Module 10: Summarizing Numbers

This module presents the standard summarizing numbers, also often called sample statistics or point estimates, that are essential to using and understanding data. Module 10 focuses on measures of central tendency, including means, medians, modes, midranges and geometric means. A discussion of Percentiles and Box plots are also included.

Measures of Central Tendency

- Mean—Average
- Median—Middle
- Mode—Most frequent
- Midrange—Halfway between smallest, largest
- Geometric Mean—Uses logarithms

Sample 1

Person	x_i
1	18
2	19
3	20
4	21
5	22

Mean or Average

- The mean or average is obtained by adding up the values for all the observations and then dividing by the number of observations
- In general, the mean is the best measure of central tendency to use, but there are exceptions

Calculating the Mean

Sample 1	Sample 2
x_1	x_2
18	90
19	4
20	3
21	2
22	1
$Sum_1 = \sum_{i=1}^{n} x_{1i}$	$Sum_2 = \sum_{i=1}^{n} x_{2i}$
$\overline{x}_1 = \sum_{i=1}^{5} x_{1i} / 5$	$\overline{x}_2 = \sum_{i=1}^{5} x_{2i} / 5$

Mean for Sample 1

Person	x_i
1	18
2	19
3	20
4	21
5	22
Sum (x_i)	100
Mean	20.0

$$\sum_{i=1}^{n} x_i = Sum(x_i)$$

$$Mean = \bar{X} = \sum_{i=1}^{n} x_i / n$$

Mean for Sample 2

Person	x_i
1	90
2	4
3	3
4	2
5	1
$\text{Sum}(x_i)$	100
Mean	20

Median

- The median is the “middle” observation when the complete list of observations is sorted in order.
- When there is a odd number of observations, the value of the middle one is the median.
- When there is a even number of observations, the value of the average of the two “middle” observations is used as the median.
- The median may be a better indication of the center of a group of numbers if there are some values that are considerably more extreme than others
- Median income is often used for this reason

Median for Sample 1

Person	x_i	
1	18	
2	19	
3	20	← **Median**
4	21	
5	22	
Sum (x_i)	100	
Mean	20	

Median for Sample 2

Person	x_i	
1	90	
2	4	
3	3	← **Median**
4	2	
5	1	
Sum (x_i)	100	
Mean	20	

Midrange

- The value of the point halfway between the smallest and the largest observations.
- Easily calculated by calculating the average of the values for the smallest and largest observations.
- Note that the value of the midrange need not be a number that is a value for one of the observations.

Midrange for Sample 1

Person	x_i	
1	18	
2	19	
3	20	← Midrange
4	21	
5	22	
Sum (x_i)	100	
Mean	20.0	

Midrange = (18 + 22)/2 = 20

Midrange for Sample 2

Person	x_i
1	90
2	4
3	3
4	2
5	1
Sum(x_i)	100
Mean	20

Midrange = (90 + 1)/2 = 45.5

Mode

- Value of observation that occurs most frequently.
- Represents a number that does occur in the observations.
- Not always well-defined since there may not be one value that occurs most frequently.

Mode for Sample 1

Person	x_i
1	18
2	19
3	20
4	21
5	22
Sum (x_i)	100
Mean	20.0

No mode since all values occur equally frequently

Mode for Sample 2

Person	x_i
1	90
2	4
3	3
4	2
5	1
Sum(x_i)	100
Mean	20

No mode since all values occur equally frequently

Geometric Mean

- First, take log for each sample point
- Second, calculate mean for log values
- Convert mean of log values back to original scale

Geometric Mean for Sample 1

Person	Age	$\log_{10}$(age)	$\log_e$(age)
1	18	1.26	2.89
2	19	1.28	2.94
3	20	1.30	3.00
4	21	1.32	3.04
5	22	1.34	3.09
$\sum$	100	6.50	14.97
$\overline{x}$	20	1.30	2.99
GM	-	19.95	19.89

$$10^{1.30} = 19.95 \qquad e^{2.99} = 19.89$$

Geometric Mean for Sample 2

Person	Age	$\log_{10}$(age)	$\log_e$(age)
1	90	1.95	4.50
2	4	0.60	1.39
3	3	0.48	1.10
4	2	0.30	0.69
5	1	0.00	0.00
$\sum$	100	3.33	7.68
$\bar{x}$	20	0.67	1.54
GM	-	4.68	4.64

$$10^{0.67} = 4.68 \qquad e^{1.54} = 4.64$$

Measures of Central Tendency

Measure	Sample 1	Sample 2
Mean	20.0 years	20.0 years
Median	20 years	3 years
Mode	none	none
Midrange	20 years	45.5 years
Geometric Mean	19.95 years	4.68 years

Knowing the Mean is not Enough

- What else would it be useful to know?
- A key issue is how alike or “unlike” each other the individual observations are
- How can we measure “unlikeness”

Percentiles

Percentiles are numbers that divide the data into 100 equal parts. For a set of observations arranged in order of magnitude, the p^{th} percentile is the value that has p percent of the observations below it and (100-p) percent above it. The most commonly used percentiles are the 25^{th}, 50^{th} and 75^{th} percentiles.

The 50^{th} percentile is that observation or number that has 50% of the observations below it and the other 50% above it ; this is simply the ‘middle’ observation when the set of observations are arranged in order of magnitude. The 50^{th} percentile is usually referred to as the median.

Example: Age Distribution from Module 9

The p^{th} percentile is the $\left(\frac{n*p}{100}\right)^{th}$ observation, when the set of observations are arranged in order or magnitude; where n is the sample size.

For the age distribution, n = 121;

The 75^{th} percentile for the age distribution is the (75 *121)/100 = 90.75 ~ 91^{st} observation when the ages are arranged in an increasing order of magnitude. The 75^{th} percentile of the ages is therefore 31 years; the 25^{th} percentile, 50^{th} and 80^{th} percentile are the 31^{st}, 61^{st}, and 97^{th} observations respectively, as shown on the next slide.

Age	Frequency	Cumulative Frequency
21	6	6
22	16	22
23	11	33
24	9	42
25	17	59
26	13	72
27	6	78
28	5	83
29	4	87
30	3	90
31	1	91
32	4	95
33	3	98
34	2	100
35+	21	121
Total	121	

25th percentile →

50th percentile →

75th percentile →

80th percentile →

The 31st observation falls in this group

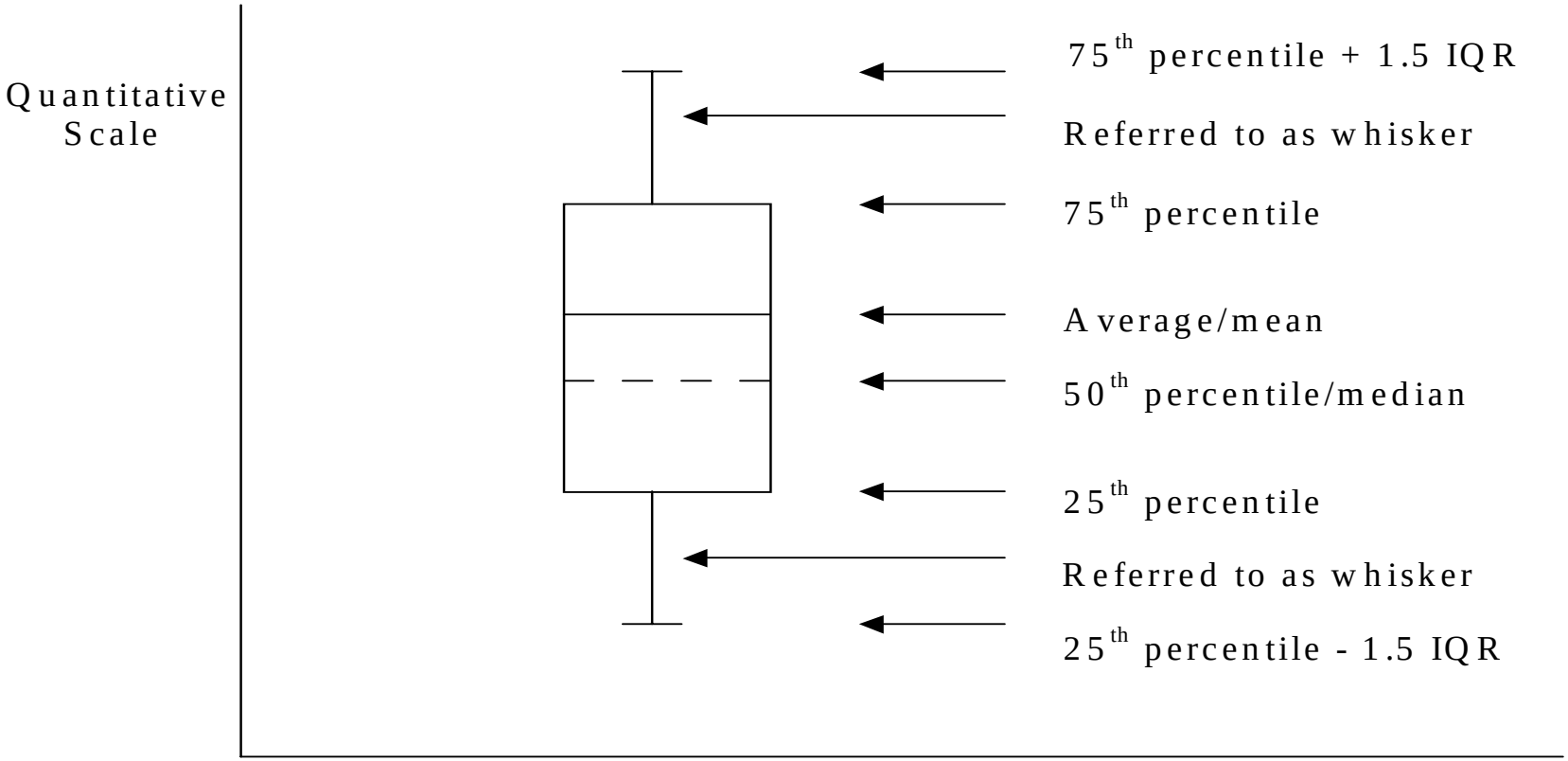
The 61st observation falls in this group

The 97th observation falls in this group

Box Plot

An individual box symbol summarizes the distribution of data within a data set. By using a box symbol, in addition to the average value, other information about the distribution of the measurements can also be displayed. As shown on the next slide, the 25^{th}, 50^{th} (Median), and 75^{th} percentile of the distribution can be displayed along with the average (mean) value of the distribution.

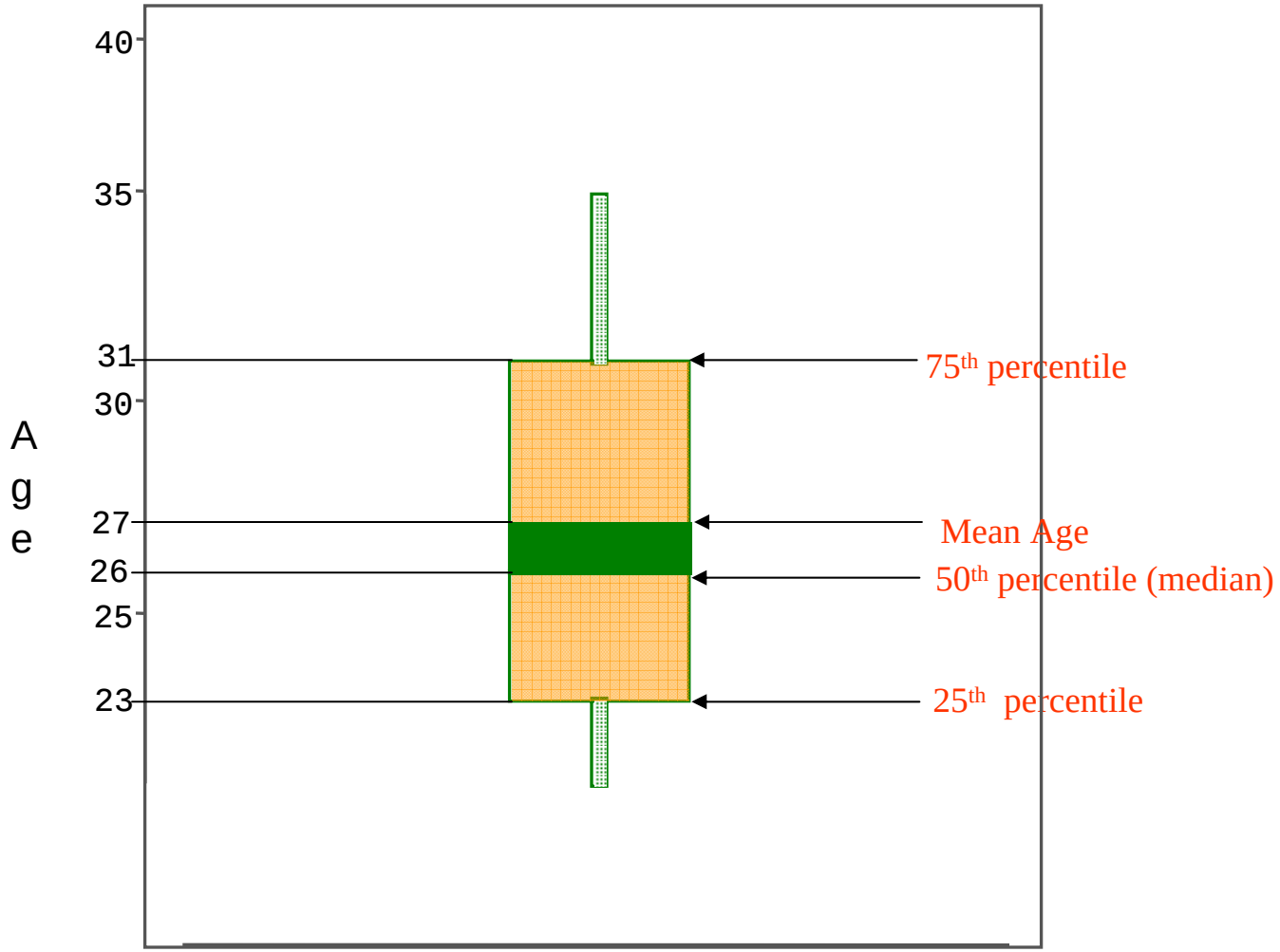
Box Plot



Individual box symbol

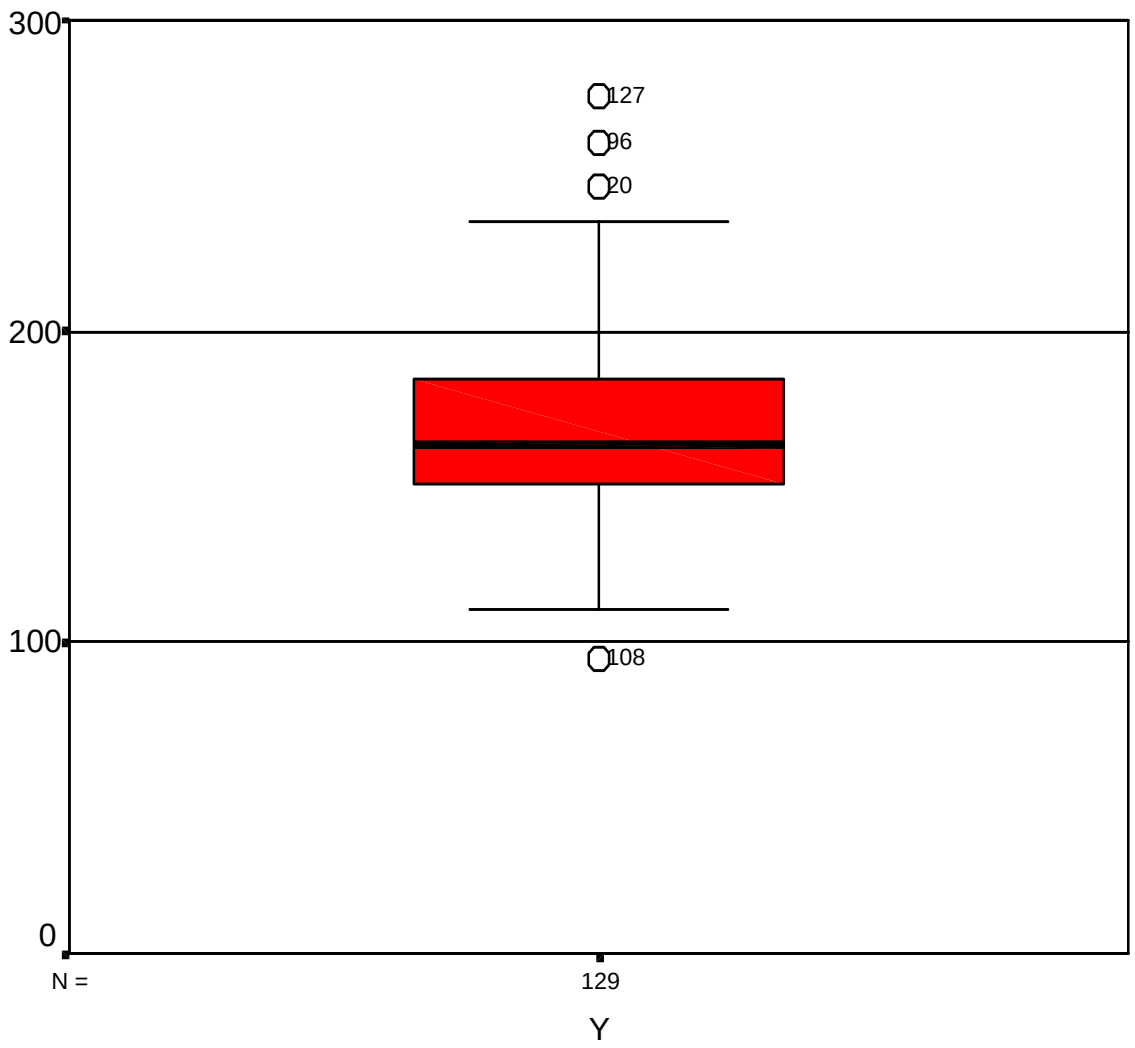
IQR: Interquartile range, which is calculated by substracting the 25th percentile of the data from 75th percentile; consequently, it contains the middle 50% of the observations.

Box Plot for Age distribution



SAS generated Box plot

Box plot from SPSS



Box plot from ViSta (The Visual Statistics System)