JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Exploratory Data Analysis

Marie Diener-West, PhD
Johns Hopkins University

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section A

Exploratory Data Analysis

Data and Variables

- **Variable**
  - A characteristic taking on different values
- **Random variable**
  - A variable taking on different possible values as a result of chance factors

Types of Variables

- **Quantitative** or numerical
  - Implies amount or quantity
- **Qualitative** or categorical
  - Implies attribute or quality

Types of Random Variables

- **Discrete**
  - Random variable with values that comprise a countable set
  - There can be gaps in its possible values
- **Continuous**
  - Random variable with values comprising an interval of real numbers
  - There are no gaps in its possible values

Measurement Scales—Quantitative Variables

- **Counts**
  - Numbers represented by whole numbers
    - For example, number of births, number of relapses
- **Interval**
  - The same distances or intervals between values are equal
    - For example, temperature, altitude
- **Ratio**
  - The same ratios of values are equal
    - For example, weight, height, time, hospital length of stay
  - A true zero point indicates the absence of the quantity being measured

Measurement Scales—Qualitative Variables

- **Nominal**
  - Classifications based on names
    - Binary or dichotomous
      - For example, gender, alive or dead
    - Polychotomous or polytomous
      - For example, marital status, ethnicity
- **Ordinal**
  - Classifications based on an ordering or ranking
    - For example, ratings, preferences

Quick Check

- What type of variable is disease status?
- What type of variable is blood pressure?

Review

- Variables may be **quantitative** (numerical) or **qualitative** (categorical)
- Variables may be **discrete** (have gaps) or **continuous** (have no gaps)
- Variables are measured on different **measurement scales:**
  - Counts
  - Interval scale
  - Ratio scale
  - Nominal scale
  - Ordinal scale

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section B

Organizing, Grouping, and Summarizing Data

Methods for Organizing Data

- Ordering data
  - Tallies
  - Stem and leaf displays
- Grouping data
  - Frequency distributions
- Summarizing data
  - Measures of central tendency
  - Measures of dispersion
  - Box-and-whiskers plots
- Displaying data
  - Tables
  - Histograms, bar diagrams
  - Pie charts
  - Scatterplots
  - Graphs

Ordering Data

- Example: ages of graduate students (n=10)
- Suppose the unordered data were:
  - **35, 40, 52, 27, 31, 42, 43, 28, 50, 35**
- Data could be ordered by hand:
  - **27, 28, 31, 35, 35, 40, 42, 43, 50, 52**
- Ordering data by hand can be tedious, especially when there is a large number of observations
- Alternatives to this method are:
  - Tallies
  - Stem and leaf displays

Tallies of Data

- Advantage
  - Provide information regarding the frequency of observations in groups or categories
- Disadvantage
  - The actual values of observations within groups are not retained

Age Group	Observations
20–29	//
30–39	///
40–49	///
50–59	//

Stem-and-Leaf Displays

- Each 10-year age group is considered a **stem**
- An individual age is denoted by a **leaf**
- Observations are assigned to an age group (stem)
- Individual observations (**leaves**) are ordered within a stem

Ordering Data with Stem-and-Leaf Displays

- If you have a set of observations, there are a number of ways to order those observations
- Example: Ages of Graduate Certificate Students
  - 35, 40, 52, 27, 31, 42, 43, 28, 50, 35
- You could order the observations by hand
- Alternatively, you could use a stem and leaf display to record and order your observations

Age Group	Observations
20-29	
30-39	
40-49	
50-59	

Ordering Data with Stem-and-Leaf Displays

- To create an unordered stem and leaf display, take each observation and place the last digit in the appropriate row on the display
- i.e. the 8 in 28 goes in the 20-29 group
- To order the observations in the stem and leaf display, all you need to do is sort the numbers in each row

35, 40, 52, 27, 31, 42, 43, 28, 50, 35

Age Group	Observations
20-29	7 8
30-39	1 5 5
40-49	0 2 3
50-59	0 2

Stem-and-Leaf Displays

- Turned on its side, the stem and leaf display forms a histogram

Age Group	Observations
20–29	78
30–39	515
40–49	023
50–59	20

Stem-and-Leaf Displays

- The ordered stem and leaf display show ages from youngest to oldest

Age Group	Ordered Observations
20–29	78
30–39	155
40–49	023
50–59	02

Stem-and-Leaf Displays

- Aid in sorting or ordering data
- Retain more information than a tally
- Use logic to determine the number of stems
- Rough guideline for the number of stems is:

$$\sqrt{2\text{Number of datapoints}}$$

Stem-and-Leaf Displays

- The previous example also could be shown as:

2	78
3	155
4	023
5	02

or as

2*	78
3*	155
4*	023
5*	02

Where 2* = 20–29
Where 3* = 30–39

Grouping Data: Frequency Distribution Table

Age Interval	Frequency
20–29	2
30–39	3
40–49	3
50–59	2
Total	10

Grouping Data: Some Definitions

- **Frequency**
  - Count or number of observations within an interval or group
- **Cumulative frequency**
  - Count within the current interval and all preceding intervals
- **Relative frequency**
  - Count within an interval divided by the total number of observations
- **Cumulative relative frequency**
  - Count within the current interval and all preceding intervals divided by the total number of observations

Grouping Data: Example

Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
20–29	2	2	.2	.2
30–39	3	5	.3	.5
40–49	3	8	.3	.8
50–59	2	10	.2	1.0
Total	10		1.0	

Summarizing Central Tendency of Data

- Measures of central tendency or location
- Mean (average) =

$$\frac{\sum x_i}{n} = \bar{x}$$

- Median = middle observation
- Mode = most frequent observation
- Percentiles, quartiles

Summarizing Dispersion or Variability

- **Range**
  - Difference between largest and smallest values
- **Variance (s^2)**
  - Dispersion measured relative to the scatter of the values about their mean
- **Standard deviation (s)**
  - Square root of the variance

Sample Variance Formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Calculating Summary Measures for the Example

- Graduate student ages: 27, 28, 31, 35, 35, 40, 42, 43, 50, 52
  - **Mean =**

$$\frac{\sum_{i=1}^{10} x_i}{10} = \frac{383}{10} = \bar{x} = 38.3 \quad \text{years}$$

  - **Mode =** 35 years
  - **Median =** (35 + 40) / 2 = 37.5 years
    - The average of the two middle observations
  - **Range =** 52 – 27 = 25 years

Example: Summary Measures

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^{10}(x_i - 38.3)^2}{10-1}$$

$$s^2 = \frac{(27-38.3)^2 + (28-38.3)^2 + ... + (52-38.3)^2}{10-1}$$

$$s^2 = 74.7$$

Summarizing Variability in Data

- Sample variance = 74.7 years^2
- Standard deviation
  = $\sqrt{}$ sample variance = s = 8.6 years

Percentiles

- The **p^{th} percentile P** is the value that is greater than or equal to p percent of the observations
- Common percentiles are
  - 25th
  - 50th
  - 75th

Percentile Formulas (Exact Formulas)

Percentile	Quartile	Formula
P_{25}	Q_1	(n+1) / 4th observation
P_{50}	Q_2	(n+1) / 2nd observation
P_{75}	Q_3	3(n+1) / 4th observation

Example: Using Exact Formulas for Percentiles

- P_{25} = Q1
  = $[(10 + 1)/4]^{th}$ observation
  = $[2.75]^{th}$ observation
  = 0.25(28) + 0.75(31)
  = 30.25 (or 31 if rounded to the 3rd observation)

- P_{50} = Q2
  = $[(10 + 1)/2]^{th}$ observation
  = $[5.5]^{th}$ observation
  = 0.5(35) + 0.5(40)
  = 37.5

- P_{75} = Q3
  = $3(10 + 1)/4^{th}$ observation
  = 8.25^{th} observation
  = 0.75(43) + 0.25(50)
  = 44.75 (or 43 if rounded to the 8th observation)

Easier Method for Calculating Percentiles

- $P_{50} = Q_2$ = middle observation
- $P_{25} = Q_1$ = middle observation of the lower half of observations
- $P_{75} = Q_3$ = middle observation of the upper half of observations

Easier Method for Calculating Percentiles

When the number of observations is even:

$P_{50} = Q_2$ = average of the middle two observations

$P_{25} = Q_1$ = middle observation of the lower half of n/2 observations

$P_{75} = Q_3$ = middle observation of the upper half of n/2 observations

Easier Method for Calculating Percentiles

When the number of observations is odd:

$P_{50} = Q_2$ = the middle observation

$P_{25} = Q_1$ = middle observation of the lower half of the observations (includes Q_2)

$P_{75} = Q_3$ = middle observation of the upper half of the observations (includes Q_2)

Calculating Percentiles for the Example

Graduate student ages: 27, 28, 31, 35, 35, 40, 42, 43, 50, 52

- $P_{50} = Q_2$ = average of the middle two observations = (35+40)/2 = 37.5 years
- $P_{25} = Q_1$ = middle observation of the lower 5 observations = 31 years
- $P_{75} = Q_3$ = middle observation of the upper 5 observations = 43 years

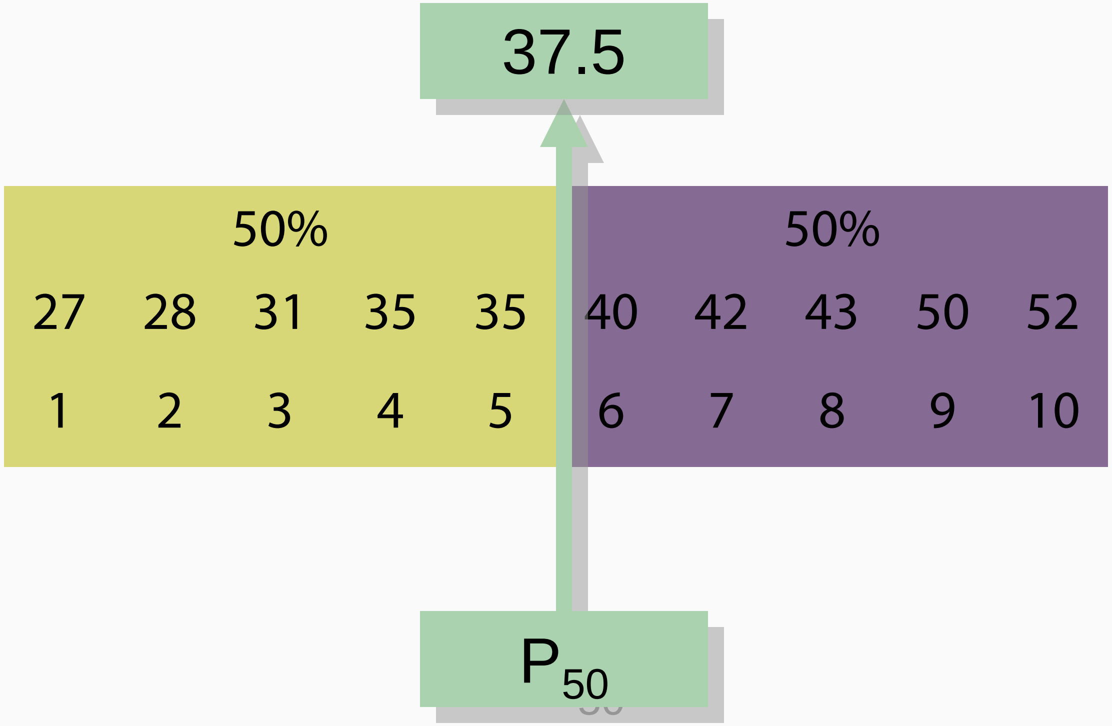
Understanding Percentiles

- Here is the data set of ages ordered from youngest to oldest:

27	28	31	35	35	40	42	43	50	52
1	2	3	4	5	6	7	8	9	10

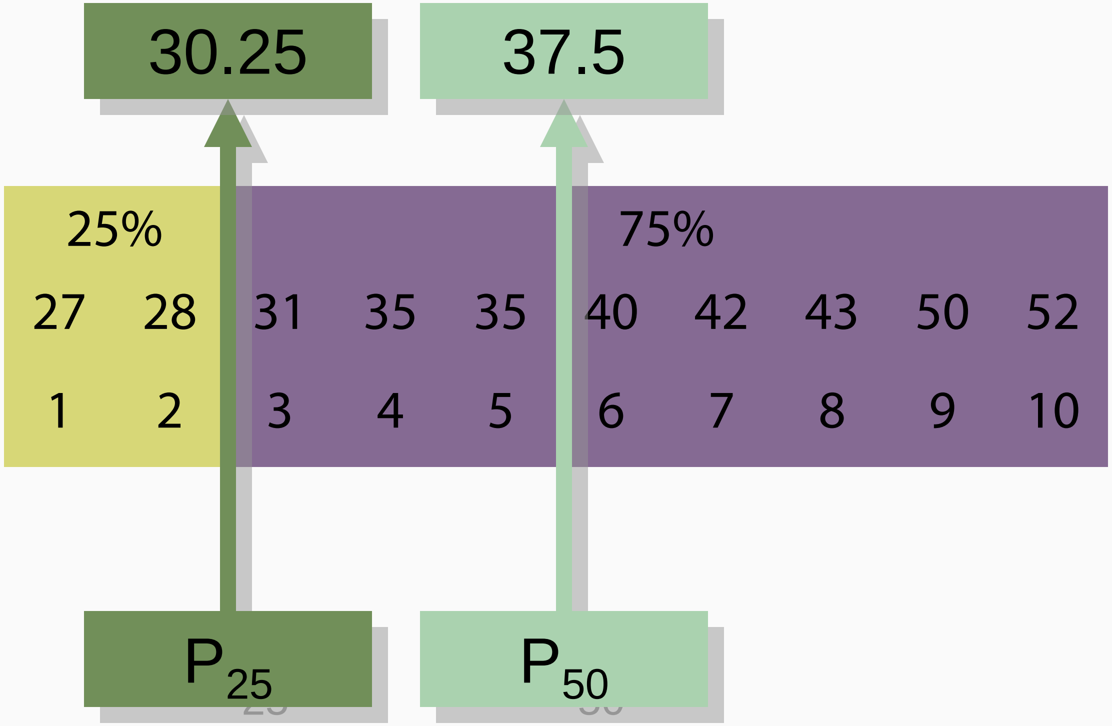
Understanding Percentiles

- The Median (P_{50}) is the value that separates the lower 50% from the upper 50% of the observations.



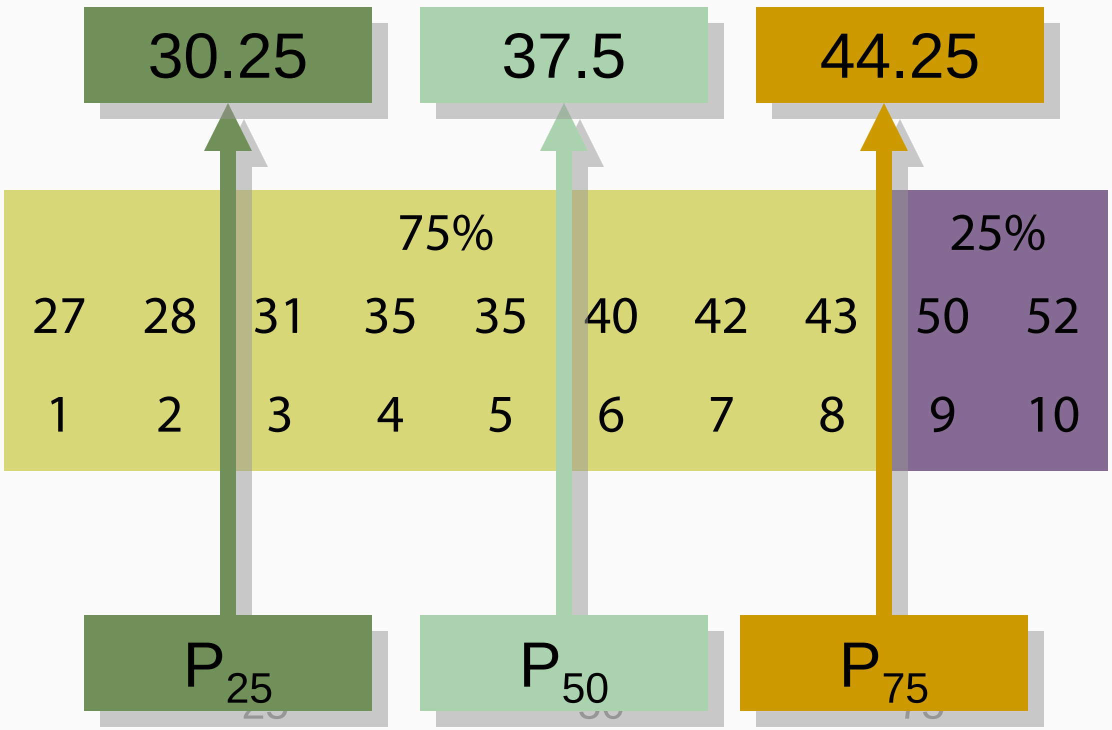
Understanding Percentiles

- The 25th percentile (P_{25}) is the value that separates the lower 25% from the upper 75% of the observations.



Understanding Percentiles

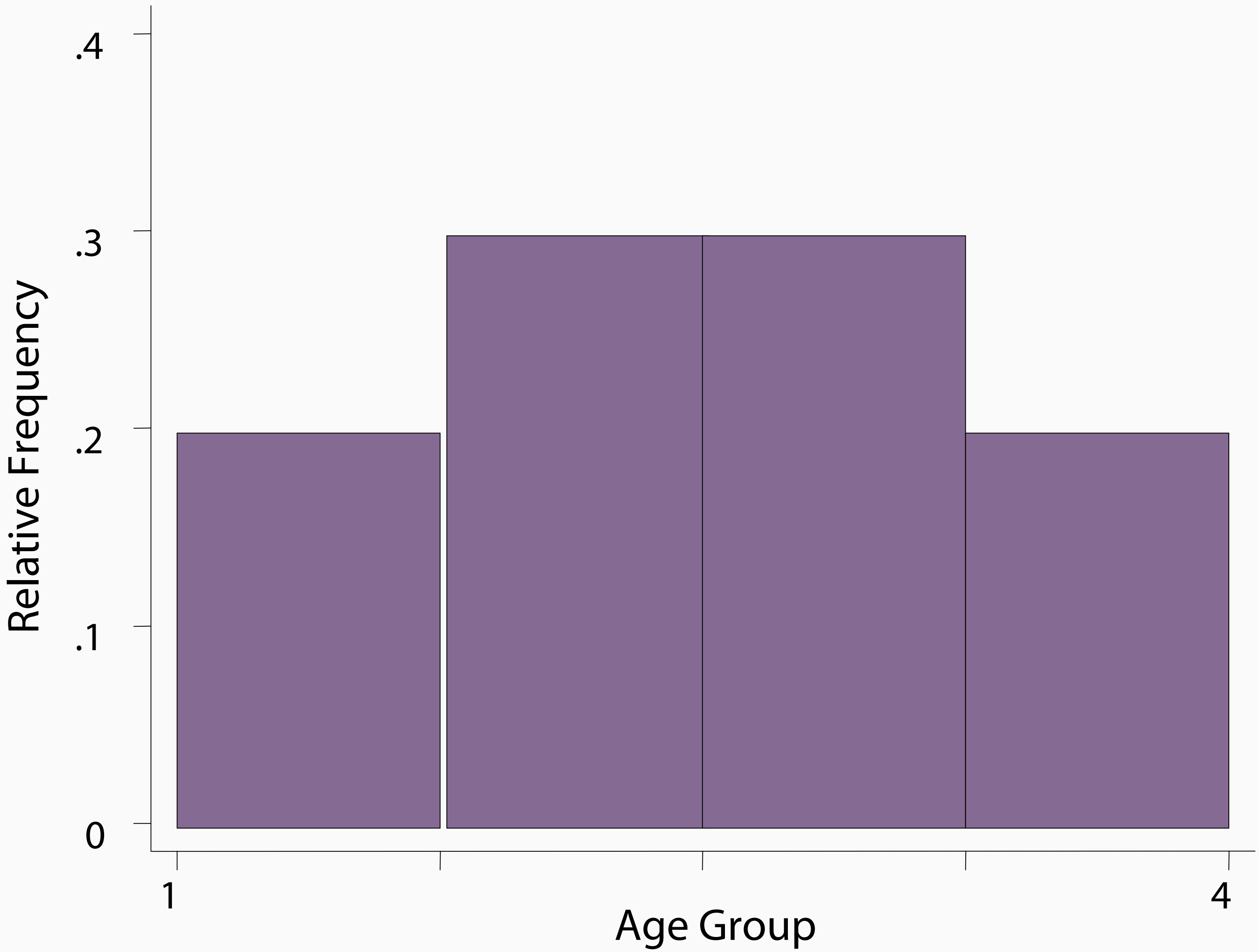
- The 75th percentile (P_{75}) is the value that separates the lower 75% from the upper 25% of the observations.



Example: Descriptive Statistics

- Minimum age is 27; maximum age is 52; range of ages is 25 years
- Mean age is 38.3 years; median is 37.5 years; mode is 35 years
- 50% of ages are greater than 37.5 years; 25% of ages are less than 31 years; 25% of ages are greater than 43 years
- These examples may be shown in a graph
  - Histogram
  - Frequency polygon
  - Cumulative relative frequency plot
  - Pie chart

Histogram of Graduate Student Ages



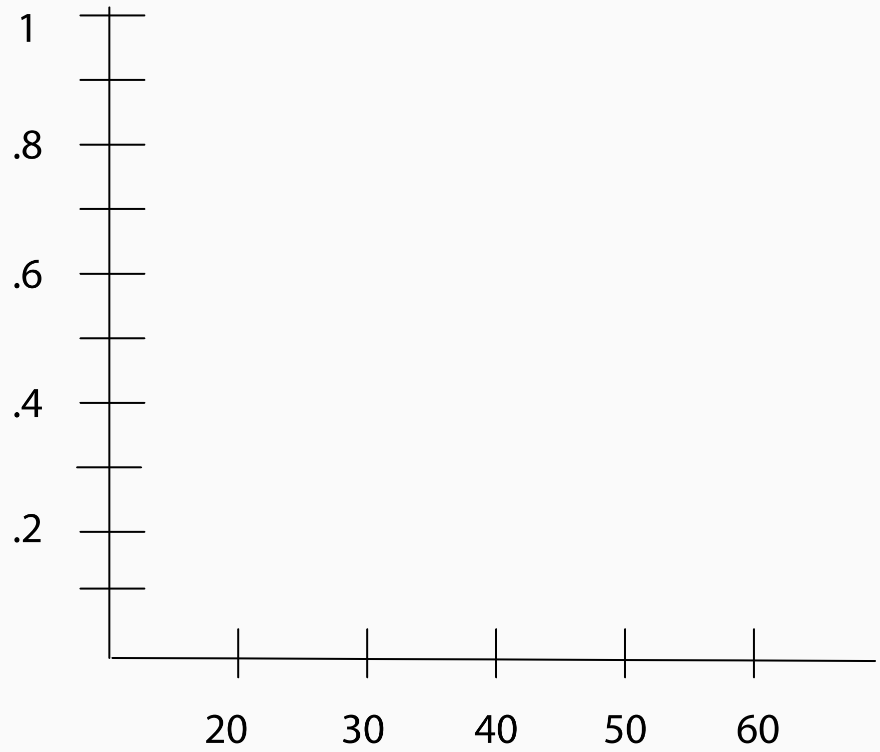
Constructing a Histogram

- The age distribution in the following table can easily be graphed as a historgram

Age Interval	Frequency	Relative Frequency
20-29	2	0.2
30-39	3	0.3
40-49	3	0.3
50-59	2	0.2
	10	1.0

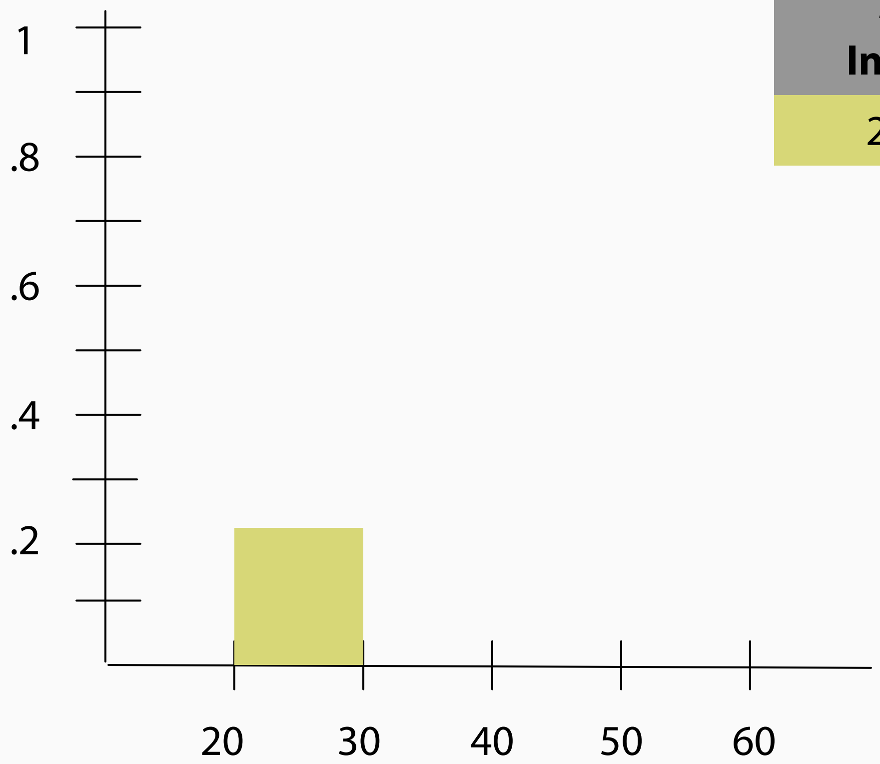
Constructing a Histogram

- The x-axis will represent the age in years ranging from 20 to 60.
- The y-axis will represent the relative frequency on percentage in each age interval ranging from 0 to 1 (0-100%)



Constructing a Histogram

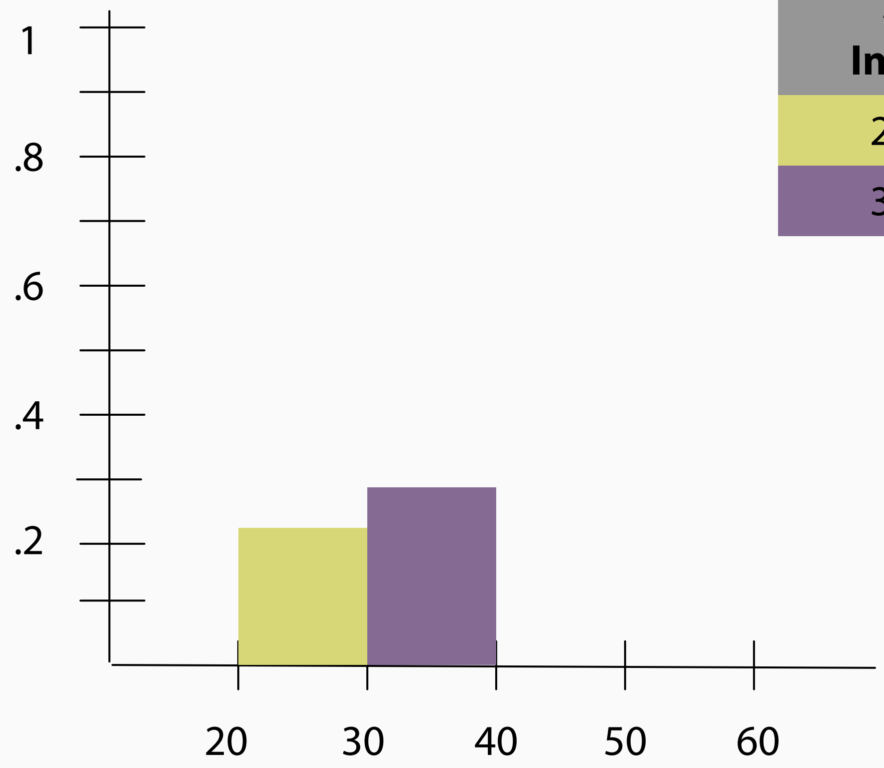
- A bar is drawn on the graph to represent the relative frequency of each age interval



Age Interval	Relative Frequency
20-29	0.2

Constructing a Histogram

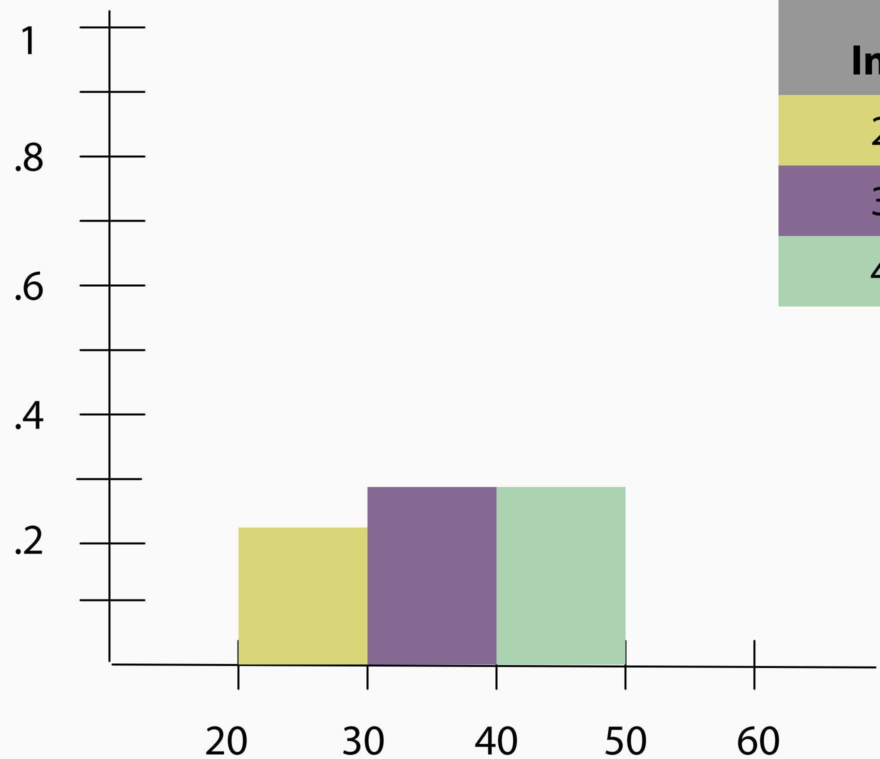
- A bar is drawn on the graph to represent the relative frequency of each age interval



Age Interval	Relative Frequency
20-29	0.2
30-39	0.3

Constructing a Histogram

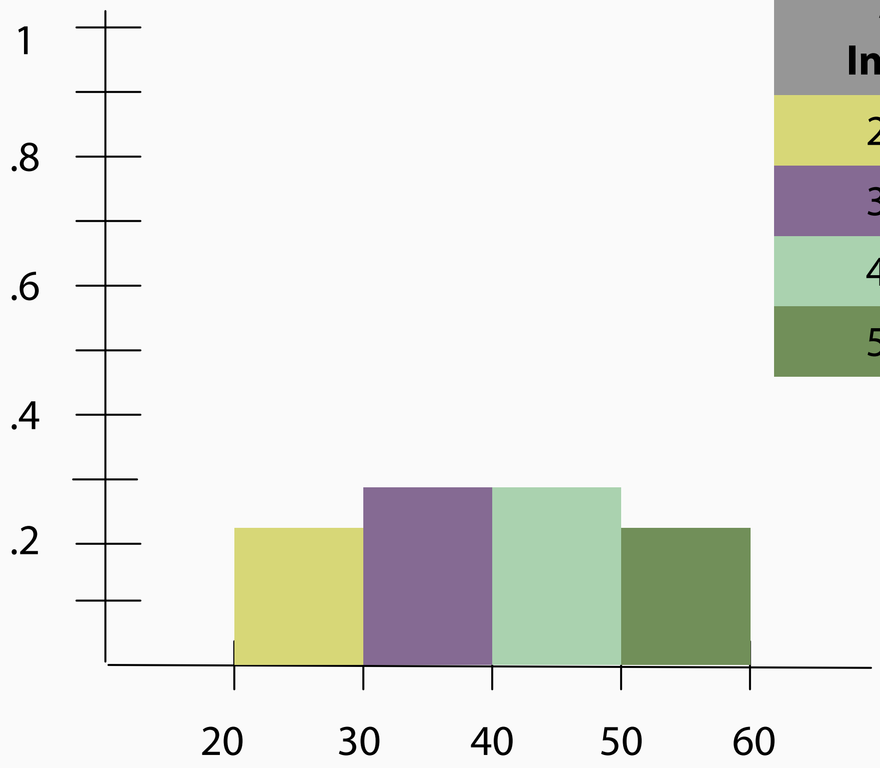
- A bar is drawn on the graph to represent the relative frequency of each age interval



Age Interval	Relative Frequency
20-29	0.2
30-39	0.3
40-49	0.3

Constructing a Histogram

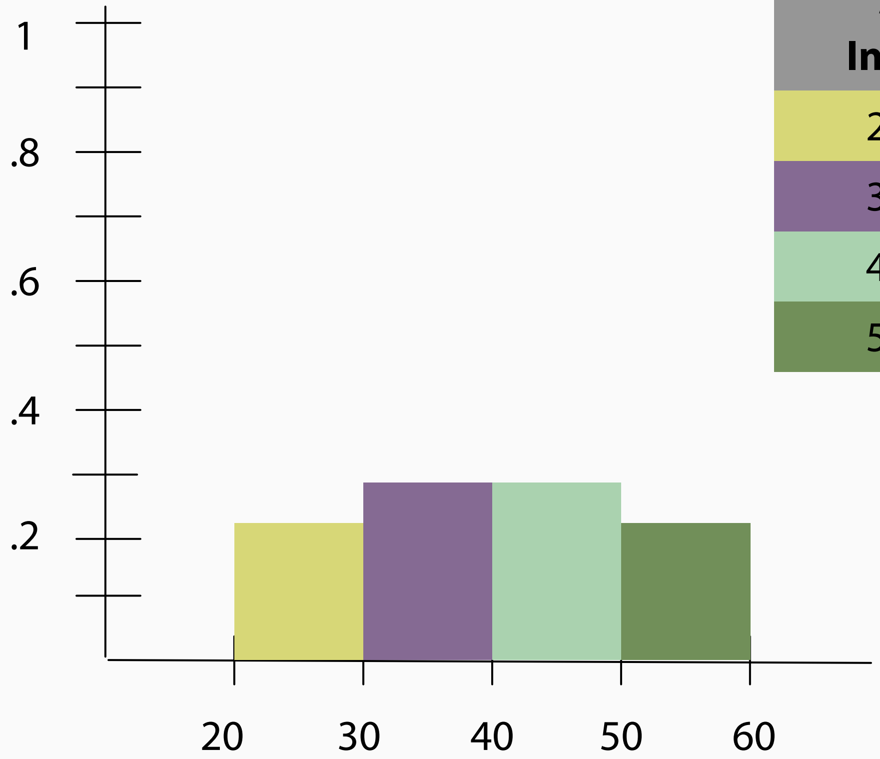
- A bar is drawn on the graph to represent the relative frequency of each age interval



Age Interval	Relative Frequency
20-29	0.2
30-39	0.3
40-49	0.3
50-59	0.2

Constructing a Histogram

- The sum of the relative frequencies in the completed histogram equals 1



Age Interval	Relative Frequency
20-29	0.2
30-39	0.3
40-49	0.3
50-59	0.2

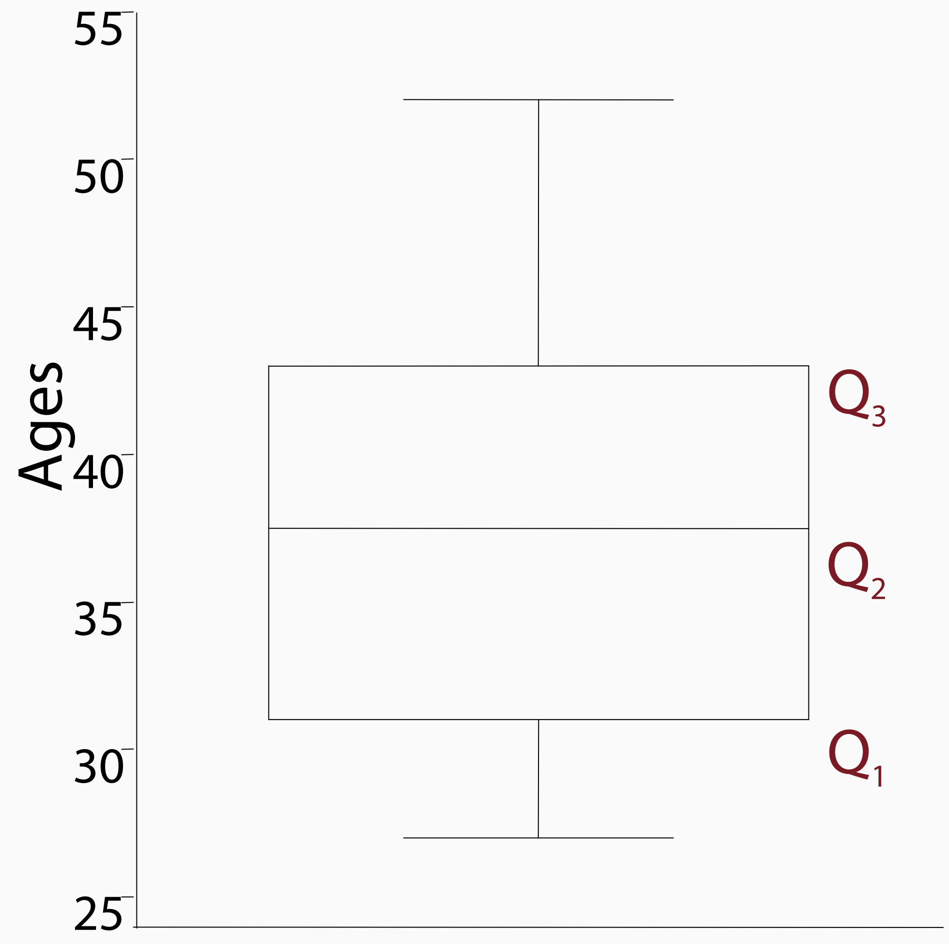
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Section C

Box-and-Whiskers Plots

Box-and-Whiskers Plot: Terminology

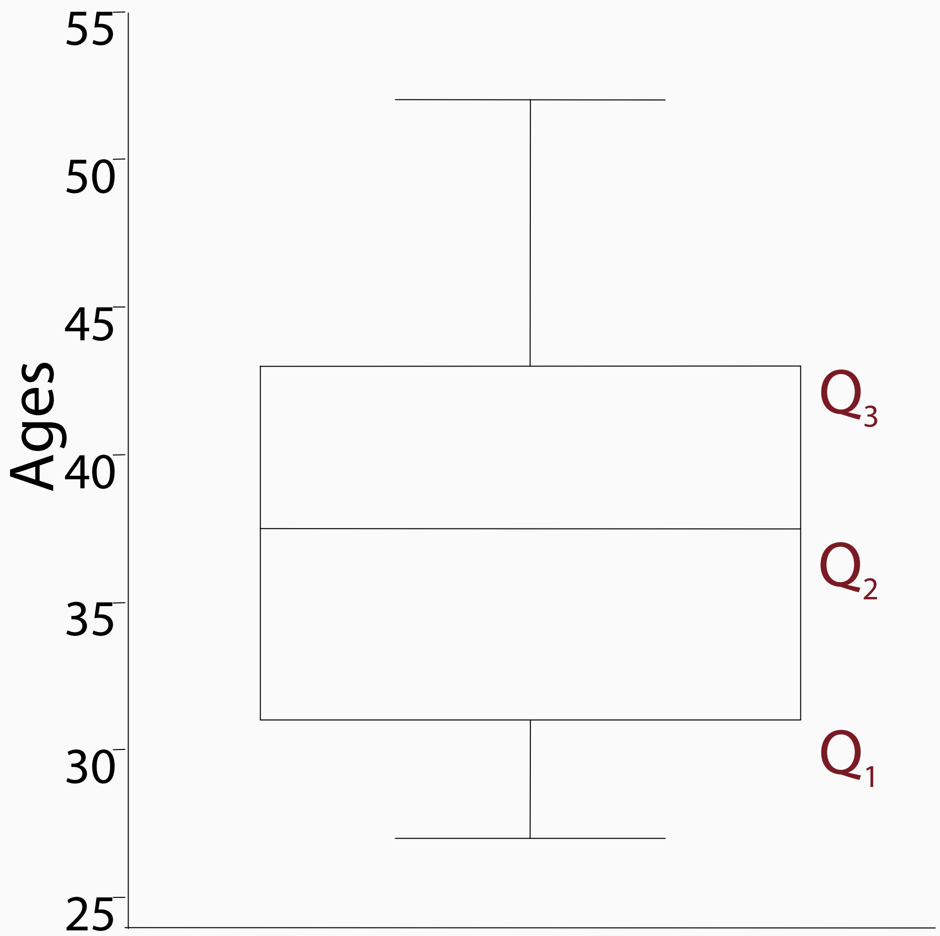
- A **box-and-whiskers plot** is a graphical display using quartiles
- Upper hinge = Q_3
- Median = Q_2
- Lower hinge = Q_1
- **H-spread** = **interquartile range** = $Q_3 - Q_1$
  - Contains 50% of the observations



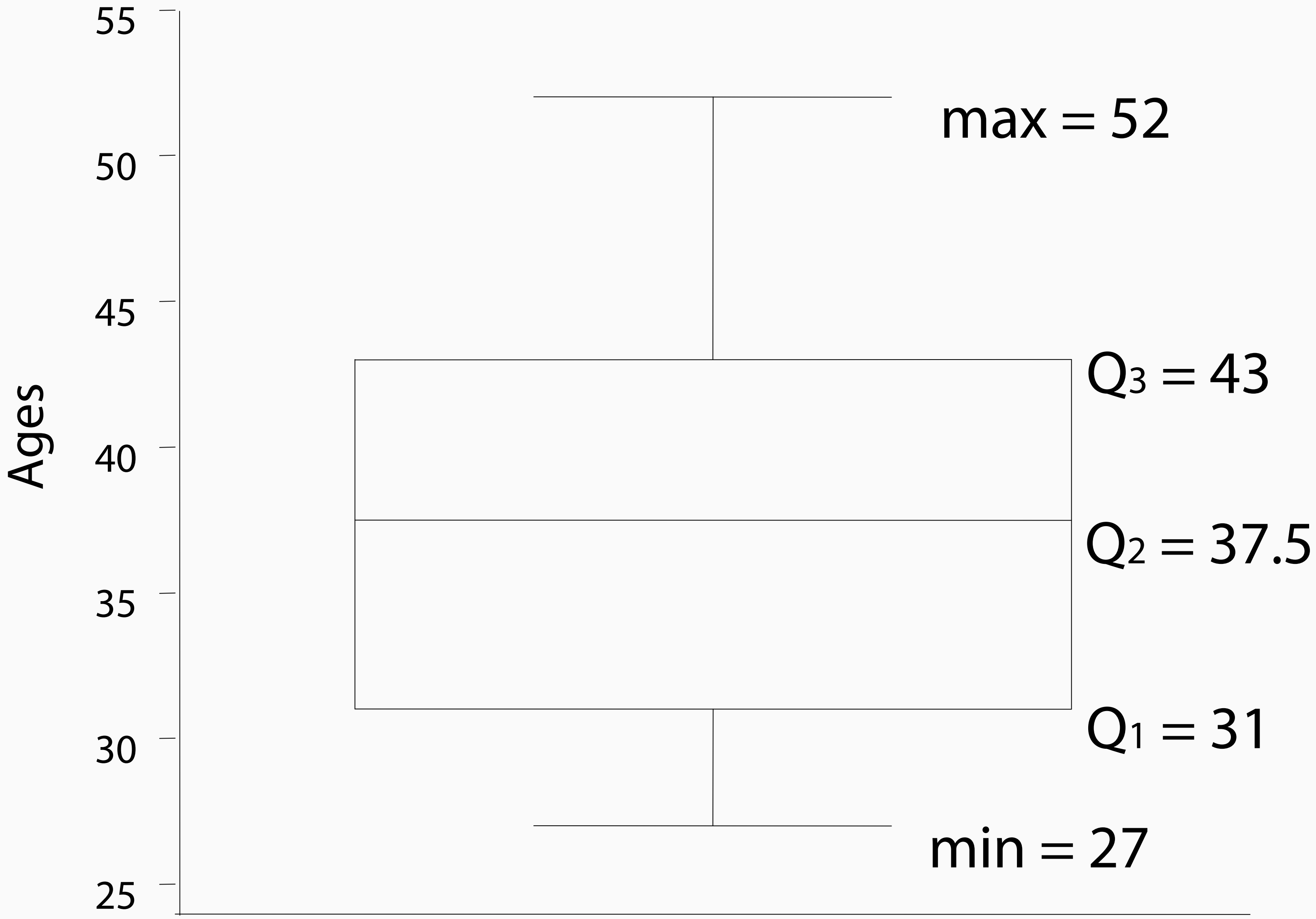
Box-and-Whiskers Plot: Terminology

- Upper fence = upper hinge +(1.5 X H-spread)
- Lower fence = Lower hinge –(1.5 X H-spread)

- The **hinges** of the box are the first and third quartiles
- The **median** (second quartile) is represented by a line drawn within the box



Box-and-Whiskers Plot of Student Ages



Box-and-Whiskers Plot: Some More Terminology

- **Whiskers** are lines drawn to the smallest and largest observations within the calculated fences

- **Outliers** are data values that lie beyond the calculated fences (high or low)

Calculated Fences in Box-and-Whiskers Plots

- The fences **are not observed values** in the data set
- The fences are calculated as **guidelines** for inspecting values which appear to be different from the majority of the observations
- Outliers require checking/validation but may be real

Constructing a Box Plot using the Example

- Minimum age = 27 years
- Maximum age = 52 years
- Median (Q_2) = 37.5 years
- Upper hinge (Q_3) = 43 years
- Lower hinge (Q_1) = 31 years

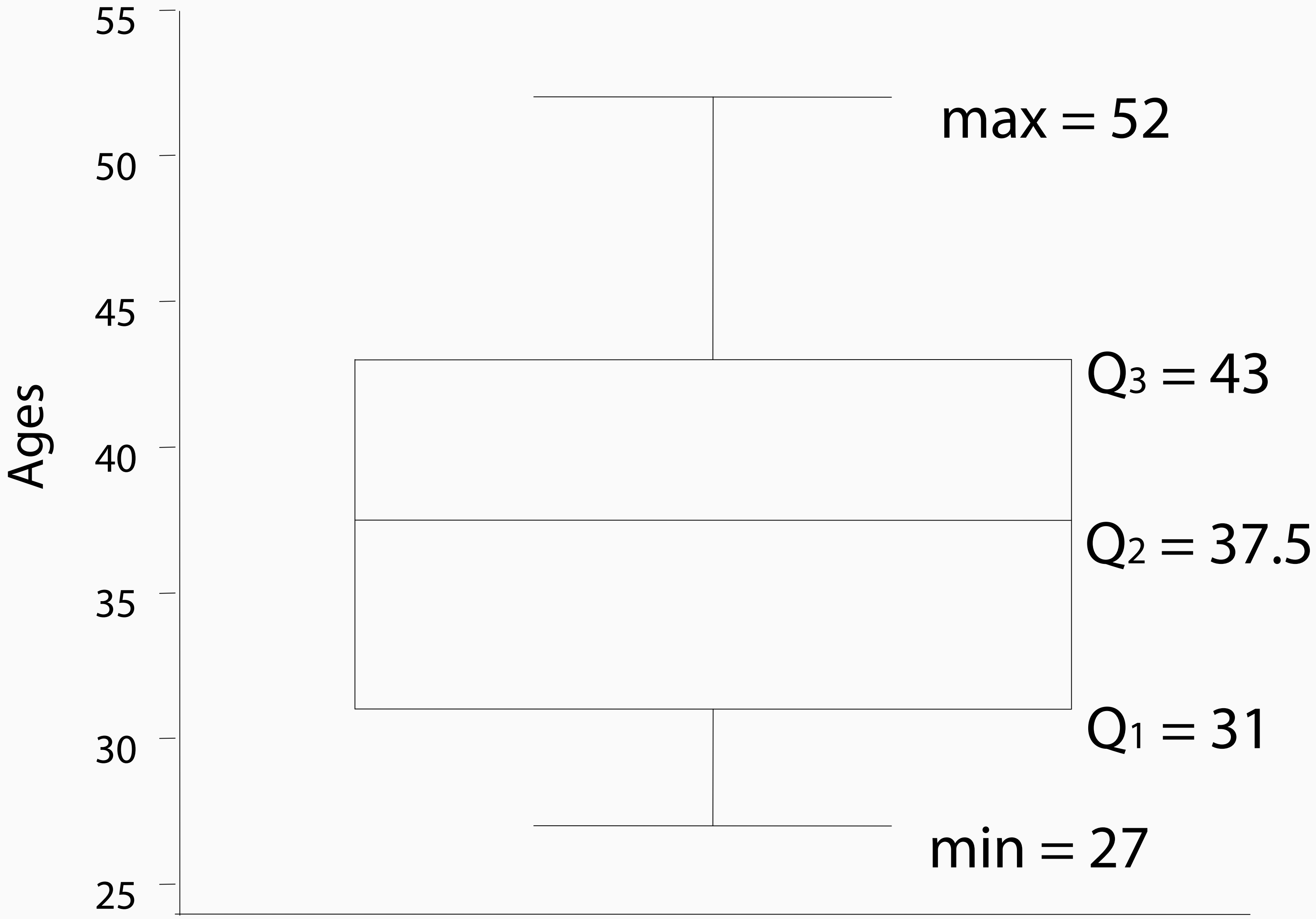
- H-spread = $Q_3 – Q_1$
  = interquartile range = 43 – 31 = 12 years
- Upper fence = Q_3 + 1.5 X (H-spread)
  = 43 + 1.5 X (12) = 61
- Lower fence = Q_1 – 1.5 X (H-spread)
  = 31 – 1.5 X (12) = 13

Example of Graduate Student Ages

- The following slide shows the box plot and associated summary values from a STATA output

- There are **no outliers** based on the calculated fences; the **whiskers** are drawn to the largest value of 52 and to the smallest value of 27 years

Box-and-Whiskers Plot of Student Ages



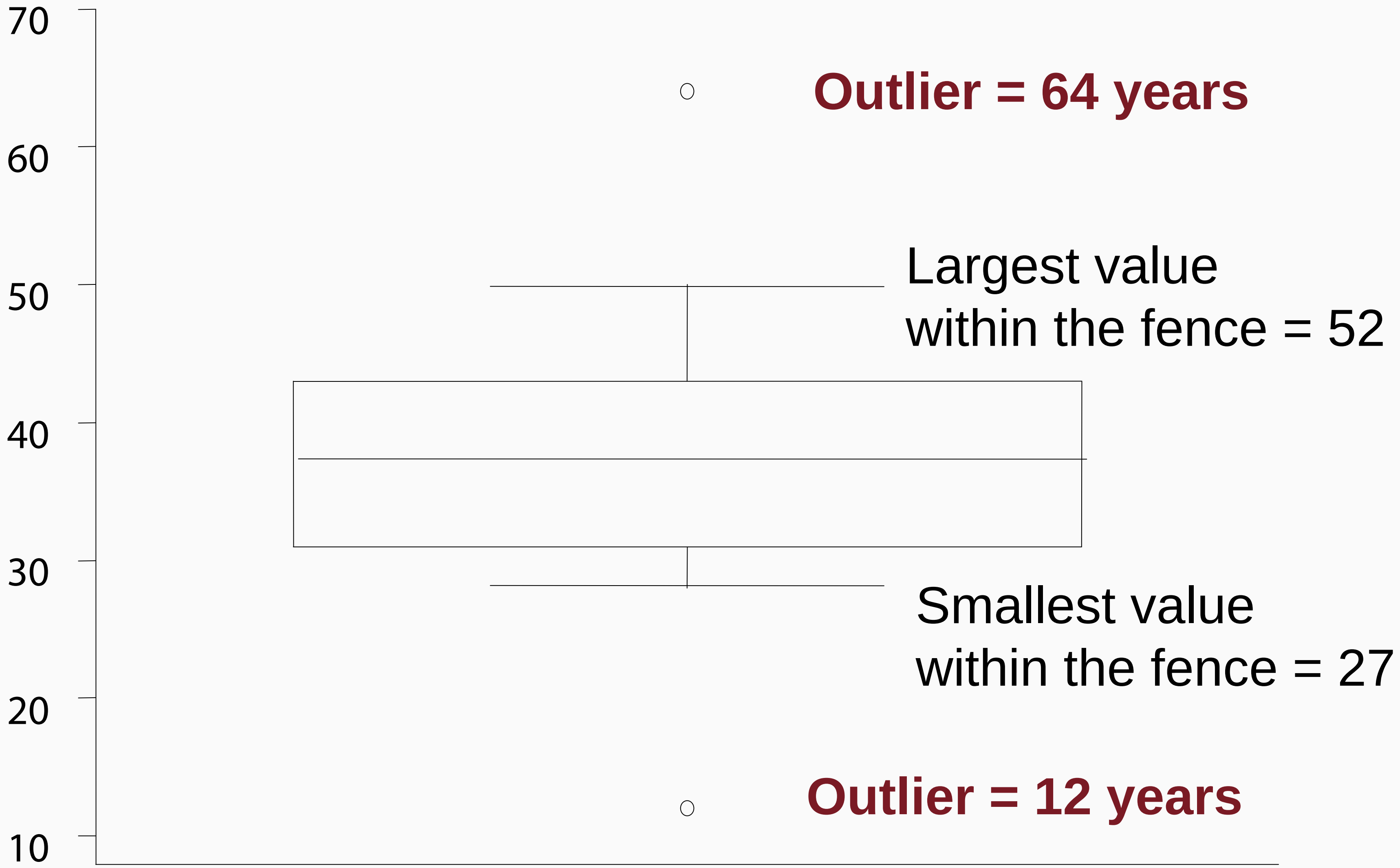
Summary Values of Student Ages

Percentiles		Smallest		
1%	27	27		
5%	27	28		
10%	27.5	31	Observation	10
25%	31	35	Sum of weight	10
50%	37.5		Mean	38.3
		Largest	Standard deviation	8.6
75%	43	42		
90%	51	43	Variance	74.7
95%	52	50	Skewness	.24
99%	52	52	Kurtosis	1.89

Example of Graduate Student Ages with Outliers

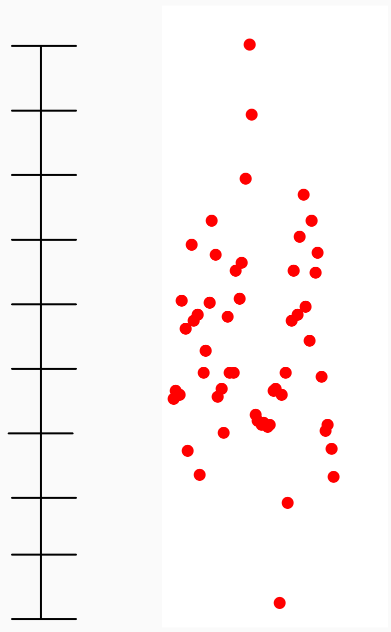
- Suppose that the data set contained individuals with ages 64 and 12 (rather than 27 and 52)
- Suppose the **fences** were now calculated as 61 and 13
- The box plot would now show the outlying values of 64 and 12 beyond the fences (**outliers**)

Box Plot of Student Ages with Outliers



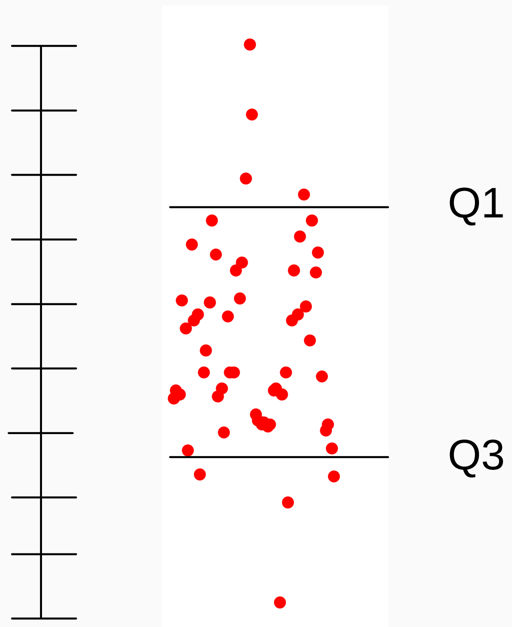
Constructing Box-and-Whisker Plots

- A box and whisker plot is a graphical display of summary measures of a set of observations



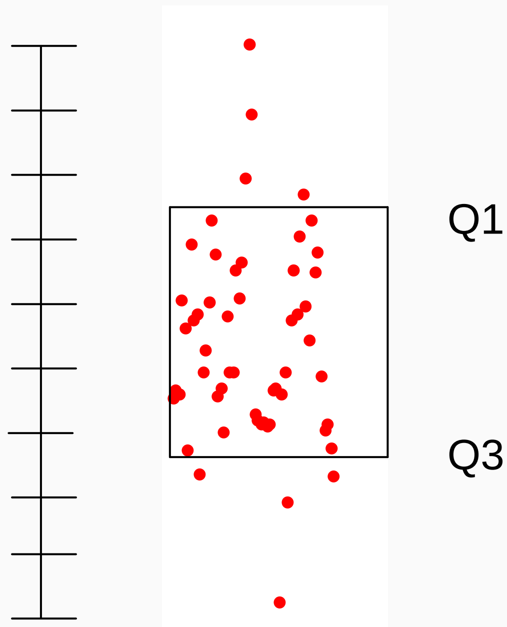
Constructing Box-and-Whisker Plots

- The hinges of the box are the first and third quartiles, Q1 and Q3, respectively

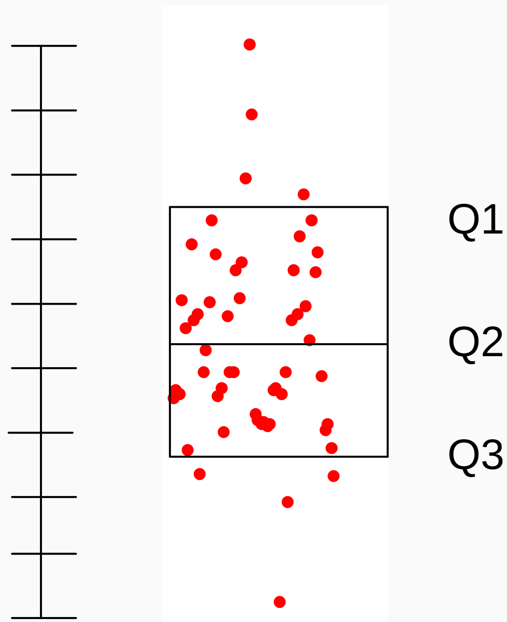


Constructing Box-and-Whisker Plots

- Fifty percent of the observations are contained within the box.

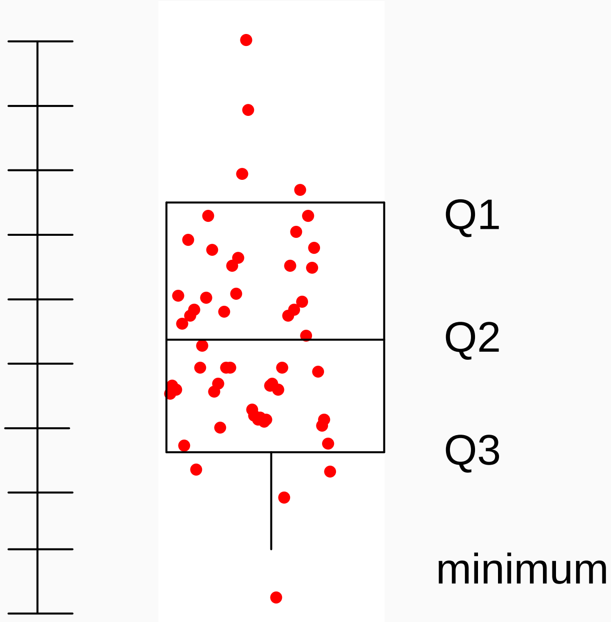


- A line drawn within the box represents the median or second quartile, Q2.



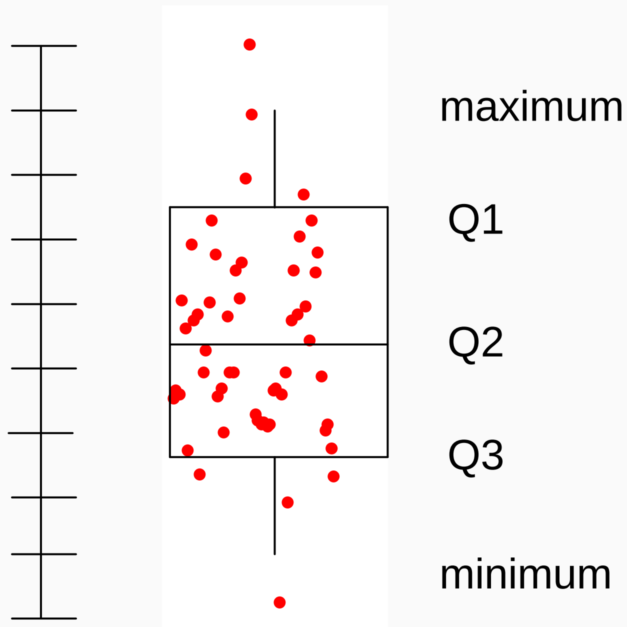
Constructing Box-and-Whisker Plots

- Whiskers are lines drawn from the bottom of the box to the smallest observation within the calculated lower fence.



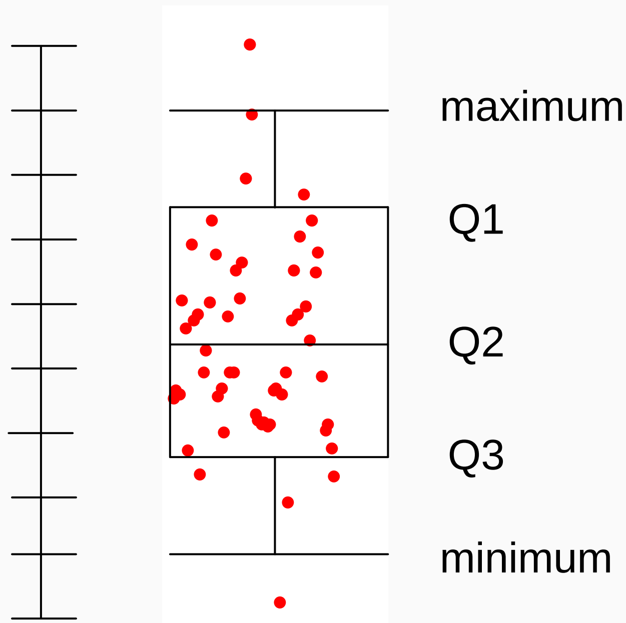
Constructing Box-and-Whisker Plots

- And from the top of the box to the largest observation within the calculated upper fence.



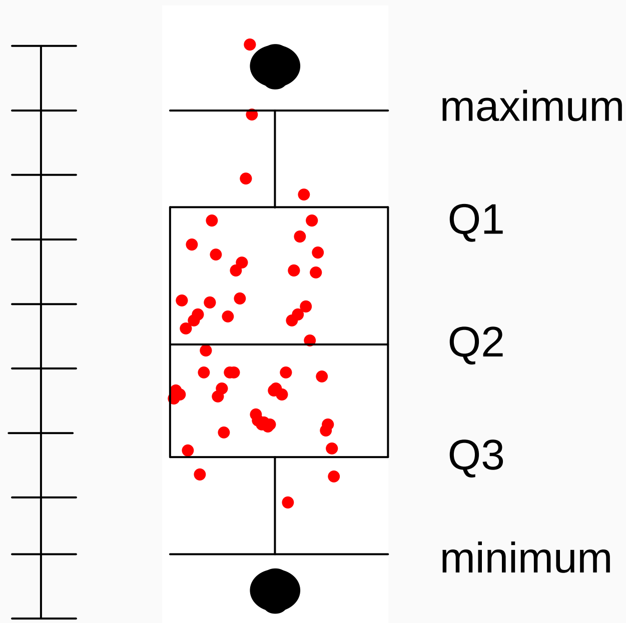
Constructing Box-and-Whisker Plots

- Some statistical computing packages, such as STATA, may draw a line segment at the end of the whisker



Constructing Box-and-Whisker Plots

- If there are observations drawn beyond the whiskers on the plot, these values are considered outliers



Constructing Box-and-Whisker Plots

- The fences are not observed values in the data set. They are calculated as guidelines for inspecting values that appear to be different from the majority of observations.