1-11-2020

Unit-4 - Modelling Data Distribution

PERCENTILES

% of data that is at or below the amount in question.

Z score :

how many σ away from μ.

$$Z\,score(x) = \frac{x - \mu}{\sigma}$$

- if Z-score > 0, data point is above avg
- if Z-score < 0, data point is below avg
- if Z-score close to 0, data point close to avg.

Effects of Linear transformation :

- Adding a constant ($\pm \alpha$)

  mean = mean + α

  std_dev = same

  median = median + α

  IQR = same

- Multiply a constant ($\times \alpha$)

  mean = mean $\times \alpha$

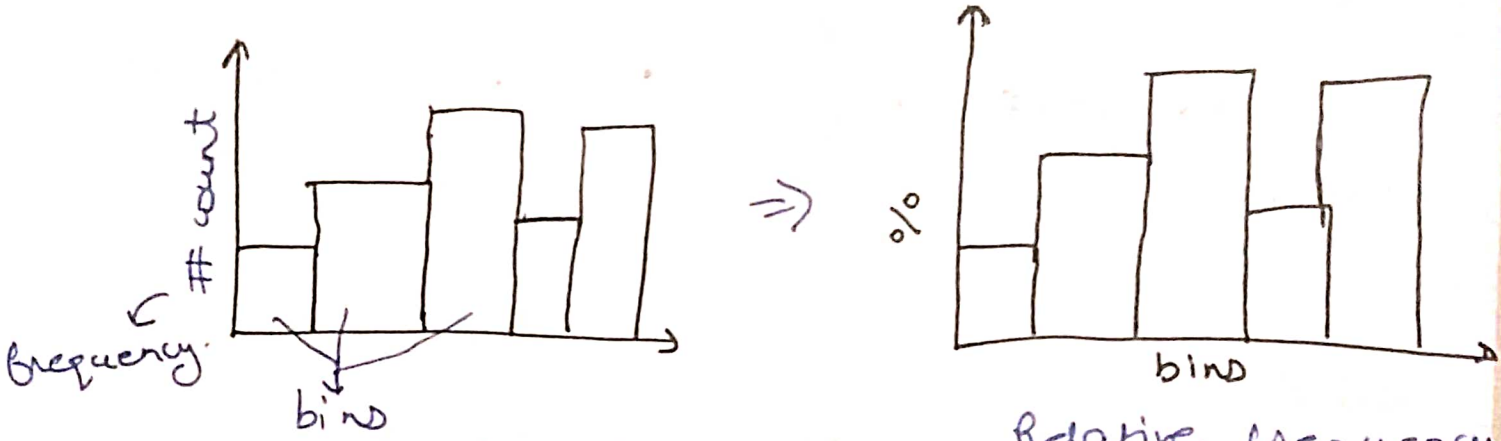
  std_dev = std_dev $\times \alpha$

  median = median $\times \alpha$

  IQR = IQR $\times \alpha$

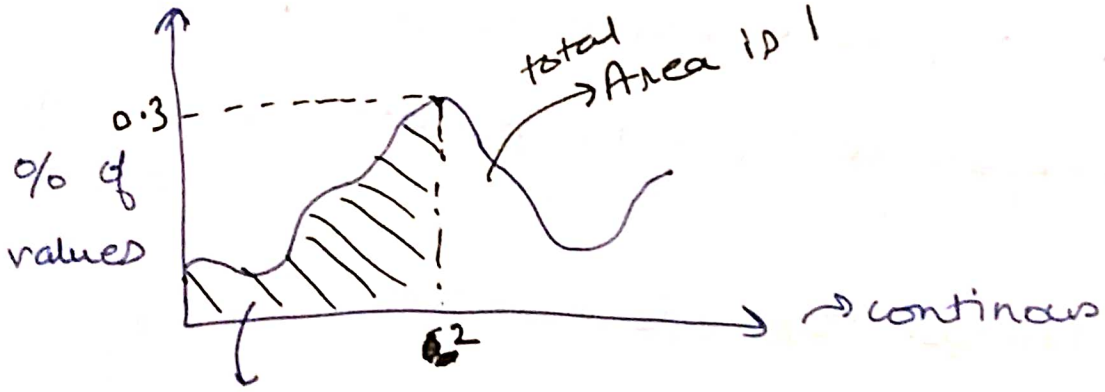
⇒ Density Curves ⇒

So far we are using histograms to visualize distribution



Relative frequency

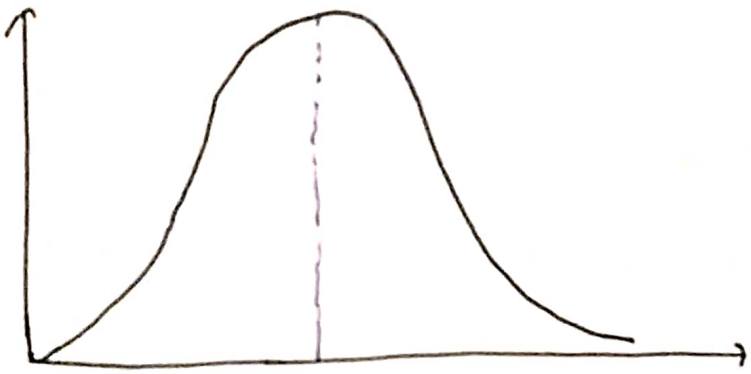
- In density graph we make infinite number of bins with approximately 0 width.



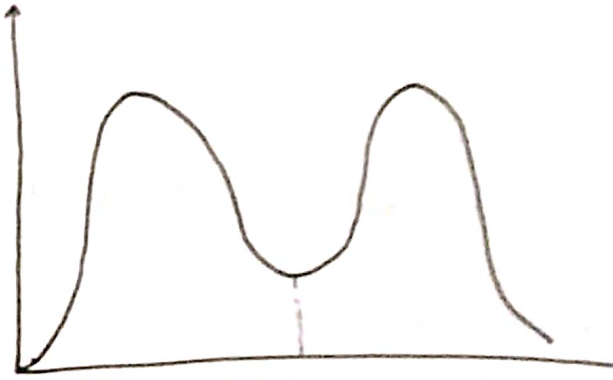
Area = % of values having value < 2.

mean, median, skew from Density curves

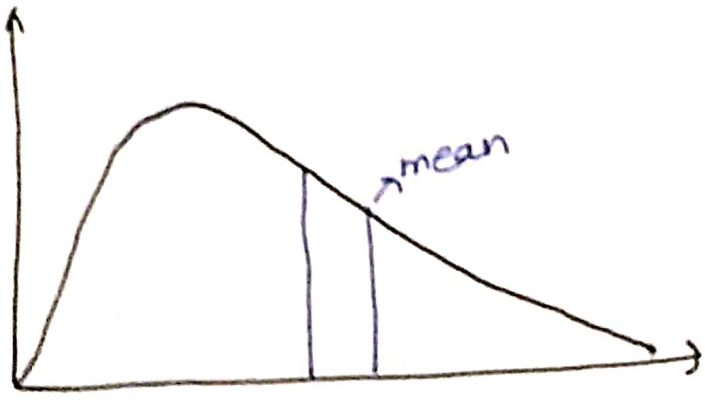
- median → Point which divides the density curve into two equal areas
- mean → take each point, multiply by their frequency and add them up.
  - Balancing point.
- For symmetrical distribution, mean = median.



mean = median



mean = median



median

(Right skewed)

(left skewed)

Normal Distribution and the Empirical Rule

68 - 95 - 99.7 rule

68% within 1 std. dev away from mean

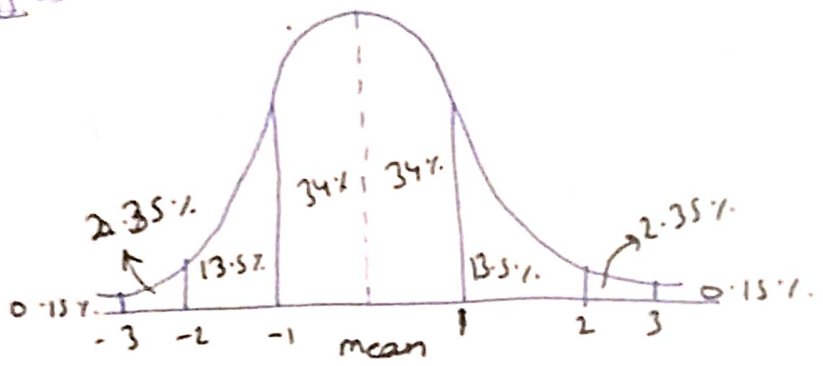
95% within 2 std dev away from mean

99.7% within 3 std dev away from mean

Standard normal distribution →

Normal dist. where mean = 0 , Std dev = 1

⇒ Summary ⇒



⇒ Normal distribution calculation

- mean & σ is given. Found the area of the shaded region

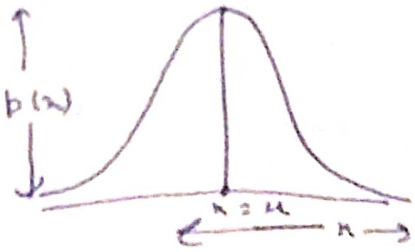
Steps →

Step-1 ÷ Find z score

Step 2 ÷ look into the table to find corresponding area.

⇒ More on Normal distribution ⇒

$$\text{pdf} \Rightarrow p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



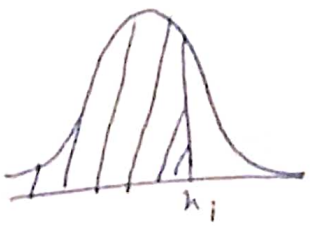
- **Note:**
  - at a given x, $p(x)$ gives the height of the function, and not the probability.
  - For a given number, let's say 5, $p(5) \neq 0$. (prob. is area under the graph), as in this case the area $= 0$.
  - To find probability we have to give some range and integrate.

$$\int_{4.9}^{5.1} p(x)\,dx$$

⇒ **cdf** – cummalative density function.

$$cdf(x) = \int_{-\infty}^{x} p(x)\,dx$$

- For a given x_i, $cdf(x)$ gives the probability the $x < x_i$,
  i.e. the area under graph ↓



- To find the prob of x lying in between $x_1 < x_2$. $P(x_1 < x < x_2)$ it will be,

$$cdf(x_2) - cdf(x_1).$$