STAT72000 – Probability and Statistics

LECTURE 1 – DESCRIPTIVE STATISTICS AND DATA SUMMARIES

Agenda

- Summarizing Data
- Visualizing Data
- Setup R language and Rstudio

Summarizing Data

- DESCRIPTIVE STATISTICS

Types of Descriptive Statistics

Summarize Data

- Measures of Location
- Measures of Dispersion

Organize Data

- Tables
- Graphs

Summarizing Data

Measures of Location

- Mean
- Median
- Mode

Measures of Dispersion

- Range
- Quartiles
- Percentiles
- Variance
- Standard Deviation

Measures of Location (Central Tendency) MEAN

- The Mean is a measure of *central tendency*
  - What most people mean by “average”
  - Sum of a set of numbers divided by the number of numbers in the set

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

Measures of Location (Central Tendency) MEDIAN & MODE

Median is the middle value when a set of numbers are arranged in ascending order. That is, it divides a set precisely in to two equal halves: 50% above the median, 50% below the median. Also known as the 50^{th} percentile.

- If number of samples n is odd, the sample median at the position (n+1)/2
- If n is even, the sample median is the average of numbers in positions of (n/2) and (n/2 +1)

Mode is the most frequently occurring number in a set.

- If several values appear with equal frequency, each one is a mode
- Given samples, it can have more than one mode

Summary of Central Tendency Measures

Type	Definition	Synonyms
Mean	Sum of all values divided by number of values.	Average
Median	The value such that one half of the data lies above and below	50% percentile
Mode	The most frequently occurring number.	

Adapted from Bruce and Bruce, 2017, Wiley Academic

Exercise: find the sample **mean**, **mode** and **median**

7 numbers X: [3, 4, 5, 3, 3, 1, 2]

Mean =

Mode =

Median =

Trimmed Mean:

- The trimmed mean is a measure of center that is designed to be unaffected by outliers.
- The trimmed mean is computed by:
  1) Arranging the sample values in order
  2) Trimming an equal number of them from the smallest end and the largest end
  3) Computing the mean of the remaining values
  4) If p% of the data are trimmed from each end, the resulting trimmed mean is called "p% trimmed mean"
     - e.g. 5% trimmed mean, 10% trimmed mean

Exercise: find trimmed mean

A list of 24 numbers:

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- Mean =
- Median =
- 5% trimmed mean =
- 10% trimmed mean =
- 20% trimmed mean =

Summarizing Data

Measures of Location

- Mean
- Median
- Mode

Measures of Dispersion

- Range
- Quartiles
- Percentiles
- Variance
- Standard Deviation

Measures of Dispersion (a.k.a. How spread out the data is)

A measure of dispersion, is used to describe the variability (*how spread out or tightly clustered*) in a sample or population. It is usually used *in conjunction with* a measure of central tendency, such as the mean or median, to provide an overall description of a set of data.

Example

- Data set 1: [1,25,50,75,100]
- Data set 2: [48,49,50,51,52]
  - Both have a mean of 50, but data set 1 clearly has greater dispersion than data set 2.

Measures of Dispersion: RANGE

The range is *one measure of dispersion*

The range is the difference between the maximum and minimum values in a set

Example

Data set 1: [1,25,50,75,100]; Range: 100-1 = 99

Data set 2: [48,49,50,51,52]; Range: 52-48 = 4

The range ignores how data are distributed and only takes the extreme scores into account

$$RANGE = (X_{largest} - X_{smallest})$$

Measures of Dispersion: Variance and Standard Deviation

Sample Variance:

The variance of a set of numbers is the average of the square of the deviations from the mean.

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

Sample Standard Deviation:

The standard deviation of a set of numbers is the positive square root of the variance.

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

Exercise: sample standard deviation

A sample of 5 people heights are collected (in inches) are:

65.51, 72.30, 68.31, 67.05, 70.68

Sample mean:

Sample variance:

Sample standard deviation:

(to find the sample std, why need to go through the above 3 steps?)

Quartiles

- Quartiles divide samples into quarters
- A sample of values has 3 quartiles:
  - Order the sample values from smallest to the largest
  - First Quartile position: 0.25*(n+1)
  - Second Quartile position (same as median): 0.5*(n+1)
  - Third Quartile position: 0.75*(n+1)

Exercise: Quartiles

Given the 24 numbers of data:

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- The first quartile:
  - position: 0.25*(24+1) = 6.25
  - the first quartile is average of 6^{th} and 7^{th} data points: (105+126)/2 = 115.5
- The second quartile:
- The third quartile:
- The IQR (interquartile range):

Percentiles

- The p-th percentile: p between 0 to 100, divides the sample so that nearly p% of the sample values are less than the p-th percentile, and (100-p)% are greater

- How to compute:
  - Order the sample values from smallest to the largest
  - Compute the position: (p/100) * (n+1)

Exercise: Percentiles

Given a sample of 24 numbers:

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

- Find the 65th percentile
  - The position: 0.65* (24+1) = 16.25
  - The 65th percentile is the average of 16th and 17th numbers = (236+240)/2 = 238

Categorical Data – Frequencies and Sample Proportions

- In most categorical data, statistics like the averaging have no physical meaning
- Frequency: the number of sample items that fall into the category
- Sample Proportion (relative frequency): is the frequency divided by the sample size

- Example: a survey on the product feedback receives 1000 responses. 810 answer 'like it', 53 'neutral', 137 'dislike'
  - Frequencies are: 810, 53, 137
  - Sample proportions are: 810/1000 = 0.810, 53/1000=0.053, 137/1000=0.137

Visualizing Data

Displaying Data

Graphical Summaries

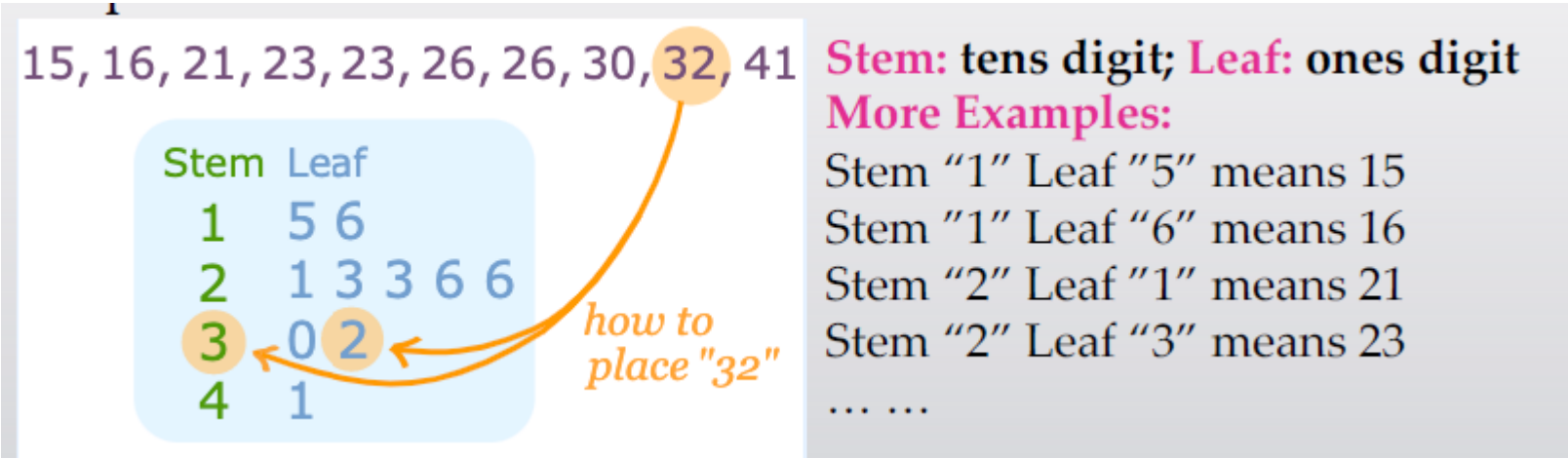
- Stem-and-leaf Plot
- Dot Plot
- Histogram
- Box Plot
- Scatter Plot

Graphical Summaries

- The mean, median, standard deviation are numerical summaries of a sample
- Graphical summaries are used to visualize a list of samples

Stem-and-leaf Plot

- Stem and Leaf Plot: a special table where each data value is split into a ‘stem’ (the first digits) and a ‘leaf’ (usually the last digit)



Example: stem-and-leaf plot

- a group of friends played a long jump and got the results:

  2.3, 2.5, 2.5, 2.7, 2.8, 3.2, 3.6, 3.6, 4.5, 5.0

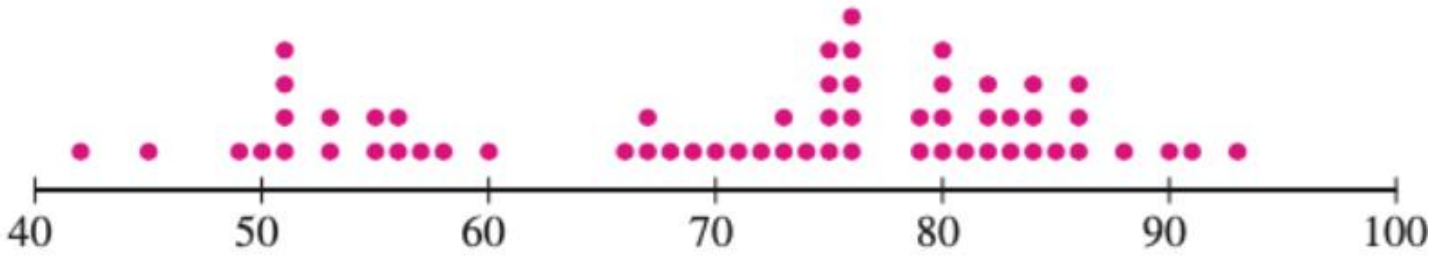
Stem	Leaf
2	3 5 5 7 8
3	2 6 6
4	5
5	0

- **Stem**: whole; **Leaf**: decimal.
- Stem "2" Leaf "3" means **2.3**)
- In this case each leaf is a decimal
- It is OK to repeat a leaf value
- 5.0 has a leaf of "0"

Dot Plot

- To plot the shape of a sample
- Large sample size and containing repeated values

42	45	49	50	51	51	51	51	53	53
55	55	56	56	57	58	60	66	67	67
68	69	70	71	72	73	73	74	75	75
75	75	76	76	76	76	76	79	79	80
80	80	80	81	82	82	82	83	83	84
84	84	85	86	86	86	88	90	91	93



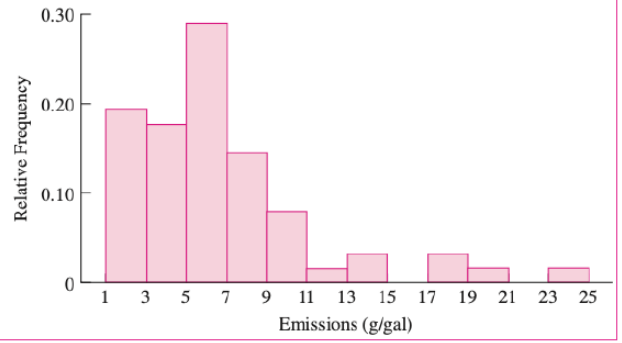
Histogram

- Histogram plots the shape of a sample, indicating regions where sample points are concentrated and regions where samples are sparse.

- Steps:
  1) Construct a frequency table
  2) Draw histograms to describe ‘frequency’ or ‘relative frequency’ for each interval

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02	23.38	9.24	3.22
2.06	4.04	17.11	12.26	19.91	8.50	7.81	7.18	6.95	18.64	7.10	6.04	5.66
8.86	4.40	3.57	4.35	3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14
8.67	9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01	5.91	5.60
5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32	1.83	7.56			

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 13	1	0.0161	0.0081
13–< 15	2	0.0323	0.0161
15–< 17	0	0.0000	0.0000
17–< 19	2	0.0323	0.0161
19–< 21	1	0.0161	0.0081
21–< 23	0	0.0000	0.0000
23–< 25	1	0.0161	0.0081



Histograms – unequal bin width

- When the bin intervals are of unequal width, the heights of the rectangles must set to the densities. The areas under the rectangles are the relative frequencies.
- density = relative-frequency / bin-interval-width

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02	23.38	9.24	3.22
2.06	4.04	17.11	12.26	19.91	8.50	7.81	7.18	6.95	18.64	7.10	6.04	5.66
8.86	4.40	3.57	4.35	3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14
8.67	9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01	5.91	5.60
5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32	1.83	7.56			

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 15	3	0.0484	0.0121
15–< 25	4	0.0645	0.0065

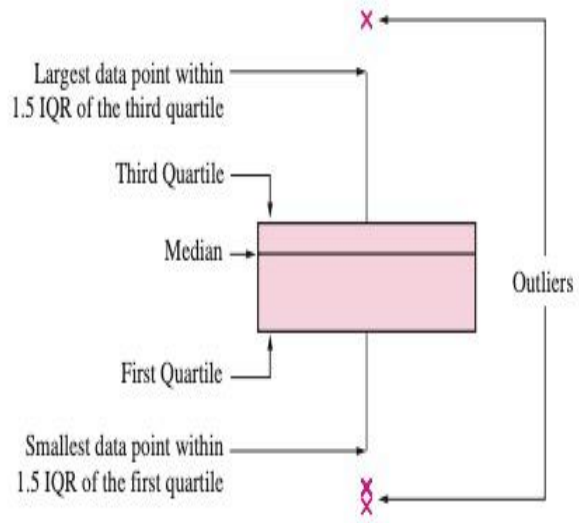
What is the **proportion** in the sample with emissions between 9 and 15 g/gal?

The proportion is the sum of the relative frequencies of the two spanning the range between 9 and 15.
So, the proportion between 9 and 15 is therefore equal to 0.0806+0.0484 = 0.129.

Box Plot

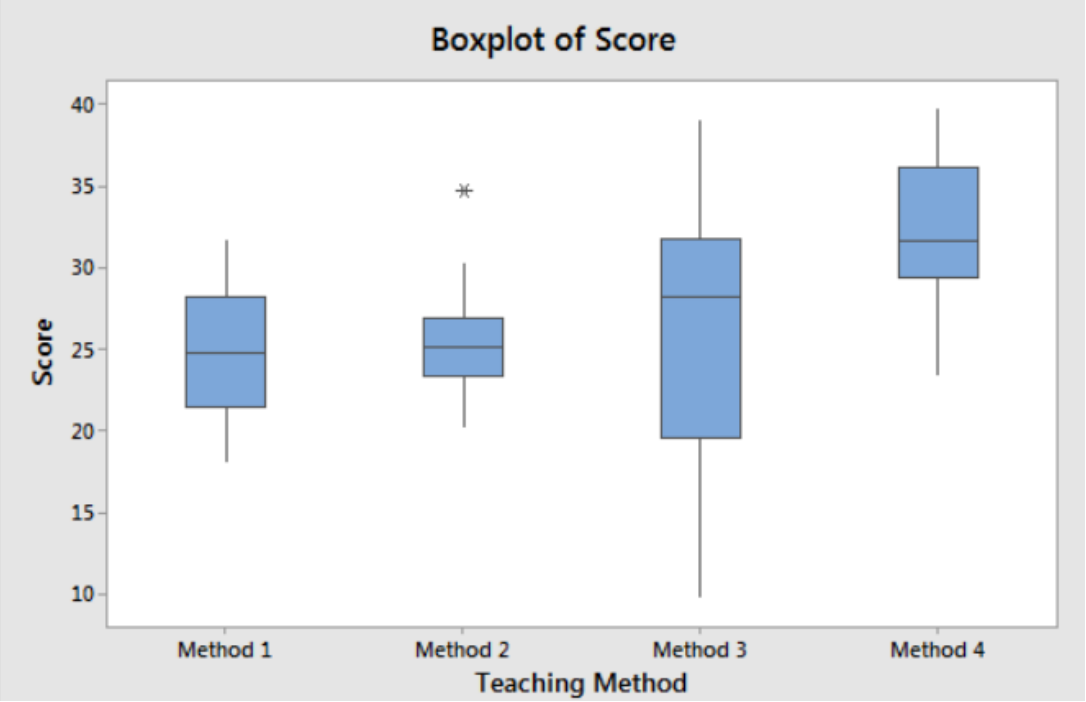
A box plot presents median, 3rd quartile, 1st quartile, minimum and maximum, and outliers.

- Interquartile Range (IQR): the difference between 3rd quartile and the 1st quartile
- Outliers: 1.5 IQR above the 3rd quartile, or 1.5 IQR below the 1st quartile is considered as outliers
- Extreme Outlier: more than 3 IQR away from the first and the third quartile.



Comparative Boxplots

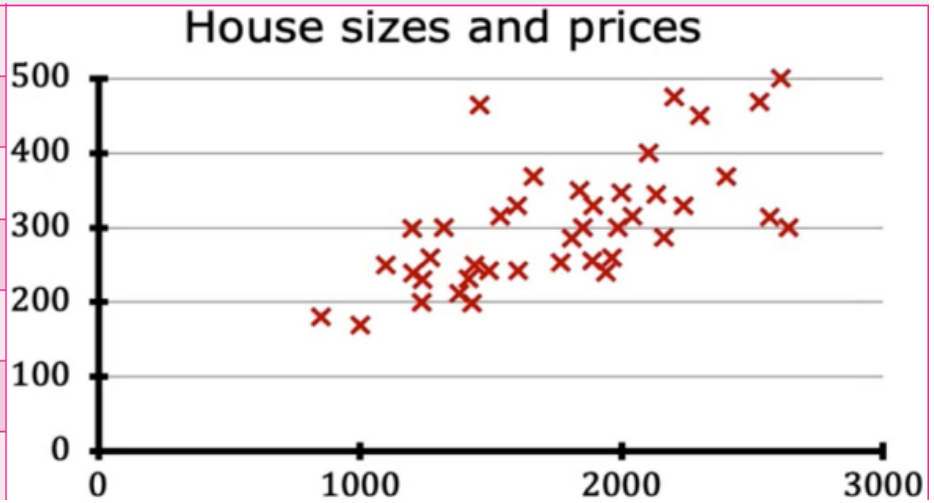
Several boxplots can be placed side by side, allowing for easy visual comparison.



Scatter Plot

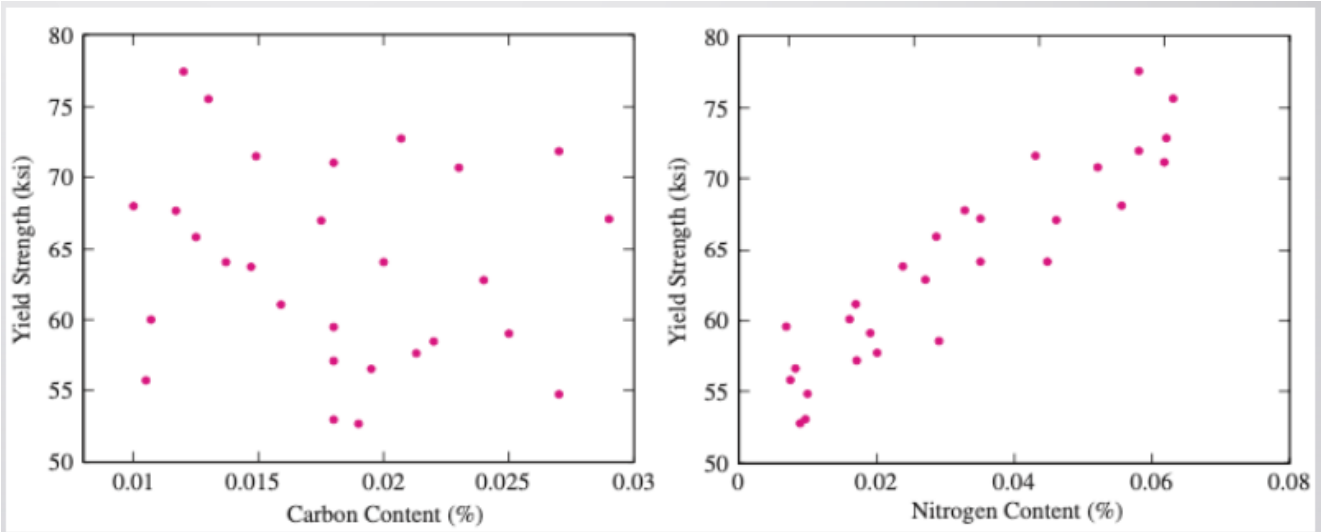
- Data for each item consists of more than one value, is called multivariate data
- when each item is a pair of values, the data are said to be bivariate
- Scatter plot is a useful graphic summaries for bivariate data

Size in Feet²	Price in $ 1000's
2104	400
1416	232
1534	315
852	178
…	…
3210	870



Scatter Plot Example

- A number of welds samples describe the chemical composition and strength characteristics
- Scatter plot: Carbon Content *vs.* Yield Strength; Nitrogen Content *vs.* Yield Strength



Not much of a relationship between carbon content and yield strength for a certain group of welds

For these same weld, higher nitrogen content is **associated** with higher yield strength

Setup R and Rstudio

- Install R: https://cran.r-project.org/bin/windows/base/
- Install Rstudio: https://posit.co/download/rstudio-desktop/
- Handon Programming with R: https://rstudio-education.github.io/hopr/