9th January 2025 (Thursday) →

Categorical → Ordinal → Histogram.
Categorical → Nominal → Pie charts
Nominal → Bar graph
Numerical → Discrete → (counted)
Numerical → Continuous → (measured)
→ Today we look at representation of numerical data.

Discrete data →

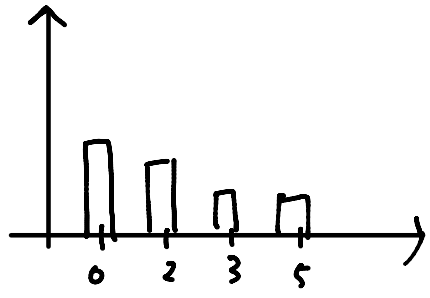
No. of phone calls recieved in a week →

2, 0, 0, 3, 2, 5, 0

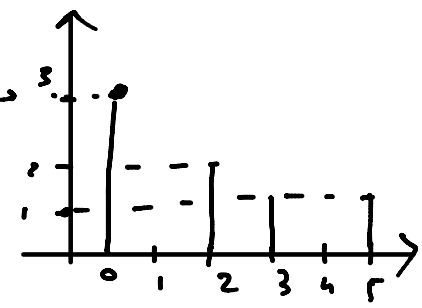
∴

# Phone calls	# days this no. of calls were recieved
0	3
2	2
3	1
5	1

To represent this,



or, a line diagram →



But these are not good representation of data — especially if there is repetition.

Stem-leaf Plot →

Marks → 78, 74, 82, 66, 94, 71, 64, 88, 55, 80, 91, 74, 82, 75, 96, 78, 84, 79, 71, 88

(Why did I copy all of this? No clue)

Stem → last digit

leaf → All other digits.

Stem	Leaves
7	8, 4, 1, 4, 5, 8, 9, 1
8	2, 8, 0, 2, 4, 3
6	6, 4
9	4, 1, 6
5	5

Now, we order by stem

Stem	Leaves
5	5
6	6, 4
7	8, 4, 1, 4, 5, 8, 9, 1
8	2, 8, 0, 2, 4, 3
9	4, 1, 6

⊛ Note: Requires # of digits does not vary a lot over the data.

Exercise → Do stem-leaves plot.

23, 26, 11, 18, 9, 21, 23, 30, 22, 11, 21, 20, 11, 13, 23, 11, 29, 25, 26, 26

S	L
2	3, 6, 1, 3, 2, 1, 0, 3, 9, 5, 6, 6
1	1, 8, 1, 1, 3, 1
0	9
3	0

ordering,

S	L
0	9
1	1, 8, 1, 1, 3, 1
2	3, 6, 1, 3, 2, 1, 0, 3, 9, 5, 6, 6
3	0

⊛ Tutorial: Tuesday, 7-8 pm (Probably from (8AM class not going to be there)

L-S plot is called exploratory analysis.

⊛ Representation of continuous data via histogram →

Suppose $x_1, x_2, \ldots, x_n$ are n observations of heights of n individuals.

Let the data value lie in the interval $[a, b]$, where $a < b$ are two real numbers.

Divide the interval $[a, b]$ into K classes of equal length, say $K > 0$

∴ The classes are,

$$[a, a_1], (a_1, a_2], (a_2, a_3], \ldots, (a_{K-1}, b]$$

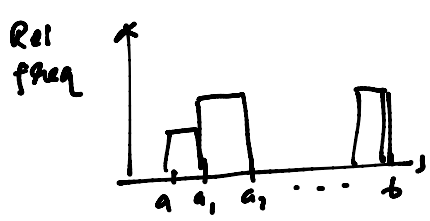
Then, we assign a frequency value to each class i, $i = 1, \ldots, K$, as,

$$f_i = \#\, x_i's \in (a_i, a_{i+1}], \text{ with } a_0 = a, a_K = b$$

Class	Frequency	Relative freq	freq. density
$[a, a_1]$	f_1	f_1/n	f_1/nh → step size
$(a_1, a_2]$	f_2	f_2/n	f_2/nh
$(a_2, a_3]$	f_3	f_3/n	f_3/nh
⋮	⋮	⋮	⋮
$(a_{K-1}, b]$	f_K	f_K/n	f_K/nh

$$n = \sum_i f_i$$

Wemay plot histogram →



So what in use of freq density?

It is an estimation of underlying frequency density function.

Remarks →

① Suppose X be a r.v. with CDF $F(x)$ and the observations are,

$x_1, x_2, \ldots, x_n$ are iid copies of X

Define another r.v,

$$Y_{i,j} = \begin{cases} 1, & x_j \in (a_i, a_{i+1}] \\ 0, & \text{otherwise.} \end{cases}$$

$\Rightarrow Y_{i,j} \sim \text{Bernoulli}(1, p_i)$, $p_i = P(a_i < x_i \leq a_{i+1})$

$Y_i = \sum_{j=1}^{n} Y_{i,j}$ → No. of success in prob p_i in n trials.

Obviously,

$$Y_i \sim \text{Binomial}(n, p_i)$$

$\Rightarrow E(Y_i) = n p_i$

Also, note, $Y_i = f_i$

$\Rightarrow E(f_i) = n p_i \Rightarrow E\left(\frac{f_i}{n}\right) = \frac{p_i}{h} = \frac{P(a_i < x \leq a_{i+1})}{h}$

We see that,

$$f_x(x) = \frac{d}{dx} F_x(x)$$

$$= \lim_{h \to 0} \frac{F_x(x+h) - F_x(x)}{h}$$

$$= \lim_{h \to 0} \frac{P(x \leq x+h) - P(x \leq x)}{h} = \lim_{h \to 0} \frac{P(x < X < x+h)}{h}$$

10th January 2025 (Friday) →

Not Audified

Correction to 2-S plot: The leaf is always the last digit, the rest is the stem.

✱ Freq distribution of the no. of peas in a pod →

Obviously, these are discreet values.

Data: A lot of numbers. I cannot fathom why writing this is useful.

I will still attempt, I guess.

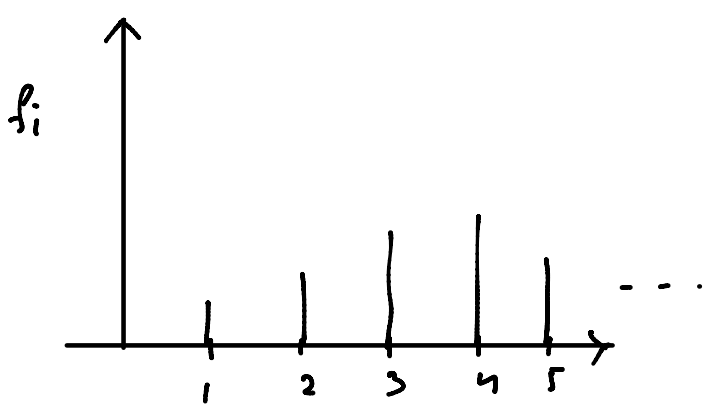
Data → 7, 1, 2, 4, 5, 2, 4, 5, 2, 4, 4, 4, 5, 3, 5, 3, 6, 3, 2, 2, 3, 4, 7, 1, 1, 1, 5, 3, 3, 4, 3, 2, 1, 6.

# Peas	Tally	fi
1	|||||	5
2	||||||	6
3	||||	.
		.
		.

Something. — cannot finish.

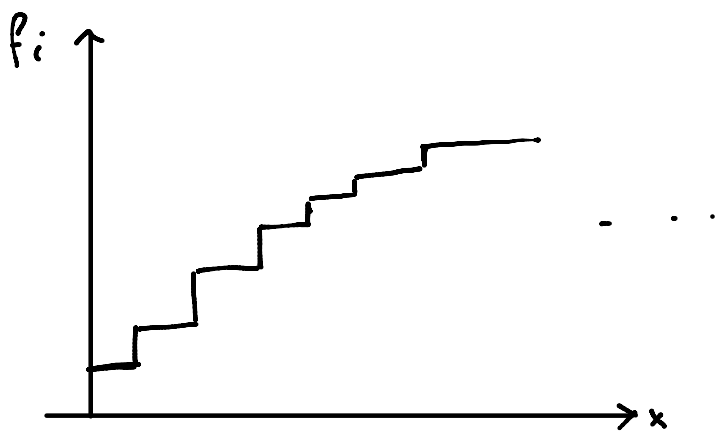
We calculate the cumulative freq > and < → The sum of all frequencies after and before some f_i

Freq Bar Diagram →



We may also draw the CF diagram.

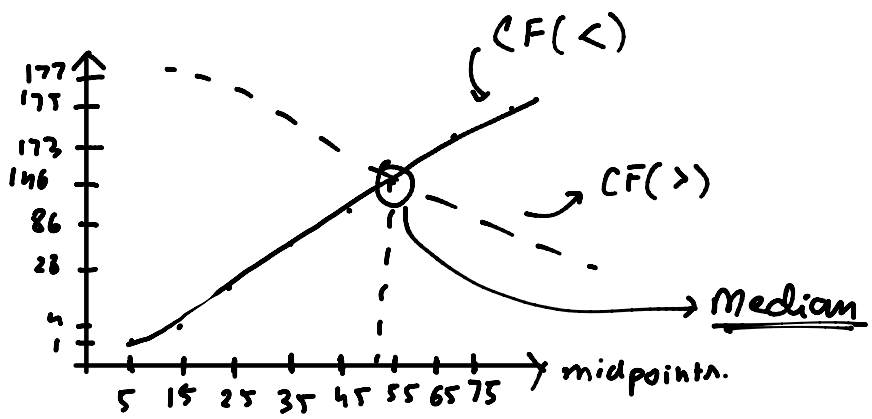
CF Bar diagram →



If this is not discrete, the continuous curve is called an ogive

— Credits to Niravrea

We draw a few ogives →



Problems (Tutorial/Exercise) → (Next to next week)

Goon Gupta, Dasgupta → Chapter-3
3.8, 3.9, 3.10, 3.11, 3.12

Measures of Central Tendency →

Tendency of data points to cluster around some mean point.

Arithmetic Mean →

Given a dataset $x_1, x_2, \ldots, x_n$,

$$A.M, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

If the data is given in freq table,

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

How to do this for continuous data in freq table,
We simply associate class mean with the freq of the class, and calculate like discrete data in freq table.
This is an approximate mean.

Some properties →

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

$\Rightarrow \sum_{i=1}^{n} (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$

⇒ Sum of the deviations of given values from its mean is necessarily 0.

(2) Mean of a variable who all have the same value must also be the same value.

Proof: $x_i = a$

$$\therefore \bar{x} = \frac{\sum_{i=1}^{n} a}{n} = \frac{na}{n} = a$$

(3) If y is a linear function of x, then $\bar{y}$ and $\bar{x}$ are related in the same way as y and x

Proof → $y = f(x)$, f is linear.

i.e, $f(x_1 + x_2) = f(x_1) + f(x_2)$, $f(ax) = af(x)$

$$\therefore \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} f(x_i)}{n} = f\left(\sum_{i=1}^{n} \frac{x_i}{n}\right) = f(\bar{x})$$

(4) If we have attached freq to each value, (x_1, n_1), $(x_2, n_2) \cdots (x_k, n_k)$

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

(5) $Z = aX + bY$

$\Rightarrow \bar{Z} = a\bar{X} + b\bar{Y}$

(6) $\bar{x}$ solves the optimization problem given data $x_1, x_2, \cdots, x_n$

$\bar{x}$ minimizes $\sum_{i=1}^{n} (x_i - c)^2$ w.r.t $c \in \mathbb{R}$.

$$\bar{x} = \text{argmin} \left\{ \sum_{i=1}^{n} (x_i - c)^2 \right\}$$

Proof → Now,

$$\sum_{i=1}^{n} \{(x_i - \bar{x}) + (\bar{x} - c)\}^2 = \sum_{i=1}^{n} [(x_i - \bar{x})^2 + (\bar{x} - c)^2 + 2(x_i - \bar{x})(\bar{x} - c)]$$

$(x_i - \bar{x})^2 \geq 0$; $(\bar{x} - c)^2 = 0$ for min. ; $2(x_i - \bar{x})(\bar{x} - c) = 0$

$\therefore \bar{x} = c$

Merits of A.M →

① Rigid Definition (?)

② Easy to understand and compute.

③ Based on all observations.

④ Amenable to algebraic manipulation.

⑤ Least affected by sampling fluctuation.

Demerits of A.M →

① Location by inspection not possible

② Affected by outliers

③ Categorical Data not possible

④ Opend interval

⑤ Skewed dist.