Displaying Quantitative Data in Tables and Graphs

Outline for Lesson 2 Part 2

- In addition to summary statistics, tables and graphs can be used to summarize and describe numerical data
- Tables and graphs for Numerical data
  - Stem-and-leaf plot
  - Frequency table
  - Histogram
  - Frequency polygon and percentage polygon
  - Cumulative relative frequency graph
  - Box plot

Stem-and-Leaf Plot

- The stem-and-leaf plot displays the shape of the data AND preserves all the individual data values.

- The plot consists of a series of rows and numbers
  - The number used to label the row is called a ***stem.***
  - The other numbers in the row are called ***leaves.***

Stem-and-Leaf Plot

- We’ll use the weight data from the 92 U of M students to illustrate a stem-and-leaf plot
  - Females

    140 120 130 138 121 125 116 145 150 112 125 130 120 130
    131 120 118 125 135 125 118 122 115 102 115 150 110 116
    108 95 125 133 110 150 108
  - Males

    140 145 160 190 155 165 150 190 195 138 160 155 153 145
    170 175 175 170 180 135 170 157 130 185 190 155 170 155
    215 150 145 155 155 150 155 150 180 160 135 160 130 155
    150 148 155 150 140 180 190 145 150 164 140 142 136 123
    155

Stem-and-Leaf: the Stem

- The **stem** is a column of numbers consisting of the weight data counted by tens (i.e. leave off the last digit)

```
 9 |
10 |
11 |
12 |
13 |
14 |
15 |
16 |
17 |
18 |
19 |
20 |
21 |
```

Stem-and-Leaf: the leaves

- Now add the final digit of each weight in the appropriate row

9	5
10	288
11	628855060
12	01553005525
13	8500850600153
14	05505580502
15	5053705505505050500500
16	050004
17	055000
18	0500
19	00500
20	
21	5

Meaning there are weights of 102, 108 and 108

Stem-and-Leaf Plot

- Finally put the “leaves” in order:

9	5
10	288
11	002556688
12	00012355555
13	0000013555688
14	00002555558
15	0000000000355555555557
16	000045
17	000055
18	0005
19	00005
20	
21	5

All the 0’s and 5’s clearly show the students’ reporting bias to round to the nearest 5 lbs.

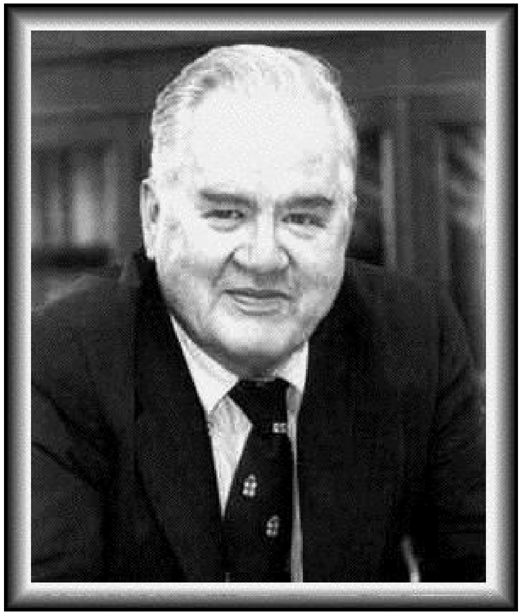
See also Table 3-6 in text:
Stem-and-leaf plot of Hebert data

Stem-and-Leaf Plot

- What do you look for in a stem-and-leaf plot?
  - Shape
  - Spread
  - Location
  - Outliers

Stem-and-Leaf Plots

- Invented in 1977 by John Tukey
  b.1915 – d. 2000
- Contributions to statistics
  - Exploratory data analysis methods
  - Time-series analysis
  - Multiple comparisons
- Tukey also coined these terms
  - 'bit' for binary digit (1948)
  - 'software' (1958)



"An appropriate answer to the right problem is worth a good deal more than an exact answer to an approximate problem"

Sources: Wikipedia
http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Tukey.html

Frequency Table

- A useful way to present data when you have a large data set is the formation of a frequency table or frequency distribution.

- Frequency – the number of observations that fall within a certain range of the data.

- A frequency table is the result of 'grouping' continuous or discrete data into categories.

- A frequency table provides information about the distribution of the data.

Example SMAF Data

- *Presenting Problem 2: page 24*
  - *Hebert and coworkers (1997) study disability and functional change measures in a community-dwelling population of people 75 years and older.*
  - *SMAF: The Functional Autonomy Measurement System, a 29 –item rating scale.*

Data for Frequency Table

Total score on the SMAF at Time 1 for 72 patients age 85 and older (from Table 3-4 in text, Hebert).

The total score is the sum of 29 functional disability items rated 0 for independent to 3 for dependent

28	8	6	22	6	9	23	12	9	5
20	3	22	20	0	30	13	47	1	3
4	12	1	13	1	35	22	1	2	3
21	2	7	8	7	11	1	12	19	21
17	27	4	9	7	38	13	4	17	23
12	30	27	26	44	21	17	10	15	4
10	18	12	37	17	14	11	4	16	5
48	9								

Raw Data

- The 72 SMAF scores on the previous slide are the ‘raw’ data. They haven’t been summarized. It’s difficult to identify any patterns in the raw data
- The next slide shows the same data summarized in a frequency table which provides information about the distribution of the SMAF scores.
- The steps for constructing the frequency table follow.

Frequency Table of SMAF scores

SMAF score interval	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0 - 4	16	16	22.2%	22.2%
5 - 9	13	29	18.1%	40.3%
10 - 14	13	42	18.1%	58.3%
15 - 19	8	50	11.1%	69.4%
20 - 24	10	60	13.9%	83.3%
25 - 29	4	64	5.6%	88.9%
30 - 34	2	66	2.8%	91.7%
35 - 39	3	69	4.2%	95.8%
40 - 44	1	70	1.4%	97.2%
45 - 49	2	72	2.8%	100%
Total	72			

Constructing A Frequency Table: Overview

1. Determine the number and width of the frequency table intervals: the classes
2. Find the frequency (*the count*) and cumulative frequency (*the cumulative count*) in each class
3. Calculate the percent and cumulative percent in each class

1. Number and width of Classes

- Decide on the number and width of the classes
- With too many classes the data may not be summarized enough for a clear visualization of how they are distributed.
- With too few classes the data may be over-summarized and some of the details of the distribution may be lost.
- This step is subjective and depends on the data being summarized. A general guideline is to have 6-14 classes

Number and Width of Classes

- Find the Minimum, Maximum and range of the data
  - minimum = 0, maximum = 48
  - Range = 48 – 0 = 48
- With 10 classes, each class has a width equal to the range divided by the number of classes = 48/10 = 4.8. Round this up to a more intuitive width of 5.
- Alternatively, we could choose the width first – a width of 5 units seems reasonable for this data. The number of classes is then equal to the range divided by the width = 48 / 5 = 9.6. Round this up to 10 classes.

Number and Width of Classes

- We’ll use 10 classes with width 5
- The minimum score = 0. The first class is 0-4
- Proceed with non-overlapping classes

CLASSES for SMAF score:

0 - 4
5 - 9
10 -14
15 -19
20 - 24
25 - 29
30 - 34
35 - 39
40 - 44
45 - 49

**See also Table 3-2 and Table 3-8 in Text:
Frequency tables of shock index
From Kline data**

2. Frequency and Cumulative Frequency

- Frequency: the number of observations in each class (or category)
- The frequency in each class can be found by tallying the observations in each class 𝍸

- Cumulative frequency: the number of observations up to and including that class
  - The cumulative frequency for each class is the sum of that class frequency and all preceding class frequencies.

Frequency and Cumulative Frequency

SMAF score interval	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0 - 4	16	16		
5 - 9	13	29		
10 - 14	13	42		
15 - 19	8	50		
20 - 24	10	60		
25 - 29	4	64		
30 - 34	2	66		
35 - 39	3	69		
40 - 44	1	70		
45 - 49	2	72		
Total	72			

16 of the SMAF scores are between 0-4, 13 are between 5-9, etc.
The cumulative frequency for the 5-9 class = 13 +16 = 29
Cumulative frequencies are the sum of the frequencies up to and including that class

3. Percent and Cumulative Percent

- Percent = $\dfrac{\text{frequency in class}}{\text{total N for data}}$

The percent is sometimes called the relative frequency

- Cumulative percent = $\dfrac{\text{Cumulative Freq. in class}}{\text{total N for data}}$

Cumulative percent is also called cumulative relative frequency

Frequency Table: percent

SMAF score interval	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0 - 4	16	16	22.2%	
5 - 9	13	29	18.1%	
10 - 14	13	42	18.1%	
15 - 19	8	50	11.1%	
20 - 24	10	60	13.9%	
25 - 29	4	64	5.6%	
30 - 34	2	66	2.8%	
35 - 39	3	69	4.2%	
40 - 44	1	70	1.4%	
45 - 49	2	72	2.8%	
Total	72			

The percent in each class = frequency divided by the total times 100

The cumulative percent for each class is the sum of the percent in that class plus the percent for all preceding classes

SMAF score interval	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0 - 4	16	16	22.2%	22.2%
5 - 9	13	29	18.1%	40.3%
10 - 14	13	42	18.1%	58.3%
15 - 19	8	50	11.1%	69.4%
20 - 24	10	60	13.9%	83.3%
25 - 29	4	64	5.6%	88.9%
30 - 34	2	66	2.8%	91.7%
35 - 39	3	69	4.2%	95.8%
40 - 44	1	70	1.4%	97.2%
45 - 49	2	72	2.8%	100%
Total	72			

Completed Frequency Table

SMAF score interval	Frequency	Cumulative Frequency	Percent	Cumulative Percent
0 - 4	16	16	22.2%	22.2%
5 - 9	13	29	18.1%	40.3%
10 - 14	13	42	18.1%	58.3%
15 - 19	8	50	11.1%	69.4%
20 - 24	10	60	13.9%	83.3%
25 - 29	4	64	5.6%	88.9%
30 - 34	2	66	2.8%	91.7%
35 - 39	3	69	4.2%	95.8%
40 - 44	1	70	1.4%	97.2%
45 - 49	2	72	2.8%	100%
Total	72			

Correction to text

- Correction on page 37 middle of column 1: The cumulative percent [not frequency] is the percentage of observations for a given value plus that for all lower values.

Mean From a Frequency Table

- If the data are presented in the grouped form of a frequency table and the raw data are not available, the mean can be approximated using a weighted average of the data
  - Multiply the midpoint of each class by the frequency in the class
  - Sum the products and divide by the total number of observations.
- Approximating the mean improves with
  - Larger data sets
  - Smaller class widths

Mean from Frequency Table

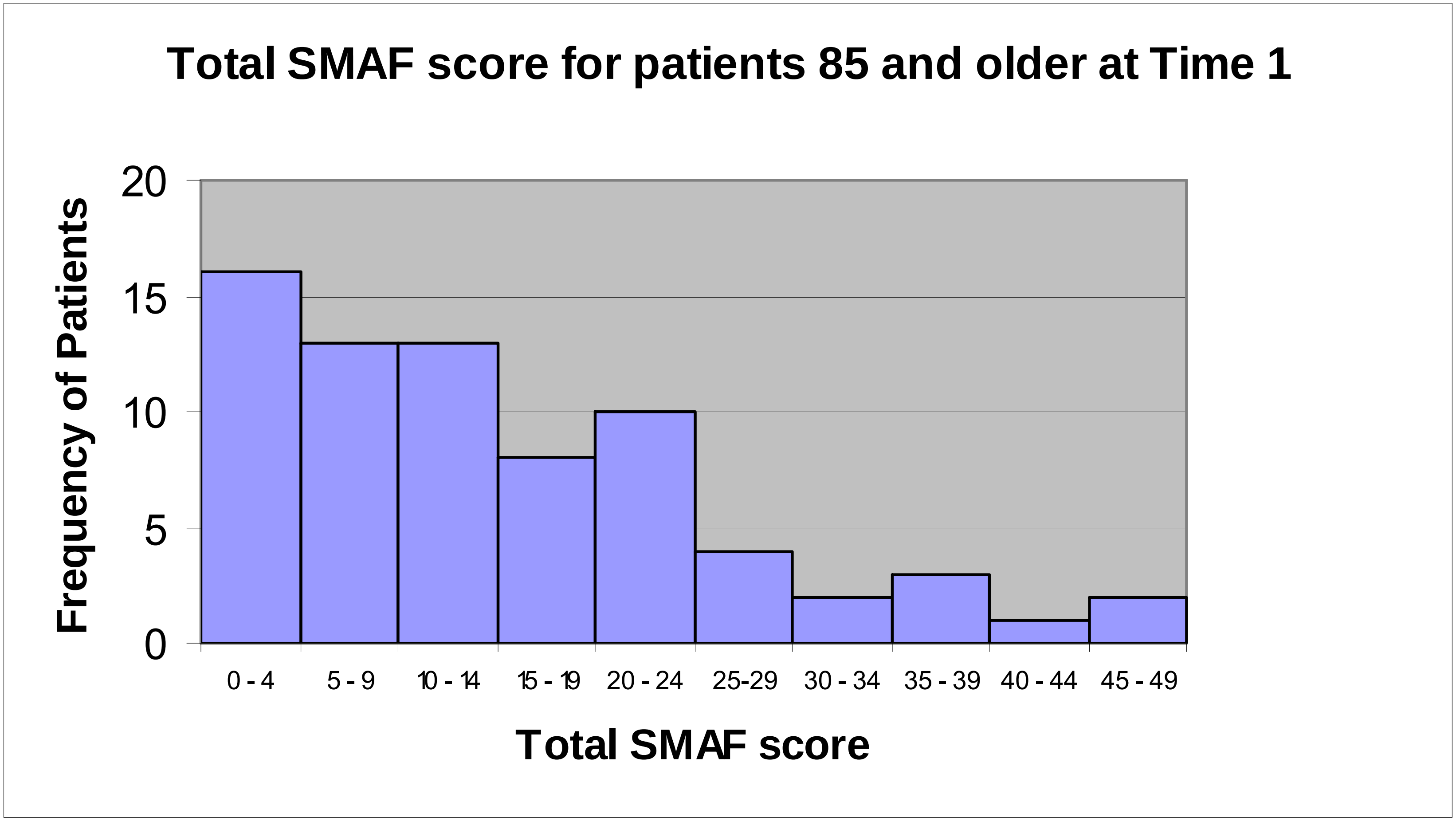
SMAF score interval	Class Midpoint	Frequency	Product
0 - 4	2	16	32
5 - 9	7	13	91
10 - 14	12	13	156
15 - 19	17	8	136
20 - 24	22	10	220
25 - 29	27	4	108
30 - 34	32	2	64
35 - 39	37	3	111
40 - 44	42	1	42
45 - 49	47	2	94
Total		72	1054
Weighted average = 1054 / 72 = 14.6			

Mean SMAF score calculated
from raw data
= 14.7

Graphs of Numerical Data

- Once the frequency table is completed, the summarized data can be illustrated graphically.
- A histogram is a plot of the frequency or percent columns in a frequency table
- A frequency polygon is a line graph of the frequency column in a frequency table
- A percentage polygon is a line graph of the percent column in a frequency table
- A cumulative relative frequency graph is a line graph of the cumulative percent column.

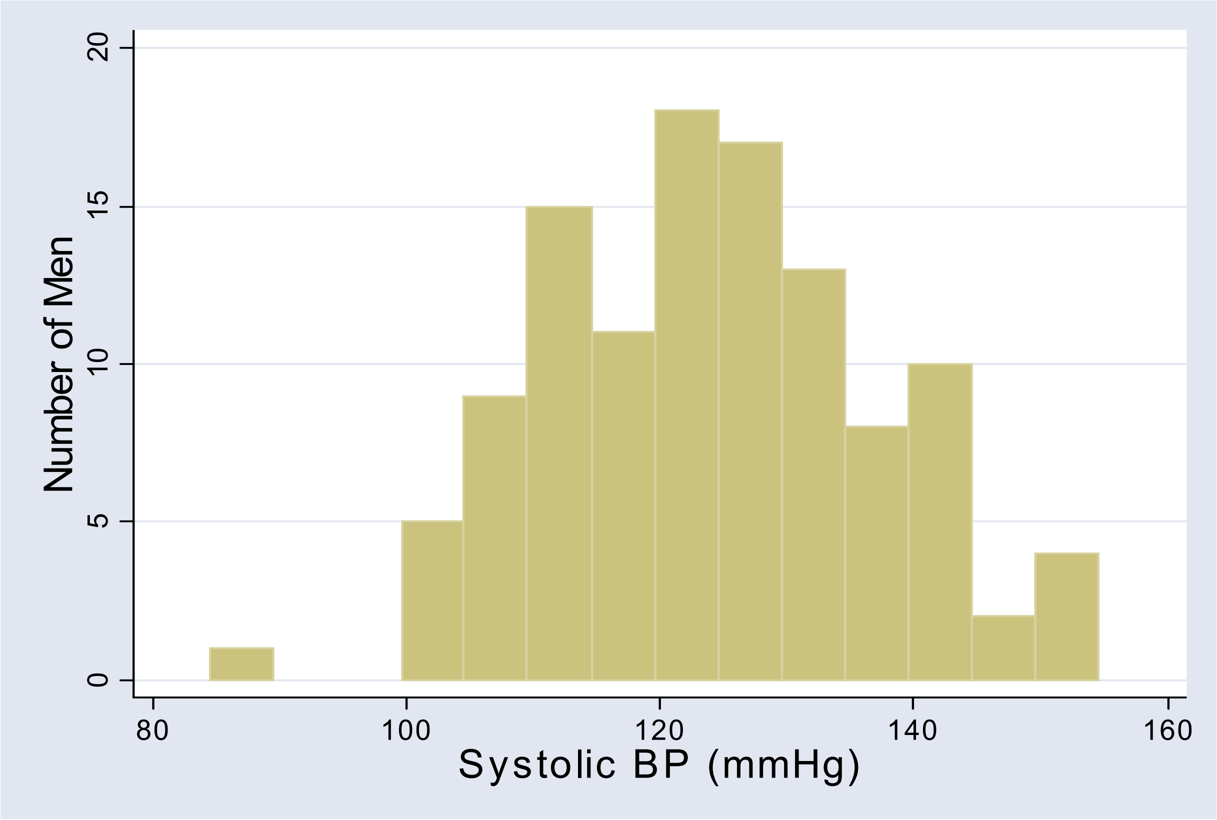
Histogram – graphical display of frequency column



Features of a Histogram

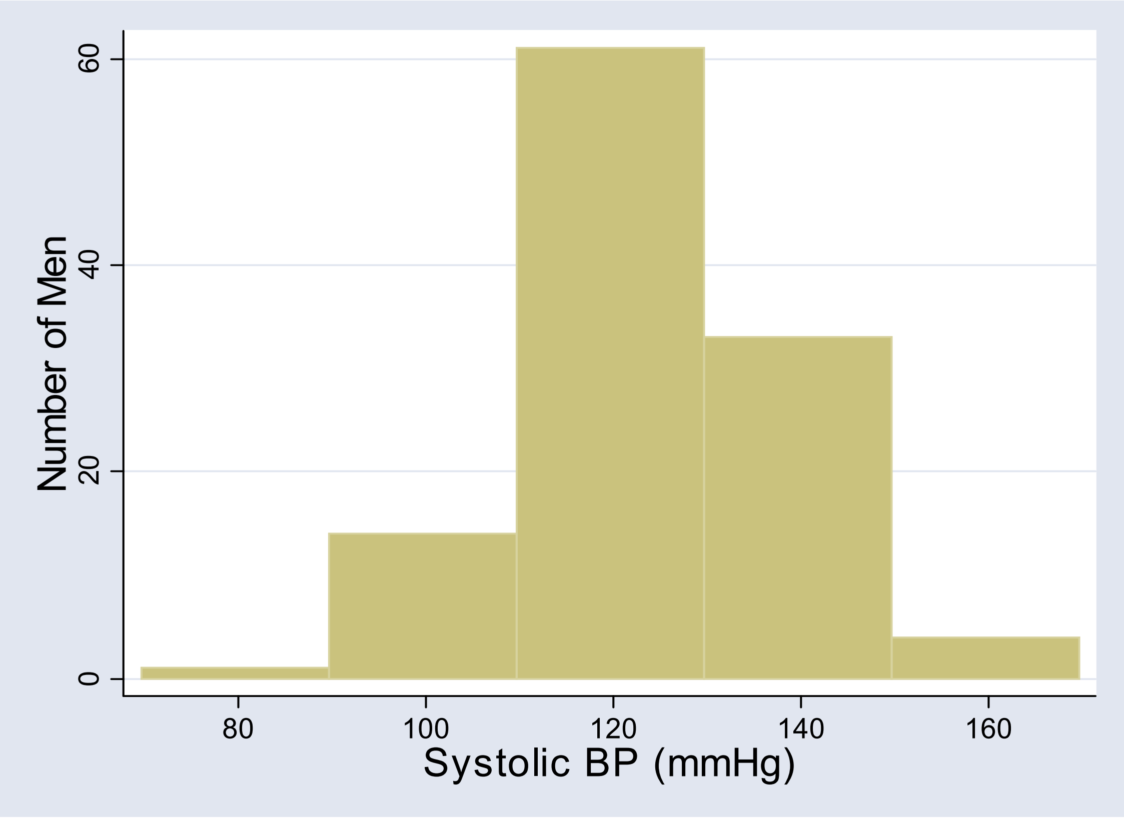
- The horizontal scale represents the classes
- The vertical scale represents either the frequency or percent in each class
  - Label the vertical axis accordingly
- Each class is represented by a bar with area proportional to the percent of observations in that class
- The rectangular bars are adjacent to each other to indicate that the underlying data is continuous

Histogram examples



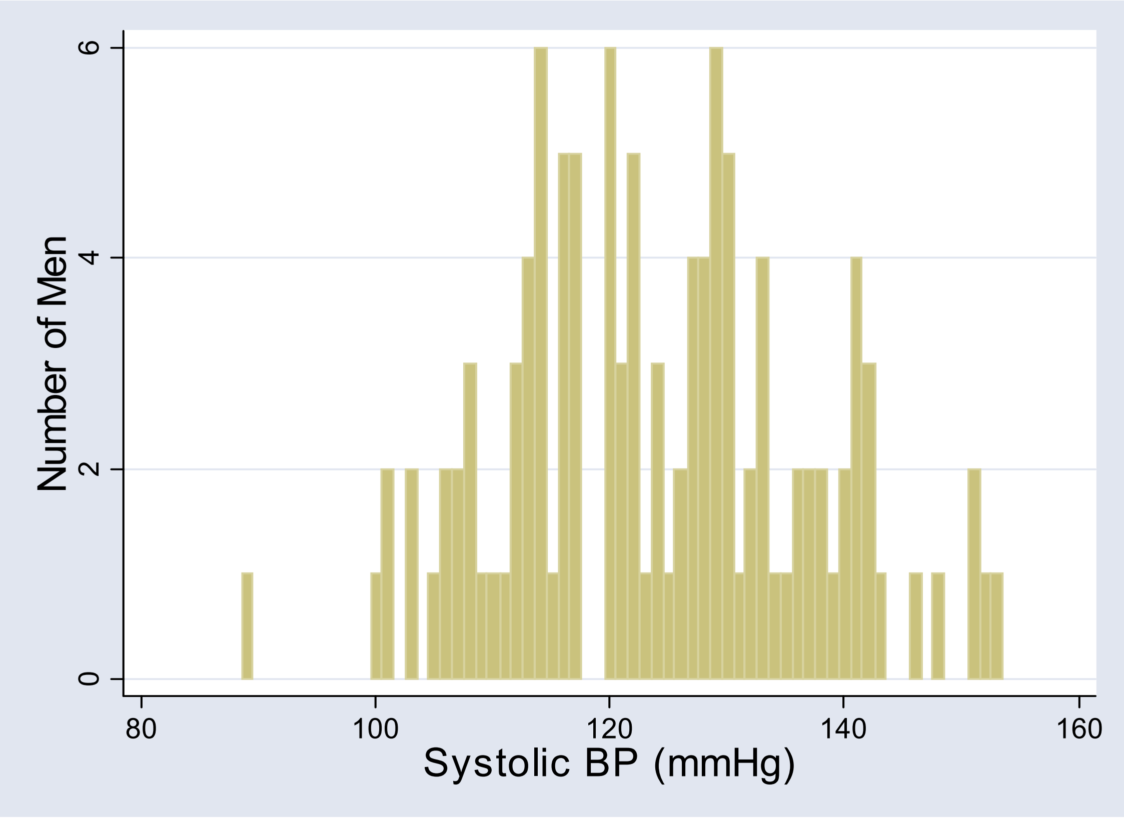
Histogram of the Systolic Blood Pressure for 113 men. Each bar spans a width of 5 mmHg on the horizontal axis. The height of each bar represents the number of individuals with SBP in that range.

Histogram: too few intervals



Another histogram of the blood pressure of 113 men. In this graph, each bar has a width of 20 mmHg, and there are a total of only 5 bars making it difficult to characterize the distribution of blood pressures in the sample.

Histogram: too many intervals



Another histogram of the same SBP information on 113 men. Here, the class width is 1 mmHg, which gives more detail than is useful in summarizing the data

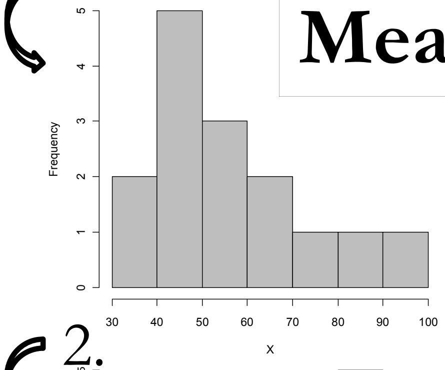
Histogram

- What do you look for in a histogram?
  - Shape
  - Spread
  - Location
  - Outliers

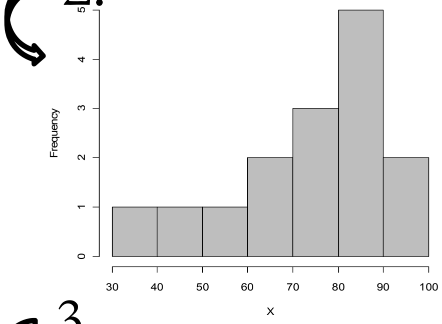
Given the mean, median and mode, what does the distribution most likely look like?

Mean = 58.8, Median = 53, Mode = 47

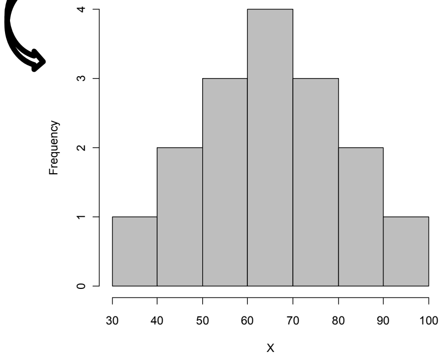
1.



2.



3.



What happens when we add ten to every number?

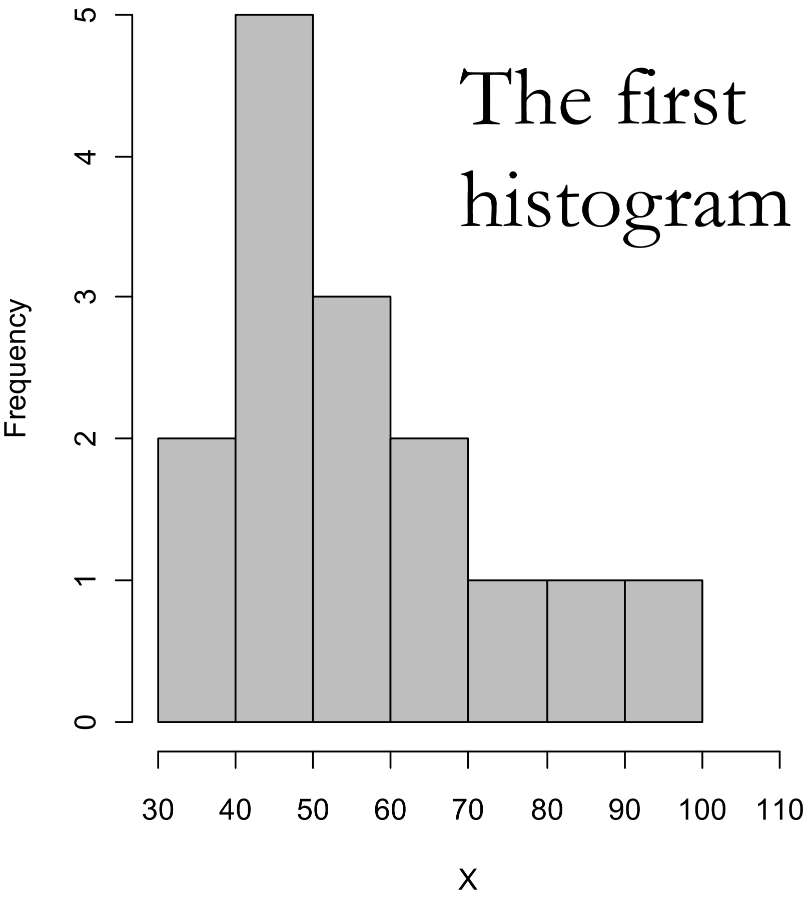
49 55 69 56 57 69 57 79

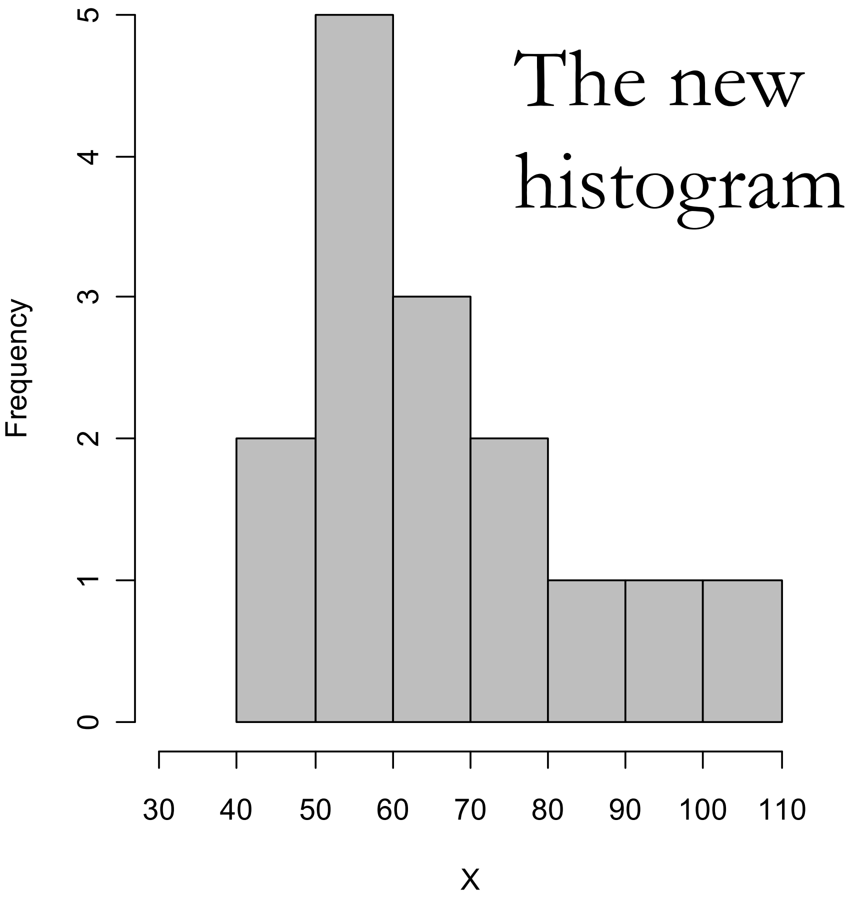
47 77 57 63 89 99 109

What happens to the histogram?

1. Shifts left
2. Shifts Right
3. Gets narrower
4. Gets wider

Let’s see.


The first histogram


The new histogram

Histogram website

www.shodor.org/interactivate/activities/histogram

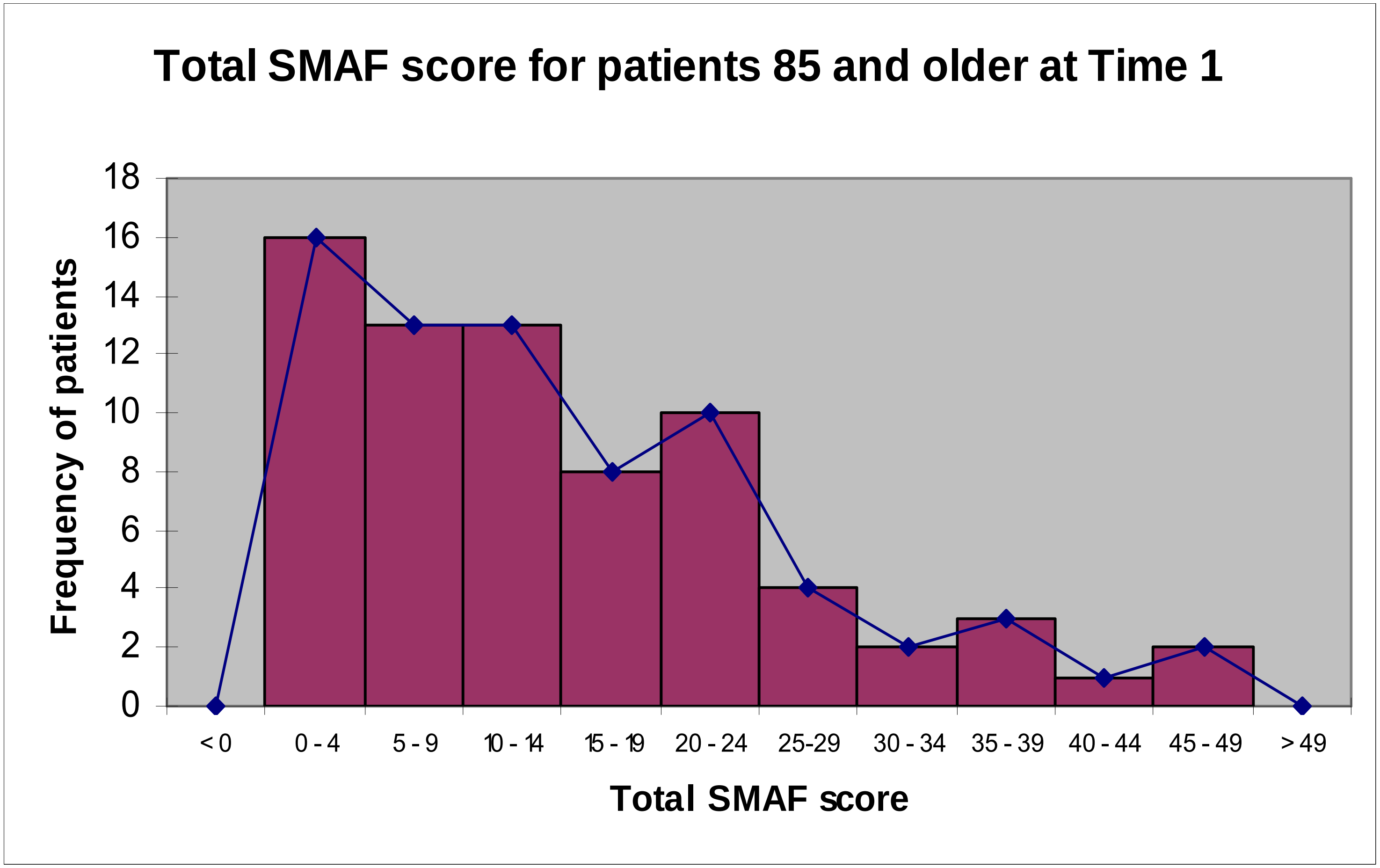
This website has several data sets and an interactive applet for creating histograms with varying interval widths

You can observe the effect of having too many intervals (the data isn’t summarized at all) or too few intervals (the summary information is lost).

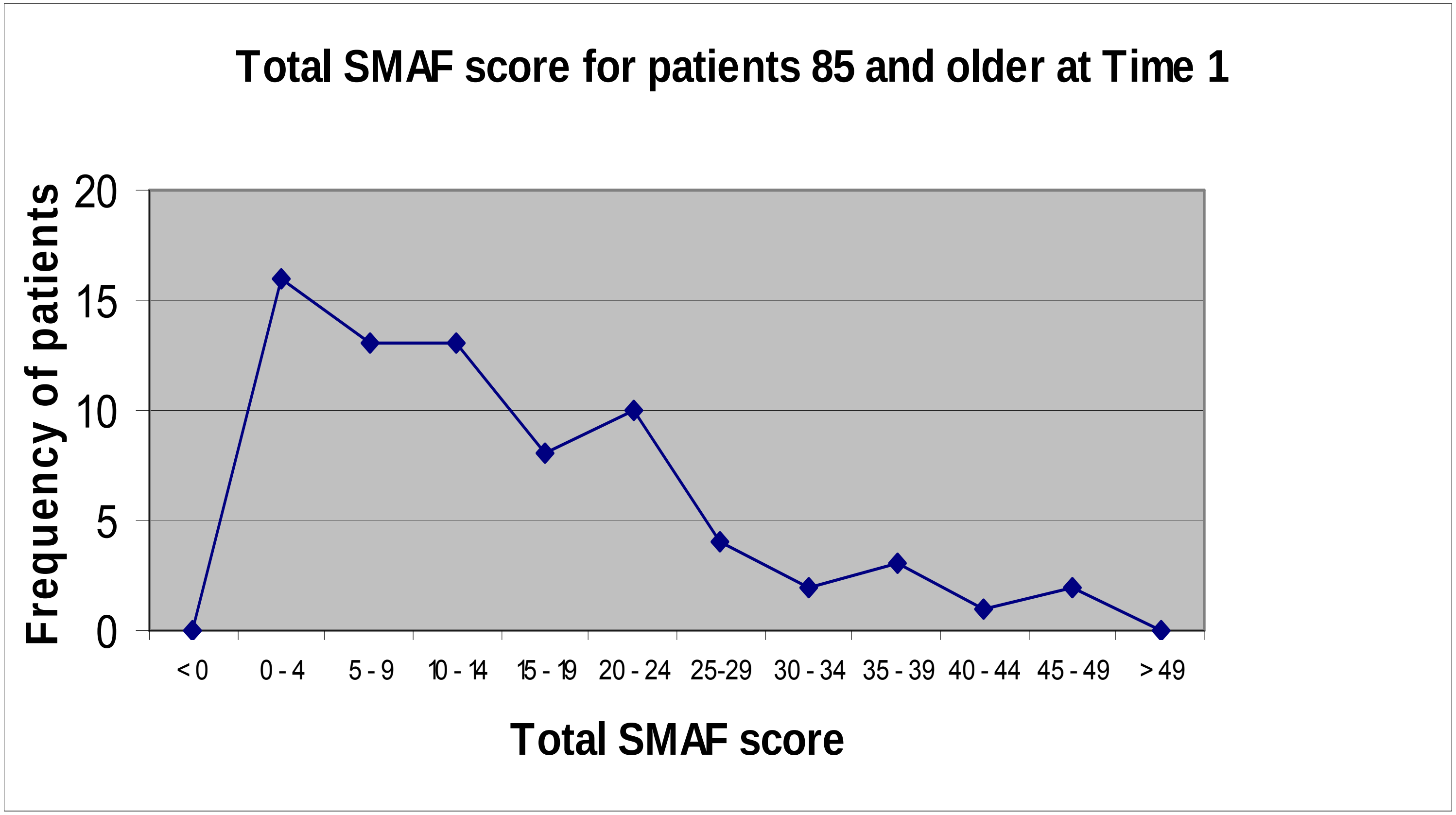
Frequency and Percentage Polygons

- A frequency polygon is a line graph that outlines the shape of the histogram of frequencies
- A percentage polygon is a line graph that outlines the shape of a histogram of percents
- The line connects the midpoints of the histogram columns
- At the ends, the points are connected to the x-axis using two additional intervals with frequency (or percent) = 0.

Frequency polygon and Histogram



Frequency Polygon

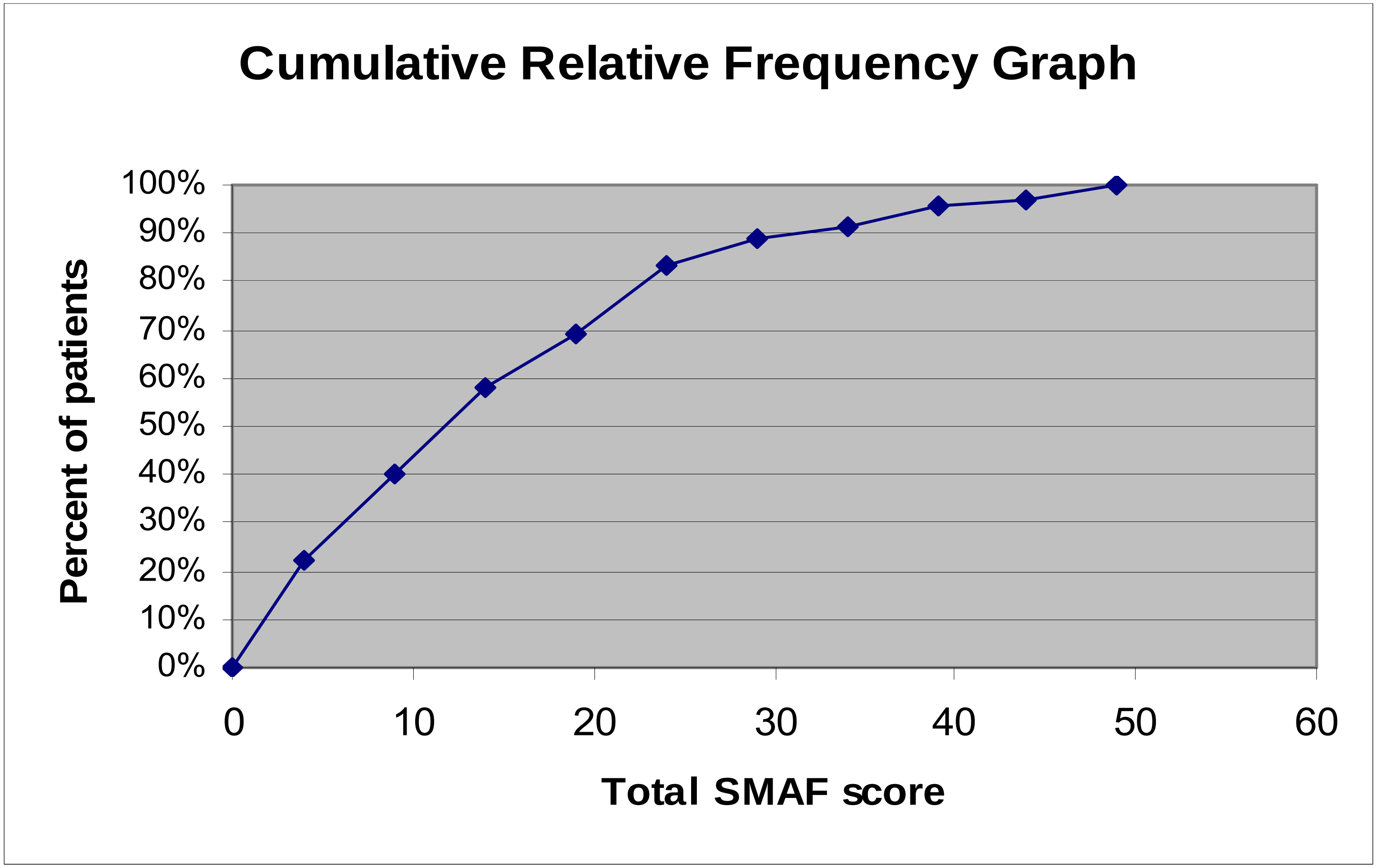


Applications for Histograms and Frequency Polygons

- Histograms and Frequency polygons provide information about data distribution
  - Is the distribution unimodal or bimodal?
  - What is the Range of the data
  - Is the distribution symmetric or skewed?
- What are some features of the SMAF score data for patients 85 and older?

The distribution is unimodal, most of the scores are less than 25. The distribution is positively skewed.

Cumulative Relative Frequency Graph: Plotting the Cumulative percents



Cumulative Relative Frequency Graph Features

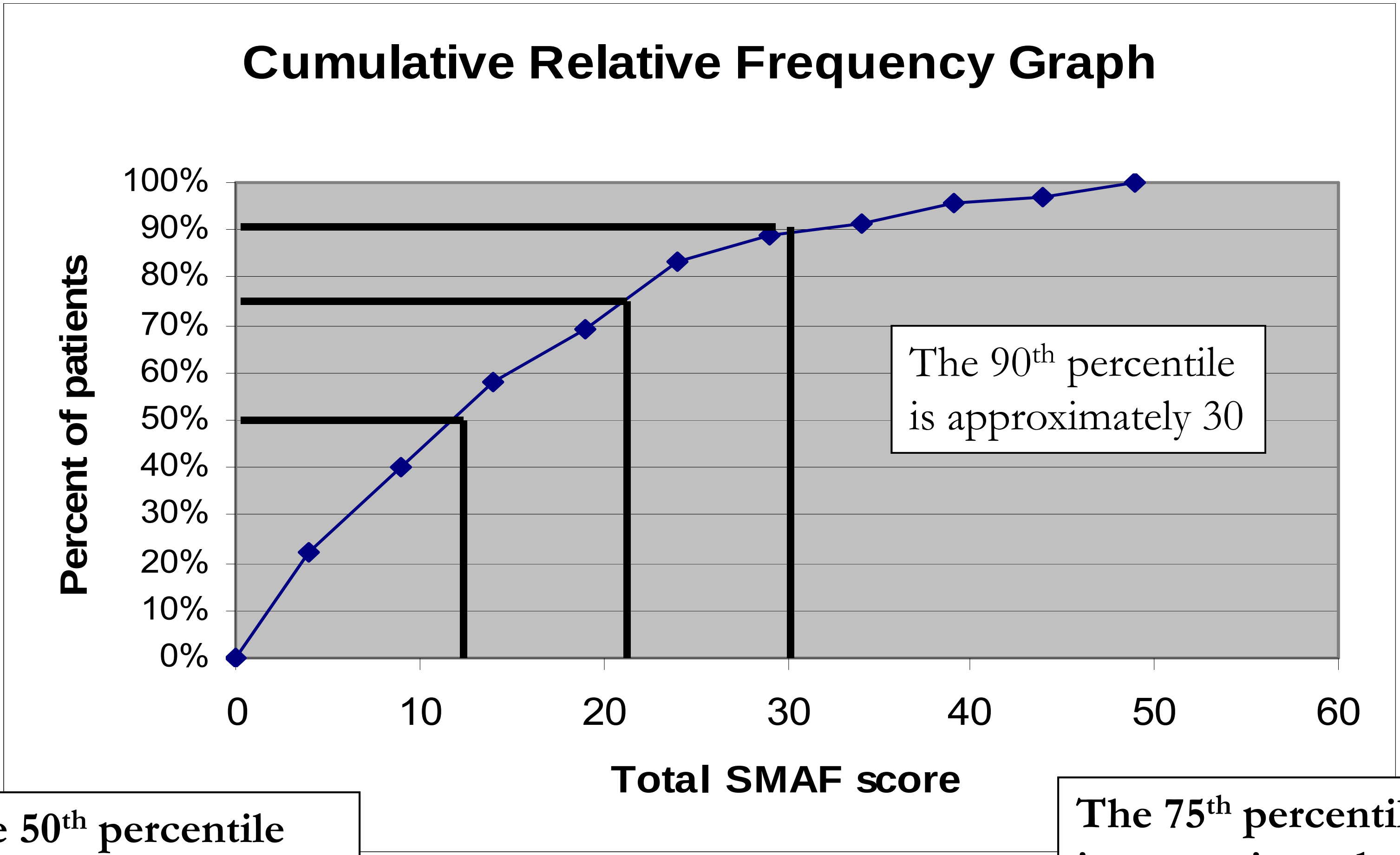
- The (x,y) points of the graph are the upper limit of each class interval (x) and the cumulative percent for that class (y).
- The points are connected with a line.
- A cumulative relative frequency graph can be used to find percentiles of the data

Percentiles

Percentiles divide a data set into 100 equal parts

- Definition of 95th percentile:
  - 95% of the observations are less than or equal to this value
  - 5% of the observations are greater than this value
- Definition of 50th percentile:
  - 50% of the observations are less than or equal to this value.
  - 50% of the observations are greater than this value
- The median is the same as the 50th percentile
- Quartile 1 = 25th percentile
- Quartile 3 = 75th percentile

Percentiles from Graph



The 90th percentile is approximately 30

The 50th percentile is approximately 12

The 75th percentile is approximately 21

Percentiles from Cumulative Relative Frequency Graph

For the SMAF score data from patients 85 or older

- 50th percentile total score = 12
  - 50% of the patients have a total score ≤ 12
- 75th percentile total score = 21
  - 75% of the patients have a total score ≤ 21
- 90th percentile of total SMAF score = 30
  - 90% of the patients have a total score ≤ 30

Box-plots

- Box-plots or box-and-whisker plots were also invented by Tukey (1977)
- A box-plot is a visual display of the distribution of a data set that illustrates the location, spread, and the degree and direction of skewness (if any).
- The Minimum, Maximum, Range, Quartile 1, Quartile 3, Median and interquartile range (IQR) are used to make box-plots.
- Box-plots can be used to compare two different data sets visually side by side.

The Box-plot: An Example

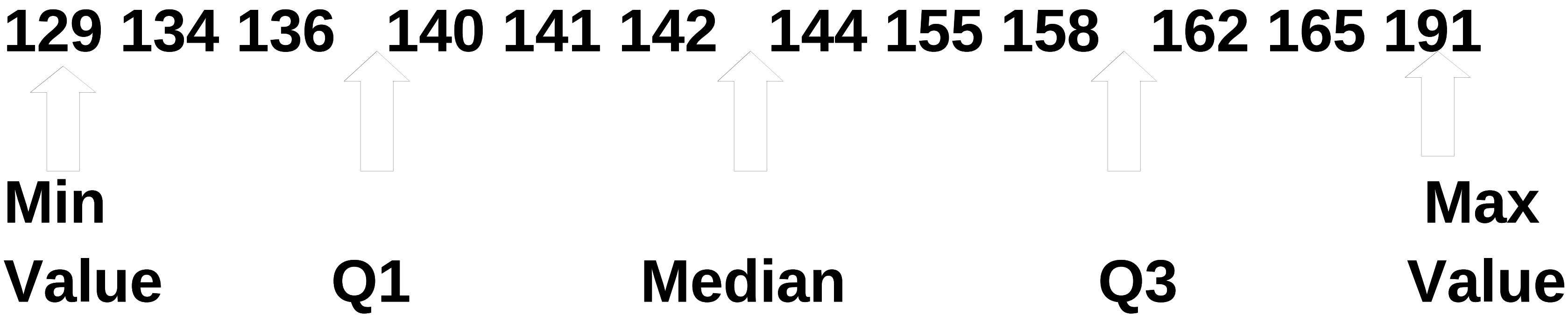
Twelve 18- year old males in a jogging club were weighed for a health study. Their weights in pounds are:

{129,134,136,140,141,142,144,155,158,162,165,191}

Elements needed for the box-plot:

Minimum, 1st quartile, Median, 3rd quartile, Maximum

The Box-plot:
An Example



129 134 136 140 141 142 144 155 158 162 165 191

Min Value **Q1** **Median** **Q3** **Max Value**

IQR = Q3 - Q1 =
160 –138 = 22

Min =129

Q1 = ½(136+140) = 138

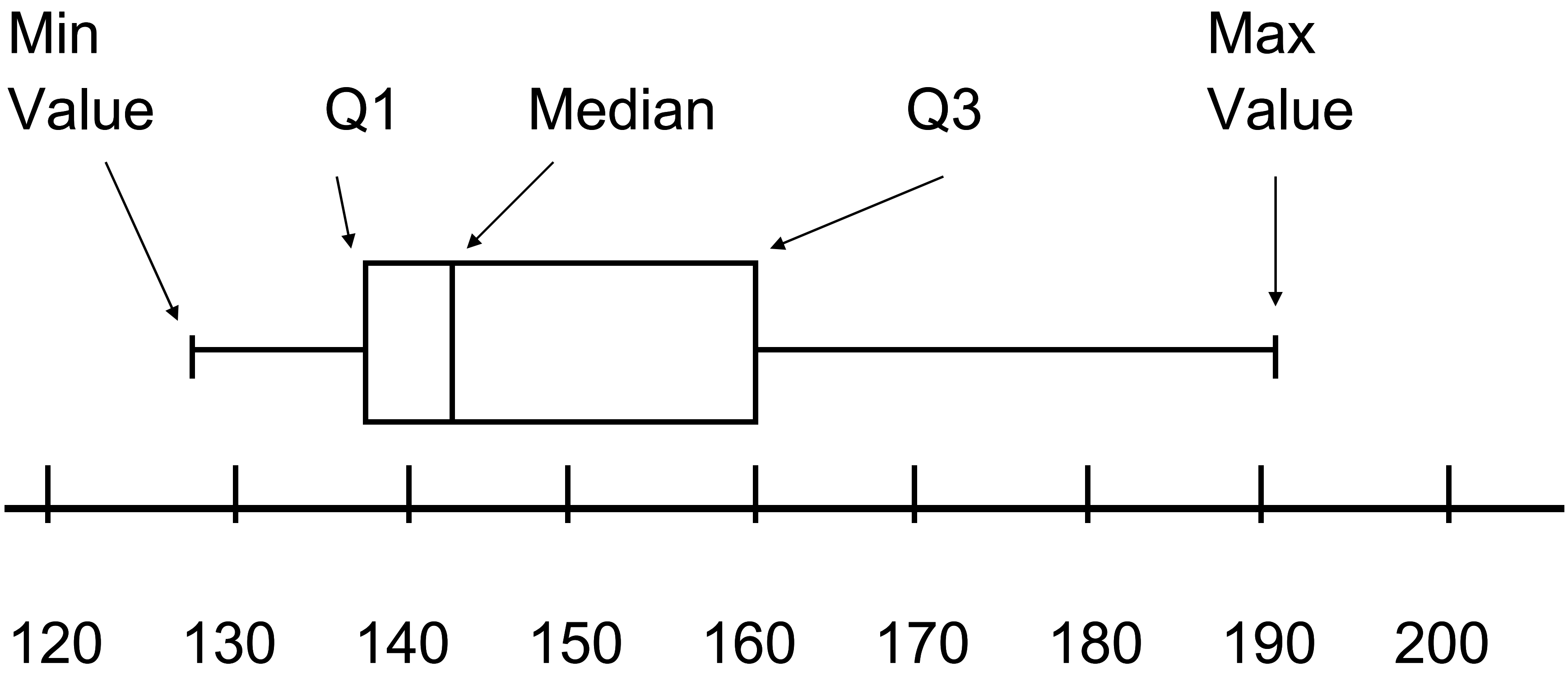
Median = ½ (142+144) = 143

Q3 = ½(158+162) = 160

Max = 191

The Box-plot: An Example

129 134 136   140 141 142   144 155 158   162 165 191



Box-plot with an Outlier

What if the data has an outlier? For example, what if the one of the weights is 220?

129 134 136 140 141 142 144 155 158 162 165 220

We might suspect 220 pounds is an outlier. One rule for identifying an outlier is if:

The Value > Q3 + 1.5 (IQR) = 160 + 33 = 193

or

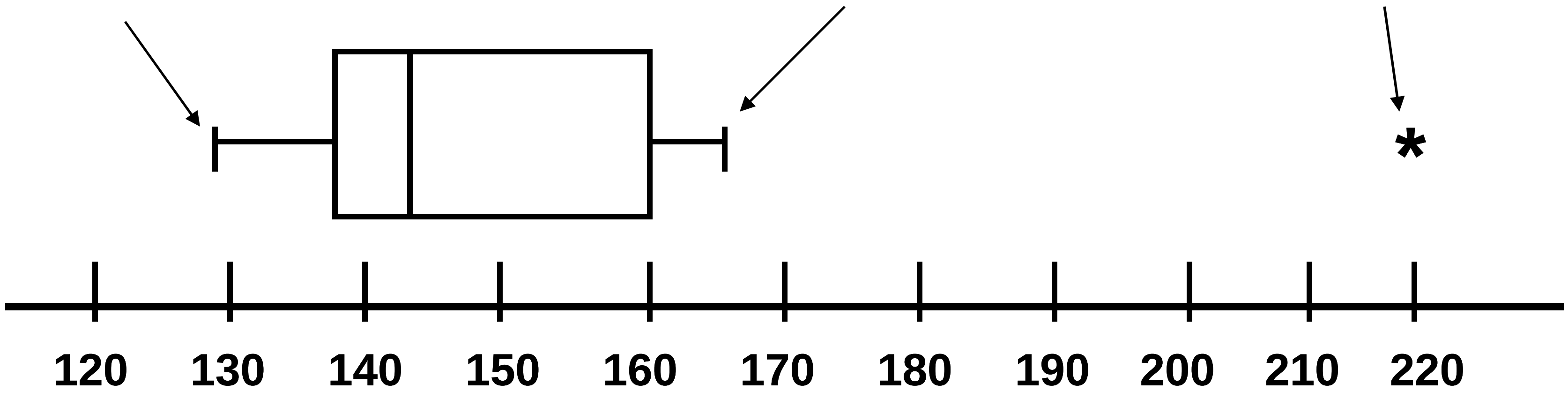
The Value < Q1 - 1.5 (IQR) = 138 – 33 = 108

Since 220 > 193, the value 220 is considered an outlier in this dataset

The Box-plot with an Outlier

When an outlier is identified, plot the outlier as an * and use the next largest value (that is not an outlier) as the end of the top whisker on the box plot

129 134 136   140 141 142   144 155 158   162 165 220



Min Value

Q1

Median

Q3

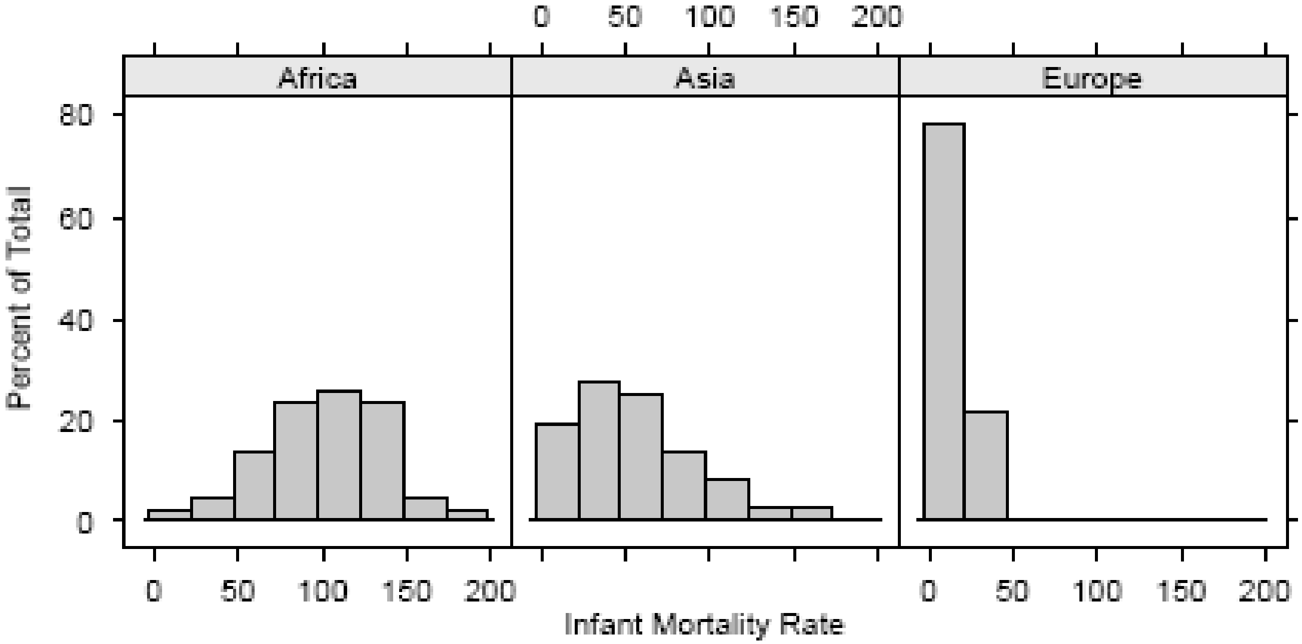
Next largest Value

Outlier

Comparing two or more groups graphically

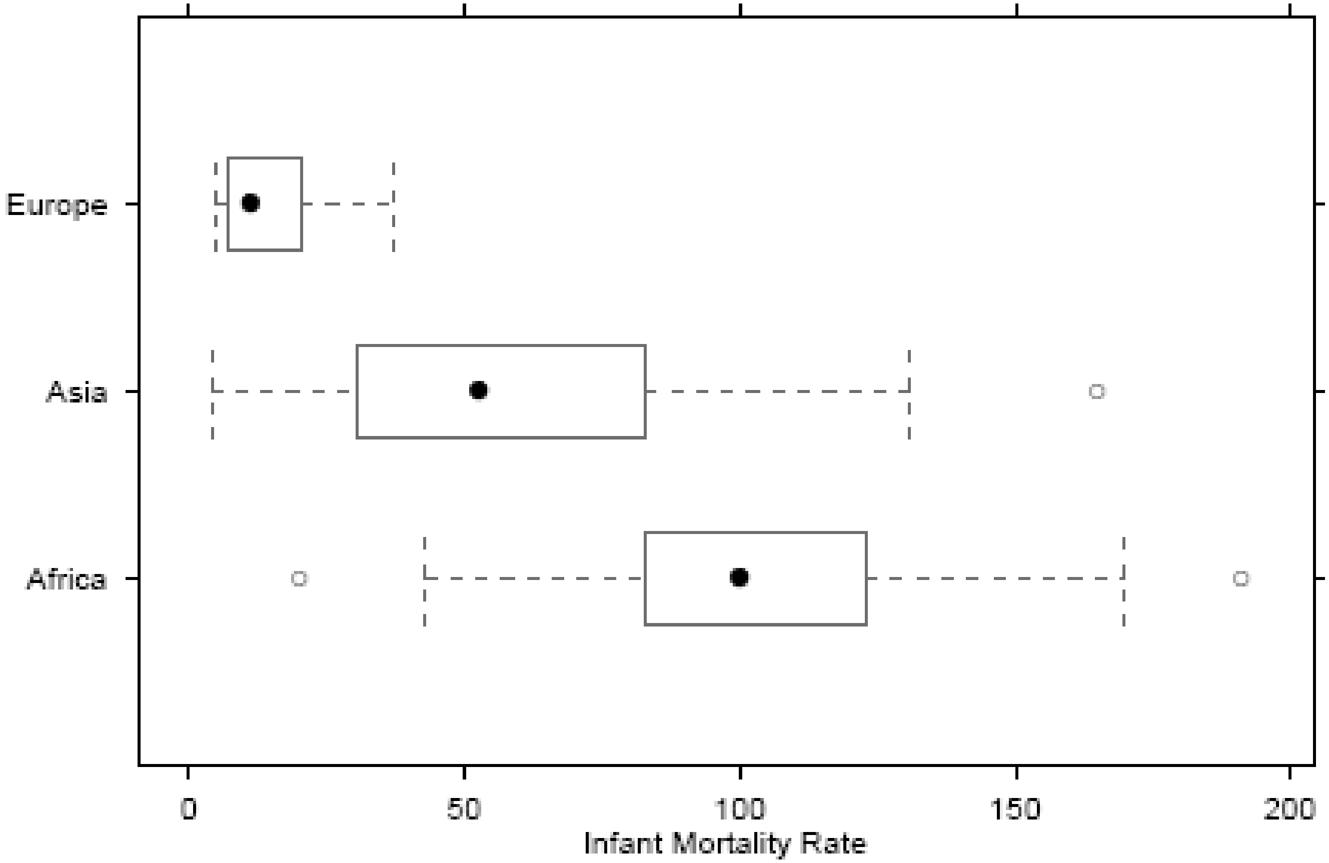
- Side by side box-plots can be used to compare distributions of two groups

Infant Mortality Rates in 1992



Deaths at <1 year of age per 1,000 live births

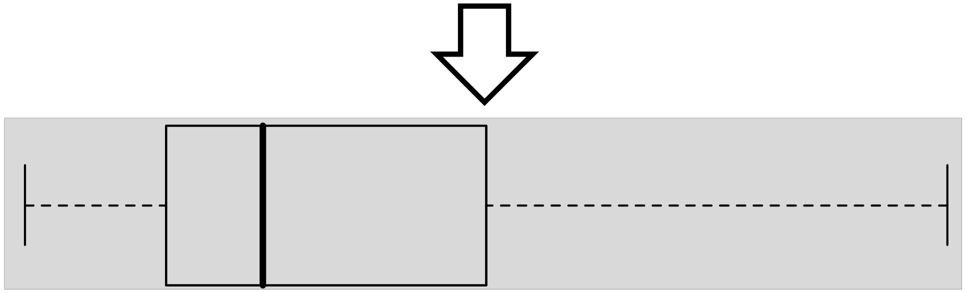
Infant Mortality Rates (cont'd)



Let build a box plot…

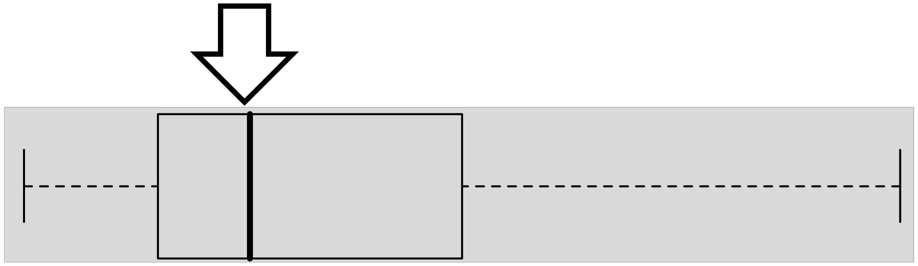
- Given data, we can calculate:
  - The Minimum
  - The Maximum
  - Q1 = 25th percentile
  - Q3 = 75th percentile
  - The Median
  - The Interquartile Range (IQR)
  - Outliers

Given the following boxplot, what is the indicated part?



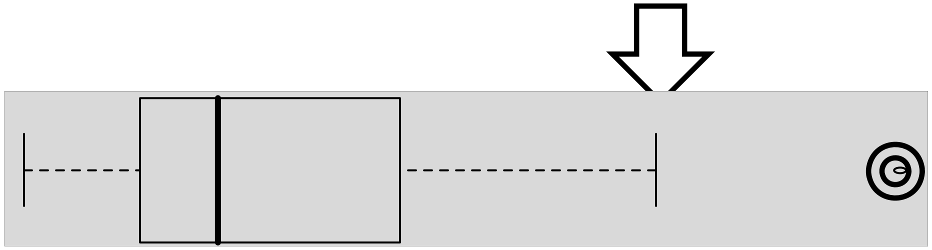
1. Maximum
2. Median
3. Q3

Given the following boxplot, what is the indicated part?



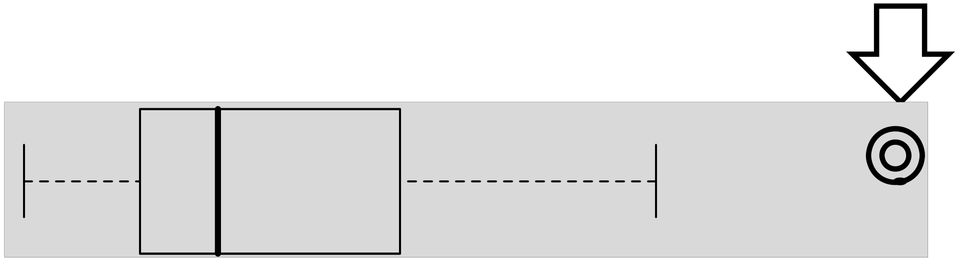
1. Minimum
2. Median
3. Q1

Given the following boxplot, what is the indicated part?



1. Maximum
2. Largest value that is not an outlier
3. IQR

Given the following boxplot, what is the indicated part?



1. Maximum
2. Outlier
3. Median

How do we calculate the IQR?

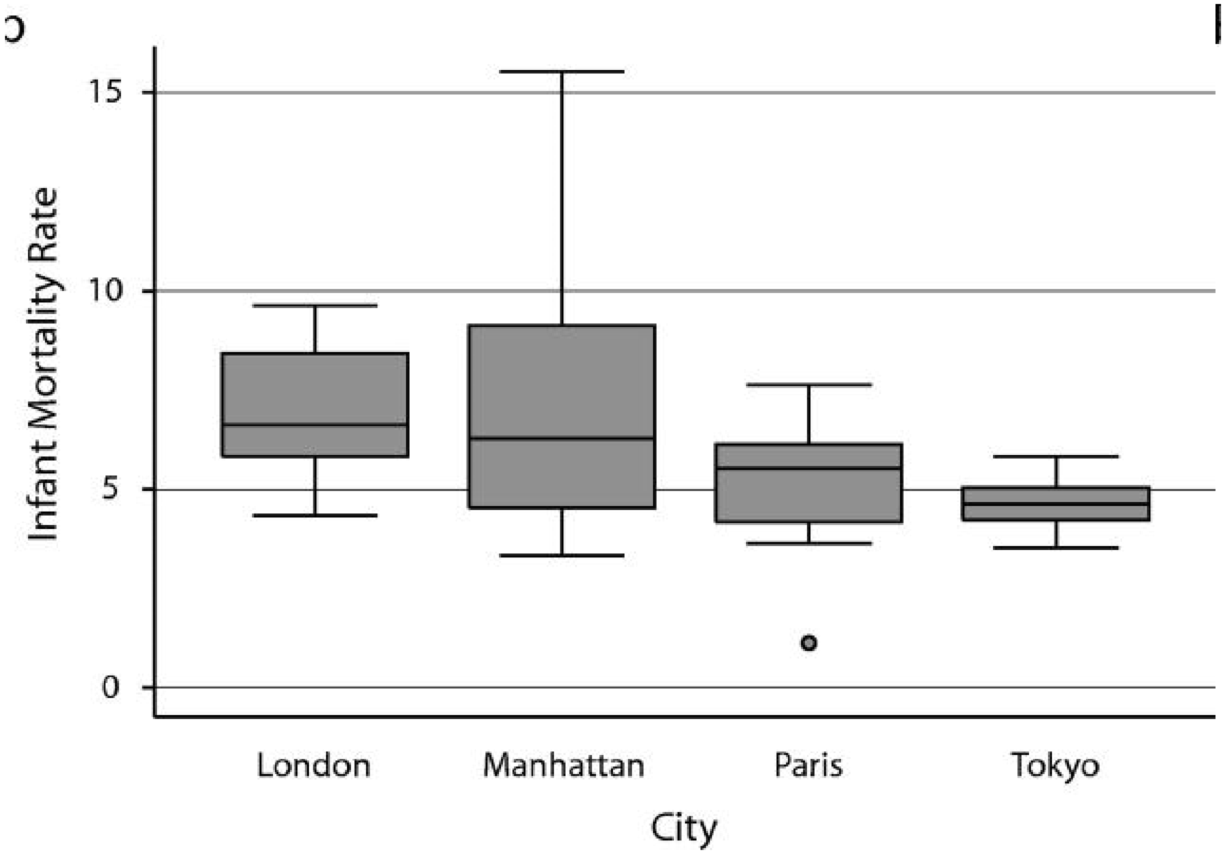
1. Q3 + Q1
2. 1.5*(Q3 – Q1)
3. ¾ * (Max – Min)
4. Q3 – Q1

Reading Quantitative Displays of Information?

****Box Plots***

Box plots of neighborhood infant mortality rate distributions for London, Manhattan, Paris, and Tokyo for 1993–1997 (Rate per 1000 live births).

Source: Am J Public Health. 2005 January; 95(1): 86–90.
doi: 10.2105/AJPH.2004.040287



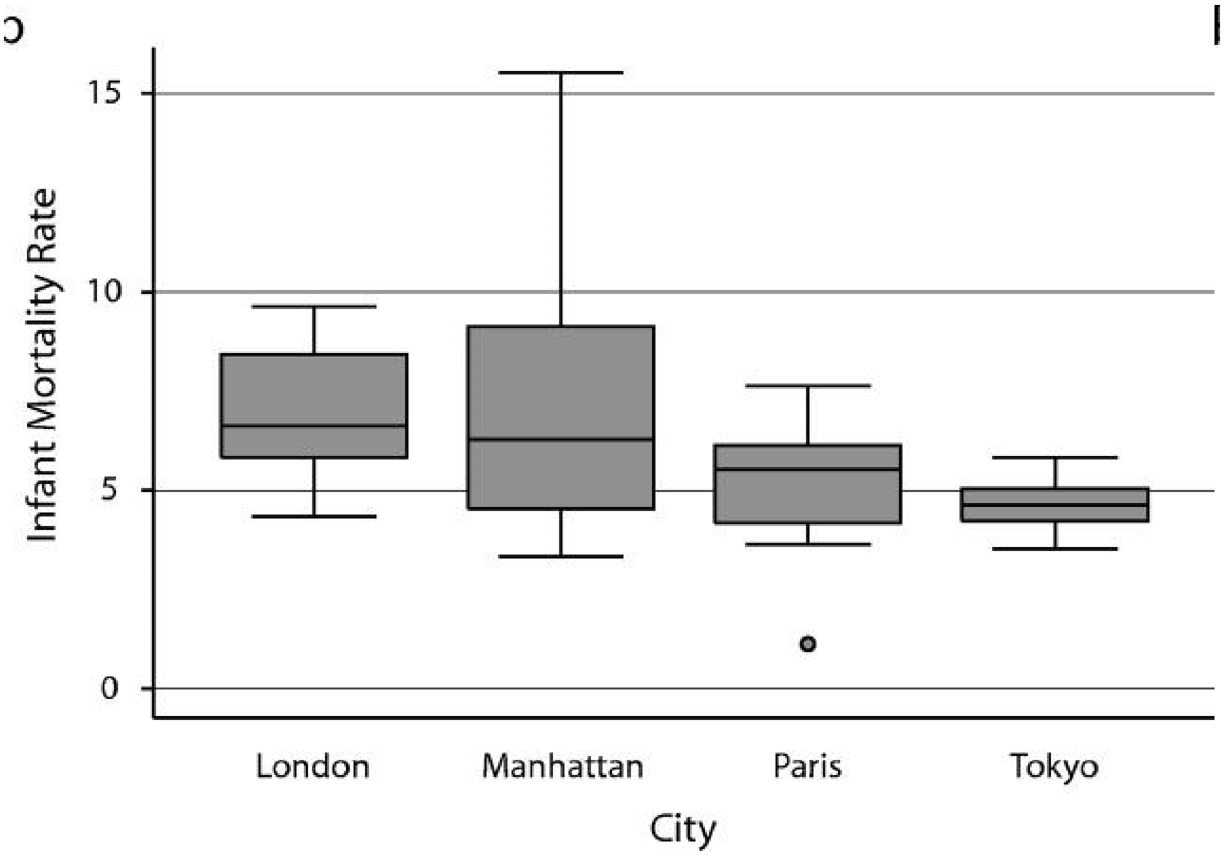
The median mortality rate is highest for which city?

The median mortality rate is highest for which city?

1. London
2. Manhattan
3. Paris
4. Tokyo

Box plots of neighborhood infant mortality rate distributions for London, Manhattan, Paris, and Tokyo for 1993–1997 (Rate per 1000 live births)..

Source: Am J Public Health. 2005 January; 95(1): 86–90.
doi: 10.2105/AJPH.2004.040287



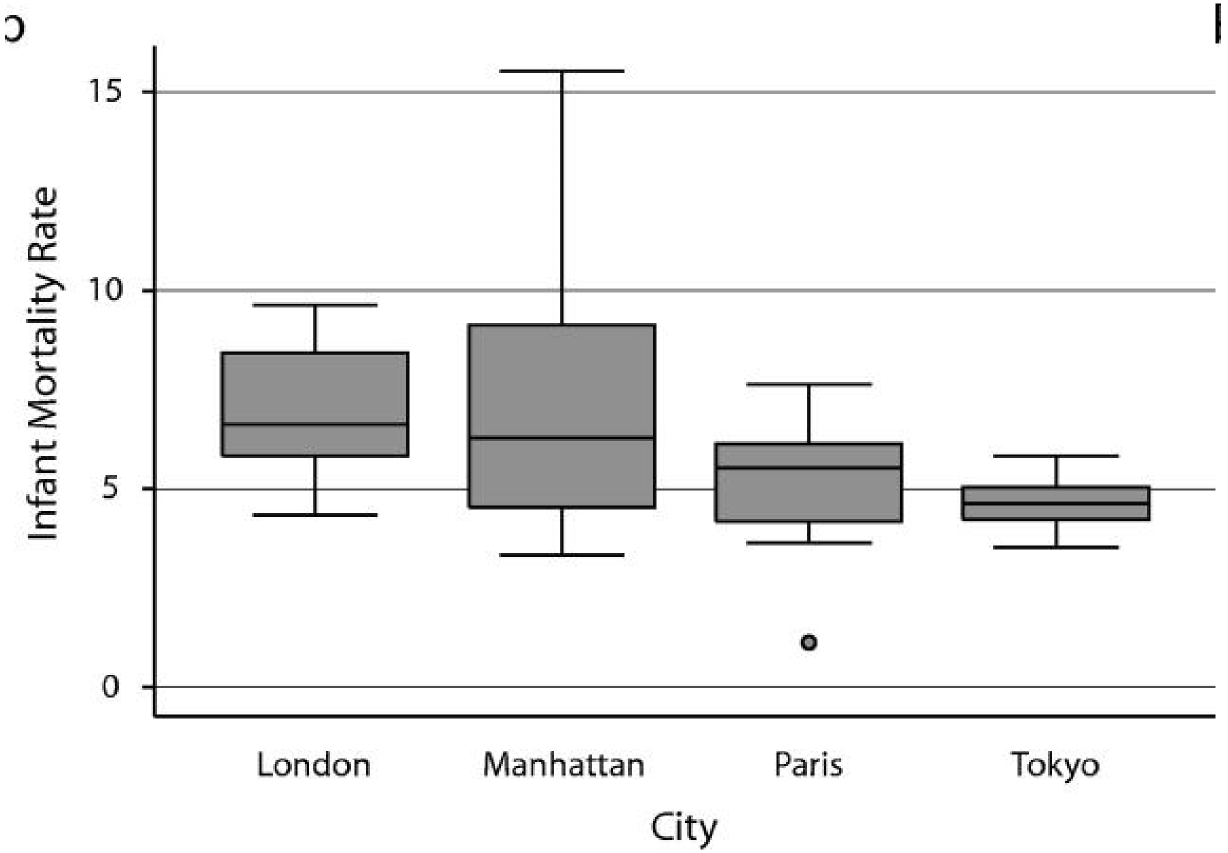
Which city has the most variability in infant mortality?

Which city has the most variability in infant mortality?

1. London
2. Manhattan
3. Paris
4. Tokyo

Box plots of neighborhood infant mortality rate distributions for London, Manhattan, Paris, and Tokyo for 1993–1997 (Rate per 1000 live births)..

Source: Am J Public Health. 2005 January; 95(1): 86–90.
doi: 10.2105/AJPH.2004.040287



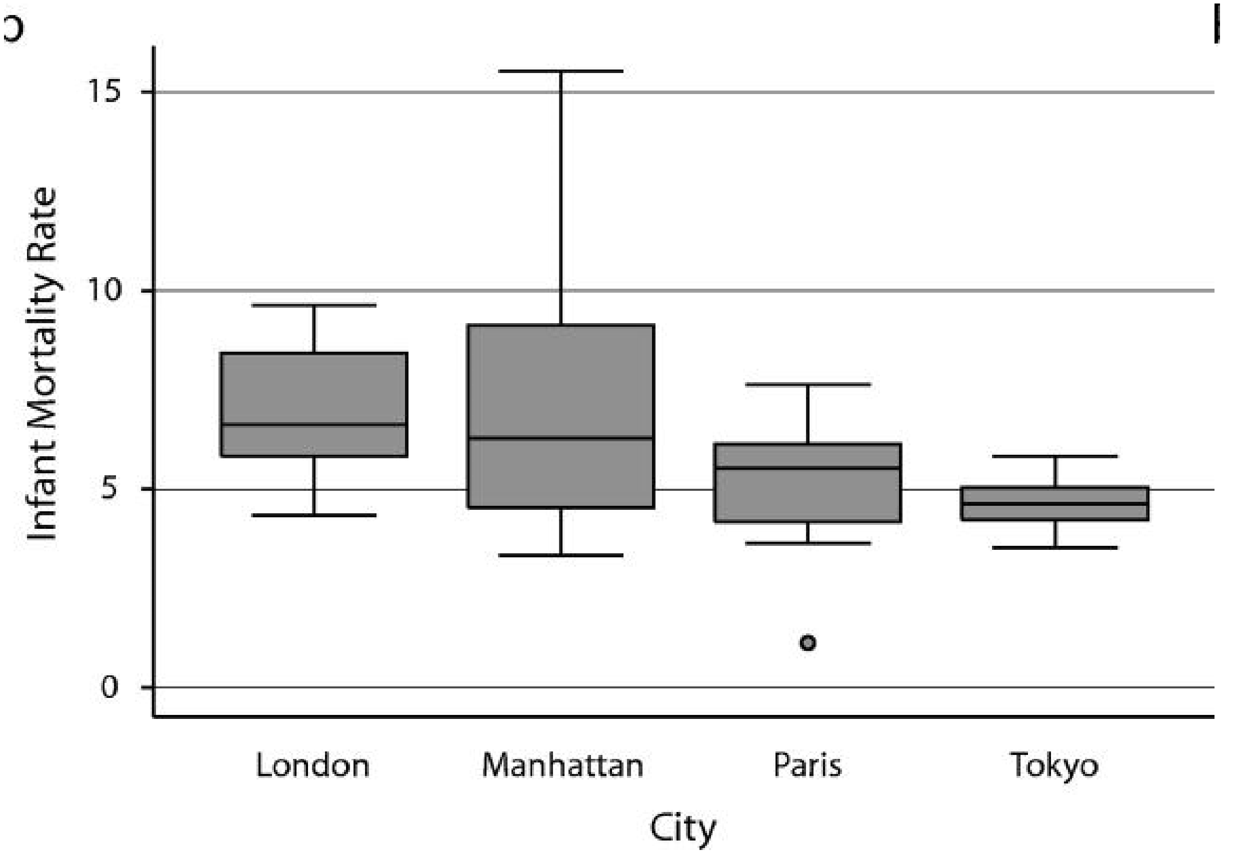
The upper quartile (Q3) for Tokyo is?

The upper quartile (Q3) for Tokyo is?

1. 7per 1,000 live births
2. 8 per 1,000 live births
3. 6 per 1,000 live births
4. 5 per 1,000 live births
5. 4 per 1,000 live births

Box plots of neighborhood infant mortality rate distributions for London, Manhattan, Paris, and Tokyo for 1993–1997 (Rate per 1000 live births)..

Source: Am J Public Health. 2005 January; 95(1): 86–90.
doi: 10.2105/AJPH.2004.040287



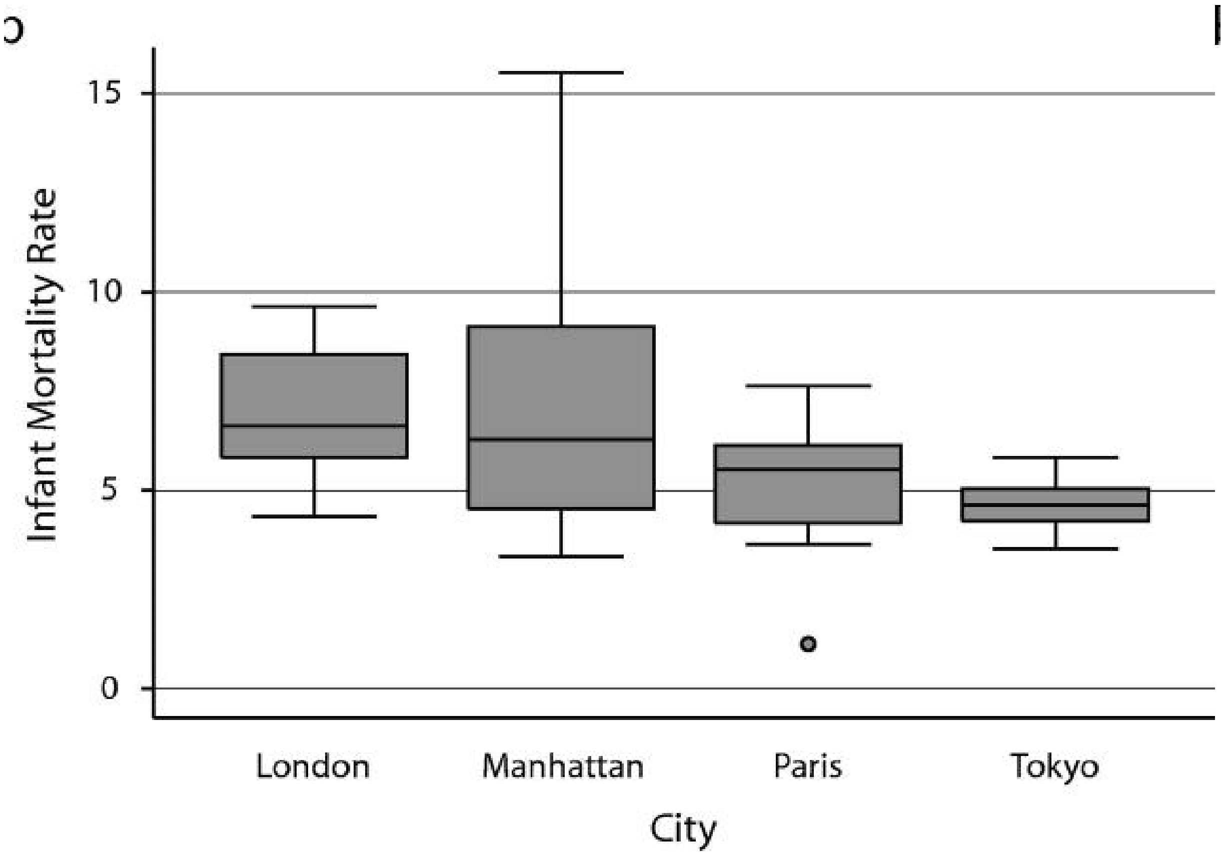
Which city has an outlier?

Which city has an outlier?

1. London
2. Manhattan
3. Paris
4. Tokyo

Box plots of neighborhood infant mortality rate distributions for London, Manhattan, Paris, and Tokyo for 1993–1997. (Rate per 1000 live births).

Source: Am J Public Health. 2005 January; 95(1): 86–90.
doi: 10.2105/AJPH.2004.040287



Can we conclude that Tokyo provides better maternal care?

Can we conclude that Tokyo provides better maternal care?

1. Yes
2. No

Overview of Exploratory Analysis for Quantitative Data

1. Summarize the data in frequency table
2. Plot the data (stem-and-leaf plot, histogram, frequency polygon, box-plot, frequency or percentage polygon).
3. Look for overall patterns (location,shape, spread, outliers). Is the distribution symmetric?
4. Investigate any outliers. Are these valid data points?
5. Calculate appropriate summary statistics of center and variability for the data.

Tables and Graphs in Excel

- Excel module 2 provides directions and examples for tables and graphs
- The FREQUENCY function can be used to generate data for a frequency table from raw data
- Use data from the frequency table to create
  - Histogram
  - Frequency or percentage polygon
  - Cumulative Relative Frequency graph
- There are no Excel functions for stem-and-leaf or box-plots

Percentiles in Excel

- The Cumulative Relative Frequency graph can be used to estimate percentiles of the data
- The PERCENTILE function in Excel can be used to calculate percentiles
- If the data are in cells A1:A100 percentiles can be found as follows
  - 95th percentile: =PERCENTILE(A1:A100, 0.95)
  - 50th percentile: =PERCENTILE(A1:A100, 0.50)
  - 5th percentile: =PERCENTILE(A1:A100, 0.05), etc

Readings and Assignments

- Reading: Chapter 3 pgs. 32 - 41
- Lesson 2 Practice Exercises: Tables and Graphs
- Excel Module 2: Tables and Graphs
- Homework 1: Problem 3 (3.2d) and Problem 4