Descriptive Statistics

What is Descriptive Statistics?

- A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of a collection of information.

- So it is a way of summarizing and presenting the data you have.

Data and Statistics

- Statistics is the science of learning from Data
- Data is essentially numbers (or text/symbols) which represent some information.
- It helps to think of data as 'values' of quantitative and qualitative variables
- What are the variable types:
  - Numerical or Quantitative: (Continuous and Discrete)
  - Categorical or Qualitative: (Always discrete)
    - Nominal
    - Ordinal

Descriptive Statistics

- Quantitatively describing the data
  - Graphical representation
  - Tabular representation
  - Summary statistics

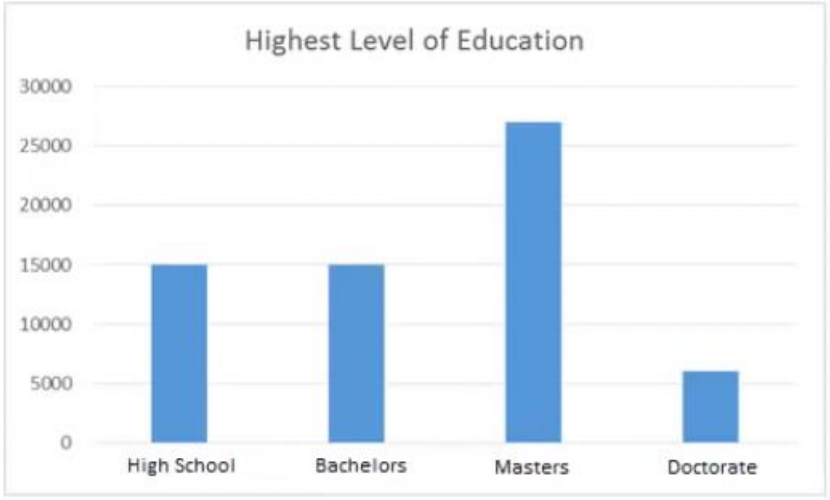
- Descriptive versus Inferential. The use of Sample and Population

- Single and Multiple variables
  - Distributions and Relationships

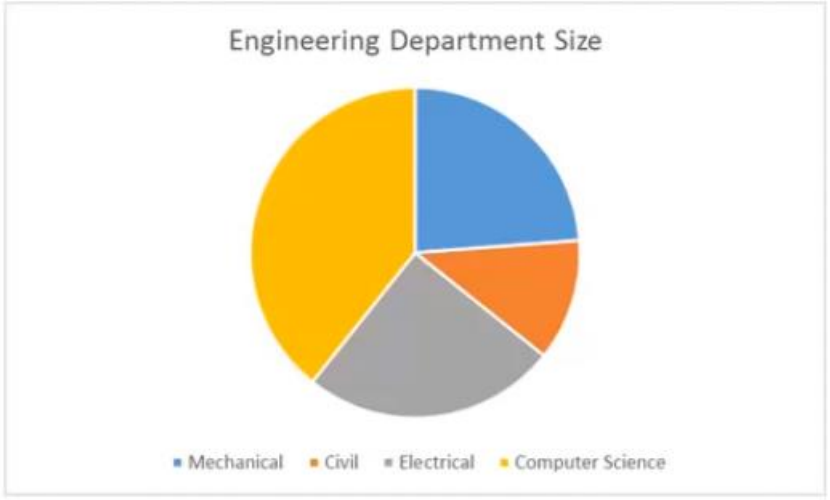
Graphical Representation: Single Variable

- For Categorical Variables

Bar Graph



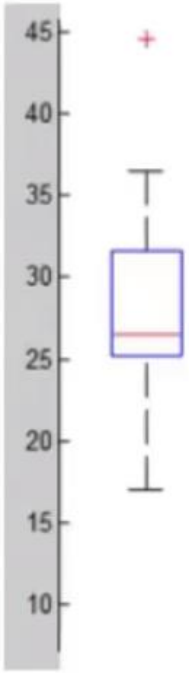
Pie Chart



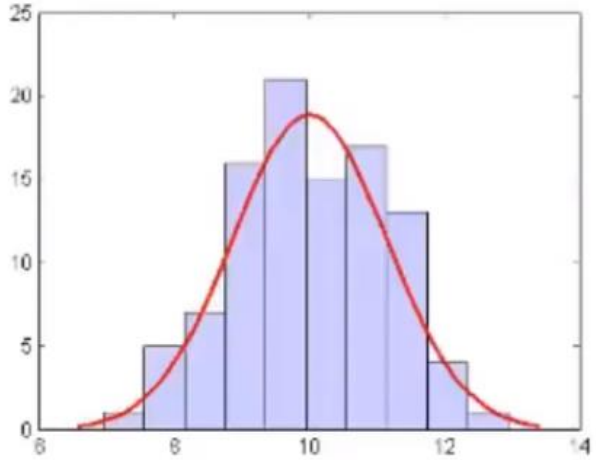
Graphical Representation: Single Variable

- For Quantitative variables

Box Plot



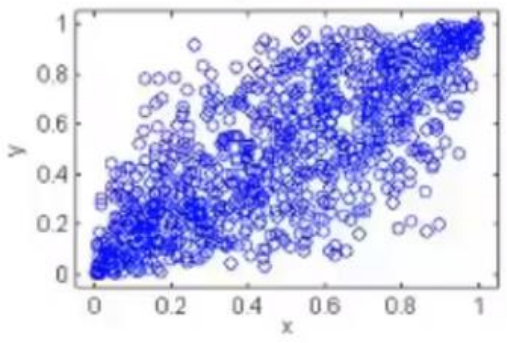
Histogram



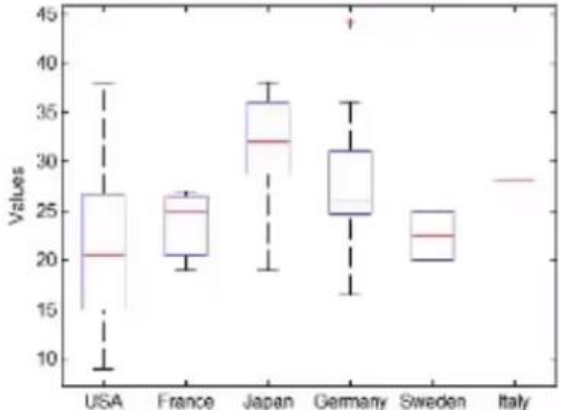
Graphical Representation: Multiple Variables

- Scatter Plots: Two quantitative variables
- Box plots – One categorical with one quantitative variable
- Contingency tables - 2 categorical variables with frequency of occurrence as the theme

Scatter Plot



Box Plots



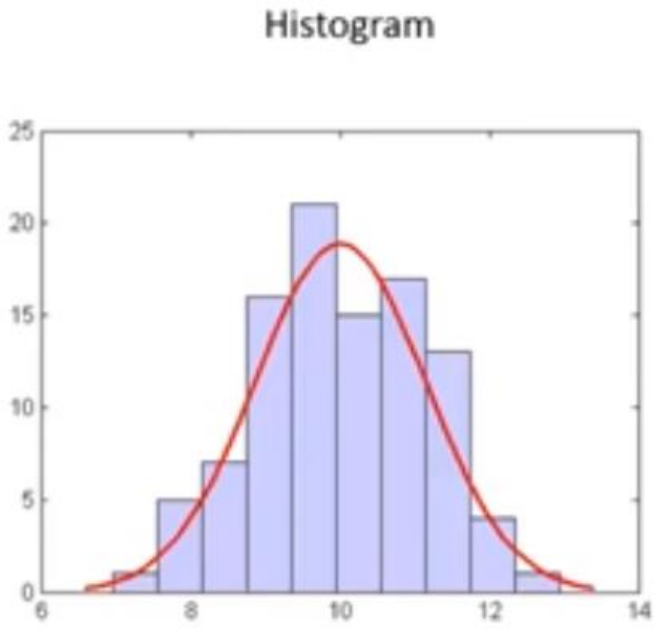
Contingency Table

Results		Work EX	
		Y	N
MBA	Y	22	7
	N	32	17

Summarizing Data through numbers

- Measures of Central Tendency
- Dispersion
- Skew and Kurtosis

Data Set
10.04
9.31
11.15
11.22
10.19
10.49
8.38
10.32
8.14
7.89
10.07
10.42
11.55
9.63
9.05
8.96
12.57



Measures of Central Tendency

- Data Set: 3,4,3,1,2,3,9,5,6,7,4,8
- Mean

$$\frac{3+4+3+1+2+3+9+5+6+7+8+4}{12} = 4.583 \qquad \frac{x_1+x_2+x_3+\cdots}{n} \text{ or } \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median

1,2,3,3,3,4,4,5,6,7,8,9 Hence Answer = 4

- Mode

The value 3 appears 3 times, and 4 appears 2 times and all other values appear once. Hence 3 is the mode

Measures of Central Tendency

- Where do we want to use Mean, Median and Mode
- Choosing between mean and median
  - Bad outliers
    - Errors
    - Do not provide a realistic picture of the story
  - Good outliers
    - The story is in the outliers
- Mode
  - Useful with nominal variables
  - Multi modal distributions

Measures of Dispersion

- Range
- Interquartile Range (75^{th} percentile – 25^{th} percentile)
- Standard Deviation and Variance
- Mean absolute deviation

Range

Example:

The two sets below have the same mean and median (7). Find the range of each set.

Set A	1	2	7	12	13
Set B	5	6	7	8	9

Range of Set A: $13 - 1 = 12$

Range of Set B: $9 - 5 = 4$

Percentile Calculation

- Step1: Order all the values in the data set from smallest to largest.
- Step2: Multiply *k* percent by the total number of values, *n*. This number is called the index.
- Step3: If the index obtained in Step 2 is not a whole number, round it up to the nearest whole number and go to Step 4a. If the index obtained in Step 2 is a whole number, go to Step 4b.

- Step 4a.Count the values in your data set from left to right (from the smallest to the largest value) until you reach the number indicated by Step 3. The corresponding value in your data set is the $k^{\mathbf{th}}$ percentile.
- Step 4b.Count the values in your data set from left to right until you reach the number indicated by Step 2. The k^{th} percentile is the average of that corresponding value in your data set and the value that directly follows it.

Percentiles

Example:

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the fortieth percentile.

40% = 0.4

0.4(30)

12

The average of the 12th and 13th items represents the 40th percentile (P_{40}).

40% of the scores were below 74.5.

Other Percentiles: Deciles and Quartiles

Deciles are the nine values (denoted D_1, D_2,…, D_9) along the scale that divide a data set into ten (approximately) equal parts.

10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%

Quartiles are the three values (Q_1, Q_2, Q_3) that divide the data set into four (approximately) equal parts.

25%, 50%, and 75%

Other Percentiles: Deciles and Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

$Q_1 = 25\%$

$25\% = 0.25$

$0.25(30)$

7.5

The 8^{th} item represents the 1^{st} quartile (Q_1)

25% of the scores were below 72.

Other Percentiles: Deciles and Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

Q_2= 50% = median

50% = 0.5

0.5(30)

15

The average of the 15th and 16th items represents the 2nd quartile (Q_2) or the median

50% of the scores were below 78.5.

Other Percentiles: Deciles and Quartiles

Example: Quartiles

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Find the three quartiles.

Q_3= 75%

75% = 0.75

0.75(30)

22.5

The 23^{rd} item represents the 3^{rd} quartile (Q_3)

75% of the scores were below 88.

Box Plots

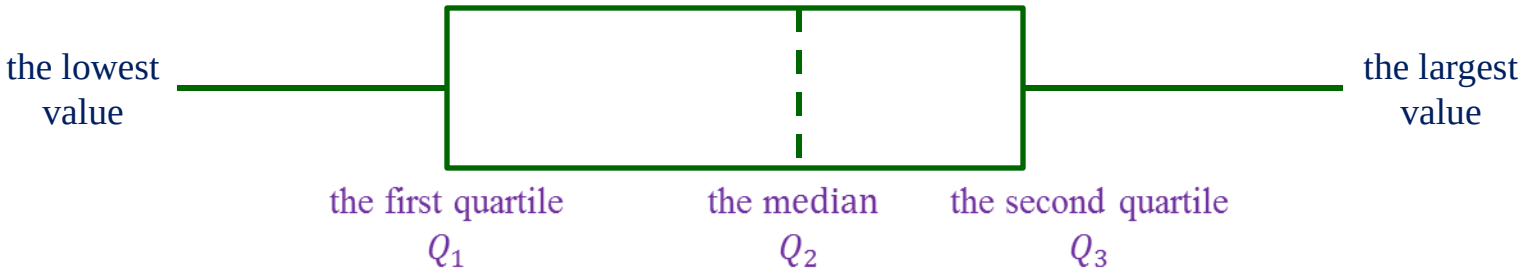
A box plot or a box and whisker plot is a visual display of five statistical measures.

The five statistical measures are:

the lowest value,

the first quartile, the median, the third quartile,

the largest value.



The Interquartile Range: $IR = Q_3 - Q_1$

Box Plots

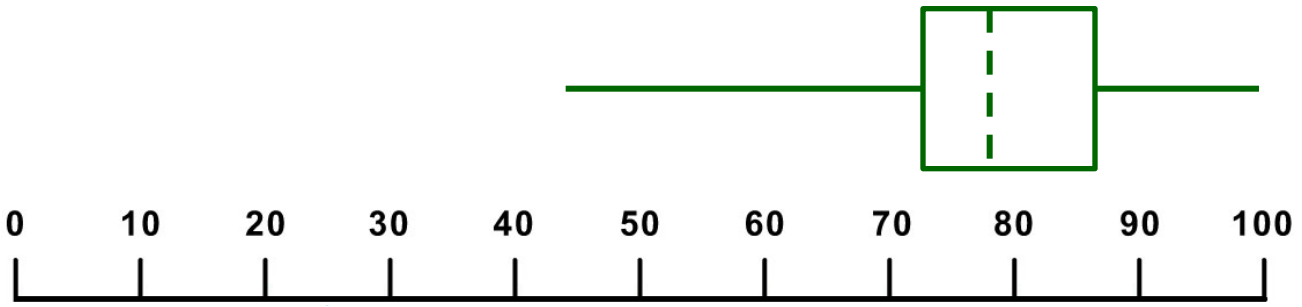
Example:

The following are test scores (out of 100) for a particular math class.

44	56	58	62	64	64	70	72	72	72
74	74	75	78	78	79	80	82	82	84
86	87	88	90	92	95	96	96	98	100

Q_1= 25% = 72 Q_2= 50% = median= 78.5 Q_3= 75%= 88

Lowest = 44 Largest = 100



The Interquartile Range: $IR = Q_3 - Q_1 = 88 - 72 = 16$

Standard Deviation

Calculating the Sample Standard Deviation

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

The **sample standard deviation** is found by calculating the square root of the **variance**.

The **variance** is found by summing the squares of the deviations and dividing that sum by $n - 1$ (since it is a sample instead of a population).

The sample standard deviation is denoted by the letter s.

The standard deviation of a population is denoted by σ.

Measures of Dispersion

- Questions that go with Standard deviation
  - Why do we use the square function on the deviations? What are its implications?
  - Why do we work on standard deviation and not the variance?
  - Why do we average by dividing by N-1 and not N?

- Mean absolute Deviation and its variants
  - Use $|x_i - \bar{x}|$ instead of $(x_i - \bar{x})^2$