BAN 602: Quantitative Fundamentals

Spring, 2020 Lecture Slides – Week 1

CAL STATE
EAST BAY

Agenda

- Introductions
- Data and Statistics
- Descriptive Statistics: Tabular and Graphical Displays
- Descriptive Statistics: Numerical Measures

Statistics and Its Applications

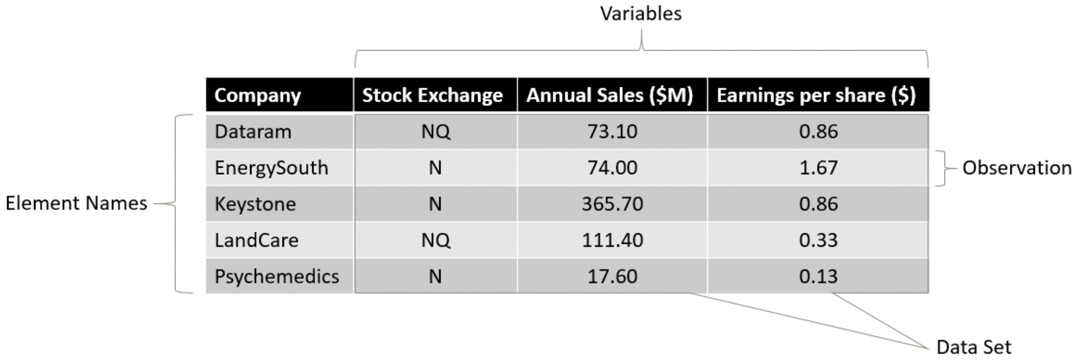
- The term statistics can refer to *numerical facts* such as averages, medians, percentages, and maximums that help us understand a variety of business and economic situations.
- Statistics can also refer to the *art and science* of collecting, analyzing, presenting, and interpreting data.

Application of Statistics:

- Accounting: Public accounting firms use statistical sampling procedures when conducting audits for their clients.
- Economics: Economists use statistical information in making forecasts about the future of the economy or some aspect of it.
- Finance: Financial advisors use price-earnings ratios and dividend yields to guide their investment advice.
- Marketing: Electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.
- Production: A variety of statistical quality control charts are used to monitor the output of a production process.
- Information Systems: A variety of statistical information helps administrators assess the performance of computer networks.

Data, Elements, Variables, Observations

- Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- All the data collected in a particular study are referred to as the data set for the study.
- Elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements obtained for a particular element is called an observation.
- A data set with *n* elements contains *n* observations.
- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.



Company	Stock Exchange	Annual Sales ($M)	Earnings per share ($)
Dataram	NQ	73.10	0.86
EnergySouth	N	74.00	1.67
Keystone	N	365.70	0.86
LandCare	NQ	111.40	0.33
Psychemedics	N	17.60	0.13

Variables

Observation

Element Names

Data Set

CAL STATE EAST BAY

Scales of Measurement

- Scales of measurement include
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

1. Nominal scale

- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or numeric code may be used.

Example: Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business, 2 denotes Humanities, 3 denotes Education, and so on).

Scales of Measurement contd.

2. Ordinal scale

- The data have the properties of nominal data and the order or rank of the data is meaningful.
- A nonnumeric label or numeric code may be used.

Example: Students of a university are classified by their class standing using a nonnumeric label such as Freshman, Sophomore, Junior, or Senior.

Alternatively, a numeric code could be used for the class standing variable (e.g. 1 denotes Freshman, 2 denotes Sophomore, and so on).

3. Interval scale

- The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.
- Interval data are always numeric.

Example: Melissa has an SAT score of 1985, while Kevin has an SAT score of 1880. Melissa scored 105 points more than Kevin.

Scales of Measurement contd.

4. Ratio scale

- Data have all the properties of interval data and the ratio of two values is meaningful.
- Ratio data are always numerical.
- Zero value is included in the scale.

Example: Price of a book at a retail store is $200, while the price of the same book sold online is $100. The ratio property shows that retail stores charge twice the online price.

Categorical and Quantitative Data

- Data can be further classified as being categorical or quantitative.
- The statistical analysis that is appropriate depends on whether the data for the variable are categorical or quantitative.
- In general, there are more alternatives for statistical analysis when the data are quantitative.

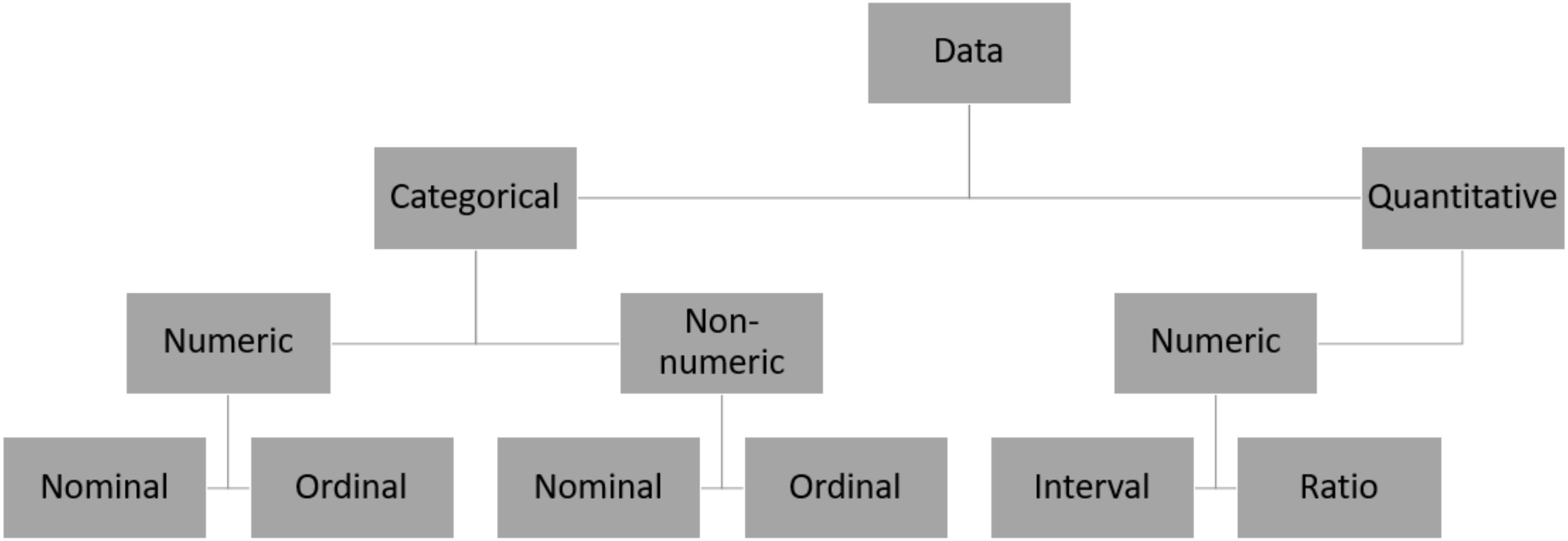
Categorical Data:

- Labels or names are used to identify an attribute of each element
- Often referred to as qualitative data
- Use either the nominal or ordinal scale of measurement
- Can be either numeric or nonnumeric
- Appropriate statistical analyses are rather limited

Quantitative Data:

- Quantitative data indicate how many or how much.
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

Scales of Measurement



Cross-Sectional vs Time Series Data

Cross-sectional data are collected at the same or approximately the same point in time.

Example: Data detailing the number of building permits issued in November 2013 in each of the counties of Ohio.

Time series data are collected over several time periods.

Example: Data detailing the number of building permits issued in Lucas County, Ohio in each of the last 36 months.

Graphs of time series data help analysts understand

- what happened in the past
- identify any trends over time, and
- project future levels for the time series

Data Sources

Existing Sources

- Internal company records – almost any department
- Business database services – Dow Jones & Co.
- Government agencies - U.S. Department of Labor
- Industry associations – Travel Industry Association of America
- Special-interest organizations – Graduate Management Admission Council (GMAT)
- Internet – more and more firms

Statistical Studies – Observational

- In observational (nonexperimental) studies no attempt is made to control or influence the variables of interest.
- Example – Survey. Studies of smokers and nonsmokers are observational studies because researchers do not determine or control who will smoke and who will not smoke.

Statistical Studies – Experimental

- In experimental studies the variable of interest is first identified. Then one or more other variables are identified and controlled so that data can be obtained about how they influence the variable of interest.
- The largest experimental study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine. Nearly two million U.S. children (grades 1- 3) were selected.

Descriptive Statistics

- Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy to understand.
- Such summaries of data, which may be tabular, graphical, or numerical, are referred to as descriptive statistics.

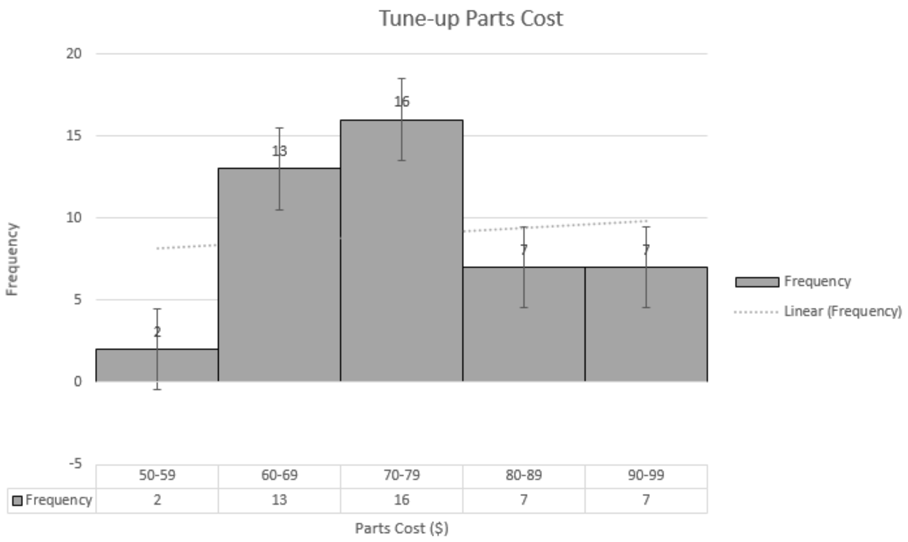
Example:

The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in her shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Tabular/Graphical/Numerical Summary: Frequency

Parts Cost ($)	Frequency	Percent Frequency
50-59	2	4%
60-69	13	26%
70-79	16	32%
80-89	7	14%
90-99	7	14%
100-109	5	10%
TOTAL	**50**	**100%**


Tune-up Parts Cost

	50-59	60-69	70-79	80-89	90-99
▨ Frequency	2	13	16	7	7

- Hudson’s mean cost of parts, based on the 50 tune-ups studied is $79 (found by summing up the 50 cost values and then dividing by 50).

Statistical Inference

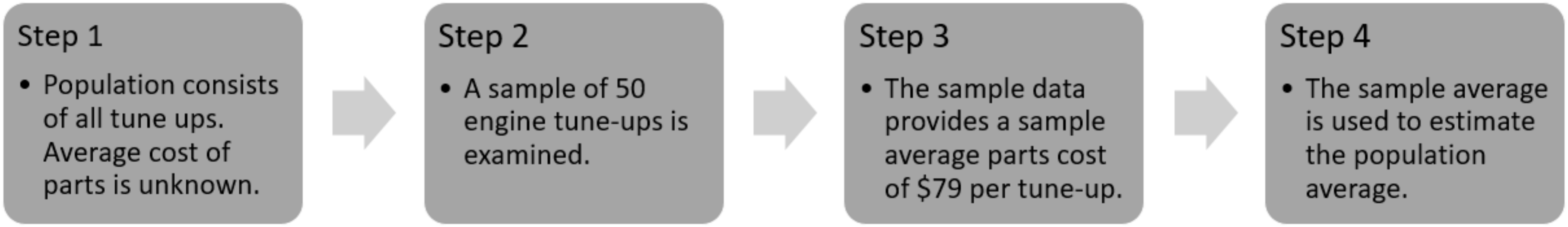
Population: The set of all elements of interest in a particular study.

Sample: A subset of the population.

Statistical inference: The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population.

Census: Collecting data for the entire population.

Sample survey: Collecting data for a sample.



Step 1

- Population consists of all tune ups. Average cost of parts is unknown.

Step 2

- A sample of 50 engine tune-ups is examined.

Step 3

- The sample data provides a sample average parts cost of $79 per tune-up.

Step 4

- The sample average is used to estimate the population average.

Analytics

Analytics is the scientific process of transforming data into insight for making better decisions.

Techniques:

- Descriptive analytics: This describes what has happened in the past.

- Predictive analytics: Use models constructed from past data to predict the future or to assess the impact of one variable on another.

- Prescriptive analytics: The set of analytical techniques that yield a best course of action.

Big Data and Data Mining:

Big data: Large and complex data set.

Three V’s of Big data:

⇒Volume : Amount of available data

⇒Velocity: Speed at which data is collected and processed

⇒Variety: Different data types

Data Warehousing and Data Mining

Data Warehousing: process of capturing, storing, and maintaining the data.

- Organizations obtain large amounts of data on a daily basis by means of magnetic card readers, bar code scanners, point of sale terminals, and touch screen monitors.
- Wal-Mart captures data on 20-30 million transactions per day.
- Visa processes 6,800 payment transactions per second.

Data Mining:

- Methods for developing useful decision-making information from large databases.
- Using a combination of procedures from statistics, mathematics, and computer science, analysts "mine the data" to convert it into useful information.
- The most effective data mining systems use automated procedures to discover relationships in the data and predict future outcomes prompted by general and even vague queries by the user.
- Statistical methodology such as multiple regression, logistic regression, and correlation are heavily used.
- Also needed are computer science technologies involving artificial intelligence and machine learning.
- A significant investment in time and money is required as well.

Ethical Guidelines for Statistical Practice

- In a statistical study, unethical behavior can take a variety of forms including:
  - Improper sampling
  - Inappropriate analysis of the data
  - Development of misleading graphs
  - Use of inappropriate summary statistics
  - Biased interpretation of the statistical results
- One should strive to be fair, thorough, objective, and neutral as you collect, analyze, and present data.
- As a consumer of statistics, one should also be aware of the possibility of unethical behavior by others.
- The American Statistical Association developed the report "Ethical Guidelines for Statistical Practice".
- It contains 67 guidelines organized into 8 topic areas:
  - Professionalism
  - Responsibilities to Funders, Clients, Employers
  - Responsibilities in Publications and Testimony
  - Responsibilities to Research Subjects
  - Responsibilities to Research Team Colleagues
  - Responsibilities to Other Statisticians/Practitioners
  - Responsibilities Regarding Allegations of Misconduct
  - Responsibilities of Employers Including Organizations, Individuals, Attorneys, or Other Clients

Summarizing Categorical Data

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Chart
- Pie Chart

Frequency Distribution

A frequency distribution is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.

Example: Marada Inn

Guests staying at Marada Inn were asked to rate the quality of their accommodations as being *excellent*, *above average*, *average*, *below average*, or *poor*.

Rating	**Frequency**
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

Relative Frequency and Percent Frequency Distributions

- The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class.

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

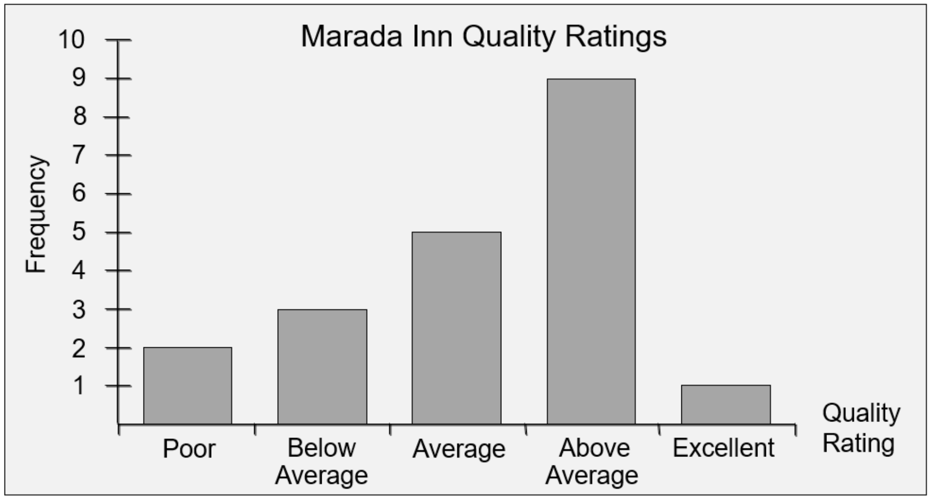
- The percent frequency of a class is the relative frequency multiplied by 100.

Example: Marada Inn

Rating	Relative Frequency	Percent Frequency
Poor	0.10	10%
Below Average	0.15	15%
Average	0.25	25%
Above Average	0.45	45%
Excellent	0.05	5%
Total	1.00	100%

Bar Chart

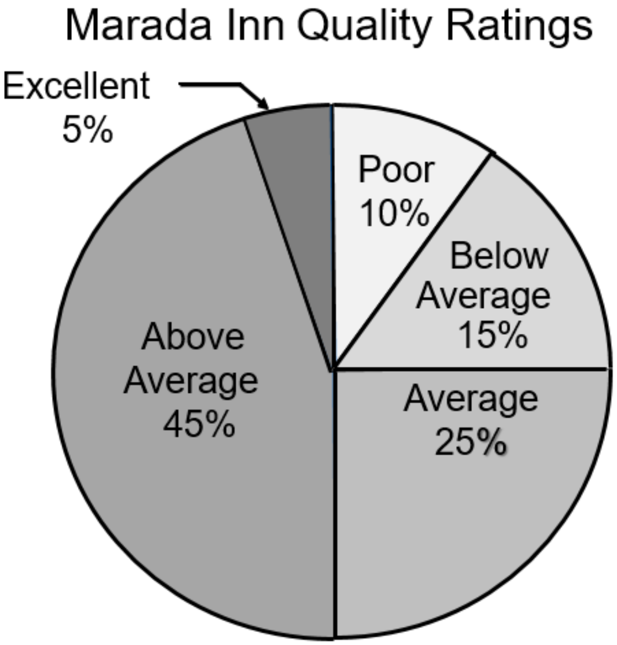
- A bar chart is a graphical display for depicting qualitative data.
- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).
- Using a bar of fixed width drawn above each class label, we extend the height appropriately.
- The bars are separated to emphasize the fact that each class is a separate category.



© Somak Paul

Pie Chart

- The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for categorical data.
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.
- Because there are 360 degrees in a circle, a class with a relative frequency of 0.25 would consume 0.25(360) = 90 degrees of the circle.
- Half of the customers surveyed gave Marada a quality rating of "above average" or "excellent" (look at the left side of the pie). This might please the manager.
- For each customer who gave an "excellent" rating, there were two customers who gave a "poor" rating (looking at the top of the pie). This should displease the manager.



© Somak Paul

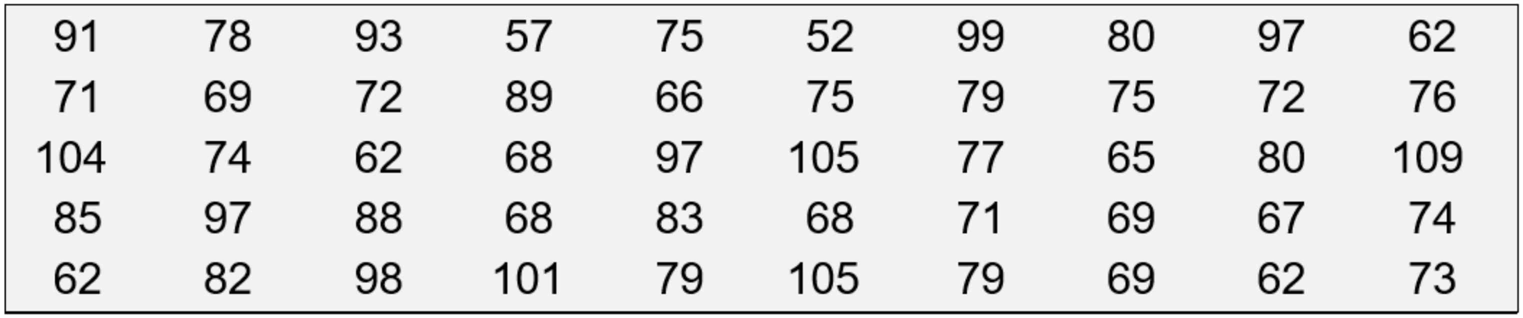
Summarizing Quantitative Data

- Frequency Distribution
- Relative Frequency and Percent Frequency Distributions
- Dot Plot
- Histogram
- Cumulative Distributions
- Stem-and-Leaf Display

Frequency Distribution – Quantitative Data

The manager of Hudson Auto would like to gain a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are shown below.

Sample of Parts Cost($) for 50 Tune-ups

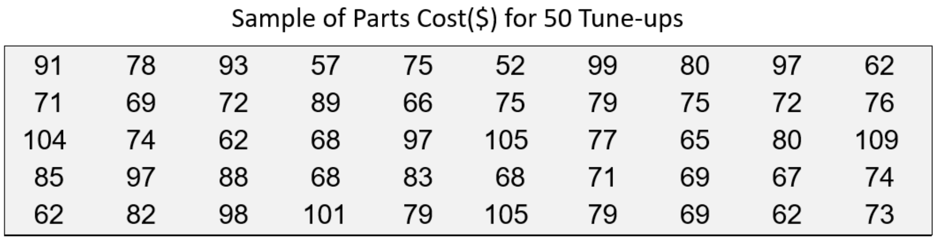


91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Frequency Distribution – Quantitative Data

Example: Hudson Auto Repair

If we choose six classes the approximate class width = (109 – 50)/6 = 9.83 or about 10.



Sample of Parts Cost($) for 50 Tune-ups

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Part Cost ($)	Frequency
50-59	2
60-69	13
70-79	16
80-89	7
90-99	7
100-109	5
Total	50

Relative Frequency and Percent Frequency Distributions

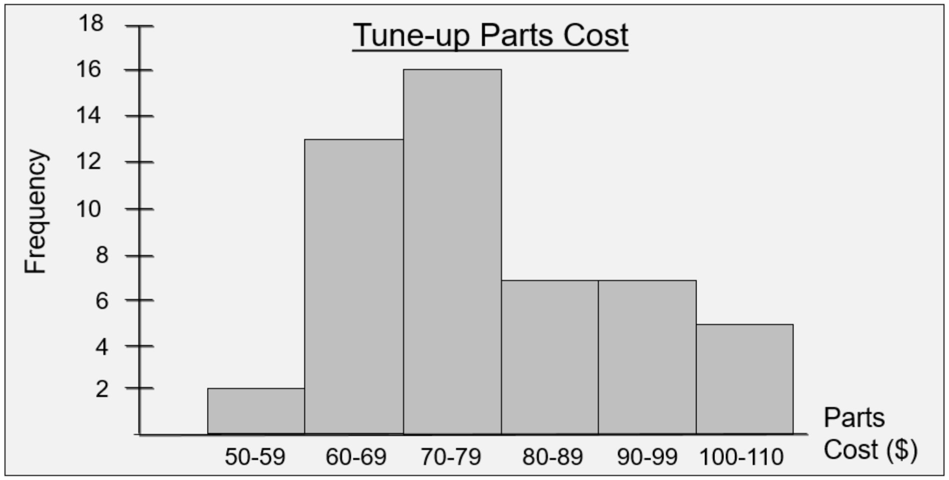
Insights

- Only 4% of the parts costs are in the $50-59 class.
- 30% of the parts costs are under $70.
- The greatest percentage (32% or almost one-third) of the parts costs are in the $70-79 class.
- 10% of the parts costs are $100 or more.

Parts Cost ($)	Relative Frequency	Percent Frequency
50-59	0.04 = 2/50	4 = .04(100)
60-69	0.26	26
70-79	0.32	32
80-89	0.14	14
90-99	0.14	14
100-109	0.10	10
Total	1.00	100

Histogram

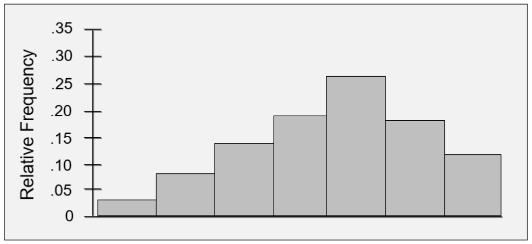
- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes.



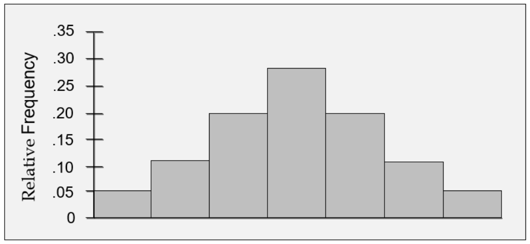
© Somak Paul

Histograms Showing Skewness

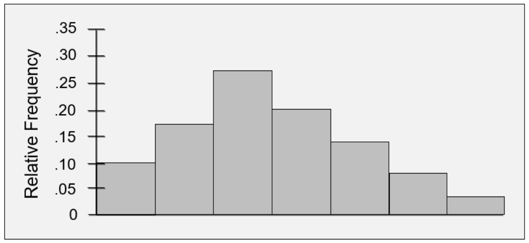
Moderately Skewed Left
A longer tail to the left
Ex: Exam Scores



Symmetric
Left tail is the mirror image of the right tail
Ex: Heights of People



Moderately Right Skewed
A Longer tail to the right
Ex: Housing Values



Cumulative Distributions

Cumulative frequency distribution - shows the *number* of items with values less than or equal to the upper limit of each class.

Cumulative relative frequency distribution – shows the *proportion* of items with values less than or equal to the upper limit of each class.

Cumulative percent frequency distribution – shows the *percentage* of items with values less than or equal to the upper limit of each class.

		Cumulative	Cumulative
	Cumulative	Relative	Percent
Cost ($)	Frequency	Frequency	Frequency
≤ 59	2	.04	4
≤ 69	15 = 2+13	.30 = 15/50	30 = .30(100)
≤ 79	31	.62	62
≤ 89	38	.76	76
≤ 99	45	.90	90
≤ 109	50	1.00	100

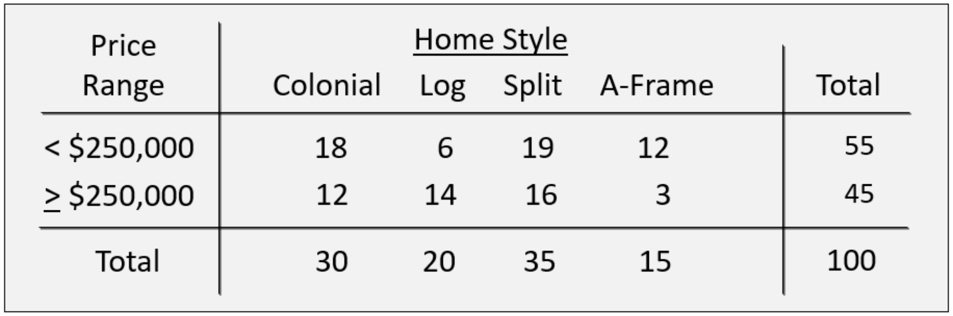
Summarizing Data for Two Variables using Tables

- Thus far we have focused on methods that are used to summarize the data for one variable at a time.
- Often a manager is interested in tabular and graphical methods that will help understand the relationship between two variables.
- Crosstabulation is a method for summarizing the data for two variables.
- Crosstabulation can be used when:
  - one variable is categorical and the other is quantitative,
  - both variables are categorical, or
  - both variables are quantitative.
- The left and top margin labels define the classes for the two variables.

Crosstabulation

Example: Finger Lakes Homes

The number of Finger Lakes homes sold for each style and price for the past two years is shown below.



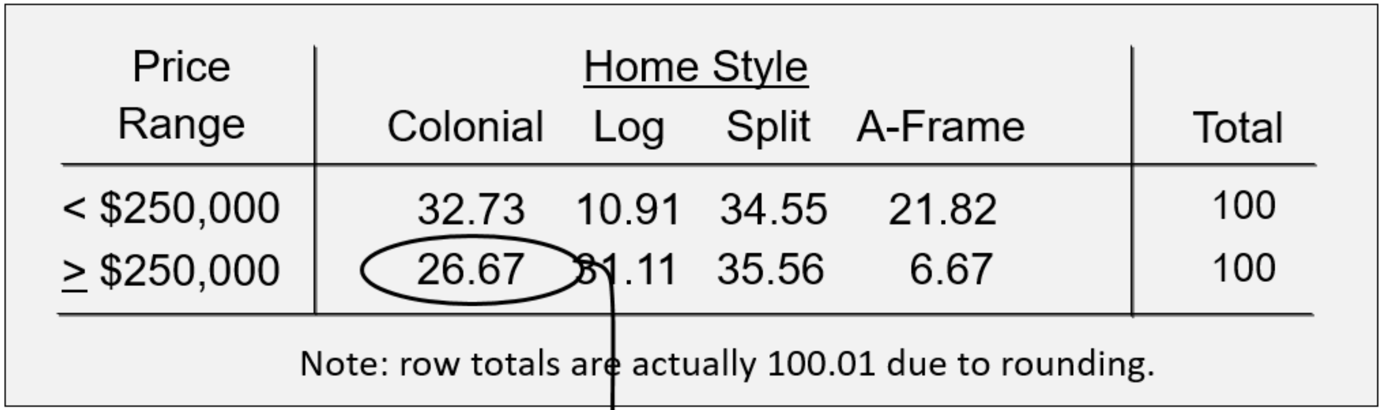
Price Range	Home Style				Total
	Colonial	Log	Split	A-Frame	
< $250,000	18	6	19	12	55
≥ $250,000	12	14	16	3	45
Total	30	20	35	15	100

Insights

- The greatest number of homes (19) in the sample are a split-level style and priced at less than $250,000.
- Only three homes in the sample are an A-Frame style and priced at $250,000 or more.

Crosstabulation: Row Percentages

Converting the entries in the table into row percentages or column percentages can provide additional insight about the relationship between the two variables.



Price	Home Style				
Range	Colonial	Log	Split	A-Frame	Total
< $250,000	32.73	10.91	34.55	21.82	100
≥ $250,000	26.67	31.11	35.56	6.67	100

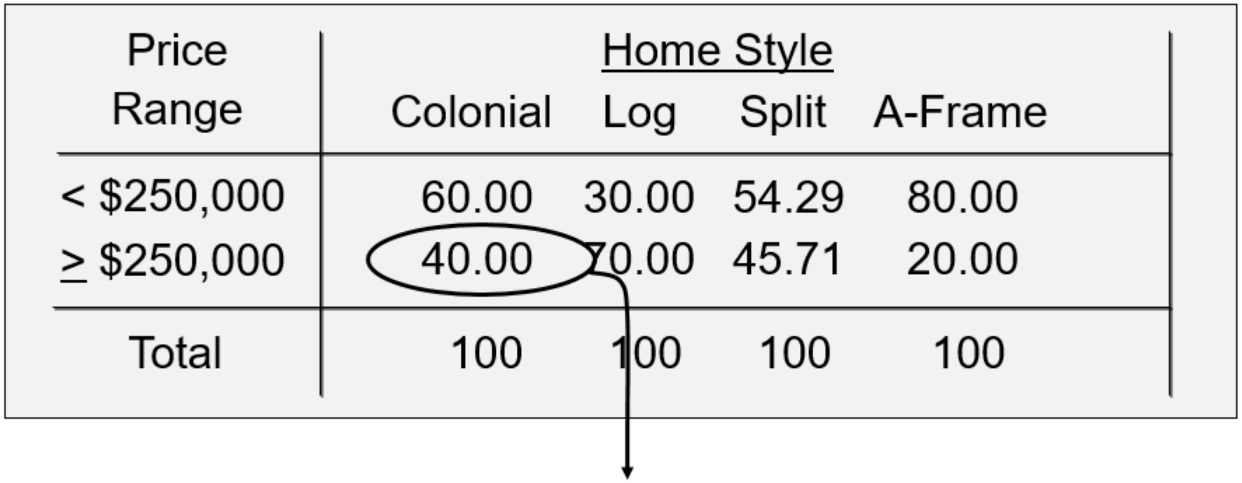
Note: row totals are actually 100.01 due to rounding.

(Colonial and ≥ $250K)/(All ≥ $250K) x 100 = (12/45) x 100

© Somak Paul

Crosstabulation: Column Percentages

Example: Finger Lakes Homes



Price Range	Home Style			
	Colonial	Log	Split	A-Frame
< $250,000	60.00	30.00	54.29	80.00
≥ $250,000	40.00	70.00	45.71	20.00
Total	100	100	100	100

(Colonial and ≥ $250K)/(All Colonial) x 100 = (12/30) x 100

Crosstabulation: Simpson’s Paradox

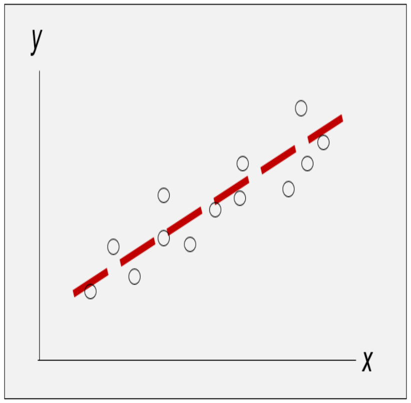
- Data in two or more crosstabulations are often aggregated to produce a summary crosstabulation.
- We must be careful in drawing conclusions about the relationship between the two variables in the aggregated crosstabulation.
- In some cases the conclusions based upon an aggregated crosstabulation can be completely reversed if we look at the unaggregated data. The reversal of conclusions based on aggregate and unaggregated data is called <u>Simpson’s paradox.</u>

Summarizing Data for Two Variables Using Graphical Displays

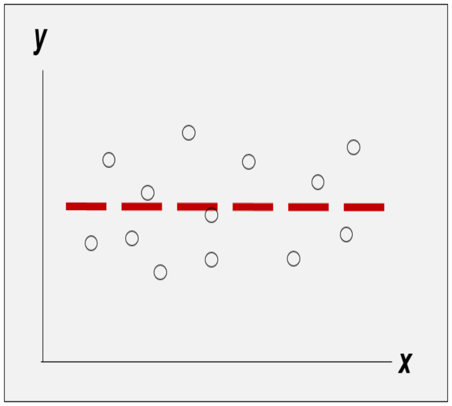
- In most cases, a graphical display is more useful than a table for recognizing patterns and trends.
- Displaying data in creative ways can lead to powerful insights.
- Scatter diagrams and trendlines are useful in exploring the relationship between two variables.
- A scatter diagram is a graphical presentation of the relationship between two quantitative variables.
  - One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
  - The general pattern of the plotted points suggests the overall relationship between the variables.
- A trendline provides an approximation of the relationship.

Scatter Diagram

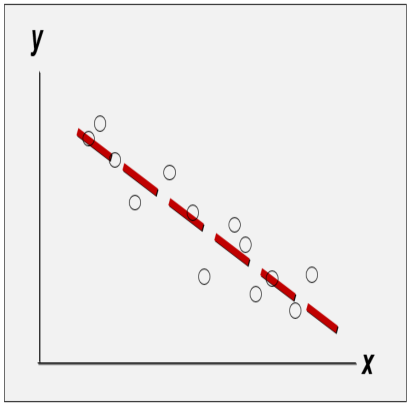
A Positive Relationship



No Apparent Relationship



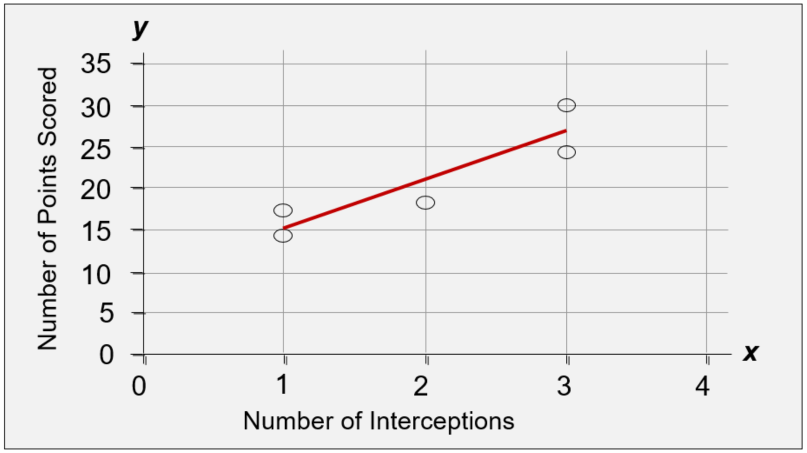
A Negative Relationship



Scatter Diagram

Example: Panthers Football Team. The Panthers football team is interested in investigating the relationship, if any, between interceptions made and points scored.

X = Number of Interceptions	Y = Number of Points Scored
1	14
3	24
2	18
1	17
3	30

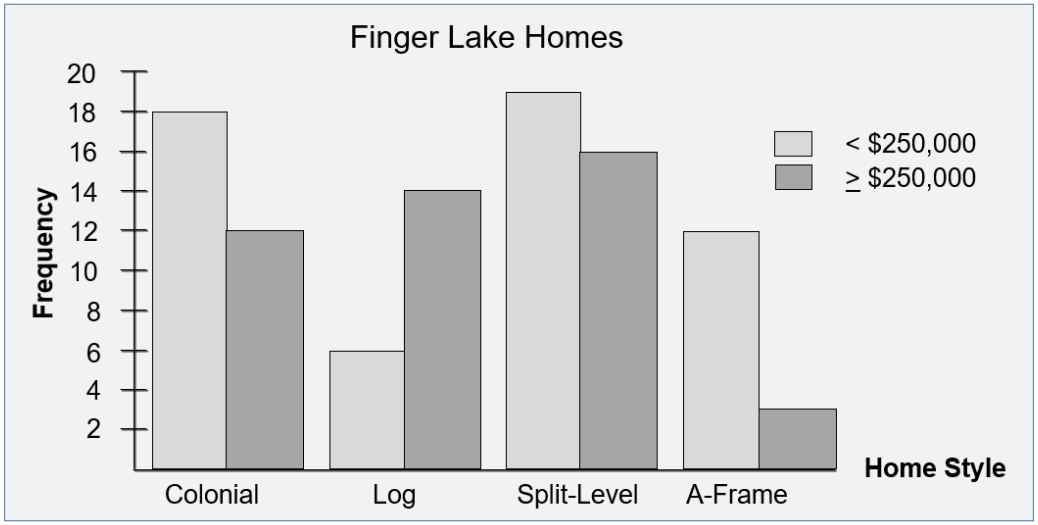


Insights Gained from the Scatter Diagram:

- The scatter diagram indicates a positive relationship between the number of interceptions and the number of points scored.
- Higher points scored are associated with a higher number of interceptions.
- The relationship is not perfect; all plotted points in the scatter diagram are not on a straight line.

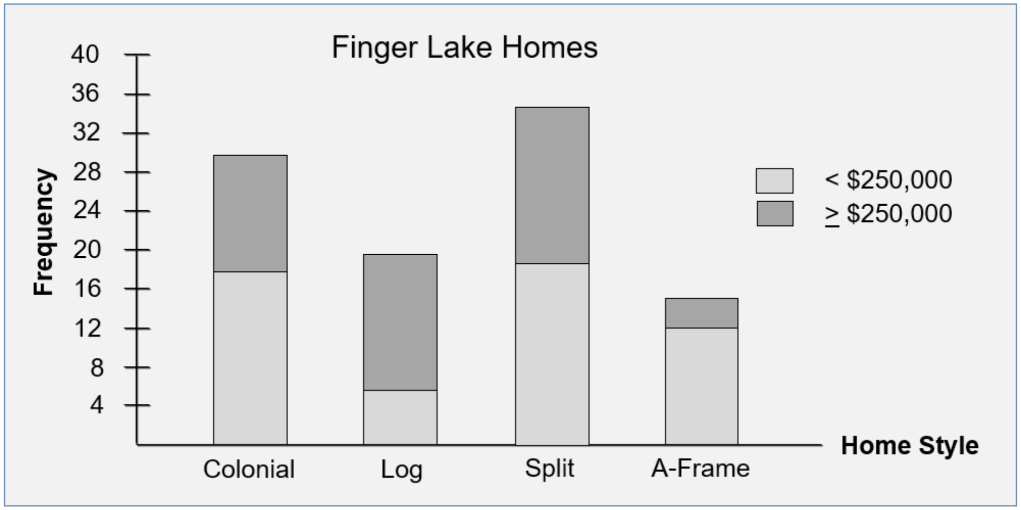
Side-by-Side Bar Chart

- A <u>side-by-side bar chart</u> is a graphical display for depicting multiple bar charts on the same display.
- Each cluster of bars represents one value of the first variable.
- Each bar within a cluster represents one value of the second variable.



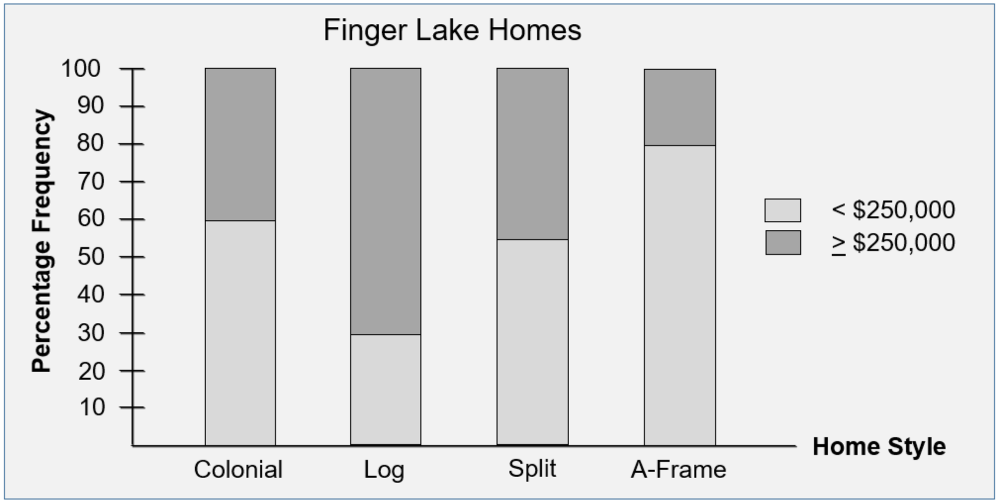
Stacked Bar Chart

- A stacked bar chart is another way to display and compare two variables on the same display.
- It is a bar chart in which each bar is broken into rectangular segments of a different color.
- If percentage frequencies are displayed, all bars will be of the same height (or length), extending to the 100% mark.



Stacked Bar Chart

If percentage frequencies are displayed, all bars will be of the same height (or length), extending to the 100% mark.



Data Visualization: Best Practices

- Data visualization is the use of graphical displays to summarize and present information about a data set.
- The goal is to communicate as effectively and clearly as possible, the key information about the data.
- Creating effective graphical displays is as much art as it is science.
- Here are some guidelines:
  1. Give the display a clear and concise title.
  2. Keep the display simple.
  3. Clearly label each axis and provide the units of measure.
  4. If colors are used, make sure they are distinct.
  5. If multiple colors or line types are used, provide a legend.

Choosing the Type of Graphical Display

Displays used to show the distribution of data:

- Bar Chart to show the frequency distribution and relative frequency distribution for categorical data
- Pie Chart to show the relative frequency and percent frequency for categorical data
- Dot Plot to show the distribution for quantitative data over the entire range of the data
- Histogram to show the frequency distribution for quantitative data over a set of class intervals
- Stem-and-Leaf Display to show both the rank order and shape of the distribution for quantitative data

Displays used to make comparisons:

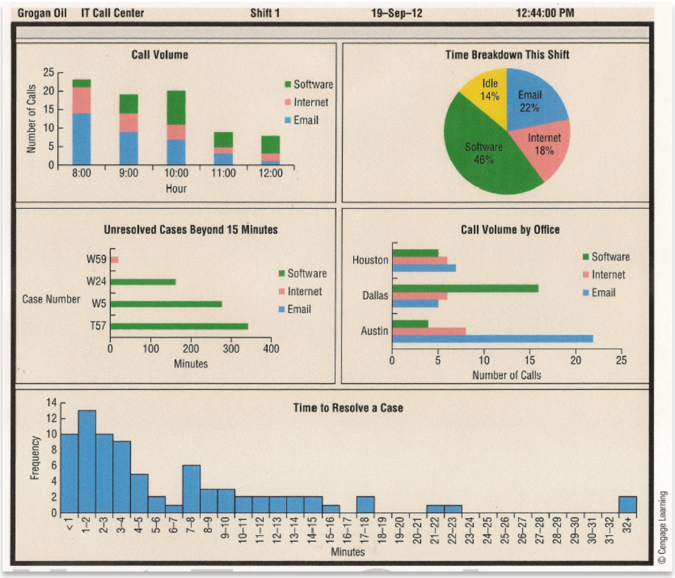
- Side-by-Side Bar Chart to compare two variables
- Stacked Bar Chart to compare the relative frequency or percent frequency of two categorical variables

Displays used to show relationships:

- Scatter Diagram to show the relationship between two quantitative variables
- Trendline to approximate the relationship of data in a scatter diagram

Data Dashboards

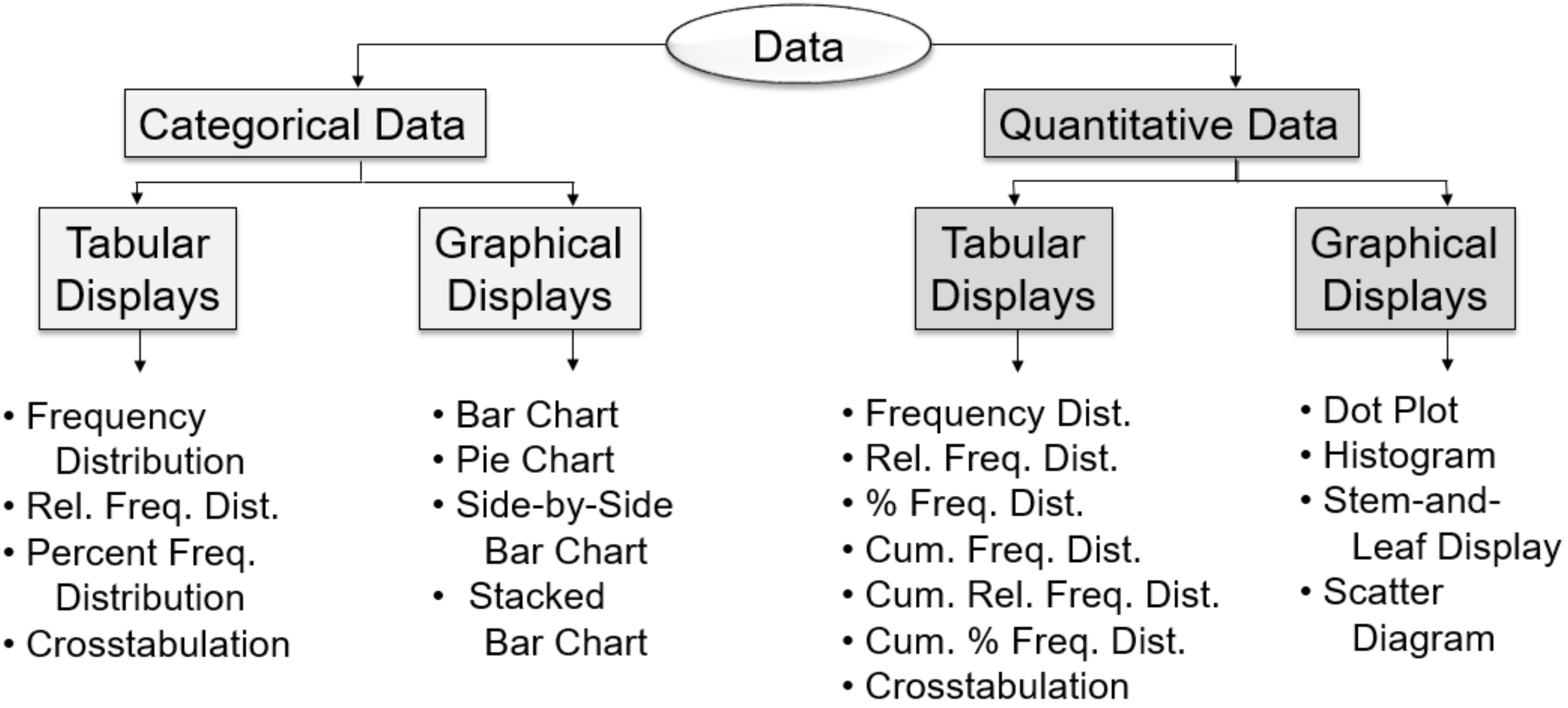
- A data dashboard is a widely used data visualization tool.
- It organizes and presents key performance indicators (KPIs) used to monitor an organization or process.
- It provides timely summary information that is easy to read, understand, and interpret.
- Some additional guidelines include:
  - Minimize the need for screen scrolling.
  - Avoid unnecessary use of color or 3D displays.
  - Use borders between charts to improve readability.



CAL STATE EAST BAY

© Somak Paul

Tabular and Graphical Displays



Data

Categorical Data

Tabular Displays

- Frequency Distribution
- Rel. Freq. Dist.
- Percent Freq. Distribution
- Crosstabulation

Graphical Displays

- Bar Chart
- Pie Chart
- Side-by-Side Bar Chart
- Stacked Bar Chart

Quantitative Data

Tabular Displays

- Frequency Dist.
- Rel. Freq. Dist.
- % Freq. Dist.
- Cum. Freq. Dist.
- Cum. Rel. Freq. Dist.
- Cum. % Freq. Dist.
- Crosstabulation

Graphical Displays

- Dot Plot
- Histogram
- Stem-and-Leaf Display
- Scatter Diagram

Numerical Measures

- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the point estimator of the corresponding population parameter.

Measures of Location

- Mean
- Median
- Mode
- Weighted Mean
- Geometric Mean
- Percentiles
- Quartiles

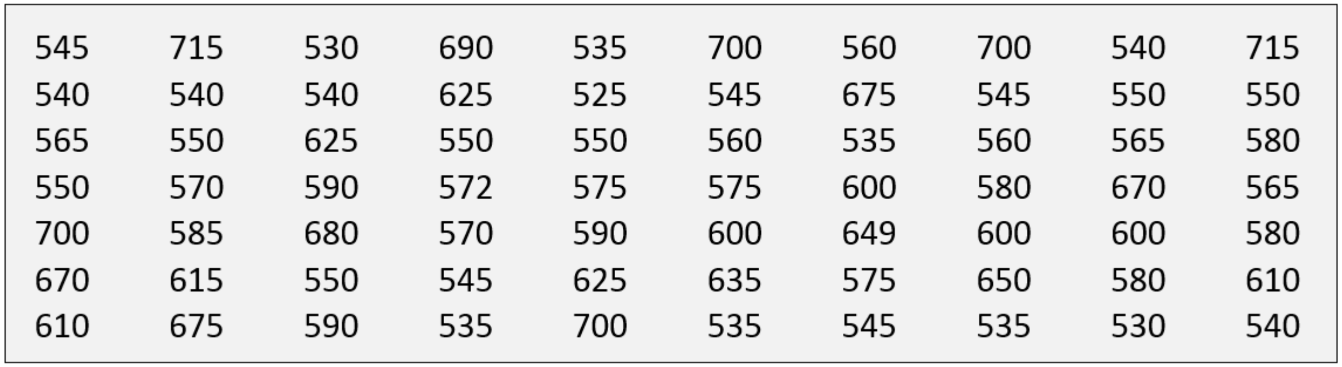
Mean

- Perhaps the most important measure of location is the mean.
- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean, μ. $\bar{x} = \frac{\sum x_i}{n}$

  where $\sum x_i$ = the sum of the values of the *n* observations and

  n = the number of observations in the sample.

Example: Seventy efficiency apartments were randomly sampled in a college town. The monthly rents for these apartments are listed below.



545	715	530	690	535	700	560	700	540	715
540	540	540	625	525	545	675	545	550	550
565	550	625	550	550	560	535	560	565	580
550	570	590	572	575	575	600	580	670	565
700	585	680	570	590	600	649	600	600	580
670	615	550	545	625	635	575	650	580	610
610	675	590	535	700	535	545	535	530	540

$$\bar{x} = \frac{\sum x_i}{n} = \frac{41{,}356}{70} = \boxed{590.80}$$

Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Example: Here we have an odd number of observations:

7 observations: 26, 18, 27, 12, 14, 27, and 19.

Rewritten in ascending order: 12, 14, 18, 19, 26, 27, and 27.

The median is the middle value in this list, so the median = 19.

Median

Example: Here we have an even number of observations:

8 observations: 26, 18, 27, 12, 14, 27, 19, and 30.

Rewritten in ascending order: 12, 14, 18, 19, 26, 27, 27, and 30.

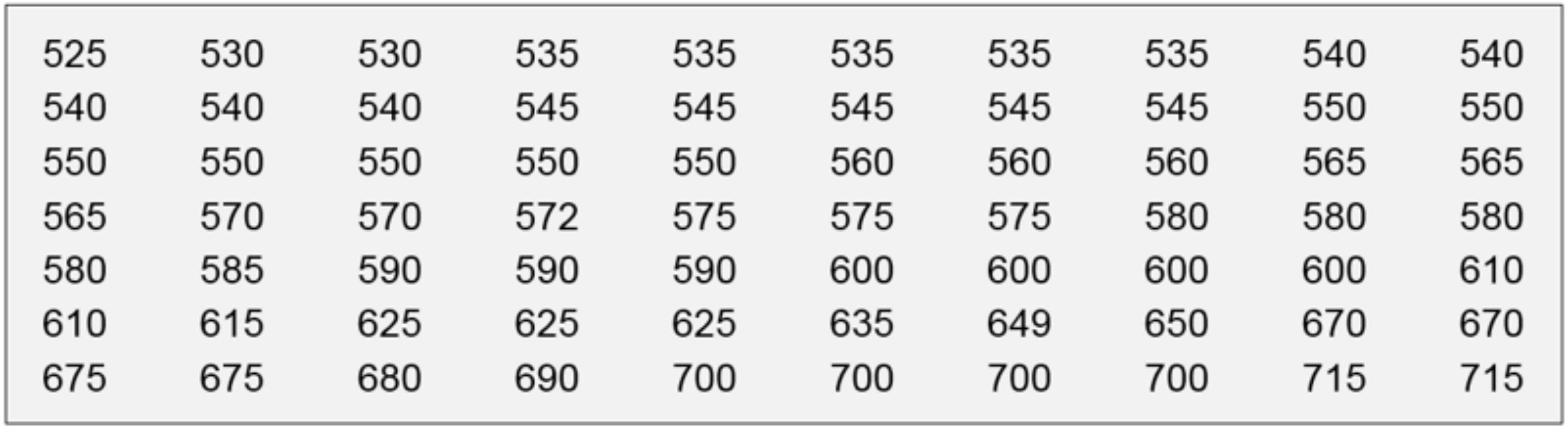
The median is the average of the two middle values in this list, so the median = (19 + 26)/2 = 22.5.

Median

Example: Apartment Rents

Notice that there are 70 values provided which are in ascending order.

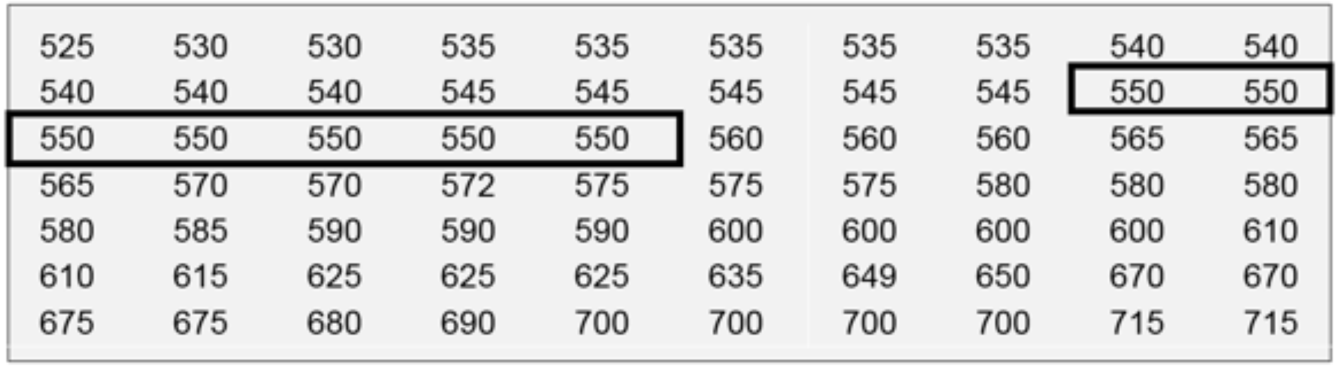
Averaging the 35th and 36th values: Median (575 + 575)/2 = 575.



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

The mode is 550.

Weighted Mean

- Sometimes mean is computed by giving each observation a weight that reflects its relative importance.
- The choice of weights depends on the application.
- The weights might be the number of credit hours earned for each grade, as in GPA.
- In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used.

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

where: x_i = value of observation *i*

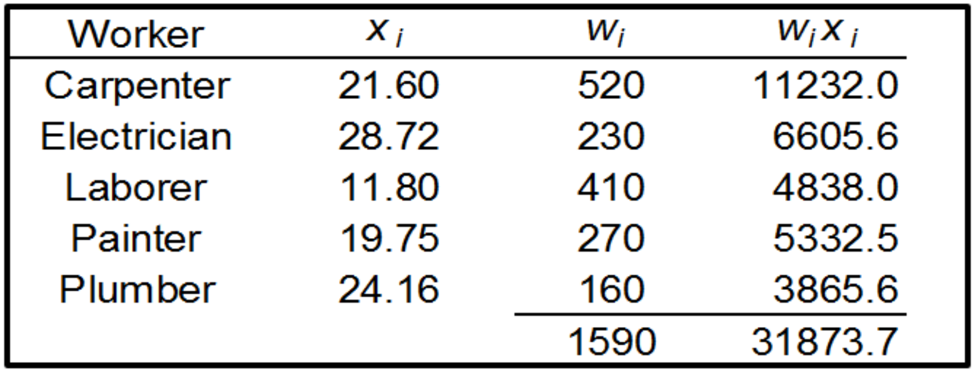
w_i = weight for observation *i*

Example: Ron Butler, a home builder, is looking over the expenses he incurred for a house he just built. For the purpose of pricing future projects, he would like to know the average wage ($/hour) he paid the workers he employed. Listed here are the categories of workers he employed, along with their respective wage and total hours worked.

Worker	Wage ($/hr)	Total Hours
Carpenter	21.60	520
Electrician	28.72	230
Laborer	11.80	410
Painter	19.75	270
Plumber	24.16	160

Weighted Mean

Example: Construction Wages



Worker	x_i	w_i	$w_i x_i$
Carpenter	21.60	520	11232.0
Electrician	28.72	230	6605.6
Laborer	11.80	410	4838.0
Painter	19.75	270	5332.5
Plumber	24.16	160	3865.6
		1590	31873.7

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{31{,}873.7}{1{,}590} = 20.0464 = \boxed{\$20.05}$$

FYI, the equally-weighted (simple) mean = $21.21

Geometric Mean

- The geometric mean is calculated by finding the *n*th root of the product of *n* values.
- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include: changes in populations of species, crop yields, pollution levels, and birth and death rates.

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)\ldots(x_n)}$$

$$= [(x_1)(x_2)\ldots(x_n)]^{1/n}$$

Example: Rate of Return

Period	Return (%)
1	-6.0
2	-8.0
3	-4.0
4	2.0
5	5.4

Growth Factor
0.940
0.920
0.960
1.020
1.054

$$\bar{x}_g = \sqrt[5]{(0.94)(0.92)(1.02)(1.054)} = (0.89254)^{1/5} = 0.97752$$

The average growth rate per period is

(0.97752 – 1)(100) = –2.248%.

Percentiles

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.
- The p^{th} percentile of a data set is a value such that at least p percent of the items take on this value or less and at least (100 – p) percent of the items take on this value or more.
- Arrange the data in ascending order.
- Compute L_p, the location of the p^{th} percentile.

$$L_p = \left(\frac{p}{100}\right)(n+1)$$

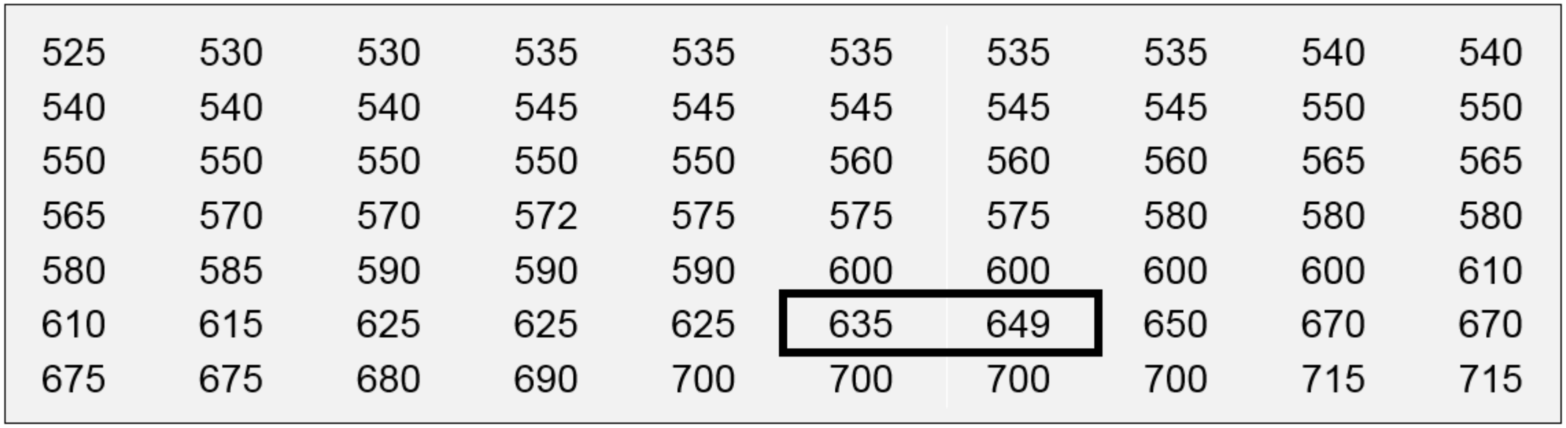
80th Percentile

Example: Apartment Rents

$$L_p = \left(\frac{p}{100}\right)(n+1) = \left(\frac{80}{100}\right)(70+1) = 56.8$$

The 80th percentile is the 56th value plus 0.8 times the difference between the 57th and 56th values.

So the 80th percentile = 635 + 0.8(649 – 635) = 646.2.



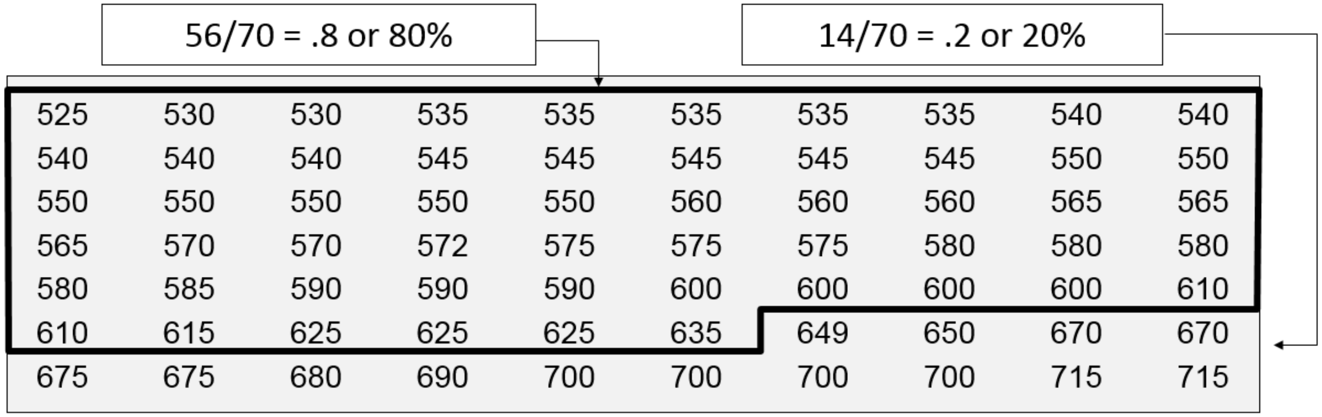
525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	$\boxed{635}$	$\boxed{649}$	650	670	670
675	675	680	690	700	700	700	700	715	715

80th Percentile, Part 2

Example: Apartment Rents

"At least 80% of the
items take on a
value of 646.2 or less."

"At least 20% of the
items take on a
value of 646.2 or more."



56/70 = .8 or 80%

14/70 = .2 or 20%

525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Quartiles

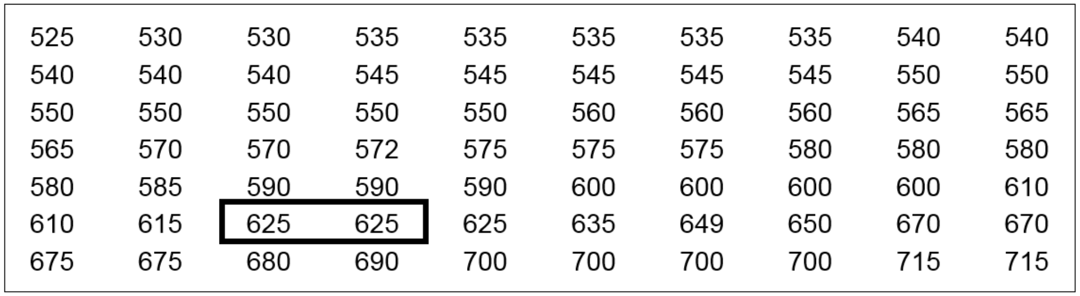
Quartiles are specific percentiles.

1. First Quartile = 25th Percentile
2. Second Quartile = 50th Percentile = Median
3. Third Quartile = 75th Percentile

Example: Third Quartile of Apartment Rents

$$L_p = \left(\frac{p}{100}\right)(n + 1) = \left(\frac{75}{100}\right)(70 + 1) = 53.25$$

The 75th percentile is the 53rd value plus 0.25 times the difference between the 54th and 53rd values.
The 75th percentile = third quartile = 625 + 0.25(625 – 625) = 625.



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

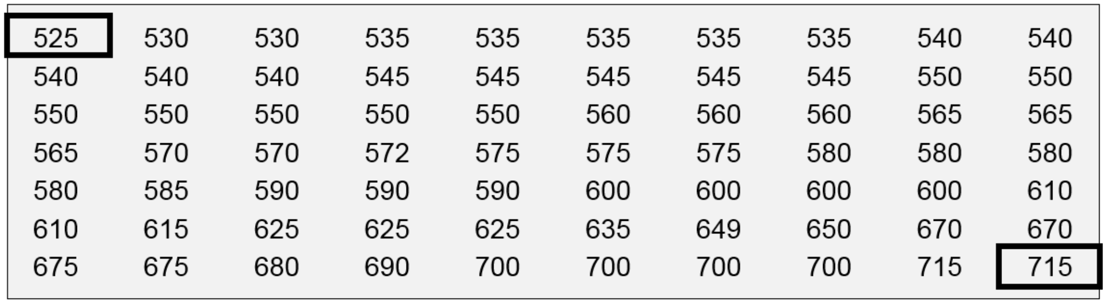
Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.
- Common measures of variability are:
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Range

- The range of a data set is the difference between the largest and smallest data value.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Range = largest value – smallest value = 715 – 525 = 190.



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

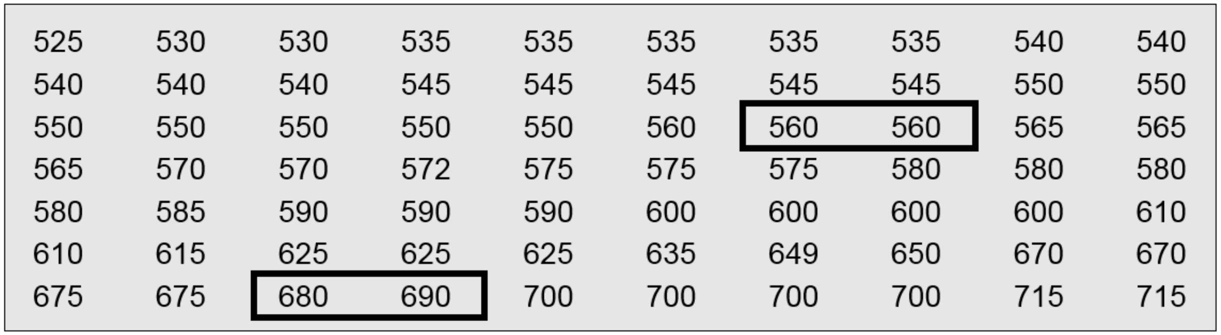
Interquartile Range (IQR)

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

3rd Quartile (Q_3) = 625

1st Quartile (Q_1) = 545

IQR = 625 – 545 = 80



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

Variance and Standard Deviation

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, *m* for a population).
- The variance is useful in comparing the variability of two or more variables.
- The variance is the average of the squared deviations between each data value and the mean.
- The variance of a **sample** is: $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$
- The variance for a **population** is: $\sigma^2 = \dfrac{\sum(x_i - \mu)^2}{N}$
- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation of a **sample** is $s = \sqrt{s^2}$
- The standard deviation of a **population** is: $\sigma = \sqrt{\sigma^2}$

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.
- The coefficient of variation of a sample is: $\left[\frac{s}{\bar{x}} \text{ x } 100\right]\%$
- The coefficient of variation of a population is: $\left[\frac{\sigma}{\mu} \text{ x } 100\right]\%$

Sample Variance, Standard Deviation, Coefficient of Variation

Example: Apartment Rents

- The sample variance is: $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \boxed{2{,}996.16}$

- The sample standard deviation is: $s = \sqrt{s^2} = \sqrt{2{,}996.16} = \boxed{54.74}$

- The sample coefficient of variation is: $\left[\frac{s}{\bar{x}} \text{ x } 100\right]\% = \left[\frac{54.74}{590.80} \text{ x } 100\right]\% = \boxed{9.27\%}$

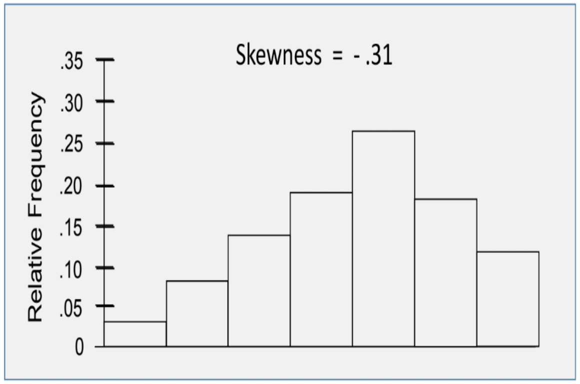
Measures of Distribution Shape, Relative Location, Detecting Outliers

- Distribution Shape
- z-Scores
- Chebyshev’s Theorem
- Empirical Rule
- Detecting Outliers

Distribution Shape: Skewness

- An important numerical measure of the shape of a distribution is called skewness.

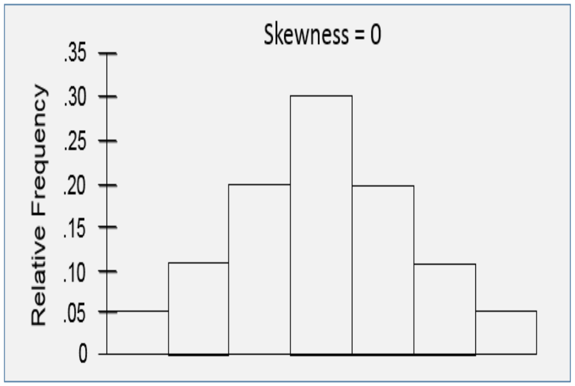
Moderately Skewed Left
Skewness is negative.
Mean will usually be
less than the median.



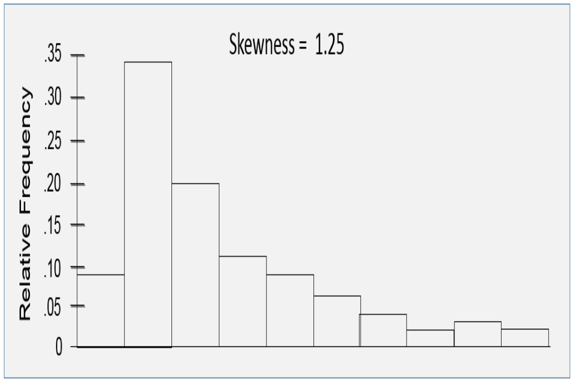
Symmetric (not skewed)

Skewness is zero.

Mean and median are equal.



Highly Skewed Right
Skewness is positive (often above 1.0)
Mean will usually be
more than the median.



CAL STATE
EAST BAY

z-Scores

- The z-score is often called the standardized value.
- It denotes the number of standard deviations a data value x_i is from the mean.
- An observation's *z*-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a *z*-score less than zero.
- A data value greater than the sample mean will have a *z*-score greater than zero.
- A data value equal to the sample mean will have a *z*-score of zero.

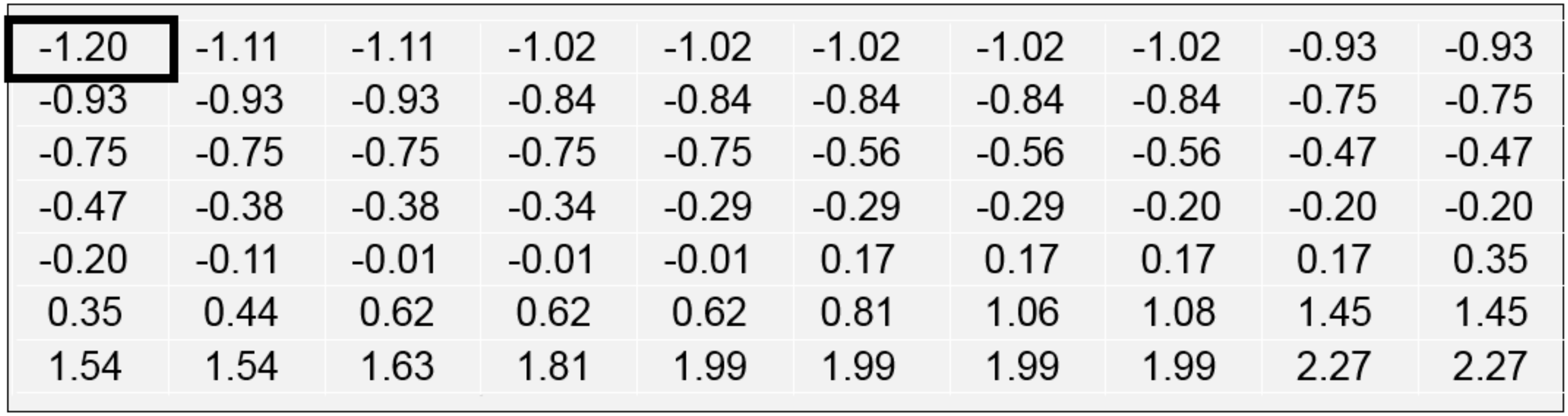
$$z_i = \frac{x_i - \bar{x}}{s}$$

z-Scores

Example: Apartment Rents

z-Score of Smallest Value (525)

Standardized Values for Apartment Rents $z_i = \frac{x_i - \bar{x}}{s} = \frac{525 - 590.80}{54.74} = \boxed{-1.20}$



$\boxed{-1.20}$	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

Chebyshev's Theorem

- At least $(1 – 1/z^2)$ of the data values must be within z standard deviations of the mean, where z is any value greater than 1.

- Chebyshev's theorem requires $z > 1$; but z need not be an integer.

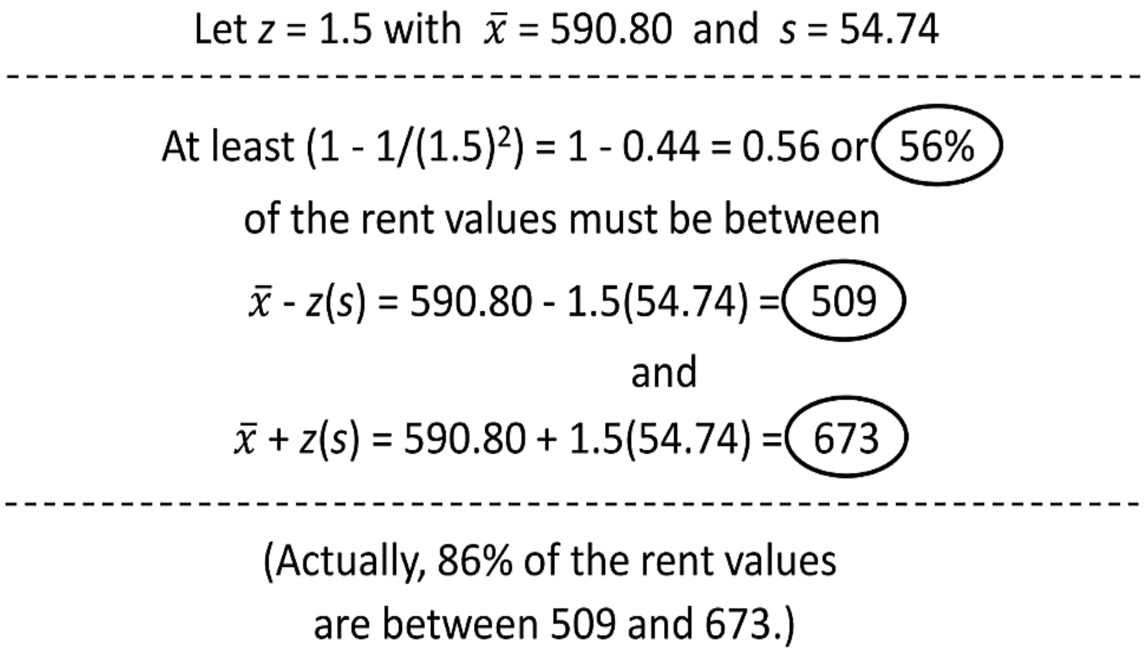
- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.

- At least 89% of the data values must be within $z = 3$ standard deviations of the mean.

- At least 94% of the data values must be within $z = 4$ standard deviations of the mean.

Chebyshev’s Theorem

Example: Apartment Rents



Let z = 1.5 with $\bar{x}$ = 590.80 and s = 54.74

At least $(1 - 1/(1.5)^2) = 1 - 0.44 = 0.56$ or $\boxed{56\%}$

of the rent values must be between

$$\bar{x} - z(s) = 590.80 - 1.5(54.74) = \boxed{509}$$

and

$$\bar{x} + z(s) = 590.80 + 1.5(54.74) = \boxed{673}$$

(Actually, 86% of the rent values
are between 509 and 673.)

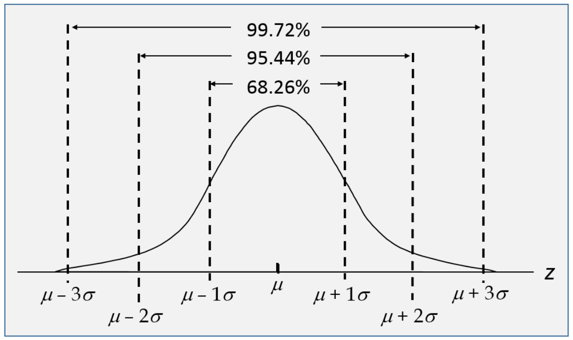
Empirical Rule

When the data are believed to approximate a bell-shaped distribution:

- The empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.
- The empirical rule is based on the normal distribution, which is covered in Chapter 6.

For data having a bell-shaped distribution:

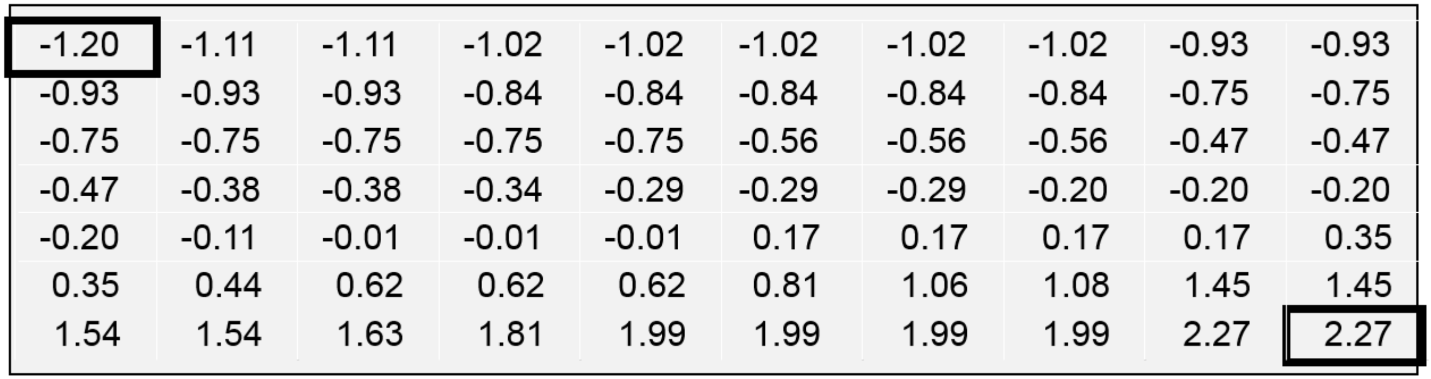
- Approximately 68% of the data values will be within one standard deviation of the mean.
- Approximately 95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.



Detecting Outliers

- An <u>outlier</u> is an unusually small or unusually large value in a data set.
- A data value with a z-score less than –3 or greater than +3 might be considered an outlier.
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded data value that belongs in the data set

Example: Apartment Rents. The most extreme z-scores are -1.20 and 2.27. Using |*z*| >_3 as the criterion for an outlier, there are no outliers in this data set.



-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

Five-Number Summaries and Box Plots

- Summary statistics and easy-to-draw graphs can be used to quickly summarize large quantities of data.
- Two tools that accomplish this are five-number summaries and box plots.

1. Smallest Value
2. First Quartile
3. Median
4. Third Quartile
5. Largest Value

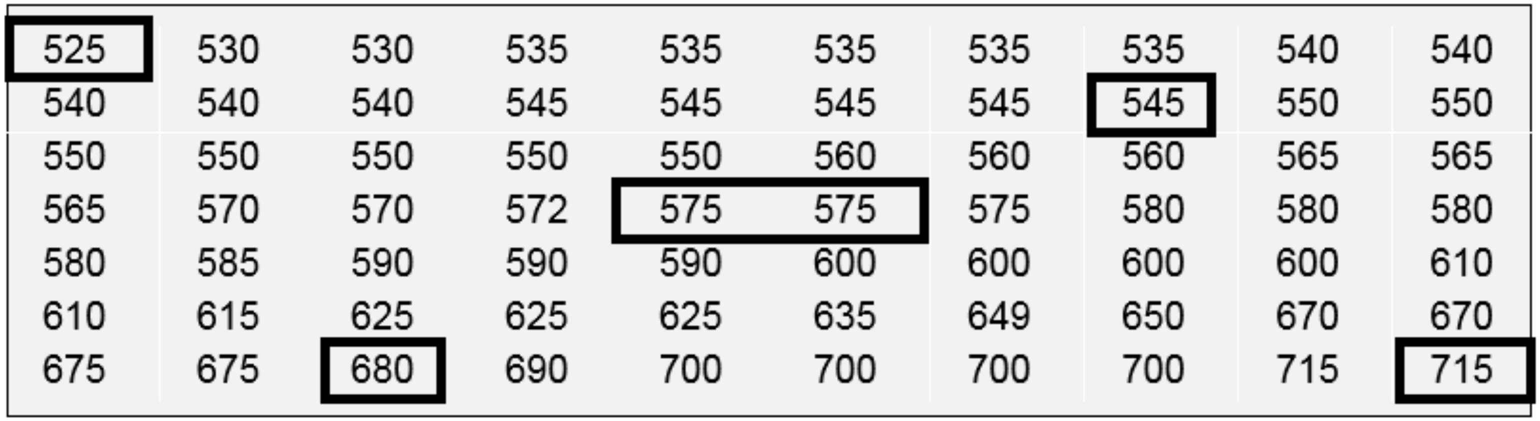
Five-Number Summary

Example: Apartment Rents

Lowest Value = 525 First Quartile = 545

Median = 575

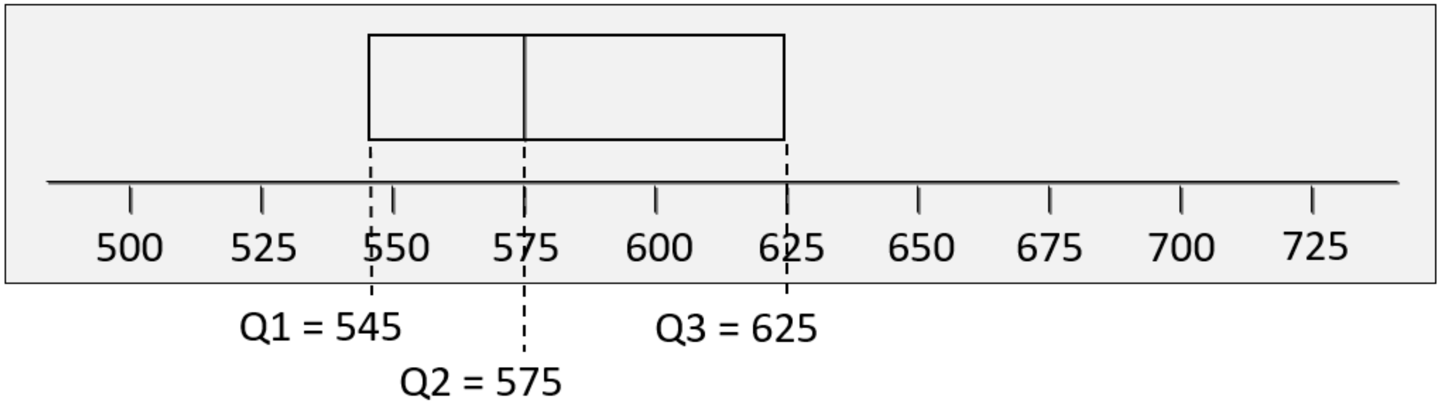
Third Quartile = 625 Largest Value = 715



525	530	530	535	535	535	535	535	540	540
540	540	540	545	545	545	545	545	550	550
550	550	550	550	550	560	560	560	565	565
565	570	570	572	575	575	575	580	580	580
580	585	590	590	590	600	600	600	600	610
610	615	625	625	625	635	649	650	670	670
675	675	680	690	700	700	700	700	715	715

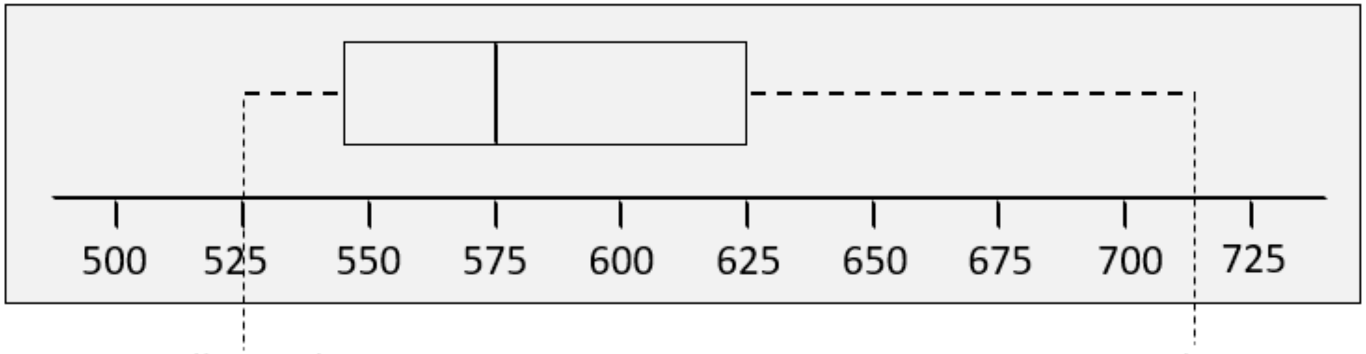
Box Plot

- A box plot is a graphical display of data that is based on a five-number summary.
- Box plots provide another way to identify outliers.
- A box is drawn with its ends located at the first and third quartiles.
- A vertical line is drawn in the box at the location of the median (second quartile).



Box Plot

- Limits are located (not drawn) using the interquartile range (IQR).
- Data outside these limits are considered outliers.
- The location of each outlier is shown with the symbol * .



Smallest value inside limits = 525

Largest value inside limits = 715

Box Plot

Example: Apartment Rents

- The lower limit is located 1.5(IQR) below Q_1.

  Lower Limit: Q_1 – 1.5(IQR) = 545 – 1.5(80) = 425

- The upper limit is located 1.5(IQR) above Q_3.

  Upper Limit: Q_3 + 1.5(IQR) = 625 + 1.5(80) = 745

- There are no outliers (values less than 425 or greater than 745) in the apartment rent data.

Covariance

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.
- The covariance is computed as follows:

For samples: $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

For populations: $\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$

Correlation Coefficient

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.
- The correlation coefficient is computed as follows:

For samples: $r_{xy}=\frac{s_{xy}}{s_x s_y}$

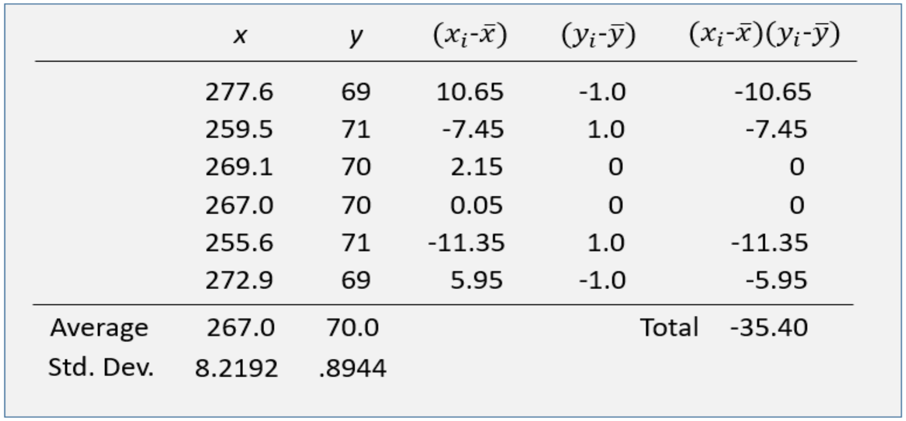
For populations: $\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x \sigma_y}$

- The coefficient can take on values between –1 and +1.
- Values near –1 indicate a <u>strong negative linear relationship.</u>
- Values near +1 indicate a <u>strong positive linear relationship.</u>
- The closer the correlation is to zero, the weaker the relationship.

Covariance and Correlation Coefficient

A golfer is interested in investigating the relationship, if any, between driving distance and 18-hole score.

Average Driving Distance (yards)	Average 18-Hole Score
277.6	69
259.5	71
269.1	70
267.0	70
255.6	71
272.9	69

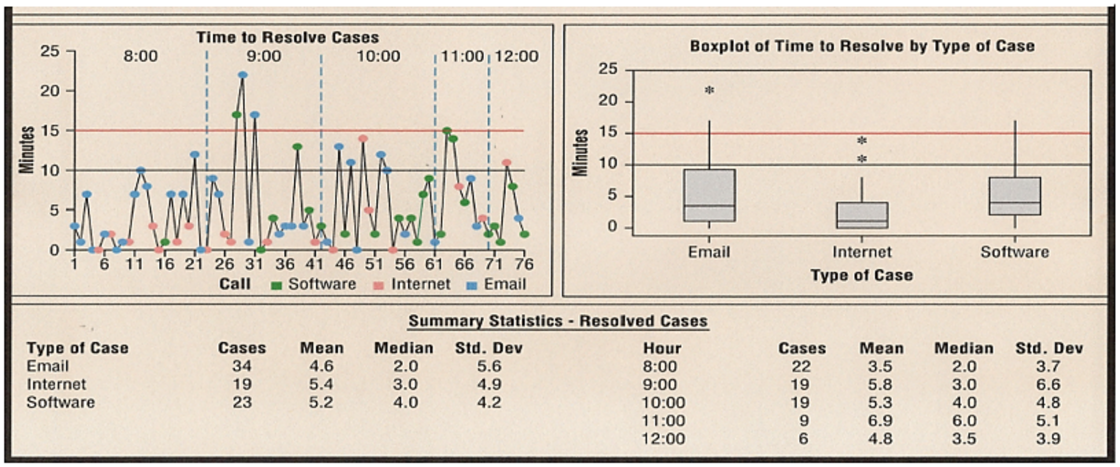


	x	y	$(x_i\text{-}\bar{x})$	$(y_i\text{-}\bar{y})$	$(x_i\text{-}\bar{x})(y_i\text{-}\bar{y})$
	277.6	69	10.65	-1.0	-10.65
	259.5	71	-7.45	1.0	-7.45
	269.1	70	2.15	0	0
	267.0	70	0.05	0	0
	255.6	71	-11.35	1.0	-11.35
	272.9	69	5.95	-1.0	-5.95
Average	267.0	70.0		Total	-35.40
Std. Dev.	8.2192	.8944			

- Sample Covariance: $s_{xy} = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \dfrac{-35.40}{6-1} = \boxed{-7.08}$
- Sample Correlation Coefficient: $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = \dfrac{-7.08}{(8.2192).8944)} = \boxed{-.9631}$

Data Dashboards: Adding Numerical Measures to Improve Effectiveness

- Data dashboards are not limited to graphical displays.
- The addition of numerical measures, such as the mean and standard deviation of KPIs, to a data dashboard is often critical.
- Dashboards are often interactive.
- <u>Drilling down</u> refers to functionality in interactive dashboards that allows the user to access information and analyses at an increasingly detailed level.



Summary Statistics - Resolved Cases

Type of Case	Cases	Mean	Median	Std. Dev
Email	34	4.6	2.0	5.6
Internet	19	5.4	3.0	4.9
Software	23	5.2	4.0	4.2

Hour	Cases	Mean	Median	Std. Dev
8:00	22	3.5	2.0	3.7
9:00	19	5.8	3.0	6.6
10:00	19	5.3	4.0	4.8
11:00	9	6.9	6.0	5.1
12:00	6	4.8	3.5	3.9

CAL STATE
EAST BAY