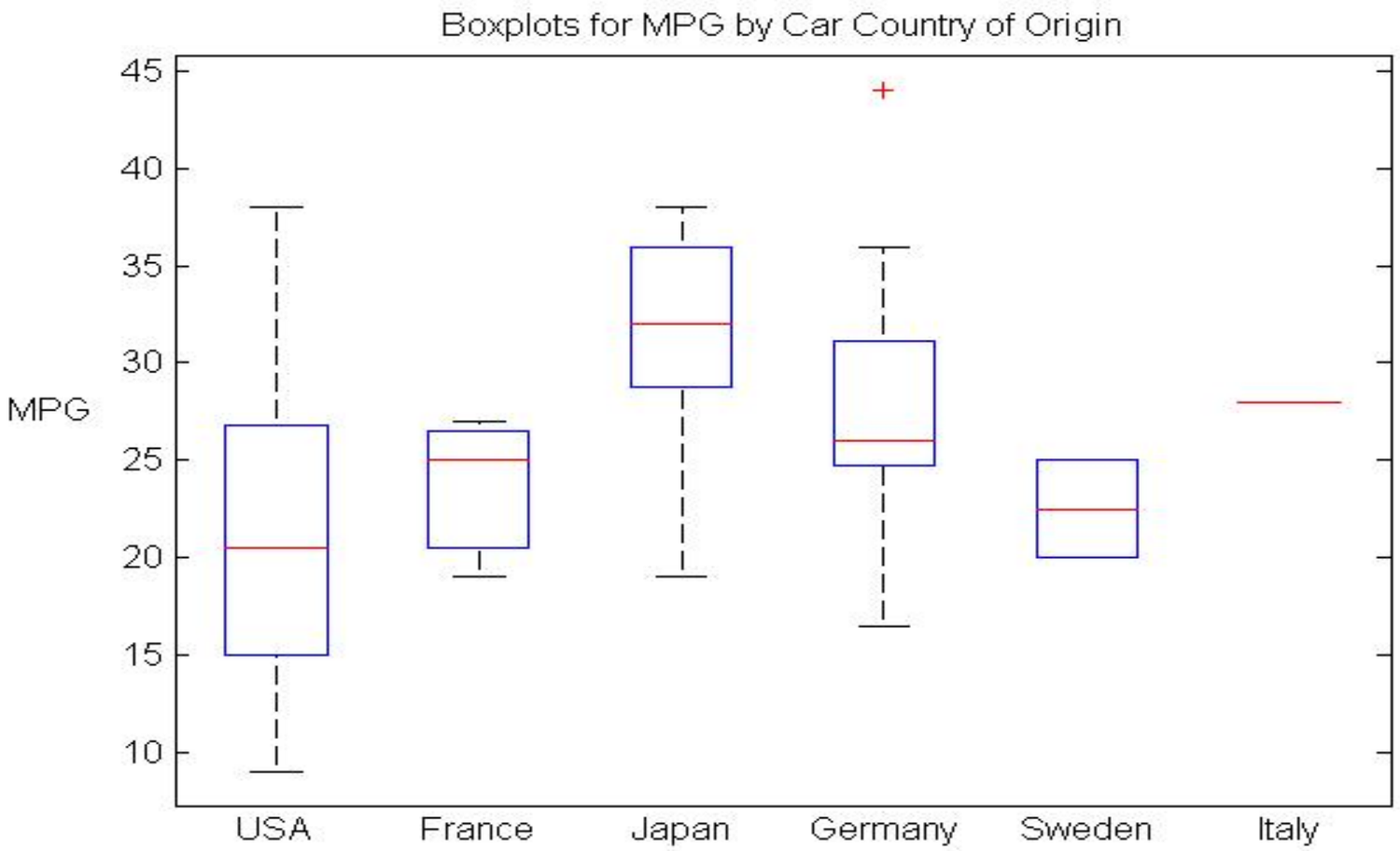
Boxplots can also be used for comparing the values of variates of two or more groups



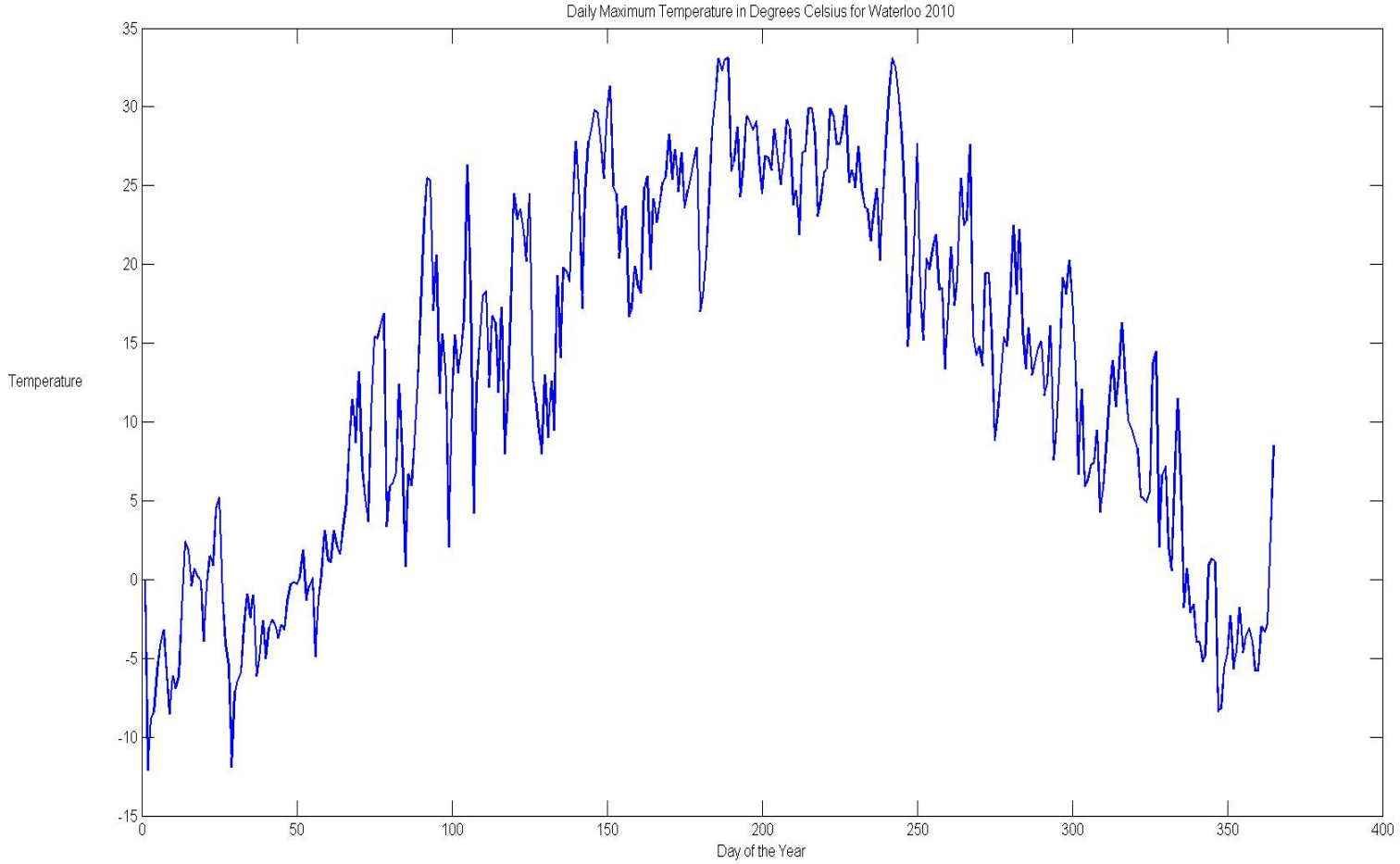
W11 Final Grades

W13 Final Grades

Comparison of MPG by Car's Country of Origin



Run Chart – A graphical summary of data which are varying over time



Scatterplots

So far we have looked at graphical and numerical summaries of data $\{y_1, y_2, \ldots, y_n\}$ where y_i is a real number.

We now wish to look at ways of summarizing datasets of the form $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ where x_i and y_i are real numbers (bivariate data).

The most obvious way of graphically summarizing these data is simply to plot the points (x_i,y_i), $i = 1,\ldots,n$.

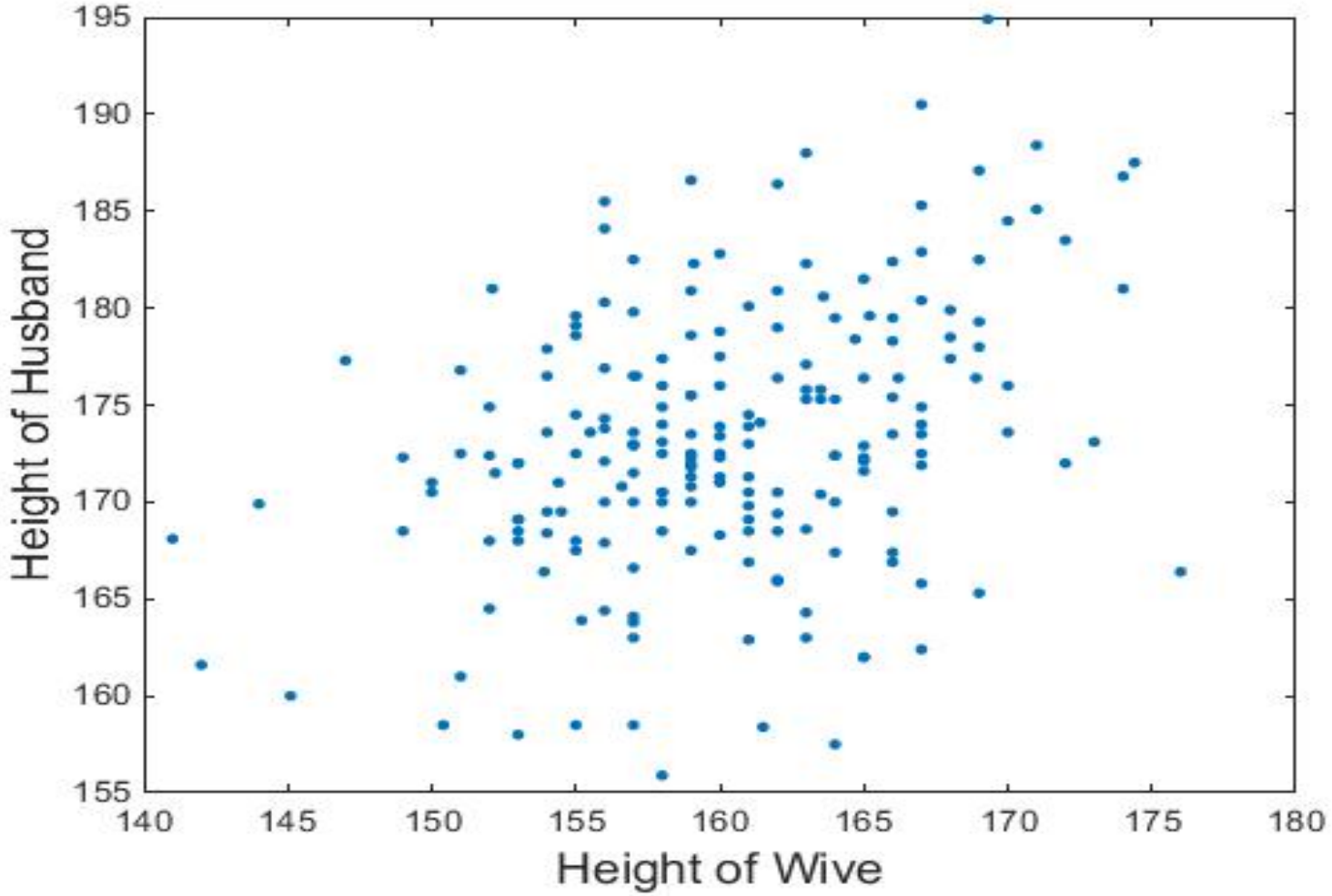
This is called a scatterplot.

Example: Random sample of 200 married men and their wives from a study of heights and weights of the adult population in Great Britain in1980.

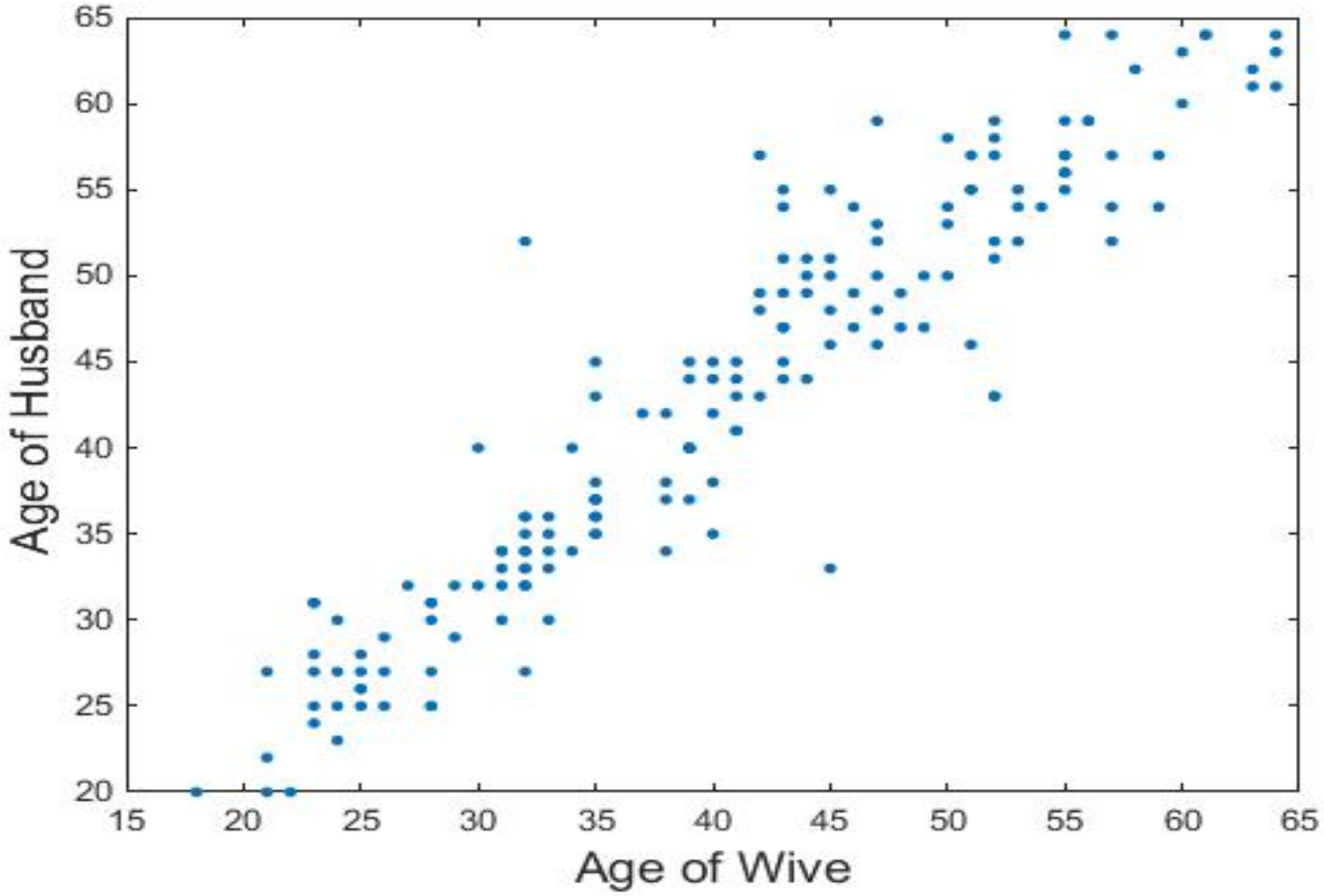
Husband Age (years)	Husband Height (cms)	Wife Age (years)	Wife Height (cms)	Husband age (years) at marriage
49	180.9	43	159.0	25
25	184.1	28	156.0	19
40	165.9	30	162.0	38
52	177.9	57	154.0	26
⋮	⋮	⋮	⋮	⋮
43	173.0	52	161.0	33
42	175.3	*	163.5	30
47	174.0	43	158.0	26
31	168.5	23	161.0	26

*** means a missing value**

Scatterplot of heights of husbands and wives (no missing data)



Scatterplot of ages of husbands and wives (ignoring 30 missing values)



Is there a relationship between x and y?

Looking at the scatterplot of heights of husbands versus their wives is there evidence of a relationship between these two variates?

Looking at the scatterplot of ages of husbands versus their wives is there evidence of a relationship between these two variates?

Sample Correlation

A numerical summary of bivariate data is the sample correlation.

Definition (Course Notes, page 14)

For data $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ the sample correlation is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$

and $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Sample Correlation Continued

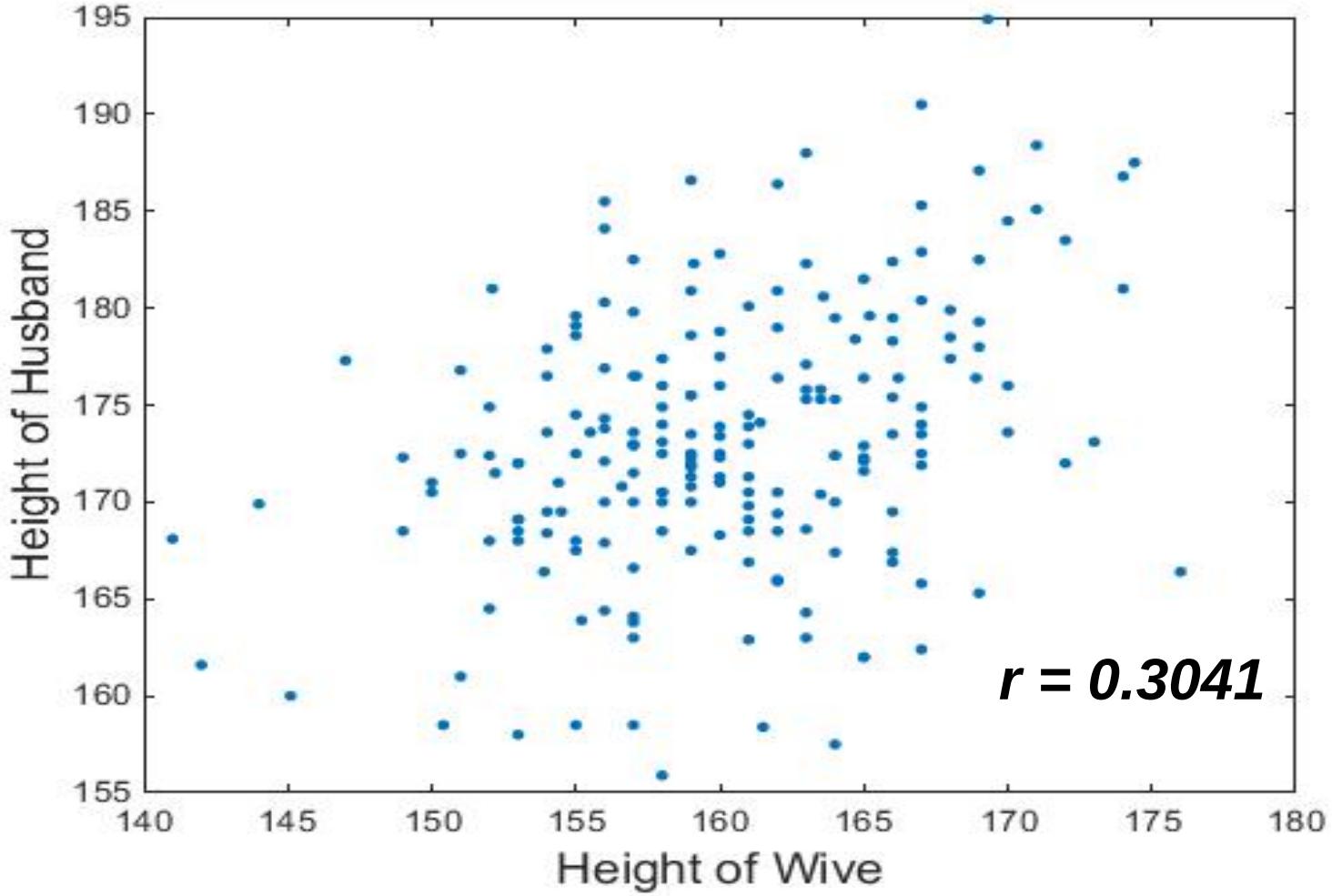
The sample correlation takes on values between *-1* and *1*. It is a measure of the linear relationship between the two variates x and y.

If the value of *r* is close to *1* then we say that there is a strong positive relationship between the two variates.

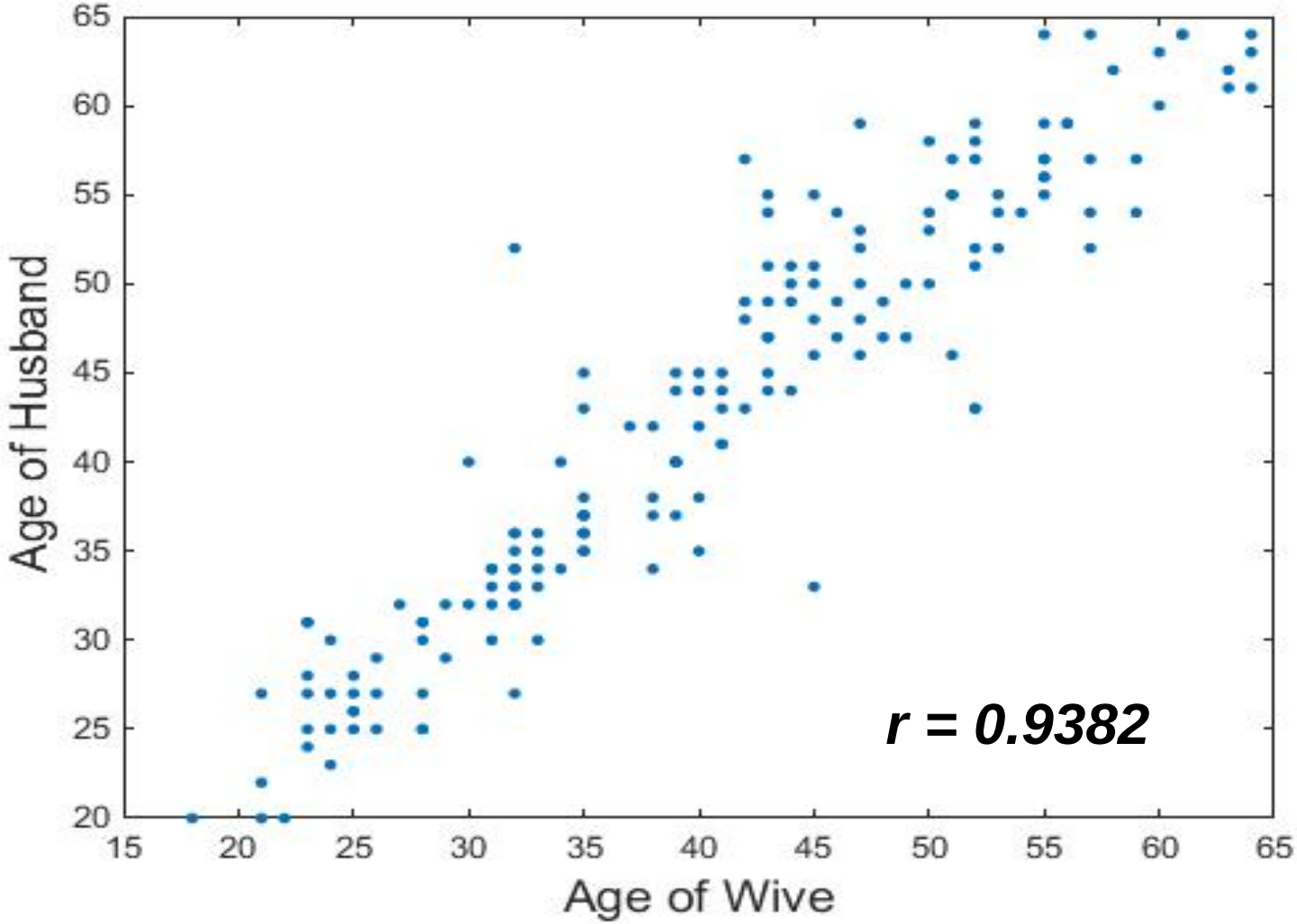
If the value of *r* is close to *-1* then we say that there is a strong negative relationship between the two variates.

If the value of *r* is close to *0* then we say that there is no relationship between the two variates.

Scatterplot of heights of husbands and wives (no missing data)

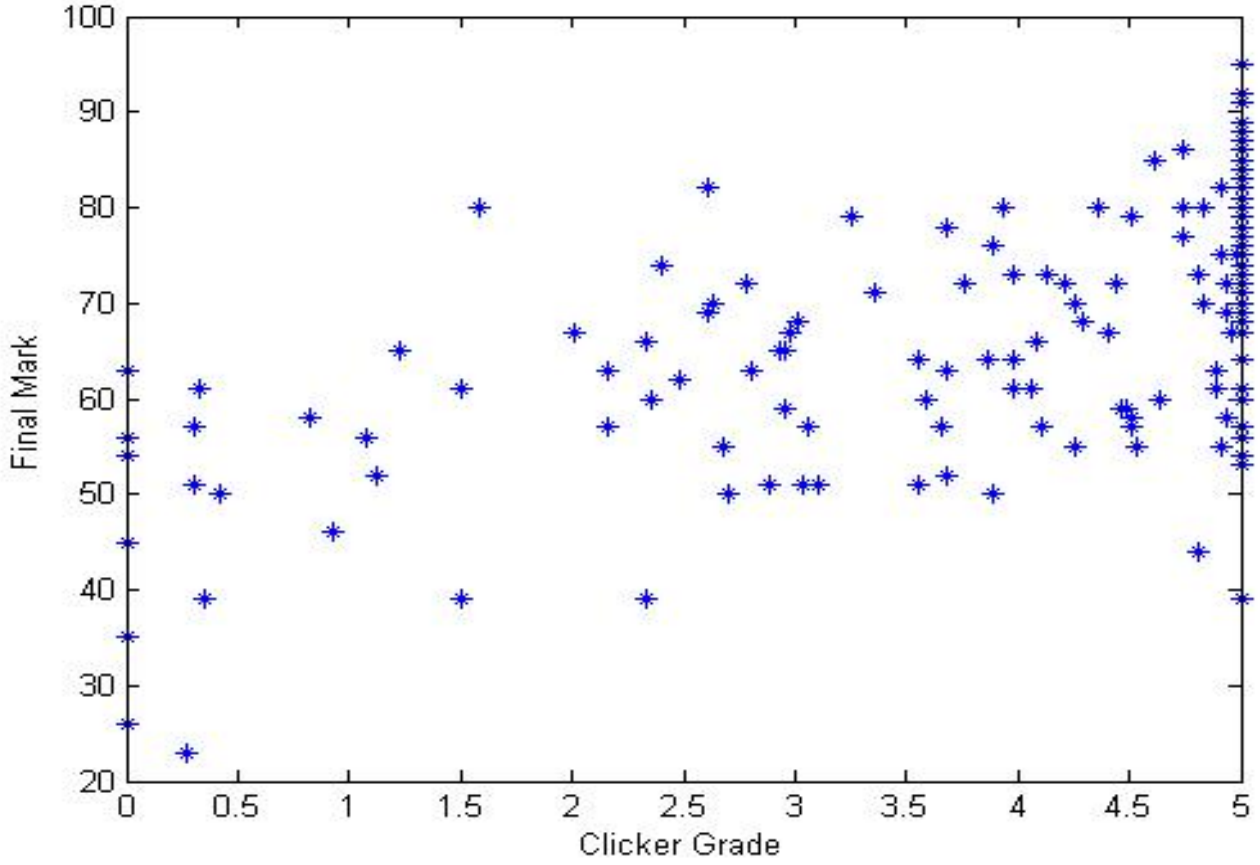


Scatterplot of ages of husbands and wives (ignoring 30 missing values)

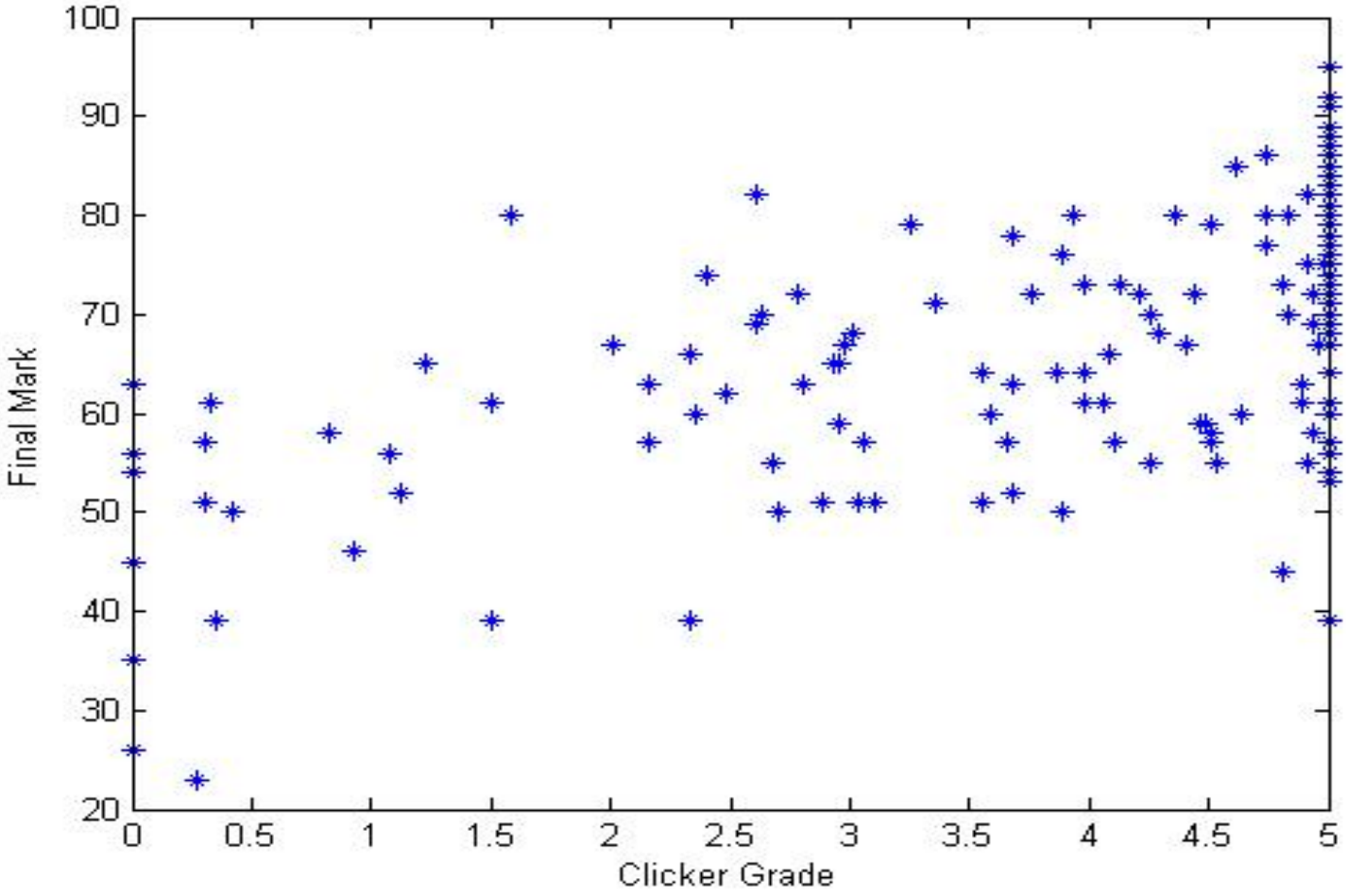


Scatterplot of STAT 231 Final Grade versus Clicker Grade - Sec 001, W16

What is the sample correlation for these data?



Scatterplot of STAT 231 Final Grade versus Clicker Grade - Sec 001, W16



r = 0.60

Response versus Explanatory Variates

In this example the problem of interest was to study the relationship between two variates (final mark in STAT 231 and clicker grade).

There is a natural division of the variates into two types: response variates and explanatory variates.

The explanatory variate *x* is in the study to partially explain or determine the distribution of *Y*.

In this example *Y* = final mark in STAT 231 is the response variate and *x* = clicker grade is the explanatory variate.

Response versus Explanatory Variates– Another Example

In a random sample of 1718 men aged 40-55, men were classified according to whether they were heavy coffee drinkers (more than 100 cups/month) or not (less than 100 cups/month) and whether they suffered from CHD (coronary heart disease) or not.

Researchers were interested in whether there was a relationship between coffee consumption and CHD.