Chapter 1 & 2

Descriptive Statistics & Graphs

Statistical Packages Course – Dr. Bashar Al-Shboul
The University of Jordan

- Population Data – everything or everyone we are studying
- Sample Data – a subset of the population

- University of Houston is interested in how many of their students buy used books as opposed to new ones. They randomly choose 100 students from the student center to interview.
- The **population** would consist of all students at the University of Houston.
- The **sample** consists of the 100 students chosen to interview.

- An elementary school is creating a new lunch menu and they want to know if it will appeal to their students. They send home questionnaires to students with last names that begin with the letters M through R. The population is all students at that elementary school.
- The **sample** consists of the students whose last names begin with the letters M through R.

- A **variable** is a characteristic of an individual that can assume more than one value. Variables can be classified as **categorical** (qualitative) or **quantitative** (numeric).

- **Categorical** variables describe qualities or characteristics that data may have. They usually represent a “type of something” such as a type of car.

- **Quantitative** variables are measurements. These will be numeric values.

- Quantitative variables can be classified as either discrete or continuous..
  - **Discrete** quantitative variables are countable.
    - For example: the number of lives given in a single play of a video game
  - **Continuous** quantitative variables can take on any value in an interval.
    - For example: the amount of time you wait in line at the driver's license office

Examples

- Political preference (categorical)
- Number of siblings (Quantitative, discrete)
- Blood type (categorical)
- Height of men on the UH basketball team (Quantitative, continuous)
- Time it takes to be on hold when calling the IRS (Quantitative, continuous)

Descriptive Statistics

- Mean, Median, Mode
  - Used to describe the location of the data
- Standard Deviation and Variance
  - Used to measure spread or dispersion
  - For example, standard deviation tells the average distance that data values fall from the mean

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- Most of the time we are not working with the entire population. instead, we are working with a sample.
- **Sample variance** –

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

- **Sample standard deviation**

Example

- A statistics teacher wants to decide whether or not to curve an exam. From her class of 300 students, she chose a sample of 10 students and their grades were:

72, 88, 85, 81, 60, 54, 70, 72, 63, 43

- Find the mean, variance and standard deviation for this sample.

- Suppose the statistics teacher decides to curve the grades by adding 10 points to each score. What is the new mean, variance and standard deviation?

- Find the variance and the standard deviation for the following set of data (whose mean is 4.5)

3, 6, 2, 7, 4, 5

- Now, multiply each value by 2. What is the new variance and the new standard deviation?

- Sometimes we want to compare the variation between two groups. The coefficient of variation can be used for this.

- The coefficient of variation is **the ratio of the standard deviation to the mean**. A **smaller** ratio will indicate **less variation** in the data.

- The following statistics were collected on two different groups of stock prices:

	Portfolio A	Portfolio B
Sample size	10	15
Sample mean	$52.65	$49.80
Sample standard deviation	$6.50	$2.95

- What can be said about the variability of each portfolio?

Range, IQR, and Finding Outliers

- Additional measures of spread (or dispersion).
- Range
- Percentiles
  - 25th percentile (or Ql)
  - 50th percentile (or Q2 or median)
  - 75th percentile (or Q3)
- Inter Quartile Range (IQR)

- The values of the minimum, Q1, Q2, Q3 and the maximum make up what is called our five number summary.

- If we are given the five number summary, we can quickly find the range and IQR.

- Twelve babies spoke for the first time at the following ages in months):

  8,9,10,11,12,13,15,15,18,20,20,26

- Find Q1, Q2, Q3, the range and the IQR

- The IQR is used to determine data classified as **outliers.**
  - An outlier is an observation that is ''distant'' from the rest of the data.
  - Outliers can occur by chance or be measurement errors so it is important to identify them.
  - Any point that falls outside the interval calculated by Q1-1.5(IQR) and Q3 + 1.5(IQR) is considered an outlier.
- Refer back to previous example. Are there any **outliers in that data set? If so,** what **are** they?

- There are other percentiles as well.
  - The kth percentile means that *k%* of the ordered data values are at or below that data value.
  - For example, if the median is 100, then 50% of the ordered data values fall at or below 100. Also, (100-k)% represents the amount of ordered data that falls above the percentile data value.
  - If you are looking for the measurement that has a desired percentile rank, the 100Pth percentile, is the measurement with rank (or position in the list) of *nP+0.5* where *n* represents the number of data values in the sample.

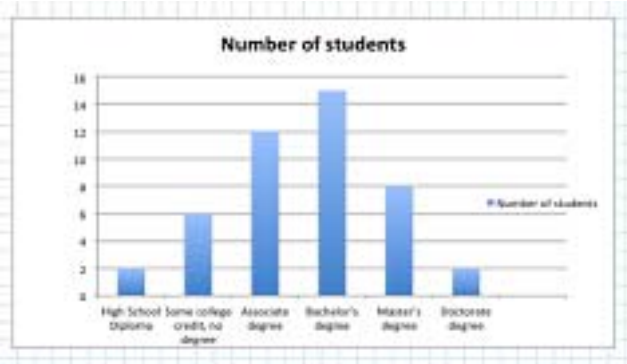
- In a collection of 30 data measurements, which measurement represents the 30^{th} percentile?

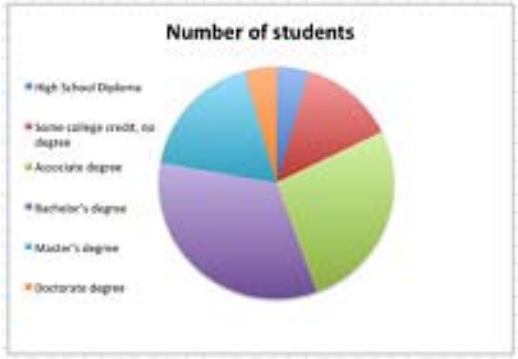
- **Suppose you know the position (the order) of a value and want to know what percentile it is ranked at.**

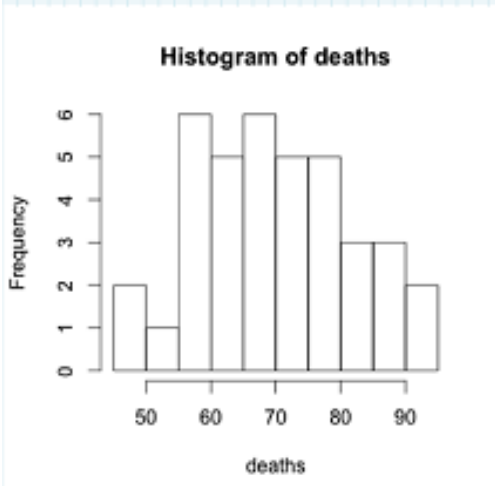
- In **general, if you have *n* data measurements,**
  - *X1* represents the 100(1- 0.5) *I n* th **percentile**,
  - *X2* represents the 100(2-0.5 *I* n th **percentile, and**
  - *Xi* represents the *100(i-0.5 I* n th **percentile.**

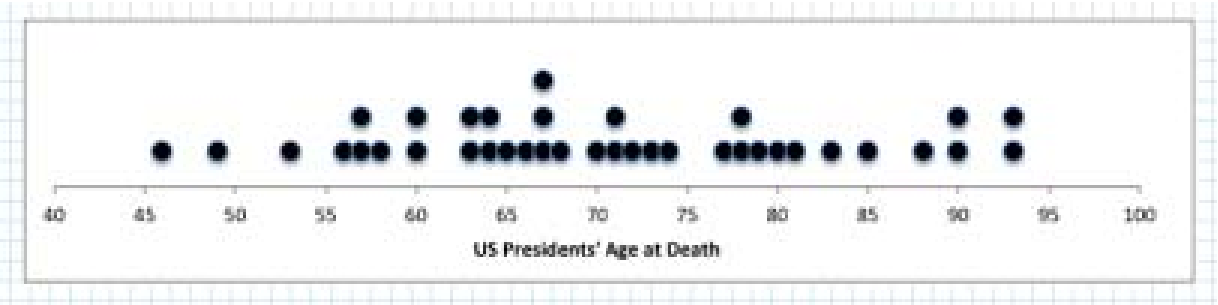
- Using the data in previous example, determine the percentile of the 4th order statistic (*X4*).

- Data can be displayed using graphs and there are several types of graphs to choose from.
- Some of the most common graphs used in statistics are:
  - Bar graph
  - Pie Chart
  - Dot plot
  - Histogram
  - Stem and leaf plot
  - Box plot
  - Cumulative Frequency plot









4 | 69

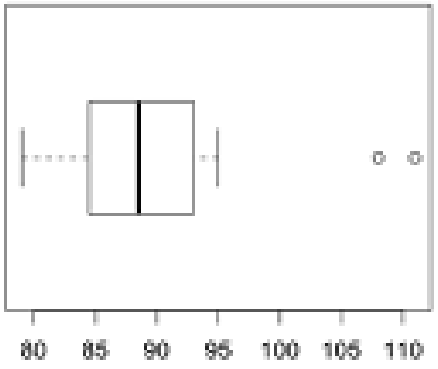
5 | 36778

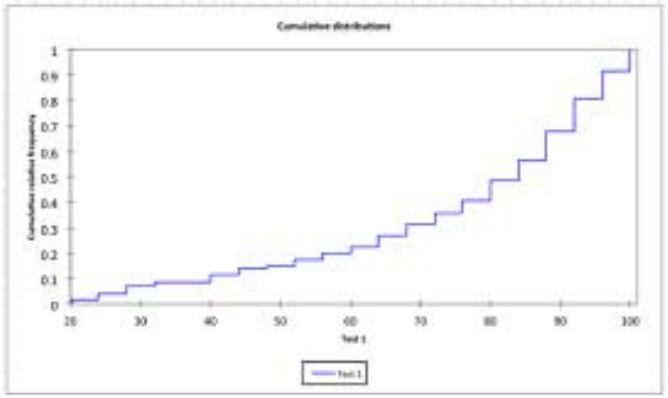
6 | 003344567778

7 | 0112347889

8 | 01358

9 | 0033





- Use the following data to make several of these types of graphs:
- Height measurements for a group of people were taken. The results are recorded below (in inches):

66, 68, 63, 71, 68, 69, 65, 70, 73, 67, 62, 59, 63, 68, 71, 63, 63, 60, 64, 66, 58

- First Step: Sort the data list

58, 59, 60, 62, 63, 63, 63, 63, 64, 65, 66, 66,
67, 68, 68, 68, 69, 70, 71, 71, 73

- A dot plot is made simply by putting dots above the values listed on a number line.

- A stem and leaf plot, the data is arranged by values. The digits in the largest place are referred to as the stem and the digits in the smallest place are referred to as the leaf (leaves). The leaves are displayed to the right of the stem. A split stemplot divides up the stems into equal groups. Back-to-back stempots can be used when comparing two sets of data.

- **Boxplots** not only help identify features about our data quickly (such as spread and location of center) but can be very helpful when comparing data sets.
- How to make a box plot:
  - Order the values in the data set in ascending order (least to greatest).
  - Find and label the median
  - Of the lower half (less than the median-do not include), find and label Q1
  - Of the upper half (greater than the median-do not include), find and label Q3.
  - Label the minimum and maximum.
  - Draw and label the scale on an axis.
  - Plot the five number summary.
  - Sketch a box starting at Q1 to Q3.
  - Sketch a segment within the box to represent the median.
  - Connect the min and max to the box with line segments.

- Note: If data contains outliers, **a box and whiskers plot** can be used instead to display: the data. In a box and whiskers p1ot, the outliers are displayed with dots above the value and the segments begin (or end) at the next data value within the outlier interval

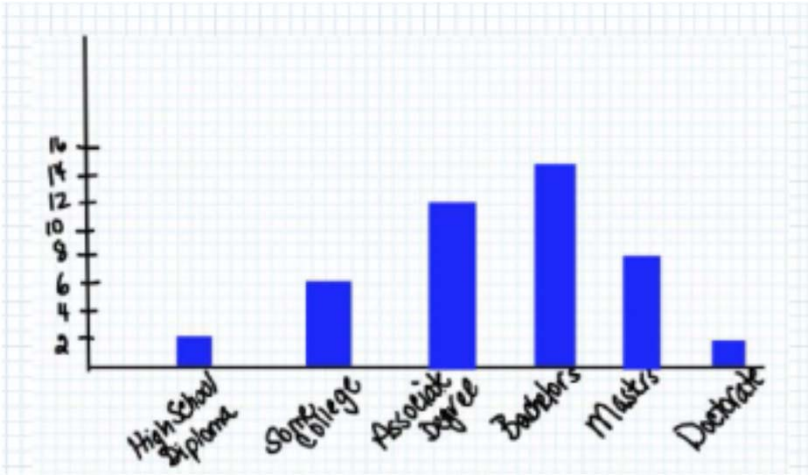
- Histograms are created by first dividing the data into classes, or bins, of equal width. Next, count the number of observations in each class. The horizontal axis will represent the variable values and the vertical axis will represent your frequency or your relative frequency.

- A cumulative frequency plot of the percentages (also called an ogive) can be used to view the total number of events that occurred up to a certain value.

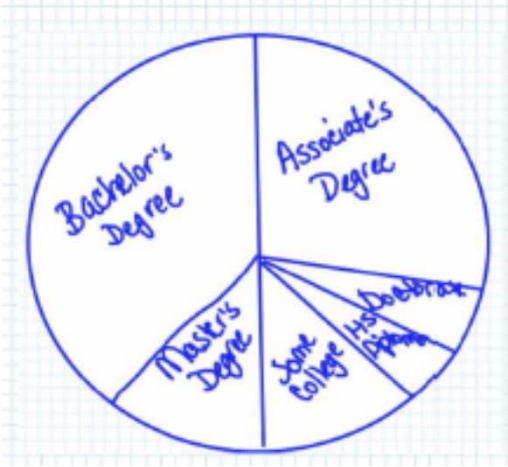
- Create a bar graph and pie chart for the
- following categorical data:

Education level	Number of students
High School Diploma	2
Some college credit, no degree	6
Associate degree	12
Bachelor's degree	15
Master's degree	8
Doctorate degree	2
Total:	45

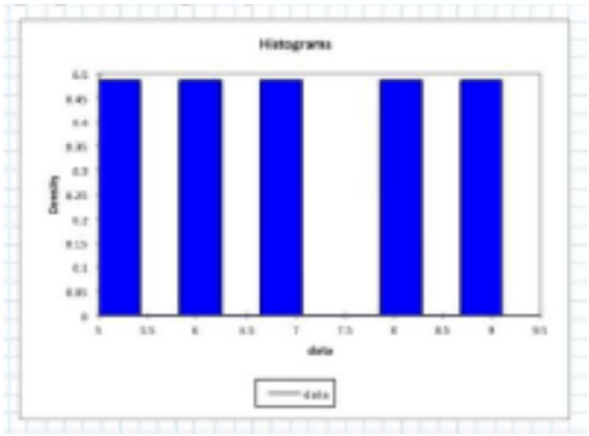
- A bar graph is created by listing the categorical data along the x-axis and the frequencies along the y-axis. Bars are drawn above each data value.

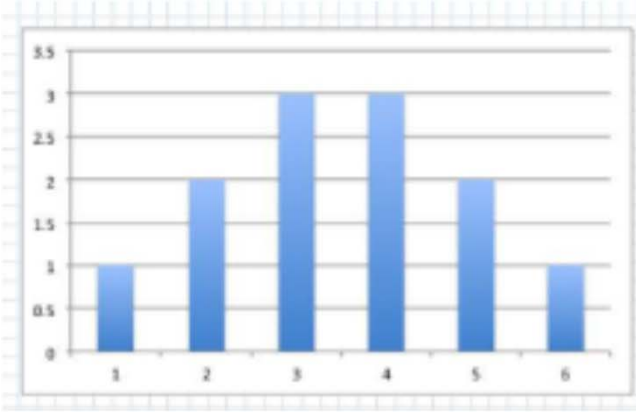


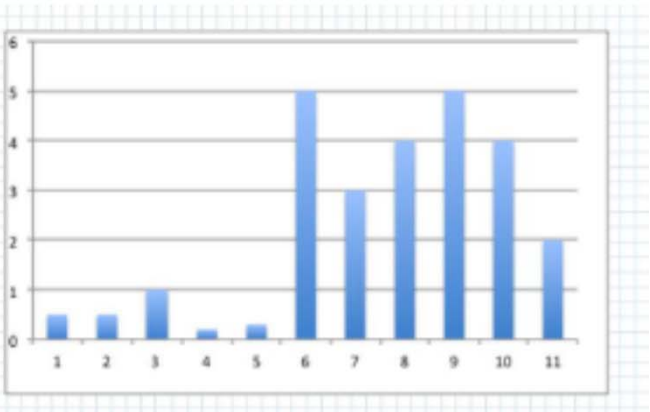
- A pie chart is a circular chart, divided into sectors, indicating the proportion of each data value compared to the entire set of values. Pie charts are good for categorical data.

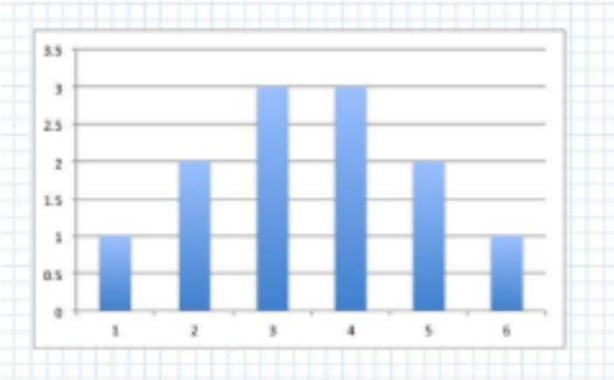


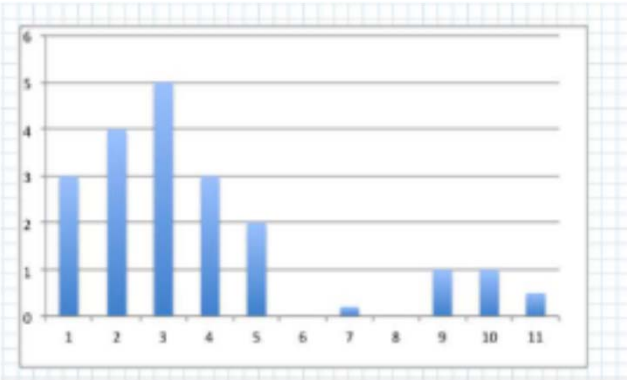
- We are also concerned with the patterns and shapes of graphs that represent our data.
- What we will look for:
  - Uniform
  - Symmetry
  - Bi-modal
  - Bell Shaped
  - Skewed Right
  - Skewed Left

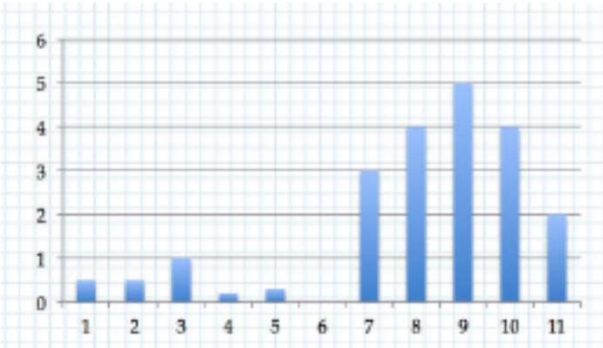

Histograms
Density
data
data


5.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6


6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11


3.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6


6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11


6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
11