Subject : .... Date : 19/8/23

Summary Measures and Decisions

The mean and standard deviation are a good way to measure central tendency.

For any mound shape, nearly symmetric distribution ~~data~~ of data;

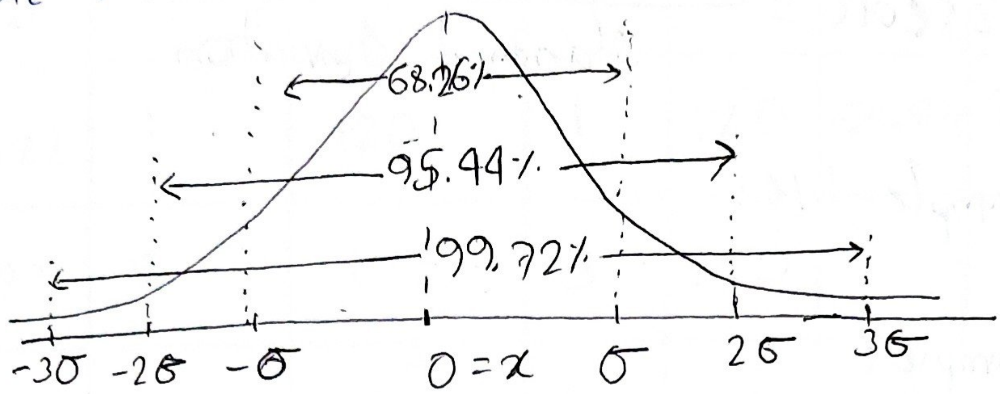
→ The interval $\bar{x} \pm s$ contains approximately 68% of data points.

→ The interval $\bar{x} \pm 2s$ contain approx. 95% of data points

→ Interval $\bar{x} \pm 3s$ contain approx. all the data points.

Subject :.. Date :..................................

Standard normal deviation: $M=0$ $S.d.=1$



Example-1.15 pg-39 (book)

~~One of the~~

[mid point also called class terms]

[$\bar{x}/\mu$ can be used for mean. μ/σ means whole population.

Sample 9σ mean we use $\bar{x}/s$]

Standardized values and z-scores:

We can calculate any data point if we know the ~~number~~ mean and standard deviation.

z-score doesn't have a unit so it is easier to ~~calculat~~ compare data point of two different datas.

Subject :.. Date :..........................

$$\text{zscore} = \frac{\text{measurement} - \text{mean}}{\text{standard deviation}}$$

Example-1.16:

Example:

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	~~176~~ 176	163	153	154	165
146	173	142	147	135	119	140	135
161	145	135	142	150	156	145	128

largest weight = 176 lb

smallest weight = 119 lb

Range = 176 − 119 = 57 lb

class interval = 5

no. of classes = $\frac{57}{5} \simeq 12$

(146.875)

Subject : $\sum f = 40$ $\qquad$ 5875 $\qquad$ Date :

Class interval	Frequency	Midpoint x_i	Cumulative frequency c.f	fx_i	$x - \bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
118-122	1	120	1	120	-26.875	722.265	722.265
123-127	2	125	3	250			
128-132	2	130	5	260			
133-137	4	135	9	540			
138-142	6	140	15	840			
143-147	8	145	23	1160			
148-152	5	150	28	750			
153-157	4	155	32	620			
158-162	2	160	34	320			
163-167	3	165	37	495			
168-172	1	170	38	170			
173-177	2	175	40	350			

Subject : .... Date : ....

	frequency	midpoint
118-126	3	122
127-135	5	131
136-144	9	140
145-153	12	149
154-162	5	158
163-171	4	167
172-180	2	176
	Total = 40	

$$\text{Mean}, \bar{x} = \frac{\Sigma f x_i}{n} = \frac{5875}{40} = 146.875$$

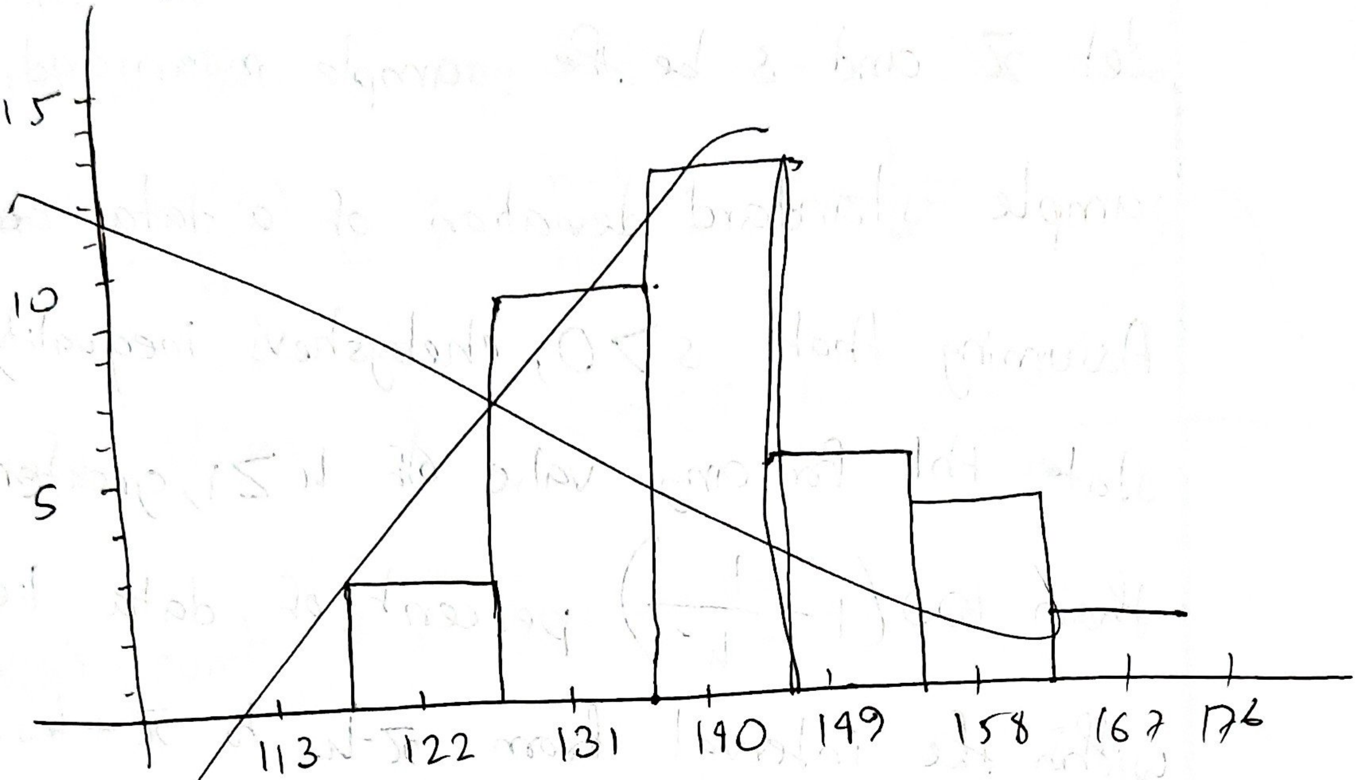
$$\text{Mode} = \cancel{L + f_m}\ 143 + \frac{8-6}{(8-6)+(8-5)} \times 5 = 145$$

$$\text{Median} = 143 + \frac{20-15}{8} \times 5 = 146.125$$

Lec-6

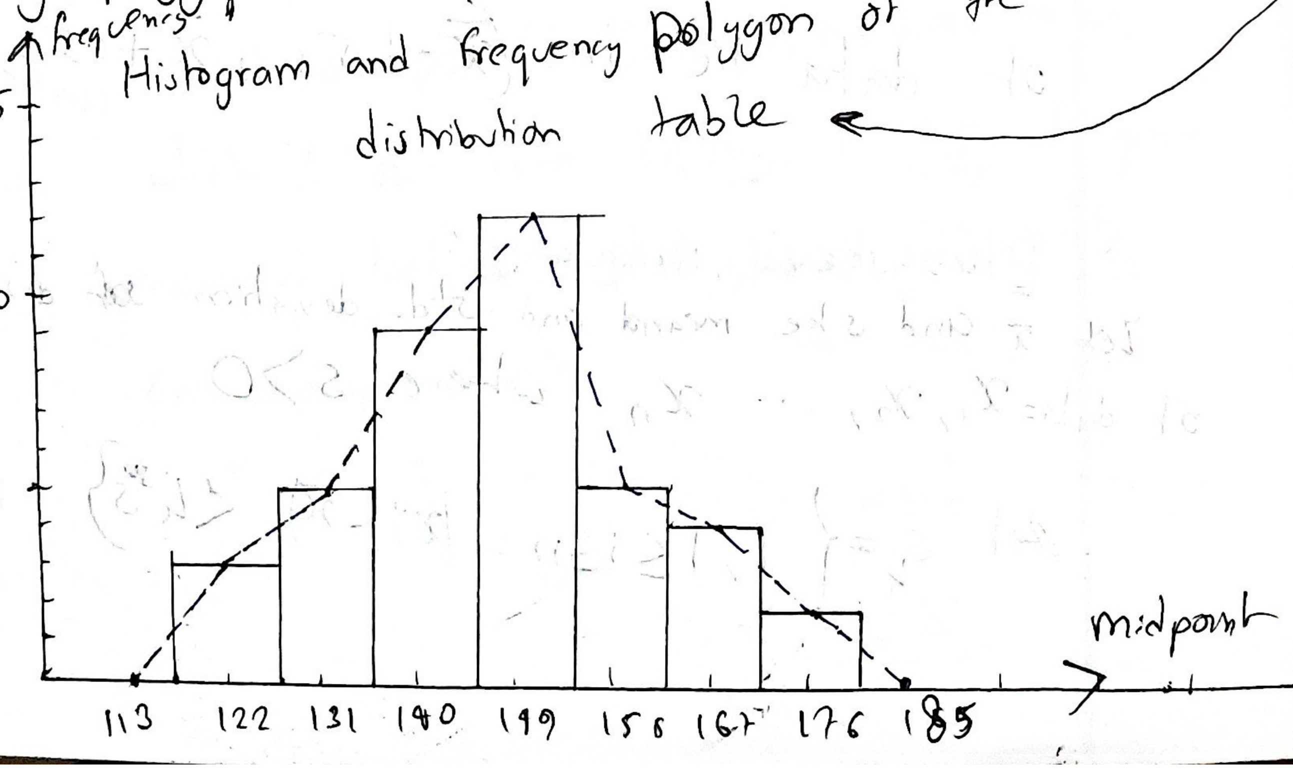
Subject: Math 207 Date: 26/8/23

Histogram and Frequency Polygon



frequency polygon ← midpoint

Histogram and frequency polygon of the distribution table



Subject : .... Date : ....

Chebyshev's ~~theorem~~ inequality : [more generalized than empirical rule]

Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data ~~and~~ set.

Assuming that $s > 0$, chebyshev's inequality states that for any value of $k \geq 1$, greater than $100\left(1 - \frac{1}{k^2}\right)$ percent of data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.

Suppose, $k = \frac{2}{3}$; $100\left(1 - \frac{4}{9}\right) = 100 \times \frac{5}{9} \approx 55.96\%$

of data lie in $(\bar{x} - 1.5s, \bar{x} + 1.5s)$

Chebyshev's inequality :

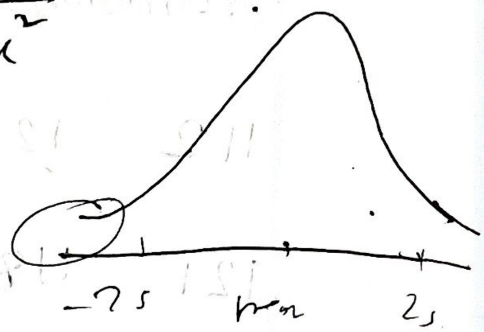
Let $\bar{x}$ and s be meand and std. deviation ~~set~~ of set consisting of data $= x_1, x_2, \ldots x_n$ where $s > 0$

Let $S_k = \{ i, 1 \leq i \leq n : |x_i - \bar{x}| < k s \}$

Subject : .... Date : ....

and let $N(S_k)$ be the number of elements in set S_k. Then for any $k \geq 1$

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$



$|x_i - \bar{x}| < ks$ $\quad \bar{x} - ks < x_i < \bar{x} + ks$

$N(k)$ = number of i : $|x_i - \bar{x}| \geq ks$,

$$\frac{N(k)}{n} \leq \frac{1}{k^2}$$

Next class: 14 data point set, mean, median, mode, what % of data belongs to 1.5s. Compare difference between chebyshev's inequality and emperical rule

Subject: Statistics Date: 3/10/23

Example: The following data represent the lifetimes (~~in hours~~) (in hours) of a sample of 40 transistors:

112	121	126	108	141	104	136	134
121	118	143	116	108	122	127	140
113	117	126	130	134	120	131	133
118	125	151	147	137	140	132	119
110	124	132	152	135	130	136	128

smallest data = 104

largest data = 152

(a) Determine the sample mean, median & mode.

(b) ~~Determine~~ Draw cummulative relative frequency plot.

(c) Are the data approximately normal?

(d) What percentage of data fall within $\bar{x} \pm 1.5$ sec?

No need to round up for human con...

Subject : .... Date : ....

(e) Compare the result in (d) to that given by emperical rule?

(f) " " " " " to the bound by Chebyshev's inequality?

a

Mean $= 127.425$ Median $= 127.5$

Mode $= 108$ Standard deviation $= 11.873$

Range $= 152 - 104 = 48$

c

	f	relative frequency	Cummulative frequency	Cummolative relative frequency
104 – 116	7	7/14	7	0.175
117 – 128	14	14/14	21	0.525
129 – 140	14	14/14	35	0.875
141 – 152	5	5/14	40	1

$\Sigma = 40$

$\Sigma = 1$

d

Chebyshev's inequality yeilds that : $100\left(1 - \frac{1}{\left(\frac{3}{2}\right)^2}\right)$ ~~400~~

$= 100\left(1 - \frac{4}{9}\right) = 55.55\%$

$3/2 = 1.5$

Subject : .... Date : ....

Stem and Leaf plot

stem → 6
leaf → 5, 9, 2

data: 65, 69, 62

data: 605, 609, 602

6 | 2, 5, 9

60 | 2, 5, 9

[most data uses single digit as stem]

Stem and leaf plot of ~~da~~ prev. data:

10	4, 8, 8	(3)
11	0, 2, 3, 6, 7, 8, 8	(7)
12	0, 1, 1, 2, 4, 5, 6, 6, 7, 8	(10)
13	0, 0, 1, 2, 2, 3, 4, 4, 5, 6, 6, 7	(12)
14	0, 0, 1, 3, 7	(5)
15	1, 2	(2)

Fairly normal distribution

~~The~~ Mode: ~~Highest frequency~~ = Multimodal

Median: avg of 20th and 21st data

Subject .... Date : ....

cummulative frequency plot → ogive

1st quartile, 2nd Quartile, 3rd Quartile

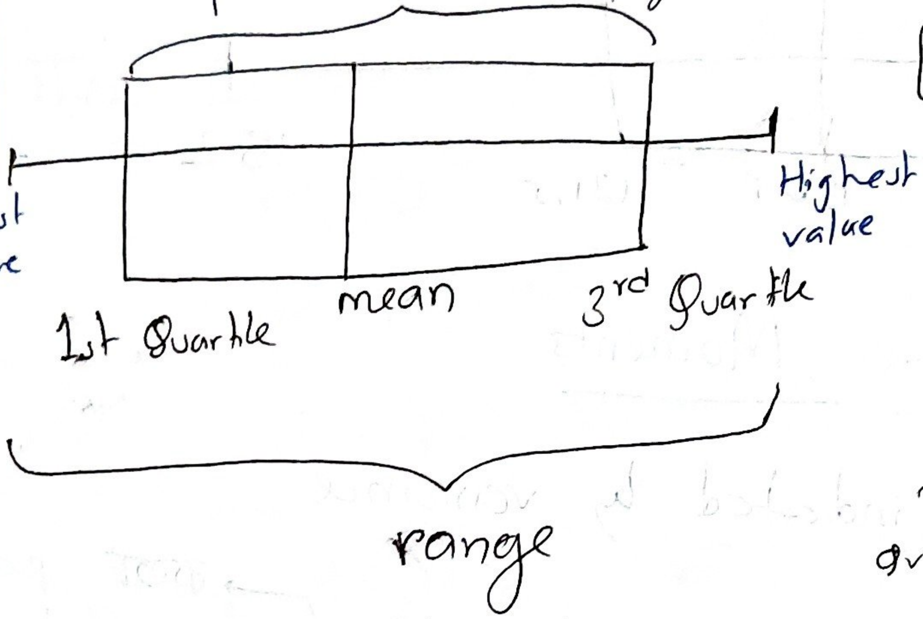
$\frac{10^{th} + 11^{th}}{2}$ avg ($\frac{20^{th} + 21^{st}}{2}$ avg $\frac{30^{th} + 31^{st}}{2}$ avg

4th Quartile

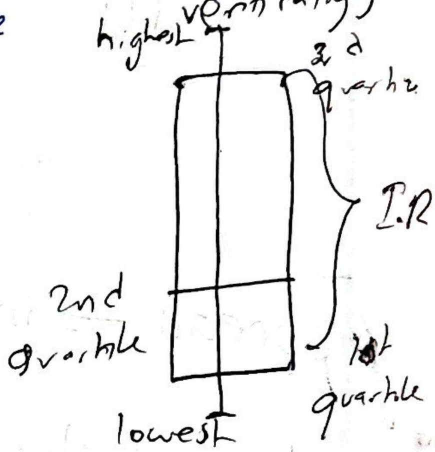
$\frac{40^{th} + 41^{st}}{2}$ avg

[If data point is odd/cannot be divided ~~into equal~~ by 4 then we take avg of two nearest value]

② Box plot interquartile range = 3rd quartile - 1st quartile



[can also be drawn vertically]



[sometimes some data points are very far from typical range of data and can drastically effect mean but not variance. These data are $\boxed{\text{Outliers.}}$ They are included in range

Subject :..Date :....................................

after dividing in percentile then data shows percentage of data in 100.

1st quartile = 111

2nd quartile = 120.5

3rd " = 131.5