Lecture 1
Introduction to Statistics

Objective and Scope of Statistics

Study Design, Types of Studies, Variables

Bias, Sample Size Issues

Data Classification

Homework problems

What is Statistics?

- Statistics is the science of learning from data.
- Statistics involves collecting, presenting, analysing and interpreting data.

Objectives and Scope of Statistical Studies

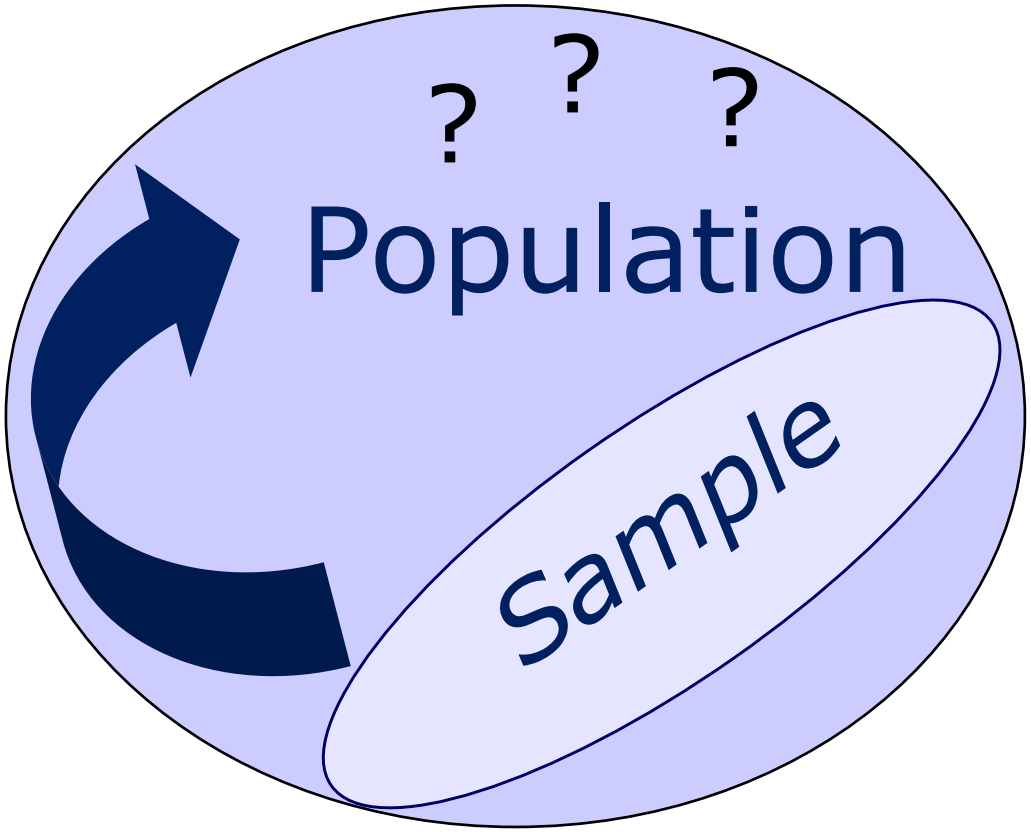
Objectives

- A primary *objective* of the statistician is to obtain information about a *target population*, using a *sample*.
- The *target population* comprises all relevant subjects of interest.
- The *sample* is a manageable subset, selected to make the study feasible.

Answering Research Questions

Research Questions

We use a SAMPLE to answer questions about a target POPULATION



A Data Set

We will look at a data set which has some variables which are numerical and some variables which are categorical. These data were recorded from the CIA World Factbook 2013. 27 countries of the world were selected at random and the following information was recorded on each country:

Variable Name	**Variable Description**
Location	Continent in which country is located
Life Expectancy	Life expectancy of birth for total population
Infant Mortality Rate	Deaths during first year per 1000 live births
Unemployment	Unemployment Rate: 1 = <10%, 2 = 10 to 20%, 3 = 20 to 30%, 4 = 30 to 40%, 5 = 40 to 50%, 6 = >50%
Olympics	Number of times country has hosted either summer or winter Olympic Games

27 Countries of the World

Country	Location	LifeExp	InfMort	Unemp	Olympics	Country	Location	LifeExp	InfMort	Unemp	Olympics
Australia	Oceania	82.0	4.49	1	2	Malaysia	Asia	74.3	14.12	1	0
Cameroon	Africa	55.0	58.51	3	0	Mozambique	Africa	52.3	74.63	2	0
Chad	Africa	49.1	91.94	3	0	Namibia	Africa	52.0	45.62	6	0
China	Asia	75.0	15.2	1	1	New Zealand	Oceania	80.8	4.65	1	0
Ethiopia	Africa	60.0	58.28	2	0	Singapore	Asia	84.1	2.59	1	0
Fiji	Oceania	71.9	10.46	1	0	Solomon Is.	Oceania	74.7	16.7	4	0
France	Europe	81.6	3.34	1	5	Spain	Europe	81.4	3.35	3	1
Germany	Europe	80.3	3.48	1	3	Sri Lanka	Asia	76.2	9.24	1	0
Greece	Europe	80.2	4.85	3	3	Sweden	Europe	81.3	2.73	1	1
Hong Kong	Asia	82.2	2.89	1	0	Tanzania	Africa	60.8	45.10	2	0
India	Asia	67.5	44.60	1	0	Tunisia	Africa	75.5	24.07	2	0
Indonesia	Asia	71.9	26.06	1	0	UK	Europe	80.3	4.50	1	3
Italy	Europe	82.0	3.33	2	3	Zambia	Africa	51.5	68.58	2	0
Japan	Asia	84.2	2.17	1	3						

Location is *nominal (categorical)*

Life expectancy is *continuous (numerical)*

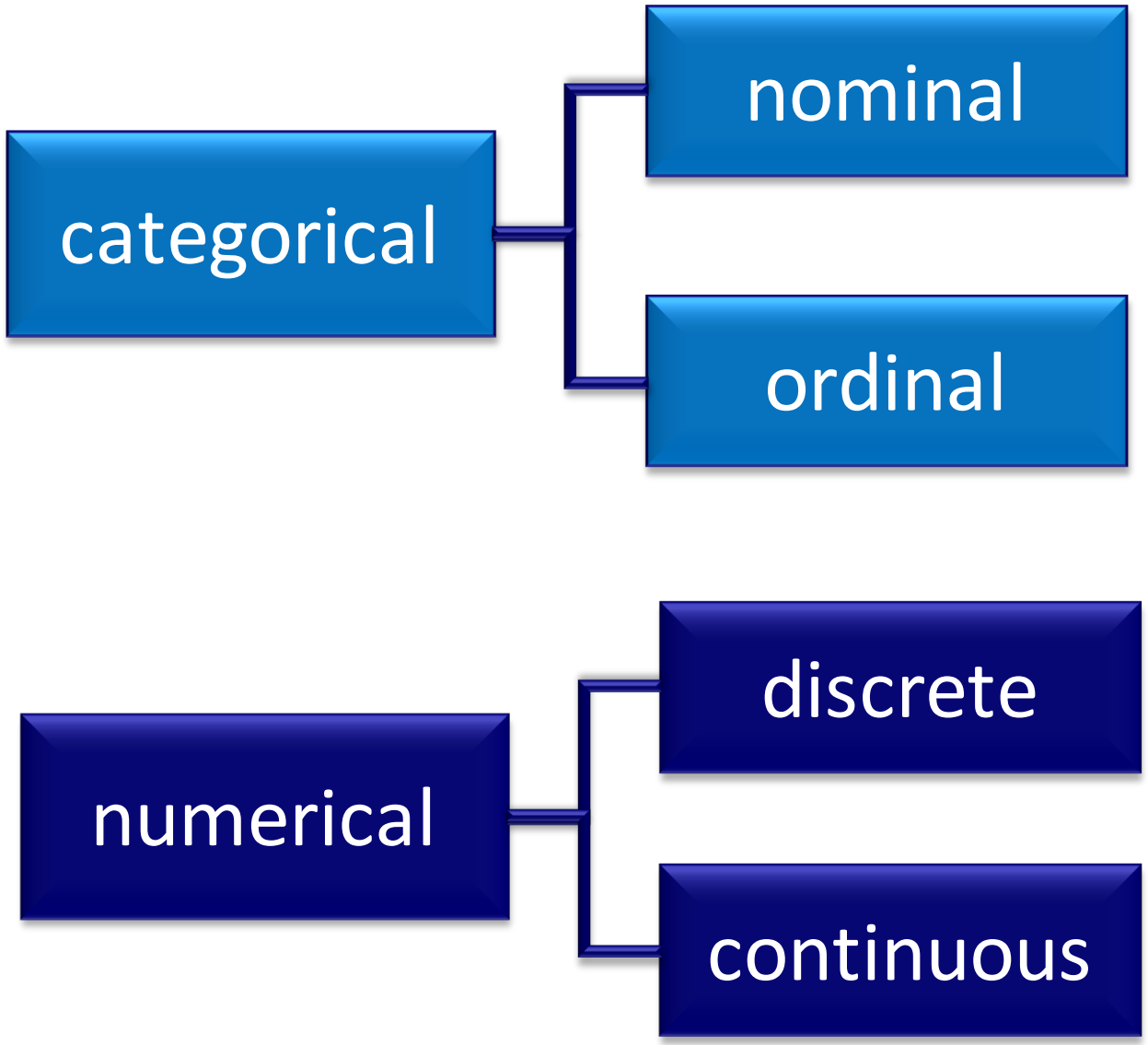
Infant Mortality is *continuous (numerical)*

Unemployment Rate is *ordinal (categorical)*

Olympics is *discrete (numerical)*

Data Classification

Data Classification

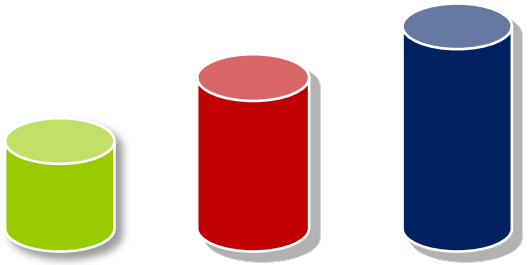


Numerical or Categorical?

Numerical

An observation can take ANY integer or non-integer value

Categorical



An observation can only belong to one of two or more groups

Categorical Variables

Categorical variables are variables for which each observation falls into one of a number of groups.

Nominal variables are named variables with no inherent ordering: eg. favourite colour.

Ordinal variables are grouped variables with some ordering: eg. grade attained in unit.

If there are **two groups** the variable may be referred to as *binary* or *dichotomous*.

Binary variables can be either

- **nominal**, eg. sex, or
- **ordinal** eg. age group - ie. < 20 years, ≥ 20 years.

Colour (*nom*inal	Size (*ord*inal)		
	Small	Medium	Large
White			
Green			
Purple			

Numerical Variables

Numerical variables are measured variables and can be either discrete or continuous.

- **Discrete** variables are variables that take discrete values:

  eg. number of students in class.

- **Continuous** variables are those that can assume any value, usually within a certain range:

  eg. height, weight.

27 Countries of the World

Country	Location	LifeExp	InfMort	Unemp	Olympics	Country	Location	LifeExp	InfMort	Unemp	Olympics
Australia	Oceania	82.0	4.49	1	2	Malaysia	Asia	74.3	14.12	1	0
Cameroon	Africa	55.0	58.51	3	0	Mozambique	Africa	52.3	74.63	2	0
Chad	Africa	49.1	91.94	3	0	Namibia	Africa	52.0	45.62	6	0
China	Asia	75.0	15.2	1	1	New Zealand	Oceania	80.8	4.65	1	0
Ethiopia	Africa	60.0	58.28	2	0	Singapore	Asia	84.1	2.59	1	0
Fiji	Oceania	71.9	10.46	1	0	Solomon Is.	Oceania	74.7	16.7	4	0
France	Europe	81.6	3.34	1	5	Spain	Europe	81.4	3.35	3	1
Germany	Europe	80.3	3.48	1	3	Sri Lanka	Asia	76.2	9.24	1	0
Greece	Europe	80.2	4.85	3	3	Sweden	Europe	81.3	2.73	1	1
Hong Kong	Asia	82.2	2.89	1	0	Tanzania	Africa	60.8	45.10	2	0
India	Asia	67.5	44.60	1	0	Tunisia	Africa	75.5	24.07	2	0
Indonesia	Asia	71.9	26.06	1	0	UK	Europe	80.3	4.50	1	3
Italy	Europe	82.0	3.33	2	3	Zambia	Africa	51.5	68.58	2	0
Japan	Asia	84.2	2.17	1	3						

Location is *nominal (categorical)*

Life expectancy is *continuous (numerical)*

Infant Mortality is *continuous (numerical)*

Unemployment Rate is *ordinal (categorical)*

Olympics is *discrete (numerical)*

Graphing Data

"Graphs allow us to explore data to see overall patterns and to see detailed behaviour; no other approach can compete in revealing the structure of data so thoroughly." **(Cleveland (1994))**

The type of display that we provide depends on the variable/s being displayed.

We shall begin with two displays for categorical data. We will also look at displays for numerical data and a display which combines information from both a categorical and a numerical variable.

Displaying Data: Graphical Displays/Graphical Summaries

DATA	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Graphing categorical data

- Bar charts
- Pie charts

2012 Olympic Medals

The table below shows the results of the first ten events listed in the Olympic medal tally for the 2012 Olympic Games held on London.

Sport	Event	Gold	Silver	Bronze
Archery	M. Individual	S. Korea	Japan	China
Archery	F. Individual	S. Korea	Mexico	Mexico
Archery	M. Team	Italy	US	S. Korea
Archery	F. Team	S. Korea	China	Japan
Athletics	M. 100m	Jamaica	Jamaica	US
Athletics	F. 100m	Jamaica	US	Jamaica
Athletics	M. 200m	Jamaica	Jamaica	Jamaica
Athletics	F. 200m	US	Jamaica	US
Athletics	M. 400m	Grenada	Dominican Republic	Trinidad & Tobago
Athletics	F. 400m	US	Great Britain	US
…….	…….	…….	…….	…….

2012 Olympic Medals by Region

There are too many countries to display individually so the countries have been grouped by region as follows:

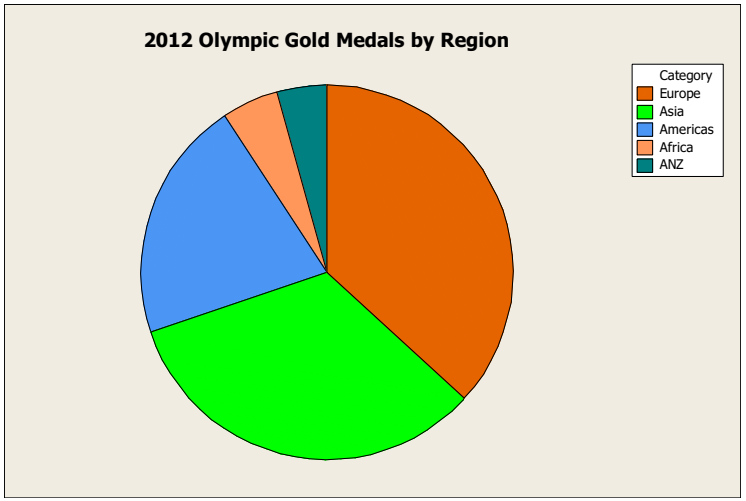
Region	Gold	Silver	Bronze	Total
Africa	15	16	15	46
The Americas	63	52	59	184
Australia/New Zealand (ANZ)	13	18	17	48
Asia	100	92	114	306
Europe	111	126	141	378

We will plot the gold medals for each region. The groups could be reordered by frequency to make the plot easier to interpret – we would only do this for nominal variables.

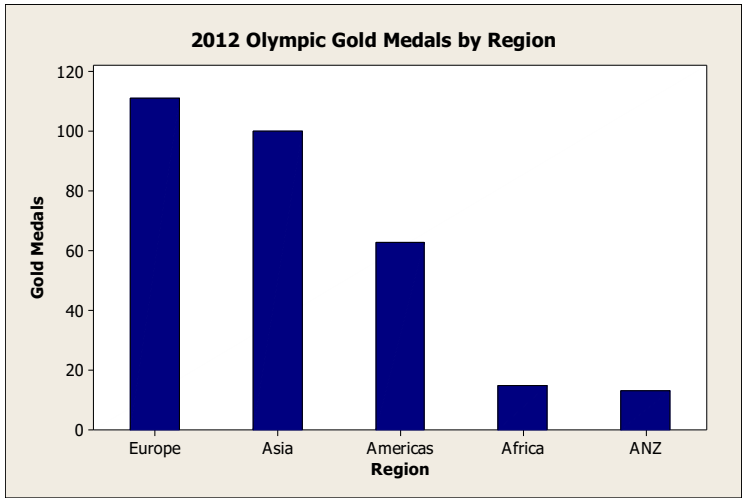
Region	Number of Gold Medals	Proportion of Gold Medals
Europe	111	0.37
Asia	100	0.33
Americas	63	0.21
Africa	15	0.05
ANZ	13	0.04
Total	302	1.00

Gold Medals: Country by Region

Pie Chart



Bar Chart



European and Asian countries won the majority of the gold medals at the 2012 Olympics. Australia and New Zealand, along with African countries won the least numbers of gold medals.

Graphs need:

- a title
- clearly labelled axes
- an explanatory comment
- to be clear and uncluttered

Displaying the Association Between Two Categorical Variables

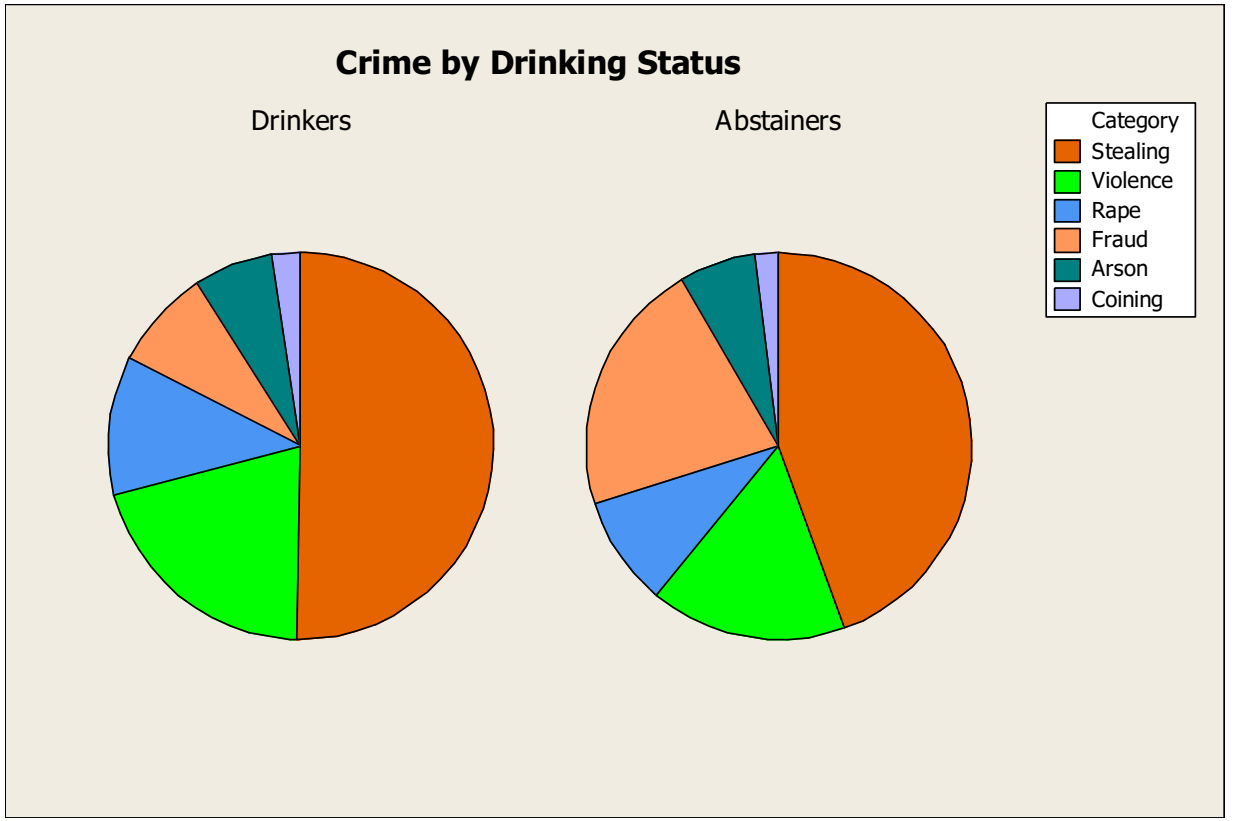
These data, collected from a gaol in England in 1909, show the numbers of people convicted of the six crimes listed.

Crime	Drinker	Abstainer
Arson	50	43
Rape	88	62
Violence	155	110
Stealing	379	300
Coining	18	14
Fraud	63	144

We'll use both pie charts and a clustered bar chart to compare the crimes committed by drinkers and abstainers.

Source: Hand, D. J. et al. *A Handbook of Small Data Sets* (1994)

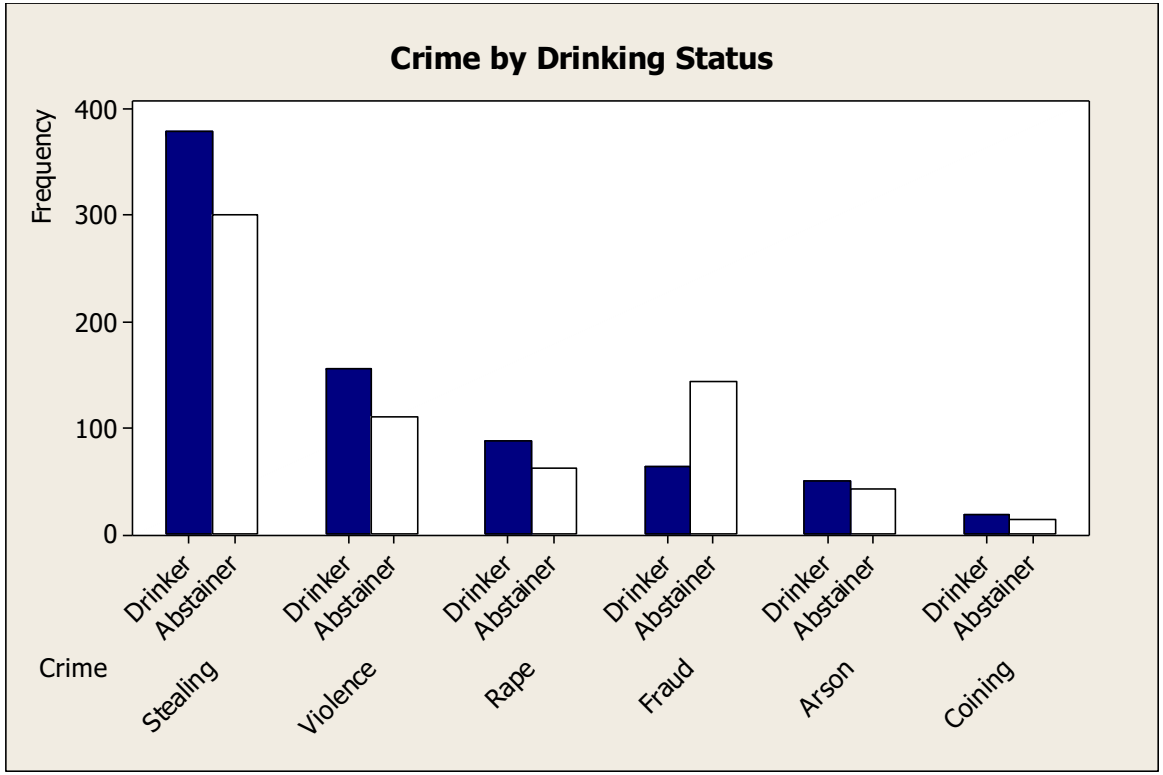
Comparative Pie Charts



It is very difficult here to compare the frequencies between the two groups. About all we can see is that stealing is the most common crime for each group.

Clustered Bar Charts

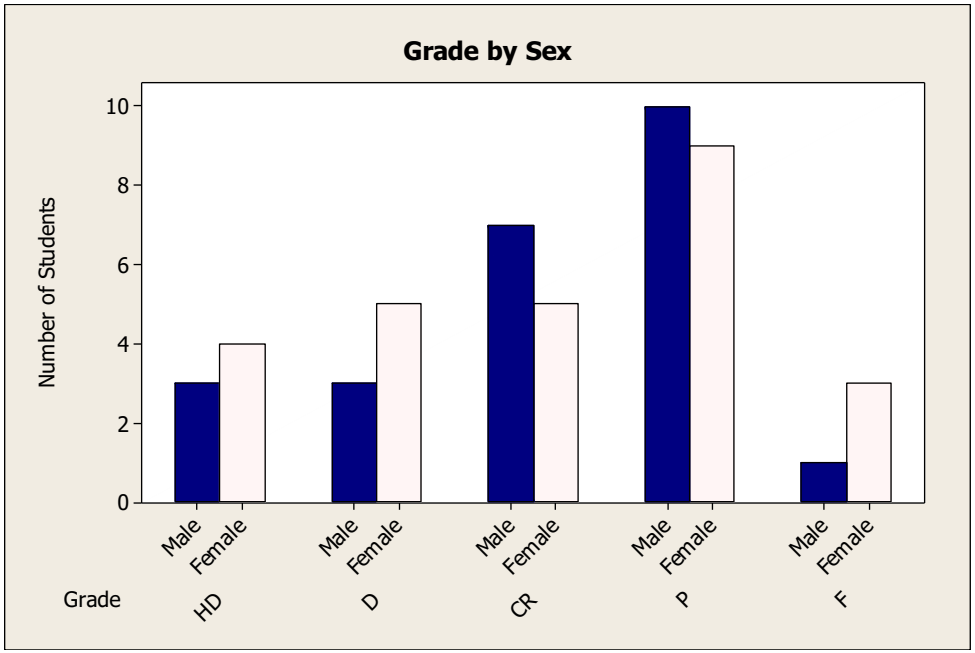
Crime is nominal: Groups have been re-ordered by frequencies for the drinkers



Now it is much easier to make comparisons. What comment would you make here about the association between the type of crime and the drinking status?

Clustered Bar Chart – Another Example

The following graph displays the results obtained by a random sample of 50 STAT170 students from 2013. The students have been categorised by sex and by grade.



Grade is an ordinal variable so the order of the grades has been maintained.

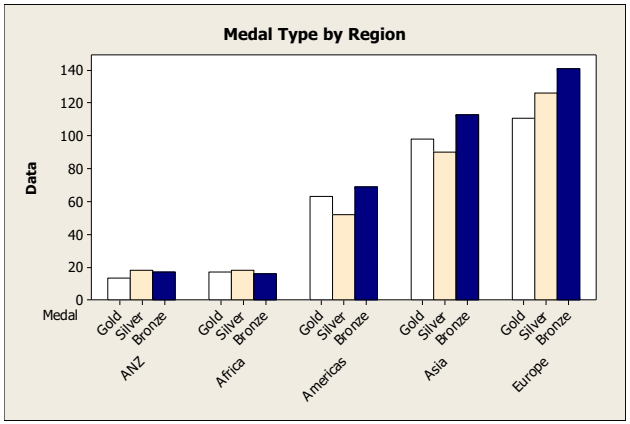
The grades appear to be similar for males and females. There does not seem to be a strong association between sex and grade.

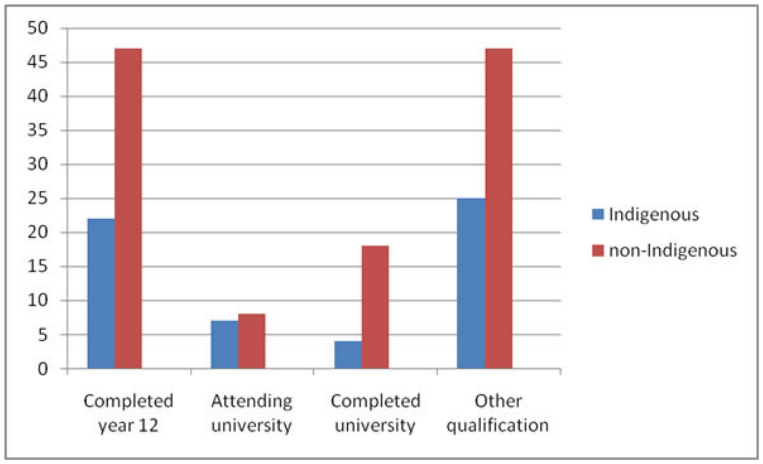
Source: *Stat170 student database* (2013)

Quiz 4

For each of these clustered bar charts, comment on any associations between variables:

Medal Type by Region
London2012.Olympics.com.au





Comparison in percentages of education levels of indigenous and non-indigenous Australians aged 15 years and older.

Healthinfonet.ecu.edu.au 2006.

Graphing numerical data

- Histograms
- Box Plots
- Scatter Plots

Constructing a Histogram

Life expectancies in African Countries (sorted in ascending order):

49.1	49.5	49.5	50.0	50.9	51.2	51.5	52.0	52.1	52.3	52.3	52.5	52.8
53.9	54.0	54.3	54.4	54.5	55.0	55.0	55.6	56.1	57.0	57.8	58.9	59.1
59.7	60.0	60.6	60.7	60.8	61.9	62.0	63.0	63.1	63.1	63.2	63.3	63.6
63.9	64.1	64.8	65.3	71.3	73.2	74.0	74.9	75.5	75.8	76.2	76.3	

Specify bins and count frequencies (number of observations) in each bin.

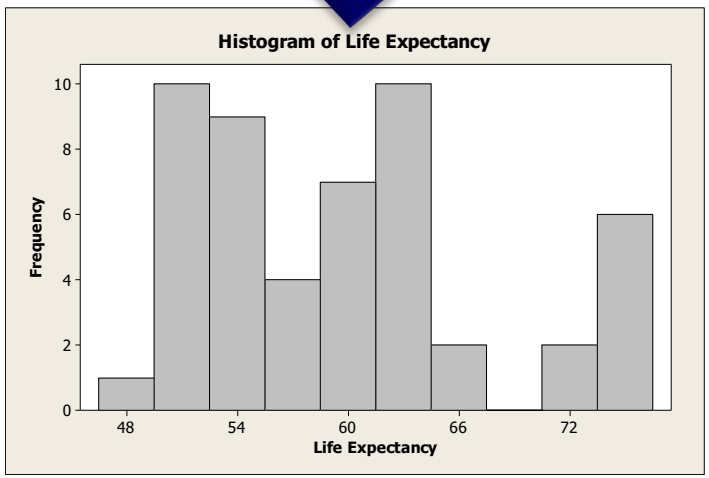
bin	frequency
45 – 50⁻	3
50 – 55⁻	15
55 – 60⁻	9
60 – 65⁻	15
65 – 70⁻	1
70 – 75⁻	4
75 – 80⁻	4

up to but not including 50

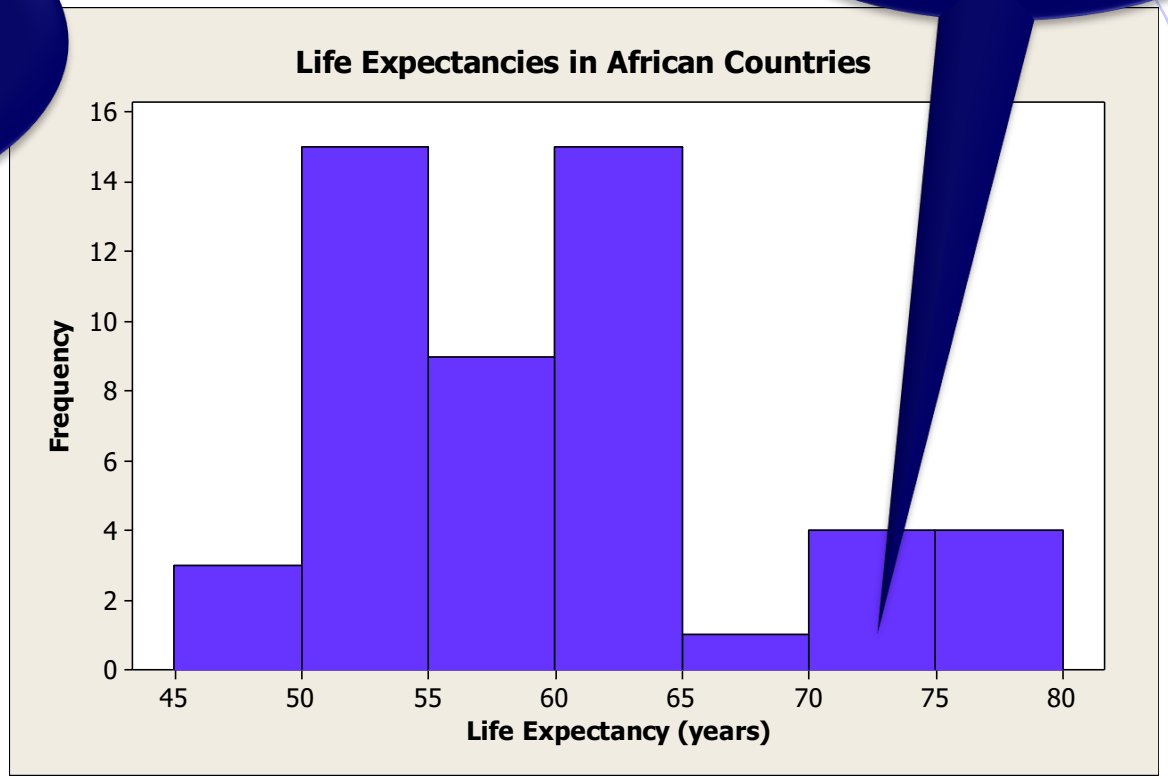
Source: *World Fact Book* (2013)

Histogram: Life Expectancies in Africa

Minitab initially produced the histogram as shown below by default. We have formatted the bins to present the graph on the right. Always present your graphs clearly and concisely



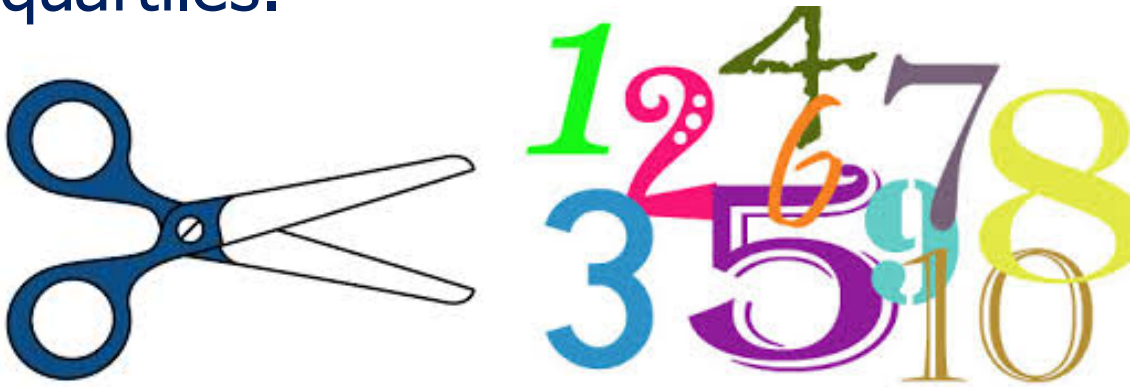
This scale reminds us that the data are continuous



Comment: Most African countries have total life expectancies below 65 years. A small number of African countries have total life expectancies between 70 and 80 years.

Constructing a Box Plot: Locating the Median and the Quartiles

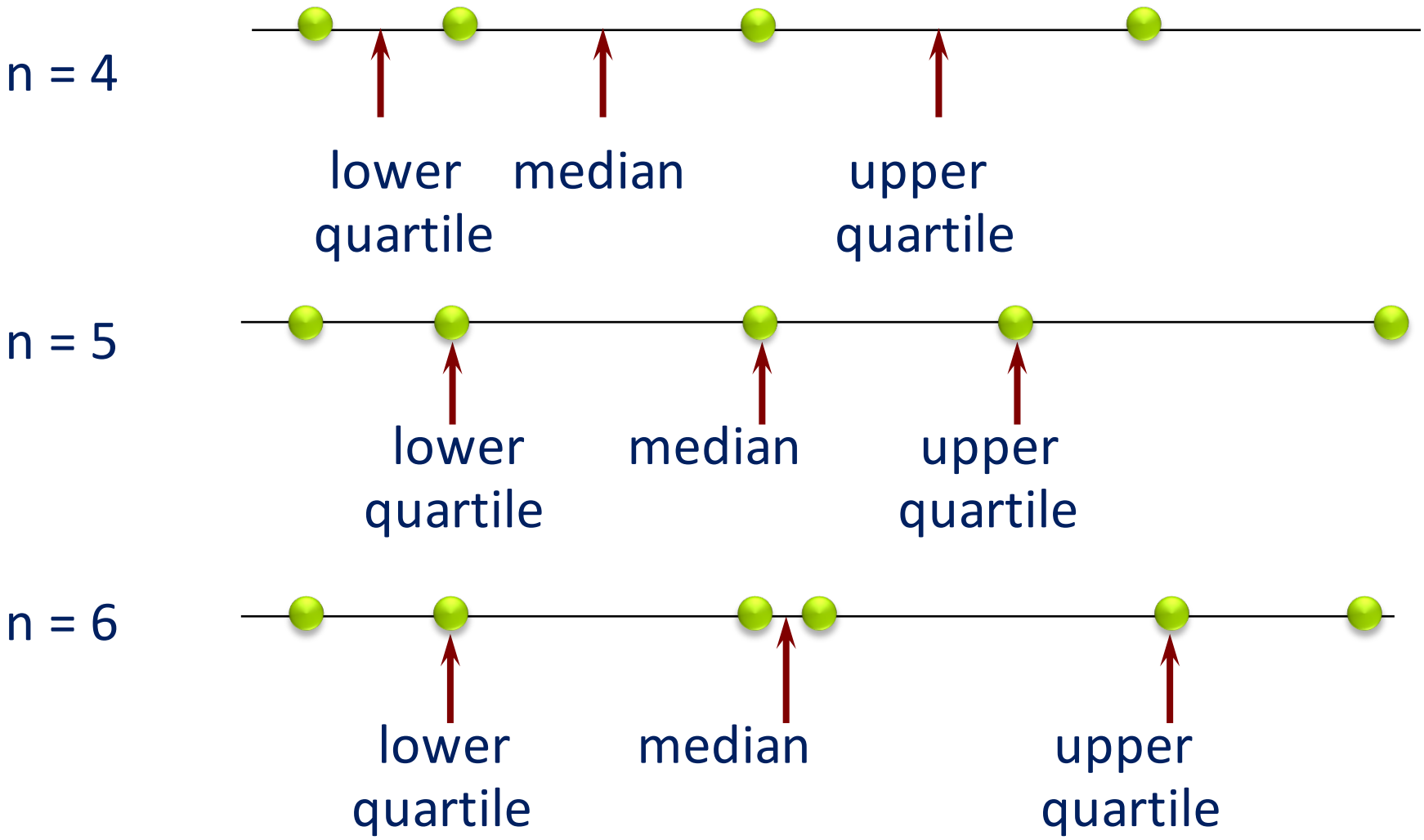
Before we draw a box plot we need to cut the sorted data set into quarters – ie. we need to locate the median and the quartiles.



Median and Quartiles

- If we sort a sample according to values of a particular variable (the *variable of interest*), then the *median*, together with the *quartiles* divide the sample into quarters.
- The ***median*** is the value which divides the sample in *half*. That is, the median separates the smallest 50% of data from the largest 50% of the data values i.e. the median is the 50^{th} percentile.
- The ***lower quartile*** **(LQ)** separates the smallest 25% of data values from the rest of the data set i.e. the lower quartile is the 25^{th} percentile.
- The ***upper quartile*** **(UQ)** separates the largest 25% of data values from the rest of the data set i.e. the upper quartile is the 75^{th} percentile.

Examples of Medians and Quartiles



Locating Medians and Quartiles

- Sort the data into ascending order.
- Find the 50^{th} percentile. This value is the median.
  - if n, the number of values in the data set is odd, the median is the middle observation
  - if n is even, the median is the average of the middle two observations
- Find the 25^{th} percentile. This value is the lower quartile.
- Find the 75^{th} percentile. This value is the upper quartile.

Locating P

Note that Minitab uses a slightly different method of locating percentiles so you may not get exactly the same results as Minitab when finding percentiles.

- n = sample size
- P = Percentile

  For the lower quartile we want the 25th ntile so P = 25, for the median we want the 50 rcentile so P = 50 and for the upper quartile we t the 75th percentile so P = 75.
- To locate the P^{th} percentile, calculate m here

  $m = n \times P/100$
- Use m to locate the P^{th} percentile as follows:

  If m is not an integer (ie. not a whole number) then round m up to the next whole number and this number gives you the position of the P^{th} percentile.

  If m is an integer (ie. a whole number), then you will find the P^{th} percentile half way between the value in the m^{th} position and the value in the $(m+1)^{th}$ position.

Location of the Lower Quartile

$\frac{n \times 25}{100}$

not an integer (m)

<u>Position:</u>

Next whole number

integer (m)

<u>Position</u>:

Half way between m^{th} & $(m+1)^{th}$ values

Location of the Median

$\frac{n \times 50}{100}$

not an integer (m)

Position:

Next whole number

integer (m)

Position:

Half way between m^{th} & $(m+1)^{th}$ values

Location of the Upper Quartile

$\frac{n \times 75}{100}$

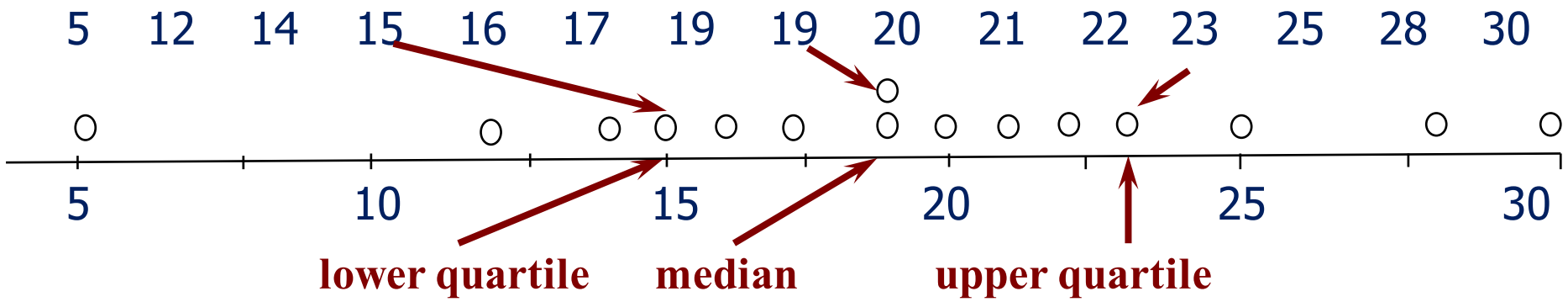
not an integer (m) → Position: Next whole number

integer (m) → Position: Half way between m^{th} & $(m+1)^{th}$ values

Note: If the lower quartile has already been located, then the location of the upper quartile can be easily determined from this eg. If the lower quartile is the 2nd smallest value in the data set, the upper quartile must be the 2nd largest value in the data set etc.

Students' Test Marks

Sample of n = 15 test marks:
(25% of 15 is 3.75, so each quarter contains 3.75 observations)



The lower quartile is 15, the **4th** value in the sorted list.

The median is the 8th value, ie. 19.

The upper quartile is 23, the 12th value in the sorted list, **or the 4th value in the sorted list counting down from the top**.

Source: *Stat170 student database* (2012)

Quiz 5

Locate the median and the quartiles of these samples:

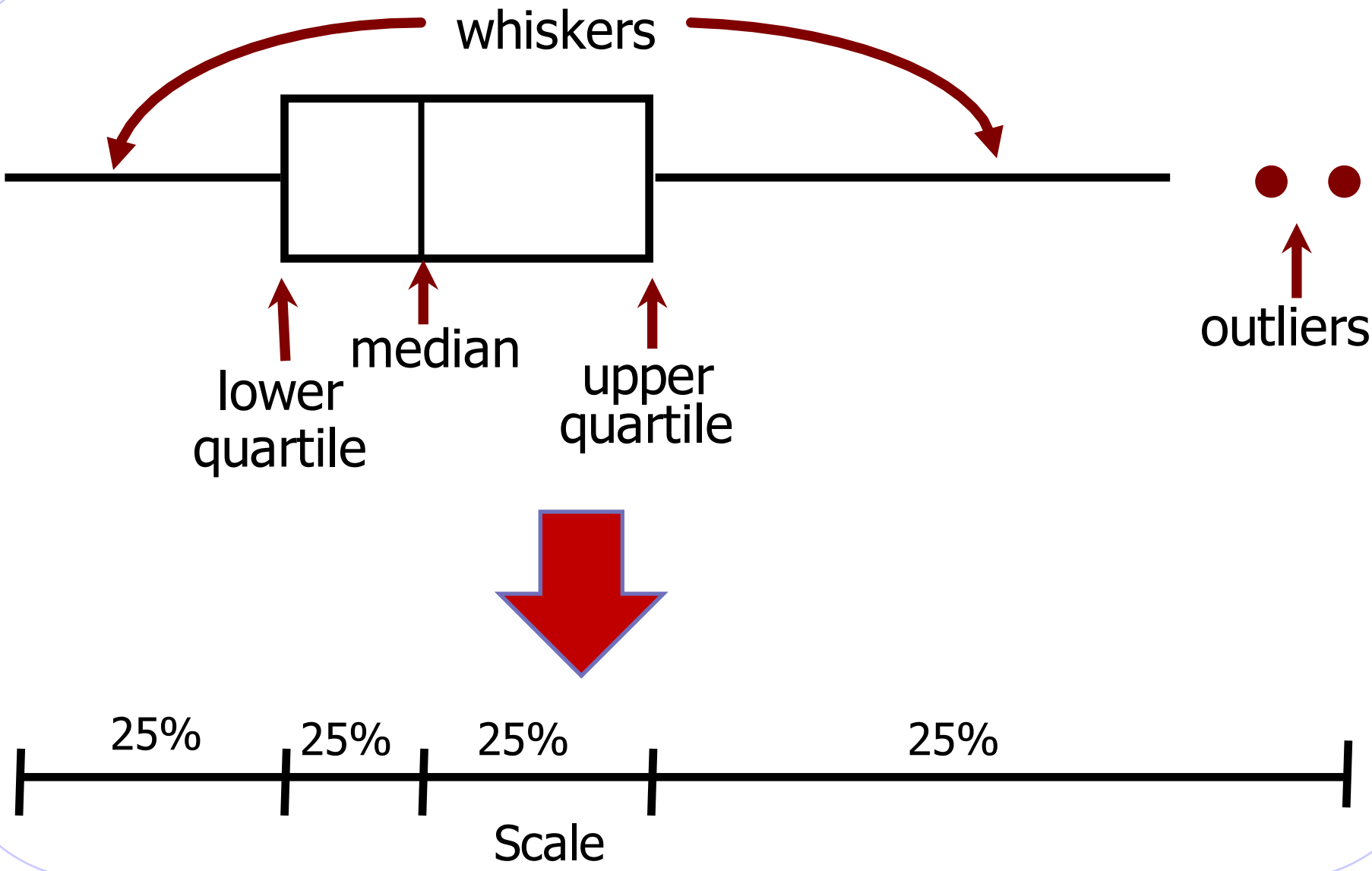
a) Ages (in years) of 9 female heart attack patients:

47 60 63 72 73 79 81 84 84

b) Time (in minutes) for 12 employees to travel to work:

13 15 20 22 25 28 36 45 58 110 120 145

Box Plots


whiskers
lower quartile
median
upper quartile
outliers
25%
25%
25%
25%
Scale

Fences, Whiskers and Outliers

Defining *fences* allows us to determine whether the data set has any outliers (unusually large or small values). Fences are defined using the **inter-quartile range (iqr):**

$$\mathbf{iqr = uq - lq}$$

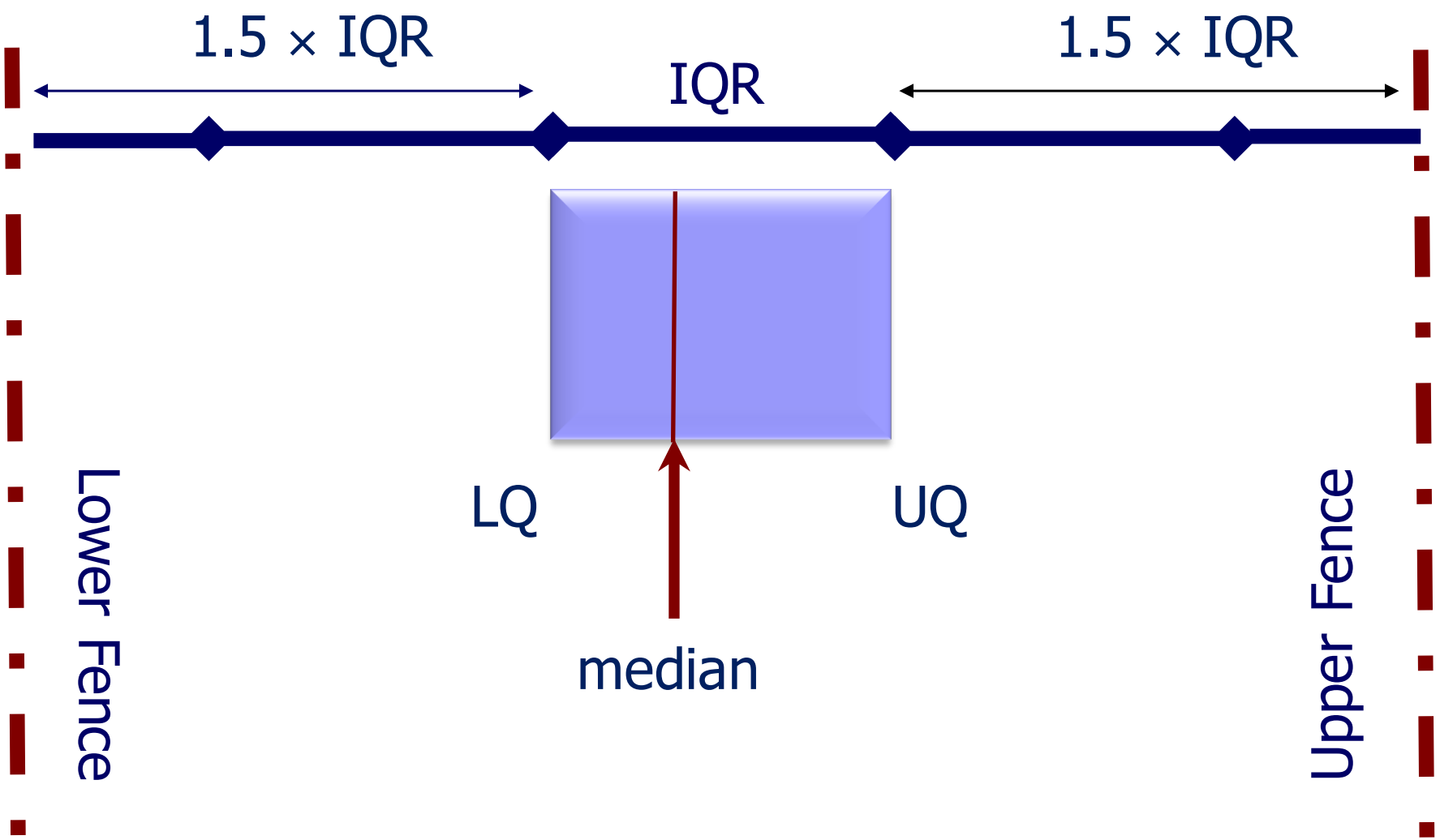
The *fences* are located at:

$$lq - 1.5 \times iqr \quad \text{(lower fence)}$$
$$uq + 1.5 \times iqr \quad \text{(upper fence)}$$

The whiskers *extend from each quartile to the furthest data value inside or on the corresponding fence.*

Outliers are data values which are outside the fences.

Fences



Unemployment Rates in African Countries

A random sample of 23 African countries were selected. Unemployment rates were recorded. We will construct a box plot to display these data.

Country	Unemployment (%)	Country	Unemployment (%)
Comoros	20.0	Mauritania	30.0
Ghana	11.0	Seychelles	2.0
Mali	30.0	Zambia	14.0
Senegal	48.0	Congo	53.0
Tunisia	17.4	Kenya	40.0
Cameroon	30.8	Mozambique	17.0
Gabon	21.0	S. Africa	25.1
Algeria	10.7	Djibouti	59.0
Lesotho	25.0	Namibia	51.2
Sudan	20.0	Botswana	17.8
Egypt	12.7	Libya	30.0
Nigeria	23.9		

Source: *World Fact Book* (2013)

Constructing a Box Plot

Unemployment Rates - Africa

2.0	25.0
10.7	25.1
11.0	30.0
12.7	30.0
14.0	30.0
17.0	30.8
17.4	40.0
17.8	48.0
20.0	51.2
20.0	53.0
21.0	59.0
23.9	

LQ

UQ

Median

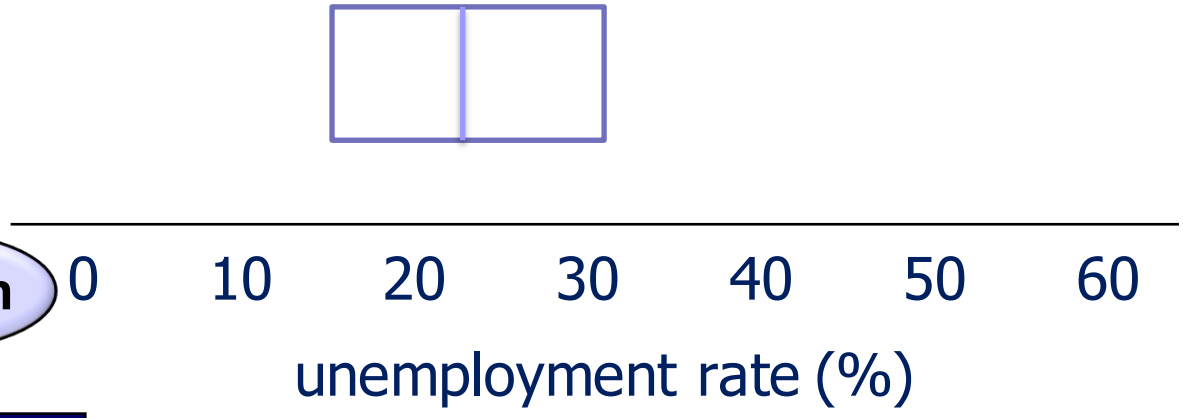
Source: *World Fact Book* (2013)

Step 1

Sort the data into ascending order

Step 2

Decide on an appropriate scale, and draw a box using the median & quartiles.



Constructing a Box Plot

Step 3

Calculate fences and trace.

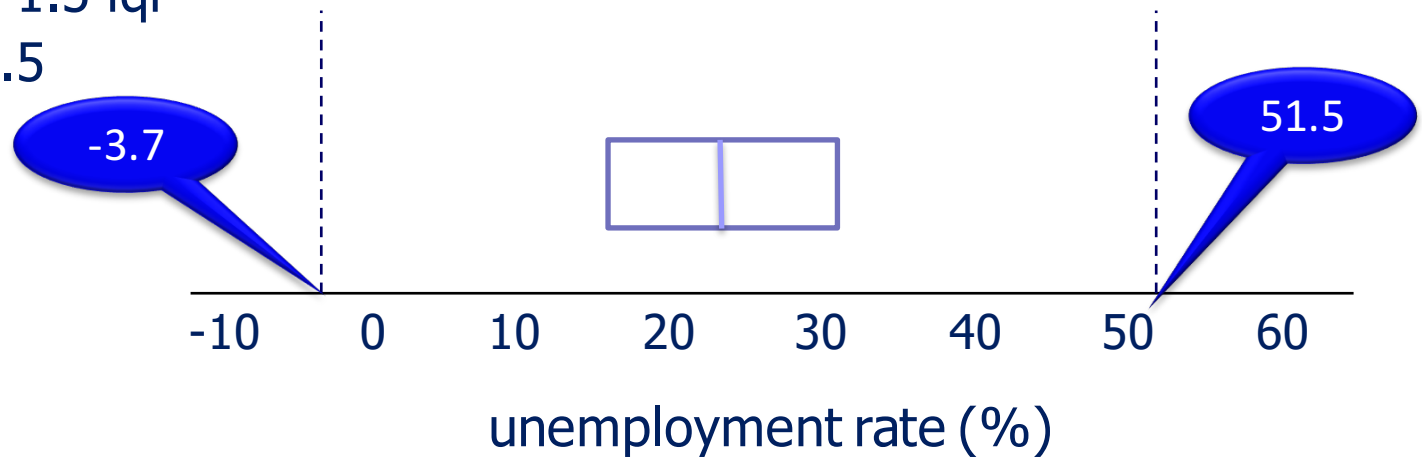
iqr=uq – lq = 30.8 – 17.0 = 13.8

lower fence = lq – 1.5 iqr

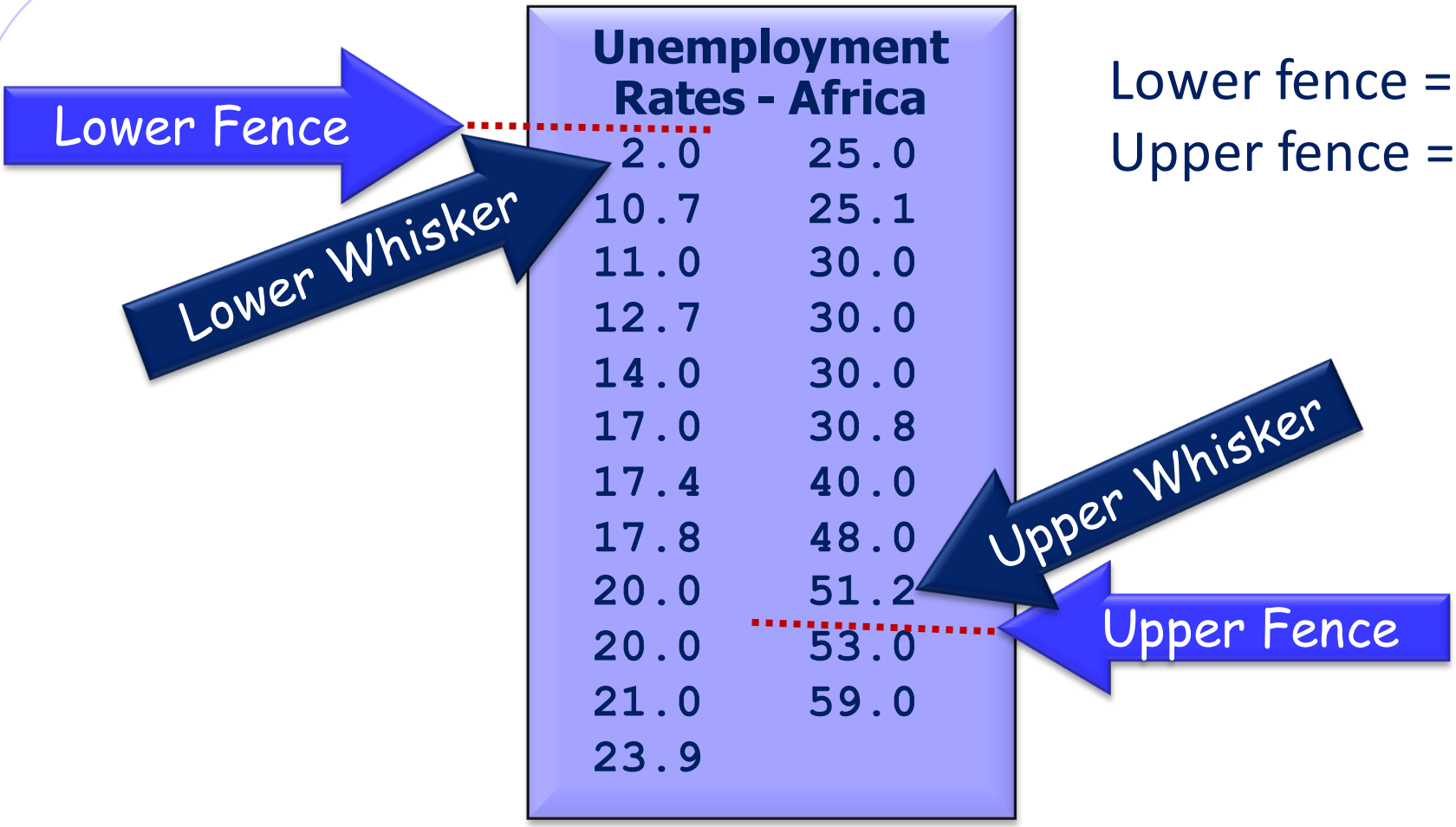
= 17 – 20.7 = – 3.7

upper fence = uq + 1.5 iqr

= 30.8 + 20.7 = 51.5



Constructing a Box Plot



Lower fence = –3.7

Upper fence = 51.5

Constructing a Box Plot

Step 3

Calculate fences and trace.

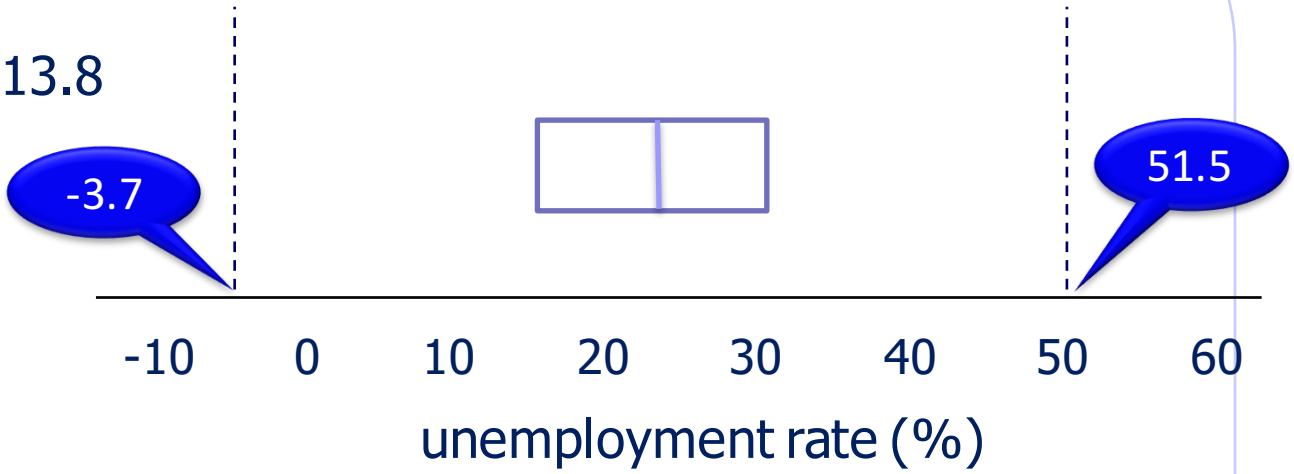
iqr=uq – lq = 30.8 – 17.0 = 13.8

lower fence = lq – 1.5 iqr

= 17 – 20.7 = – 3.7

upper fence = uq + 1.5 iqr

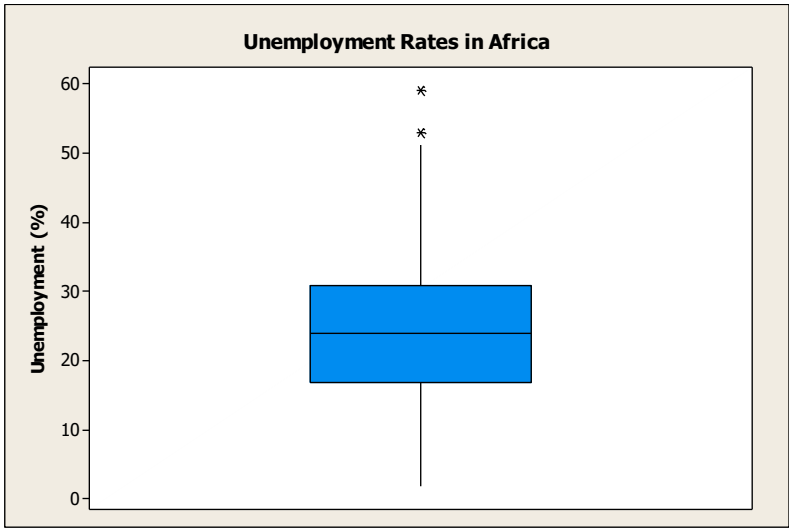
= 30.8 + 20.7 = 51.5

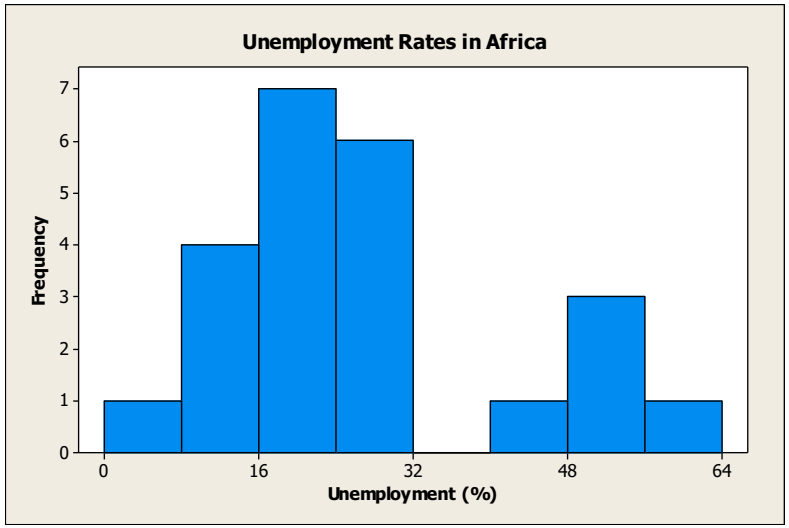


Step 4

Draw whiskers and mark outliers (**data outside fences**)

Box Plot vs Histogram





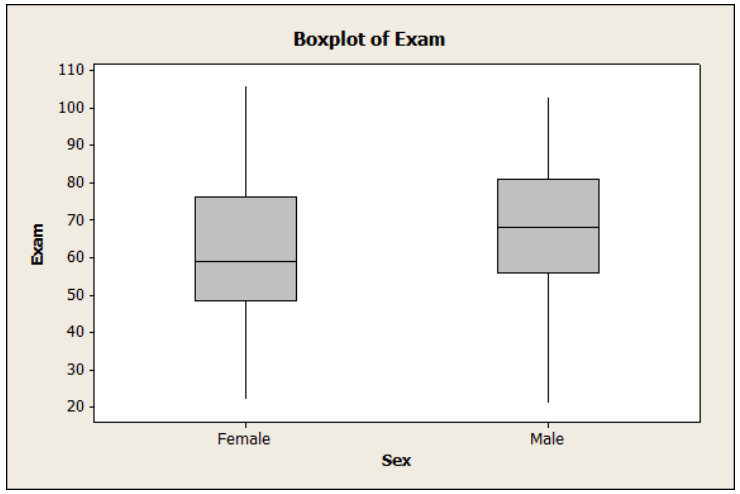
How does each graph show you the main features of the data?

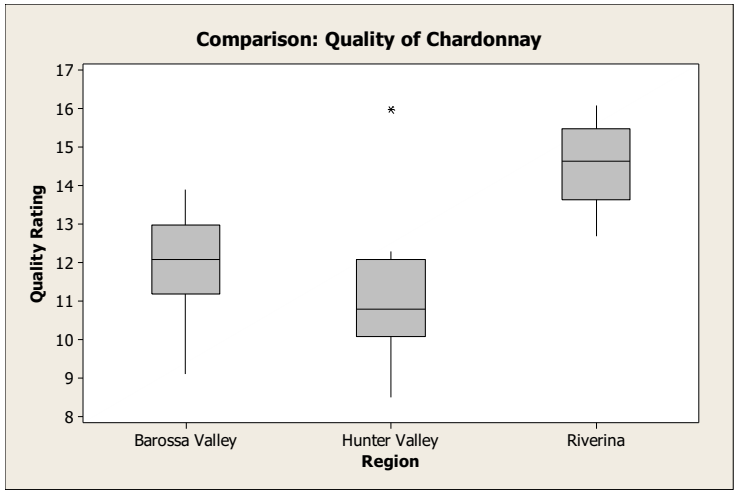
Which graph do you prefer?

Why?

Multiple/Comparative Box Plots

- Box plots are useful for graphing several samples simultaneously, facilitating *comparisons*.
- This is more effective than showing histograms, because it gives a *more concise* graph.



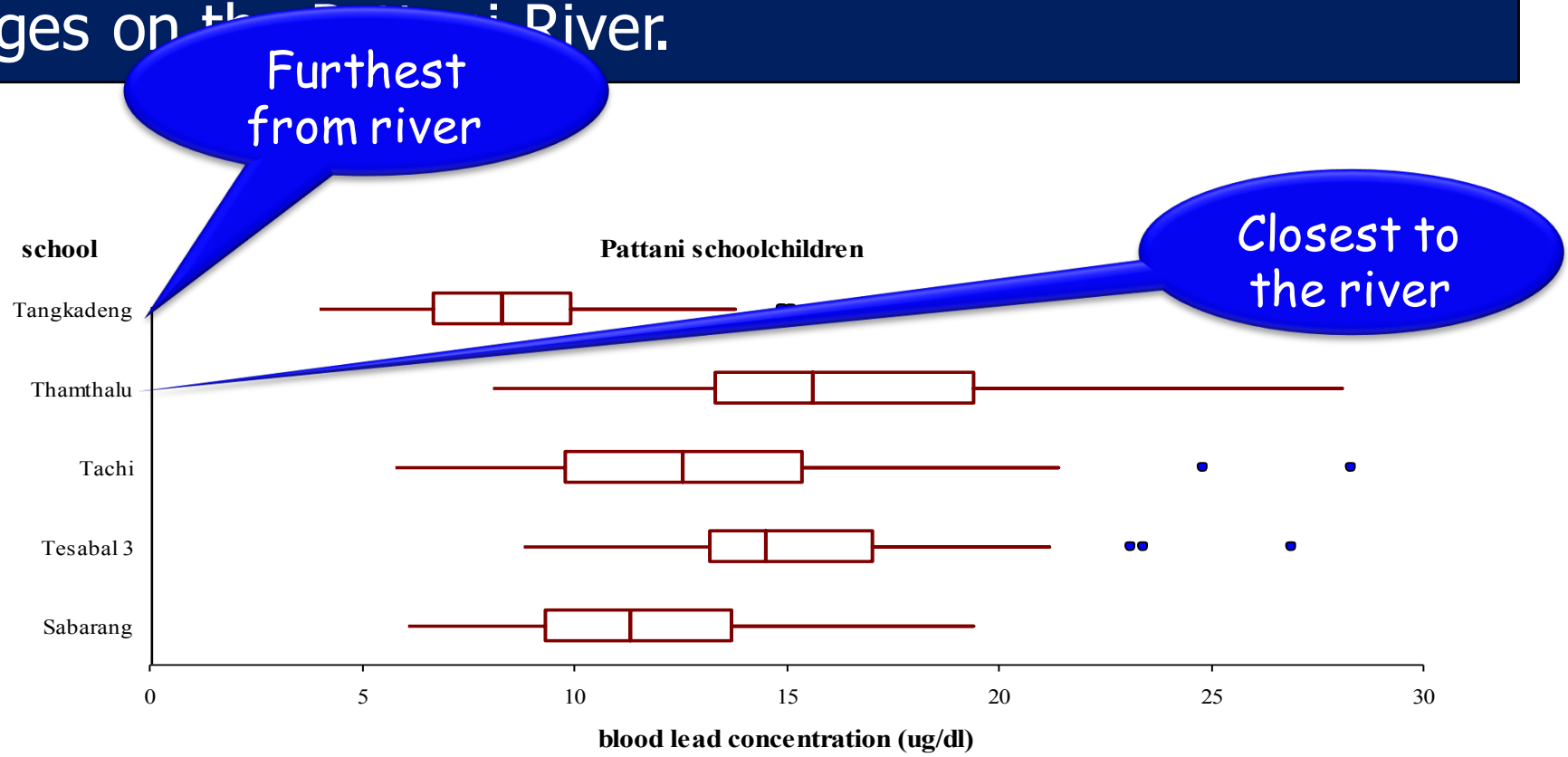


Comment: Median exam marks are slightly higher for males than females. The spread of exam marks is similar for males and females.

Comment: Typically, Chardonnay from the Riverina had the highest quality rating. The variation in ratings was similar for all three regions. One Chardonnay from the Hunter had an unusually high rating.

Blood Lead Levels: Thai School Children

Compare the blood lead concentrations (micrograms/decilitre) of children from schools in five villages on the Pattani River.



What comment would you make here?

Source: Geater, Alan and Chompikul, Jiraporn (1995)

Quiz 6

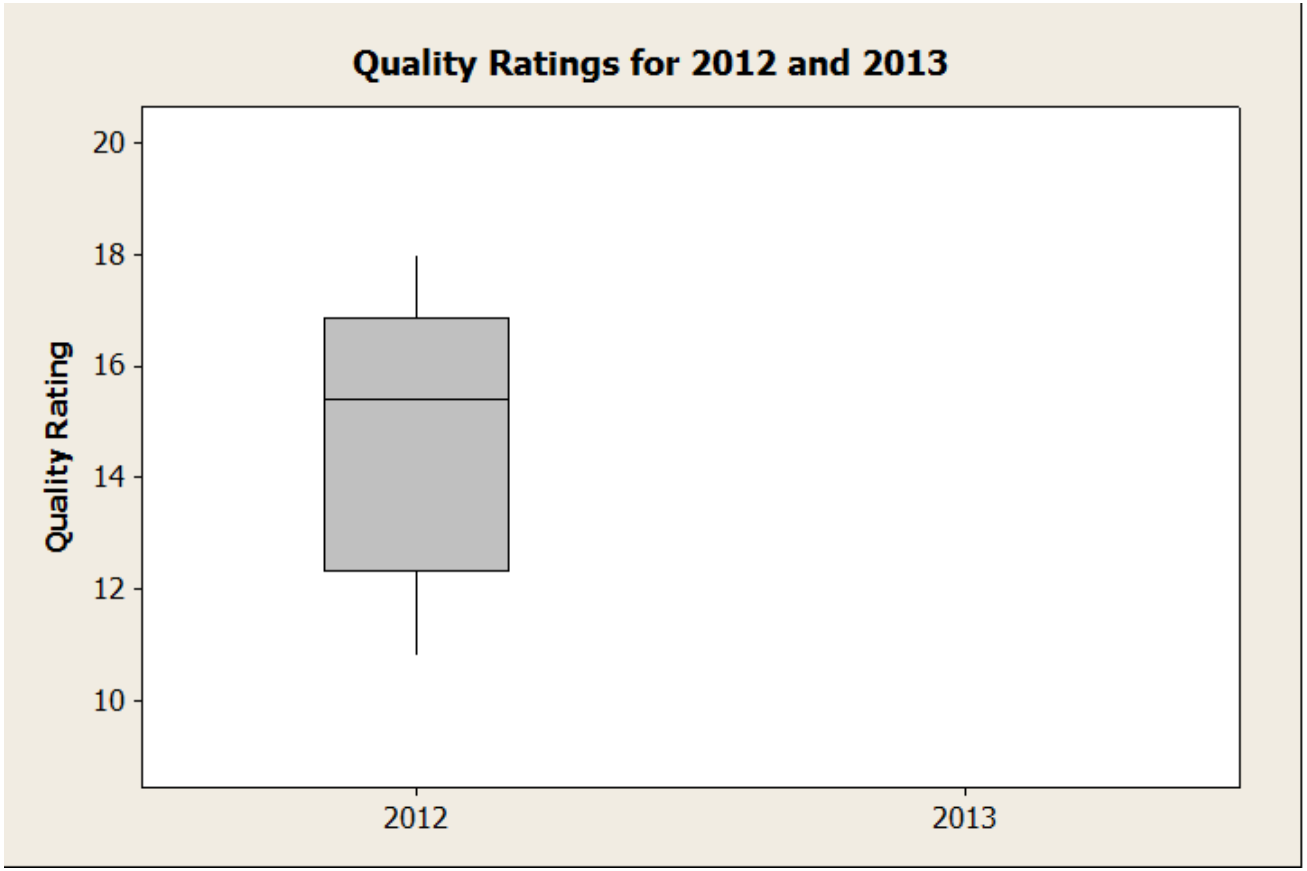
A winery in the Riverina takes part in the Good Food and Wine Show each year. The winery submits 20 bottles of Shiraz wine to be rated on quality. The scores (out of a possible 20 points) for both 2012 and 2013 are given below. The box plot on the following slide displays the scores for 2012. Use the 2013 data below to add a box plot of the scores for 2013 and comment on this display.

Quality Ratings for Shiraz									
2012									
10.8	11.2	11.4	11.9	12.0	13.3	13.4	13.4	14.8	15.0
15.8	15.9	16.2	16.4	16.4	17.0	17.3	17.4	17.7	18.8
2013									
9.1	14.0	14.8	14.9	15.2	15.8	15.9	15.9	16.0	16.2
16.3	16.4	16.9	17.0	18.2	19.1	19.5	19.7	19.9	20.0

Solution to Quiz 6

2013: 9.1 14.0 14.8 14.9 15.2 15.8 15.9 15.9 16.0 16.2
16.3 16.4 16.9 17.0 18.2 19.1 19.5 19.7 19.9 20.0

lower quartile: *median*: *upper quartile*:
iq range: *lower fence*: *upper fence*:



Comment:

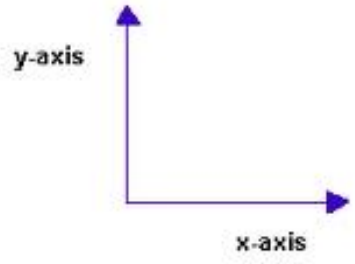
Scatter Plots

A scatter plot is used to determine whether there is a relation between **two** numerical variables. If the relation is used to determine whether one of the variables may be a predictor of the other variable, the X variable is the predictor variable and the Y variable is the response variable respectively. To construct a scatter plot:

- X is plotted on the horizontal axis and Y is plotted on the vertical axis.
- The X and Y axes must cover the ranges of the two variables
- Appropriate scales and labels are required
- One point is plotted for each observation
- Appropriate comments should be provided

Scatter Plots

A scatter plot shows the relation between two numerical variables. These two variables, X and Y, are referred to as the predictor and response variable, although they do have other names.



X

predictor

determinant

independent

Y

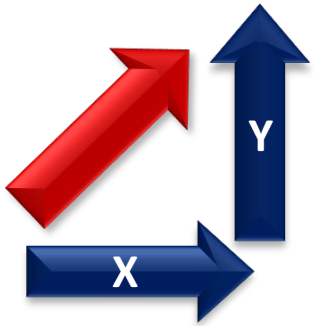
response

outcome

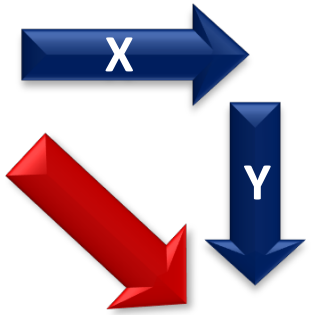
dependent

Positive or Negative Relations

If Y increases as X increases then
a POSITIVE relation exists.



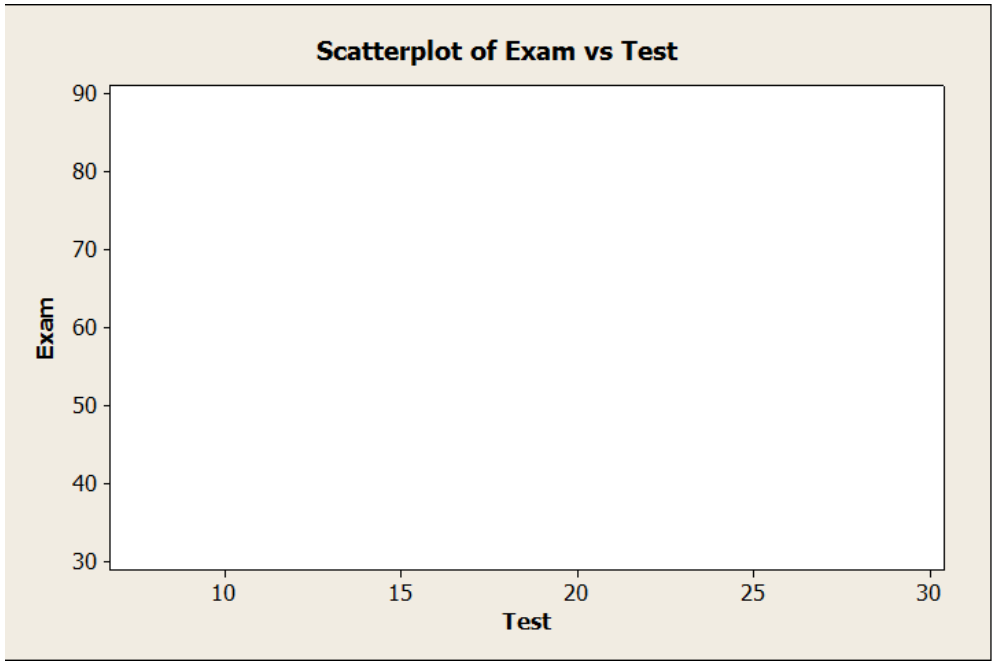
If Y decreases as X increases then
a NEGATIVE relation exists.



Constructing a Scatter Plot

Can we predict students' exam marks from their test marks?

Student	1	2	3	4	5	6	7	8	9	10	11	12	13
Test (X)	18	28	18	14	19	25	8	19	29	20	17	16	12
Exam (Y)	66	72	62	52	80	76	36	61	87	59	55	60	56

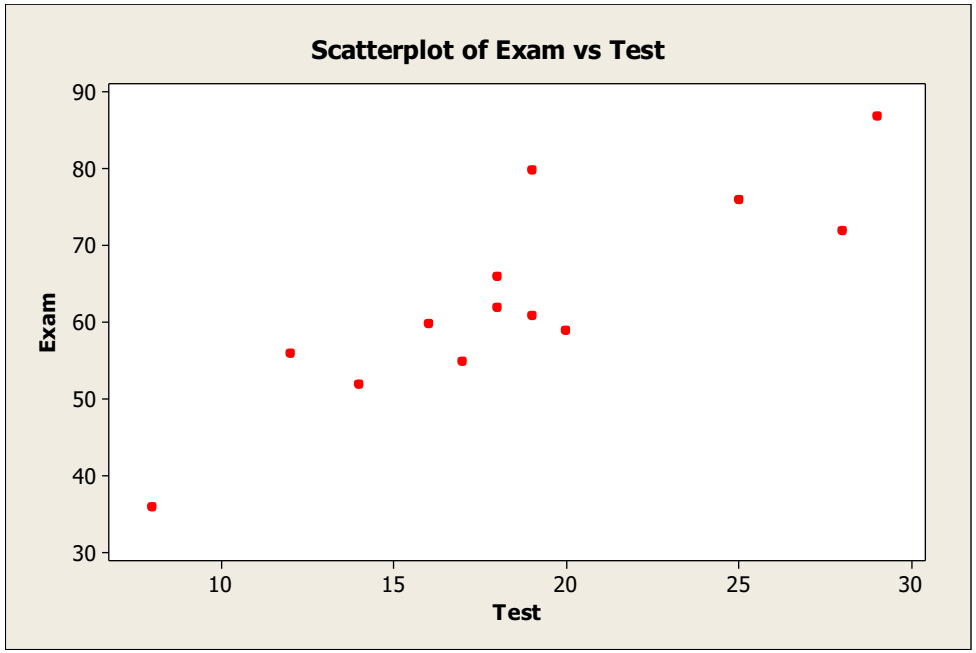


Comment:

Constructing a Scatter Plot

Can we predict students' exam marks from their test marks?

Student	1	2	3	4	5	6	7	8	9	10	11	12	13
Test (X)	18	28	18	14	19	25	8	19	29	20	17	16	12
Exam (Y)	66	72	62	52	80	76	36	61	87	59	55	60	56



Comment: There appears to be a fairly strong positive relation between test marks and exam marks. Students with higher test marks generally tend to get higher exam marks. Test marks may be a fairly good predictor of exam marks.

Displaying Data: Graphical Displays/Graphical Summaries

DATA	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Displaying Data: Graphical Displays/Graphical Summaries

DATA	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Homework problems

Homework Question 1

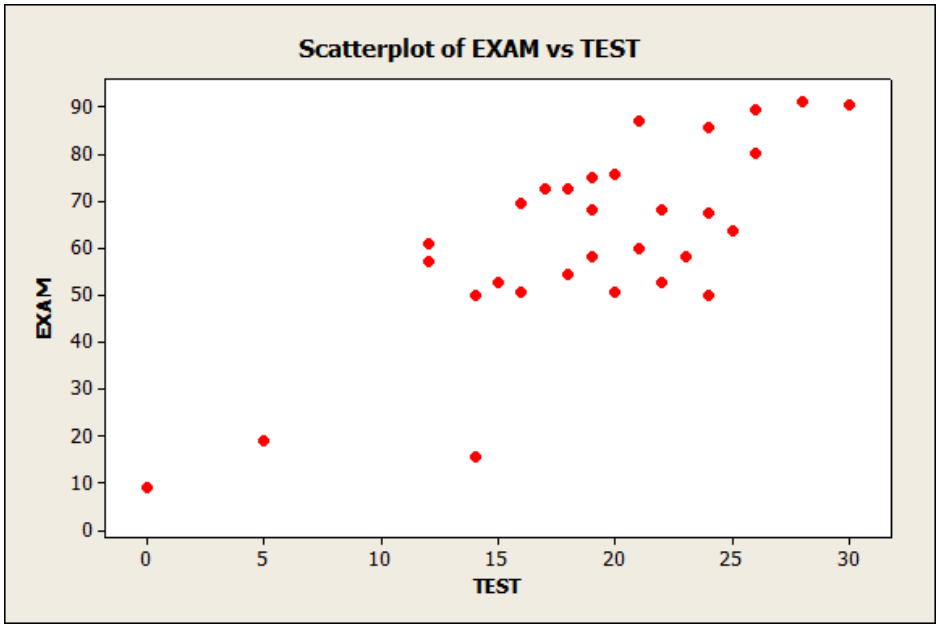
Classify the following variables, which recorded information on a sample of patients at a veterinary clinic. The patients were all female dogs which had given birth to puppies at the clinic in the past year.

a. Breed
b. Vaccinated (vaccinations up to date: yes/no)
c. Age
d. Weight
e. General health (poor, average, good, excellent)
f. Puppies (number of puppies in litter)

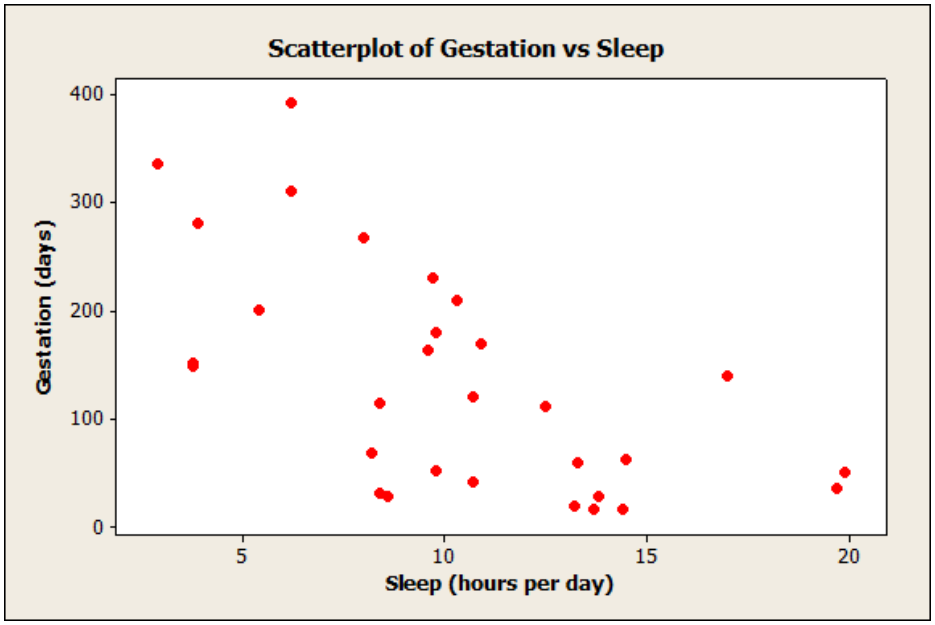
Homework Question 2

Comment on the following two scatter plots.

a. Students' Exam Marks vs Test Marks



b. Mammals' Gestation Periods vs Sleep



Homework Question 3

Find the median and the quartiles of the following small samples.

i. 8 9 9 12 15

ii. 4 7 10 10 14 18

iii. 1 2 2 4 4 6 7 8 9 9 12 18

Lecture 1 Summary

- *Studies* are needed to resolve questions of interest.
- *Statistics* involves determining *population* characteristics, using data from *samples*.
- Samples should be *unbiased* (they should represent the target population).
- Variables may be *continuous*, *discrete*, *nominal* or *ordinal*.

Lecture 1 Summary

- *Bar charts* and *pie charts* display a categorical variable.
- A *histogram* displays a numerical variable and shows the shape of the data.
- The *median* and the *quartiles* divide a set of data into quarters.
- *Box plots* are useful for graphing and comparing sets of numerical data from two or more categorical groups.
- *Scatter plots* show the relation between two numerical variables.

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 1: Pages 2 to 25 (bits, some parts will be covered in lectures 3-6)
- Chapter 2: Pages 28 – 50