Collecting & Organizing Univariate Data

Univariate - having only 1 variable

- Data can either be Qualitative or Quantitative

Quantitative - numerical data focused on Quantity
- Can be Discrete or Continuous

Qualitative - non-numerical data focused Quality (i.e. Categorical)

lets think of some examples of each type of data

Quantitative

- Height of pro basketball players
- # of world wide COVID cases
- Test Scores
- Track times in 100m dash

Qualitative

- Ice cream flavors
- Happiness Rating
- Pass / Fail
- Eye Color
- Interview Transcript

Central Tendency

aka "averages" (mean, mode, median)

Σ - "the sum of"

"X bar"

Mean ($\bar{x}$) : average value in data set $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, where n - # of values

Set: 6 8 9 10 12

x_1 x_2 x_3 x_4 x_5

n = 5 terms

$$\bar{x} = \frac{1}{5} \sum_{i=1}^{5} x_i = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$$

$$= \frac{1}{5}(6 + 8 + 9 + 10 + 12) = \frac{1}{5}(45) = 9$$

Mode : The most frequently ocurring value in the data set

* there can be 0, 1, or 2 modes (namely, bimodal)

Set_1: 1 2 3 3 4 4 5 6 — mode: 3, 4

Set_2: 1 2 3 4 5 6 — mode: none

Set_3: 1 2 3 4 5 6 6 — mode: 6

Median : The middle value in a sorted data set

Set_1: $\overbrace{1 \quad 2 \quad \boxed{3 \quad 4} \quad 5 \quad 6}^{n \text{ terms}}$

* If n is even (there's no exact middle term)
then median is the avg of the $\boxed{\text{two middle terms}}$ = 3.5

Set_2: 8 10 12 13 14

median: 12

Ex) Grades for a History test for 14 students are shown below

69 58 67 66 58 79 83 76 49 35 58 88 91 47

When a 15^{th} student took the test the mean became 66.2
Calculate the grade for the 15^{th} student

$Mean_i = 66$, $Mean_f = 66.2$ => $Student_G = \frac{924 + x}{15} = 66.2$

$$\boxed{x = 15 \cdot 66.2 - 924 = 69}$$

Frequency Tables

The lengths, in minutes, of 20 telephone calls are shown below

4.2	6.8	10.4	8.2	11.5	1.6	5.8	7.6	3.1	21.5
13.5	5.8	4.1	22.8	13.6	11.2	9.5	1.8	12.4	4.9

Organize the data into a Frequency Table

Length (mins)	Frequency
$0 \le t < 5$	6
$5 \le t < 10$	6
$10 \le t < 15$	6
$15 \le t < 20$	0
$20 \le t < 25$	2

☆ "Grouped data" because it has continuous ranges for each class

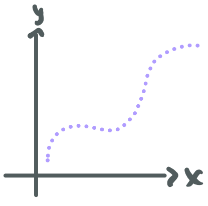
IS this data Continuous or Discrete ?

Discrete data - can be counted (ex: Shoe size, # of cars in a parking lot)

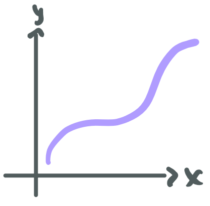
Continuous data - can be measured (height, weight, time)

☆ must contain full range of values

How can you tell if a graph is continuous or discrete ?



Discrete



Continuous

ex| Men's Jeans are sized by waist measurements. Here are the the jean sizes of 10 men:

28 30 28 34 32 30 36 28 30 30

ⓐ Find the mean, median, mode, Range

ⓑ Create a frequency table

ⓒ Is this data continuous or discrete? why?

ⓐ Sorted: 28 28 28 30 30 30 30 32 34 36

Mean = 30.6

Median = 30

Mode = 30

Range = 8

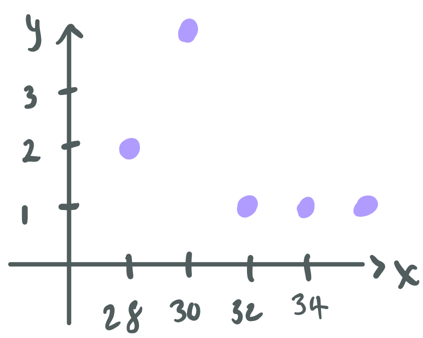
ⓑ

Jean Size	Frequency
28	3
30	4
32	1
34	1
36	1

★ Modal Group/Class - class or category that contains the highest frequency

⇒ Jean Size 30 is modal class

ⓒ Discrete because jean sizes are only whole even values



Outliers

ex| The ages of 15 cats are below:

10 10 11 11 11 12 12 12 12 13 13 14 14 24 25

Find the mean, mode, and median for this data

(a) Mean = 13.6

(b) Median = 12

(c) Mode = 12

Are there any singular data points that affect the calculation of the mean more so than the others?

Remove these values and compare recalculate the mean.

★ 24, 25 are examples of Outliers: extreme data values that can distort the results of statistical processes

- They don't fit with the rest of the data

★ Outliers are often the result of errors in collecting/reading data

Outliers ex | Find the mean, median, and mode for each data set & comment on any data values you think are outliers:

ⓐ The heights of 15 sunflowers:

1.1	2.2	2.5	2.5	2.5	3.1	3.5	3.6
3.9	4.0	4.1	4.4	4.6	4.9	6.1	

Mean = 3.53
Median = 3.6
Mode = 2.5

Outliers: 1.1, 6.1

ⓑ 20 Students' Geography test scores:

22	39	45	46	46	52	54	58	62	62
62	67	70	75	78	82	89	91	95	98

Mean = 64.65
Median = 62
Mode = 62

Outliers: 22

Sigma Notation Warm Up

Recall: Σ - upper case "sigma"
σ - lower case "sigma"

$$\sum_{i=a}^{n} i$$

n → Index upper bound (ending point)
a → Index lower bound (starting point)
i → Index

① Find $\sum_{i=1}^{4}(i+i^2) = (1+1^2)+(2+4)+(3+9)+(4+16)$
$= 2 + 6 + 12 + 20 \Rightarrow \boxed{40}$

② Find $\sum_{i=1}^{4} i + \sum_{i=1}^{4} i^2 = (1+2+3+4)+(1^2+2^2+3^2+4^2) \Rightarrow \boxed{40}$

③ Find $\sum_{i=1}^{3} y = y+y+y = \boxed{3y}$

Measures of Dispersion

(Range, standard deviation, Interquartile Range)

- Dispersion - the action / process of distributing things over a wide area
- Measures how spread out the data is
- We already learned a measure of Dispersion: Range

ex Consider the two data sets:

$S_1 = -10, 0, 10, 20, 30$ $\qquad$ $S_2 = 8, 9, 10, 11, 12$

mean: $\frac{50}{5} = 10$ $\qquad$ mean: $\frac{50}{5} = 10$

Range: $30 + 10 = 40$ $\qquad$ Range: $12 - 8 = 4$

Thus S_1 is more dispersed (more spread out)

Standard Deviation (σ): measure of how data values are spread out in relation to the mean

aka "root-mean-squared deviation"

* σ - "sigma"

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

* mean of sample

$$\sigma_1^2 = \frac{(-10-10)^2 + (0-10)^2 + (10-10)^2 + (20-10)^2 + (30-10)^2}{5} = 200$$

$$\sigma_2^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = 2$$

$\Rightarrow$ $\sigma_1 = \sqrt{200} = 10\sqrt{2}$ $\longrightarrow$ more dispersed than σ_2

$\sigma_2 = \sqrt{2}$

ex) The # of ice creams sold over a period of 13 weeks:

146 151 158 158 161 149 160 147 158 160 216 225 238

Find the Standard Deviation

$$\bar{X} = \frac{2227}{13} = 171.3$$

$$\sigma^2 = \frac{(146-171.3)^2 + (149-171.3)^2 + (151-171.3)^2 + 3(158-171.3)^2 + 2(160-171.3)^2 + (161-171.3)^2 + (216-171.3)^2 + (225-171.3)^2 + (238-171.3)^2 + (147-171.3)^2}{13} = 950.98$$

$$\Rightarrow \boxed{\sigma = \sqrt{950.98} = 30.8}$$

This is EXHAUSTING

We can do some algebraic manipulation to Rewrite this formula and make it simpler

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]}_{=\bar{x}} + \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\right]}_{=\frac{1}{n}(n\bar{x}^2)}$$

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

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

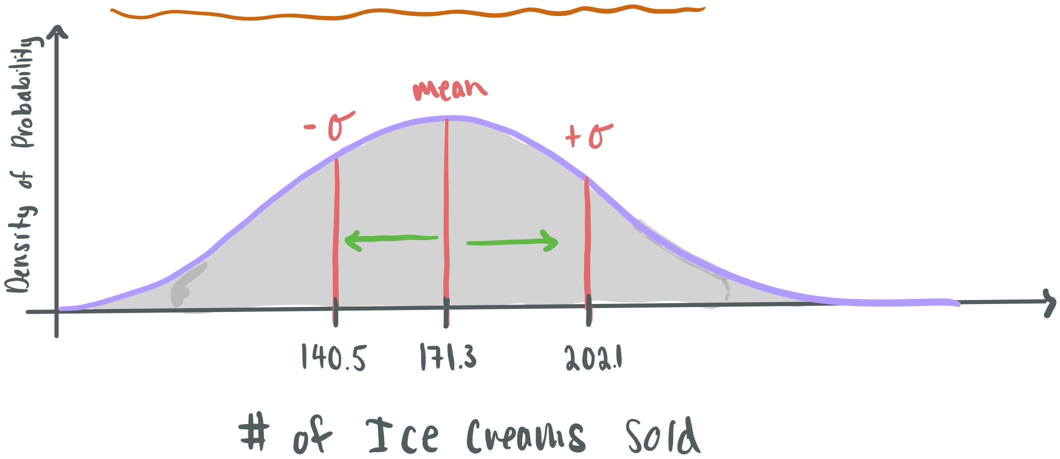
☆ Easier and should be used

Recalculate the Standard Deviation using our newly formulated definition

$$\sigma^2 = \frac{146^2 + 151^2 + 158^2 + 158^2 + 161^2 + 149^2 + 160^2 + 147^2 + 158^2 + 160^2 + 216^2 + 225^2 + 238^2}{13} - 171.3$$

$$\boxed{\sigma = \sqrt{953.6} = 30.8}$$

Normal Distribution Curve

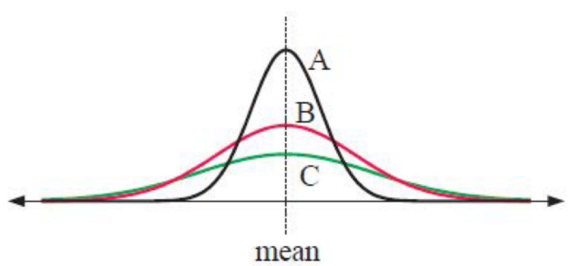


*Density of Probability - likelihood of obtaining a value corresponding to the x-axis

For Example

There's a high probability that the ice cream shop will sell 171.3 ice cream cones in 13 weeks because it has the highest y-value

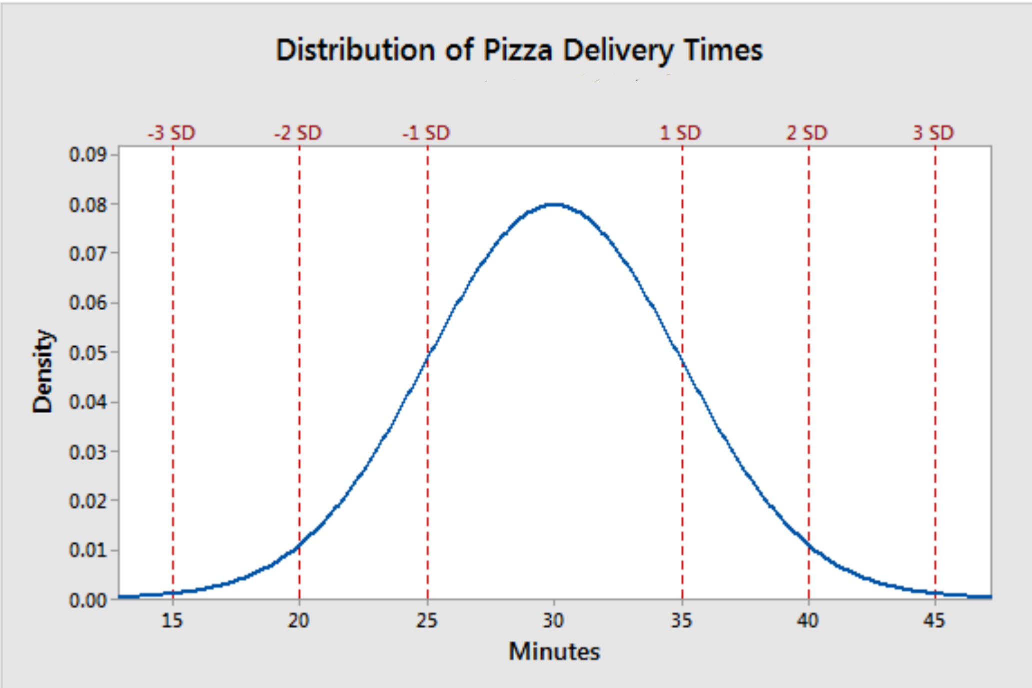
The given distributions have the same mean, but clearly they have different spreads. The A distribution has most scores close to the mean whereas the C distribution has the greatest spread.



ex) Use the graph below to find:

(a) Mean ($\bar{x}$) : 30 mins

(b) Standard Deviation (σ) : 5 mins



Interquartile Range (IQR)

better represents the dispersion of data compared to Range

ex) The following data set shows the # of animal crackers in each kid's lunch box

4 4 10 11 15 7 14 12 6

Step 1 Find the Median : 10

Sort list : $\underbrace{4 \quad 4 \quad 6 \quad 7}_{\text{lower Quartile } Q_1}$ $\boxed{10}$ $\underbrace{11 \quad 12 \quad 14 \quad 15}_{\text{Upper Quartile } Q_3}$

Step 2 Find median of Q_1 and Q_3

$$Q_1 = \frac{4+6}{2} = 5 \qquad Q_3 = \frac{12+14}{2} = 13$$

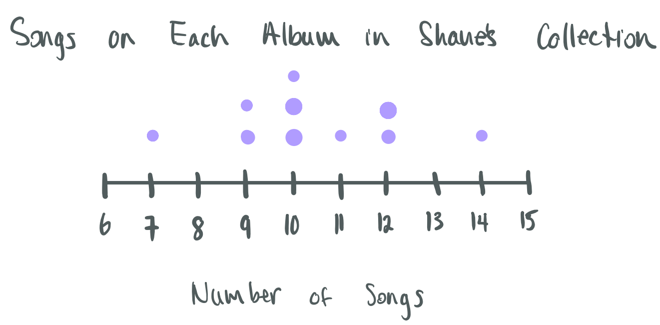
Step 3 Find the difference btw the two Ranges

$$Q_3 - Q_1 = 13 - 5 = 8$$

$$\Rightarrow \boxed{IQR = 8}$$

$$IQR = \underbrace{\text{Upper Quartile } (Q_3)}_{\text{data point at the 75\% percentile}} - \underbrace{\text{Lower Quartile } (Q_1)}_{\text{data point at the 25\% percentile}}$$

ex) Find the IQR of the data shown in the dot plot below



7 9 9 10 10 | 10 11 12 12 14

Median: $\frac{10+10}{2} = 10$

Q_1 median: 9

Q_2 median: 12

$IQR = Q_3 - Q_1 = 12 - 9 = \boxed{3}$

A **percentile** is the score below which a certain percentage of the data lies.

For example:
- the 85th percentile is the score below which 85% of the data lies.
- If your score in a test is the 95th percentile, then 95% of the class have scored less than you.

Notice that:
- the **lower quartile** (Q_1) is the 25th percentile
- the **median** (Q_2) is the 50th percentile
- the **upper quartile** (Q_3) is the 75th percentile.

A cumulative frequency graph provides a convenient way to find percentiles.

Revisiting Outliers

- an extremely high or extremely low value in our data

A data value (z) is an outlier if:

$$z > Q_3 + 1.5(IQR)$$

or

$$z < Q_1 - 1.5(IQR)$$

ex) Phone Calls Received in a day

10	12	11	15
11	14	13	17
12	22	14	11

Find any outliers

Sort: $\boxed{10 \quad 11 \quad 11 \quad 11 \quad 12 \quad \boxed{12}}\ \boxed{\boxed{13} \quad 14 \quad 14 \quad 15 \quad 17 \quad 22}$

Median: 12.5
Q_1: 11
Q_3: 14.5
IQR: 3.5

$\Rightarrow$

Outliers < 11 − 1.5(3.5)

14.5 + 1.5(3.5) > Outliers

$\Downarrow$

Outliers < 5.75 $\longrightarrow$ None

19.75 > Outliers $\longrightarrow$ $\boxed{22}$

Central Tendencies of Grouped Data

* Note if only given a grouped frequency table then you cannot find the exact values for mean, median, and mode but we can approximate them

ex) Consider the grouped data of student test scores:

Score	Frequency	Midpoint	(Midpoint)(Frequency)
$0 \le x < 60$	5	30	150
$60 \le x < 70$	4	65	260
$70 \le x < 80$	10	75	750
$80 \le x < 90$	15	85	1275
$90 \le x \le 100$	4	95	380

Find

ⓐ Mean

Step 1: Find Midpoint of each class

Step 2: Find product of Midpoints and Frequencies

Step 3: Find n - total # data points $= \sum \text{Frequencies} = 38$

Step 4: Find sum of all (Midpoint)(Frequency) products $= 2815$

Step 5: $\bar{X} = \dfrac{\sum(\text{Midpoint})(\text{Frequency})}{n} = \dfrac{2815}{38} = \boxed{76.1}$

ⓑ Median Class

$\frac{n}{2} = \frac{38}{2} = 19$

$\Rightarrow$ 19th and 20th values

$\boxed{\text{Either } 70 \le x < 80 \quad 80 \le x < 90}$

ⓒ Modal Class

$\boxed{80 \le x < 90}$ - Class with highest frequency

Summary

Quantitative - numerical data focused on Quantity

Qualitative - non-numerical data focused Quality (i.e. Categorical)

Univariate - having only 1 variable

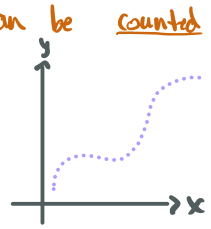
Central Tendency - Values that describe the middle of a data set
(mean, median, and mode are all measures of central tendency)

Mean ($\bar{x}$) : average value in data set. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$, where n - # of values

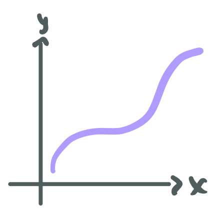
Mode : The most frequently ocurring value in the data set

Median : The middle value in a sorted data set

Discrete data - can be counted (ex: shoe size, # of cars in a parking lot)



Continuous data - can be measured (height, weight, time)



Outliers : extreme data values that can distort the results of statistical processes
A data value (z) is an outlier if:

$$z > Q_3 + 1.5(IQR)$$

or

$$z < Q_1 - 1.5(IQR)$$

Standard Deviation (σ): measure of how data values are spread out in relation to the mean

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

Normal Distribution Curve

Measures of Dispersion: values that describe how spread out the data is (Range, and IQR are measures of Dispersion)

Interquartile Range (IQR) = $Q_3 - Q_1$
(better represents dispersion)

Range: (highest value) − (Lowest Value)

Practice Problems

Pg 99, 3A, Q1
Pg 100, Ex2, Q.2
Pg 107, Ex5, Q.b,c
Pg 109, Ex6, Q. a,b,c
Pg 110, Investigation 6