EIN 5226

Chapter 3
Measurements and
S4/IEE Measure Phase

Measurements I
Sections 3.5-8

Karen E. Schmahl Ph.D., P.E.

3.5 Sampling

- Population parameters
- Sample statistics
- Simple Random Sampling
- Sampling error
- Confidence interval

Population vs. Sample

Population

– entire group of interest in an analysis

Sample

– subset of items selected from population

- Sample of size n is taken

3

Sampling

Sample

- Subset of n items
- Should represent the population

Simple random sample

- sample chosen by a method in which each collection of population items is equally likely to comprise the sample.

Sampling

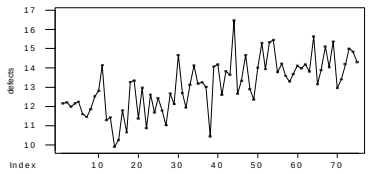
- Samples must come from a well-defined and stable population
  - Determine by initial examination of data
  - If taken over time, must not show cycle or trends

Line Graphs or (Run Charts)

- Generally used to depict time related trends
- Look for stability within system over time prior to doing statistical analysis

6

Run Chart/Time Series Chart



Is this a stable population?
Yes / No

7

3.6 Simple Graphic Presentations

- Stem and Leaf Diagram
  - Method of plotting data which displays data values as well as frequency
- Histogram
  - Bar graph displaying frequency of observations in a given bar or interval
- Dot Plot
  - Along a numbered line, a dot plot displays a dot for each observation.

Graphical Presentations

Graphical displays of data are important tools for investigating samples and populations.

Use to summarize data for easy understanding
- Location or central tendency
- Spread or variability
- Departure from symmetry, shape
- Identification of "outliers"

Stem-and-leaf Plot

- Method of plotting data which displays data values as well as frequency
- Each item in the sample is divided into two parts: a **stem**, consisting of the leftmost one or two digits, and the **leaf**, which consists of the next digits.

10

Stem & Leaf Plot

Sample observations

79	82
109	59
91	102
100	78
75	86
93	85
89	65
63	77
69	102
121	78
74	92
77	73
82	95

Ordered Data

59	82
63	85
65	86
69	89
73	91
74	92
75	93
77	95
77	100
78	102
78	102
79	109
82	121

Stem and Leaf Plot

"Stem"	"Leaf"
5	9
6	359
7	34577889
8	22569
9	1235
10	0229
11	
12	1

11

Frequency Table

Summarizes the data into groupings that show the frequency of data in each group

n=26

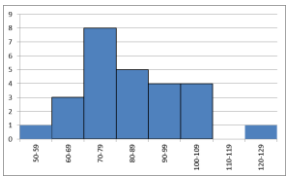
Stem & Leaf

Stem	Leaf
5	9
6	359
7	34577889
8	22569
9	1235
10	0229
11	
12	1

Class Interval	Frequency	Relative Frequency
50-59	1	0.0385
60-69	3	0.1154
70-79	8	0.3077
80-89	5	0.1923
90-99	4	0.1538
100-109	4	0.1538
110-119	0	0.0000
120-129	1	0.0385

Histograms

Stem	Leaf
5	9
6	359
7	34577889
8	22569
9	1235
10	0229
11	
12	1



A histogram is a graphical summary of the frequency of observations in a set of data placed into defined intervals.

Creating a Histogram

- Determine the number of classes to use, and construct class intervals of equal width.
  - Rule of thumb – at least 5 and no more than 15
  - Larger sample sizes -typically have more intervals
- Compute the frequency and relative frequency for each class.
- Draw a rectangle for each class. The heights of the rectangles may be set equal to the frequencies or to the relative frequencies

Creating a Histogram

- Determine the number of classes to use, and construct class intervals of equal width.
  - Rule of thumb – at least 5 and no more than 15
  - Larger sample sizes -typically have more intervals
- Compute the frequency and relative frequency for each class.
- Draw a rectangle for each class. The heights of the rectangles may be set equal to the frequencies or to the relative frequencies

True or False: The chart/bars will look exactly the same whether you use the frequency or relative frequency.

Creating a histogram

Gather data

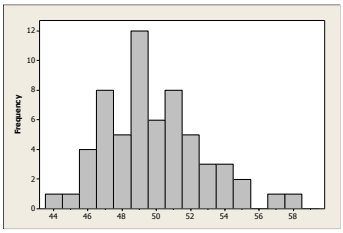
52	47	54	55	49	51	51	50	51	52	49	49	49	48	52
47	49	49	50	51	51	58	51	47	47	49	48	50	49	49
45	49	47	52	46	48	49	52	47	47	50	48	46	57	54
46	47	51	50	44	48	54	55	53	51	46	53	50	53	49

Determine interval size
Often helpful to calculate Descriptive Statistics:

N	Mean	StDev	Median	Min	Max	Range
60	49.850	2.893	49.000	44	58	14

Logical interval options: 1 (14 classes) or 2 (7 classes)

Histogram with Interval=1

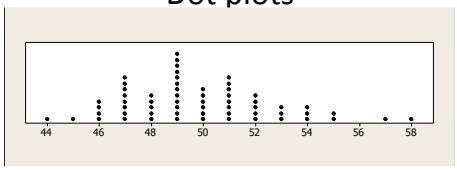


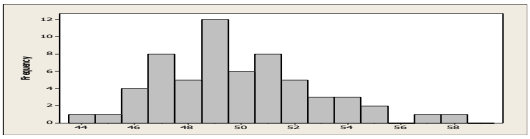
Dotplot

- Along a numbered line, a dot plot displays a dot for each observation
- It is useful when the sample size is not too large and when the sample contains some repeated values.
- Not generally used in formal presentations.

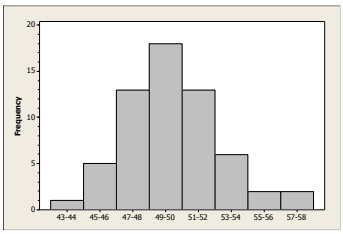
18

Dot plots





Histogram with Interval = 2



3.8 Sample Statistics

Measures of central tendency

- Sample Mean ($\bar{x}$)
- Sample Median (x_{50}, $\tilde{x}$)
- Mode

Measures of dispersion

- Range (R)
- Standard Deviation (s)
- Variance (s^2)

Sample statistics vs. population parameters

Population – entire group of interest in an analysis

Parameter – descriptive number calculated from entire populations values

Sample – subset of items selected from population

Statistic – any descriptive value calculated from the sample group's observations

22

The Arithmetic Mean

Sample Mean = $\bar{x} = \frac{\sum_{n}^{i=1} x_i}{n}$

Where x_i = value of the i^{th} individual observation

n = number of observations in sample

Population Mean = $u = \frac{\sum_{N}^{i=1} x_i}{N}$

Where x_i = value of the i^{th} individual observation

N = number of observations in population

23

Mean

Observation i	Value x
1	13
2	21
3	24
4	12
5	12
6	15
7	19
8	11
9	13
10	13
11	18
12	17
sum	188

- Calculate the mean of the sample

$$\bar{x} = \frac{\sum_{n}^{i=1} x_i}{n}$$

$$\bar{x} = \frac{188}{12} = 15.67$$

Median

Median – middle most value

- Arrange data order from smallest to largest, count to middle number
- If even number, average the two middle numbers

Example:

Previous set of data is arranged in ascending order

Average 2 middle value

$$\bar{x} = \frac{13 + 15}{2} = 14$$

11	1
12	2
12	3
13	4
13	5
13	6
15	6
17	5
18	4
19	3
21	2
24	1

25

Median using the Stem and Leaf

Median – middle most value

- Arrange data order from smallest to largest, count to middle number
- If even number, average the two middle numbers

In the stem and leaf plot, we have a set of ordered data.

For this data set, where n=26, what is the median?

A. 78.5 B. 79 C. 80.5 D. 82

		# in class
5	9	1
6	359	3
7	34577889	8
8	22569	5
9	1235	4
10	0229	4
11		0
12	1	1

Quartiles

50% Quartile – Median

25% Quartile –

Order data
Compute value .25 (n+1)
Count to this observation
If between observations, average the two values

Example - .25 (12+1) = 3.25

$$\bar{x} = \frac{12 + 13}{2} = 12.5$$

order	
1	11
2	12
3	12
4	13
5	13
6	13
7	15
8	17
9	18
10	19
11	21
12	24

25% 12.5

50% 14

27

Percentiles

50% Quartile = Median
= 50th percentile

pth percentile

Order data
Compute value (p/100) (n+1)
Count to this observation
If between observations,
average the two values

11
12
12
13
13
13
15
17
18
19
21
24

25%
12.5

50%
14

28

Percentiles

pth percentile

Order data
Compute value (p/100) (n+1)
Count to this observation
If between observations,
average the two values

What is the 60th percentile?

A. 15.5 B. 16 C. 16.5
D. 17 E. 17.5 F. 18

order	
1	11
2	12
3	12
4	13
5	13
6	13
7	15
8	17
9	18
10	19
11	21
12	24

29

Mean or Median

Housing Market for Miami

Average listing price for homes: $556,568
(Week ending Jun 26 2013, 8094 homes for sale)

Median Sales Price: $160,000
(March 13-May 13, 9289 recently sold homes)

Which is more reflective of the cost of housing?

Mode

Mode

– most frequent observation

- may also be used with non-numeric data

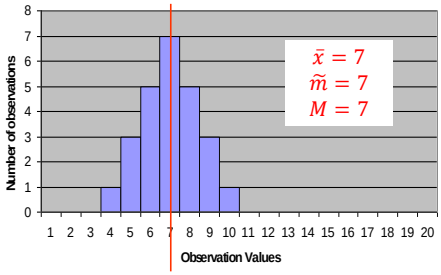
- may not exist or may have multiple

11
12
12
13
13
13
15
17
18
19
21
24

Mode

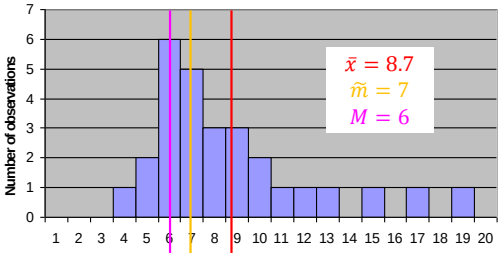
31

Shape of the distribution



With a symmetrical distribution,
the mean, medium and mode are the same

32

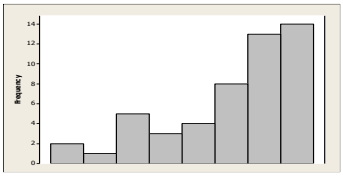


Skewed data – non-symmetrical - more data to one side of mean

Skewed to the right or positively skewed – more data on the right
Mean > Medium

33

Skewed Distributions



The above distribution is skewed left, or negatively skewed.

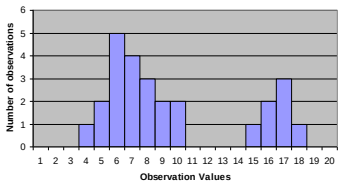
For this distribution,

A. The mean will be greater than the medium
B. The mean will be less than the medium

Mode – most frequent observation

Unimodal histogram– has only one peak

Bimodal histogram – has two clearly distinct modes



35

15.4 Box Plot

- A box plot (or box-and-whisker plot) is useful for describing various aspects of data pictorially.
- Box plots can visually show differences between characteristics of a data set.
- Common characteristics of a box plot:
  - the lower and upper quartiles (25th and 75th percentiles),
  - the median (the 50th percentile),
  - the interquartile range (IQR),
  - the minimum and maximum within 1.5 IQR of quartiles,
  - and the outliers.

Boxplots

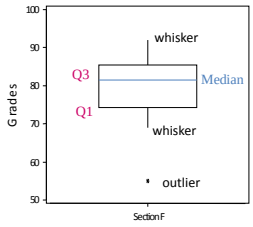
Data Set	
15	92
14	92
13	87
12	85
11	85
10	85
9	83
8	82
7	81
6	81
5	80
4	75
3	72
2	69
1	55

Q3= 85

Median = 82

Q1=75

Boxplot - graphical display of important quantitative information about a data set.



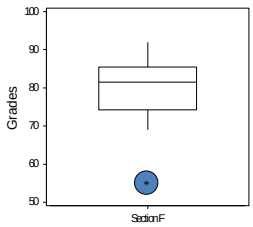
The Box Plot

Outliers - points beyond a distance of 1.5 (Q3-Q1) further than the 1st and 3rd quartiles

Data Set
92
92
87
85
85
85
83
82
81
81
80
75
72
69

Q3=85, Q1=75
Interquartile range = Q3-Q1=10

Outlier limit – low end
Q1 - 1.5 (Q3-Q1)
= 75-1.5(10)
= 75-15 = 60



55 Outlier

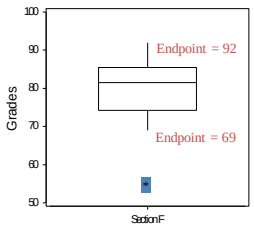
The Box Plot

Outliers - points beyond a distance of 1.5 (Q3-Q1)

Whisker end points - min and max points without outliers

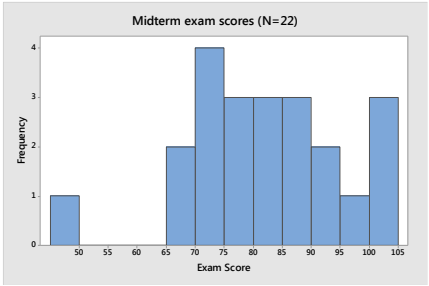
Data Set	
92	Endpoint
92	
87	
85	
85	
85	
83	
82	
81	
81	
80	
75	
72	
69	Endpoint

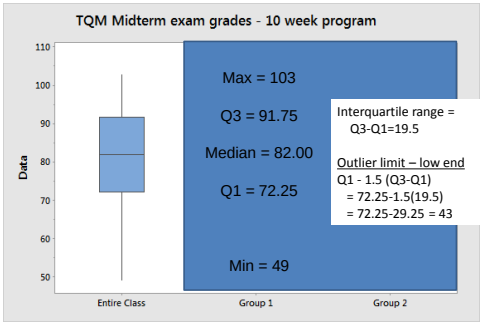
55 Outlier

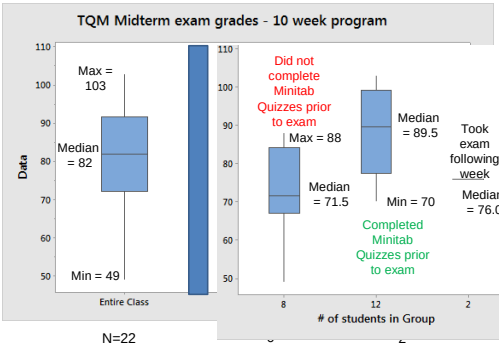


5226 Midterm grades – Spring 2015

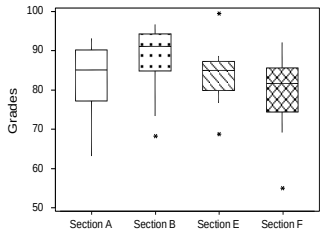
(10 week course format)







Comparison of grades for different sections



Boxplot Comprehension Questions

Answer the following questions using the box plot provided.

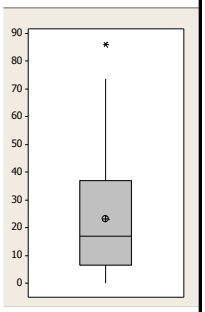
T / F The first quartile is at approximately 37.

T / F Interquartile range is approximately 30.

T / F There are no outliers in the data.

T / F The upper whisker is a bit longer than the lower one, indicating that the data has a slightly longer upper tail than lower tail.

T / F The boxplot suggests that the data are skewed to the left.



Measures of Variation

- Range = Largest – smallest
- Standard Deviation and Variance
  - Essentially a measure of the average difference of value in the sample from the sample mean
  - Formulas vary for sample and population
  - Standard deviation is the square of the variance

45

Standard Deviation

Sample: standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

46

x	$x - \bar{x}$	$(x - \bar{x})^2$
13	-2.67	7.11
21	5.33	28.44
24	8.33	69.44
12	-3.67	13.44
12	-3.67	13.44
15	-0.67	0.44
19	3.33	11.11
11	-4.67	21.78
13	-2.67	7.11
13	-2.67	7.11
18	2.33	5.44
17	1.33	1.78
188	0.00	186.67

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

n=12

$$\bar{x} = \frac{\sum x}{n} = \frac{188}{12} = 15.67$$

$$s = \sqrt{\frac{186.67}{12-1}} = 4.12$$

Standard Deviation

Population: standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Sample: standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Standard Deviation and Variance

Population:

standard deviation $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

variance $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$

Sample:

standard deviation $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$

variance $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$

49

Related Assignments

Please see Blackboard for related assignments.