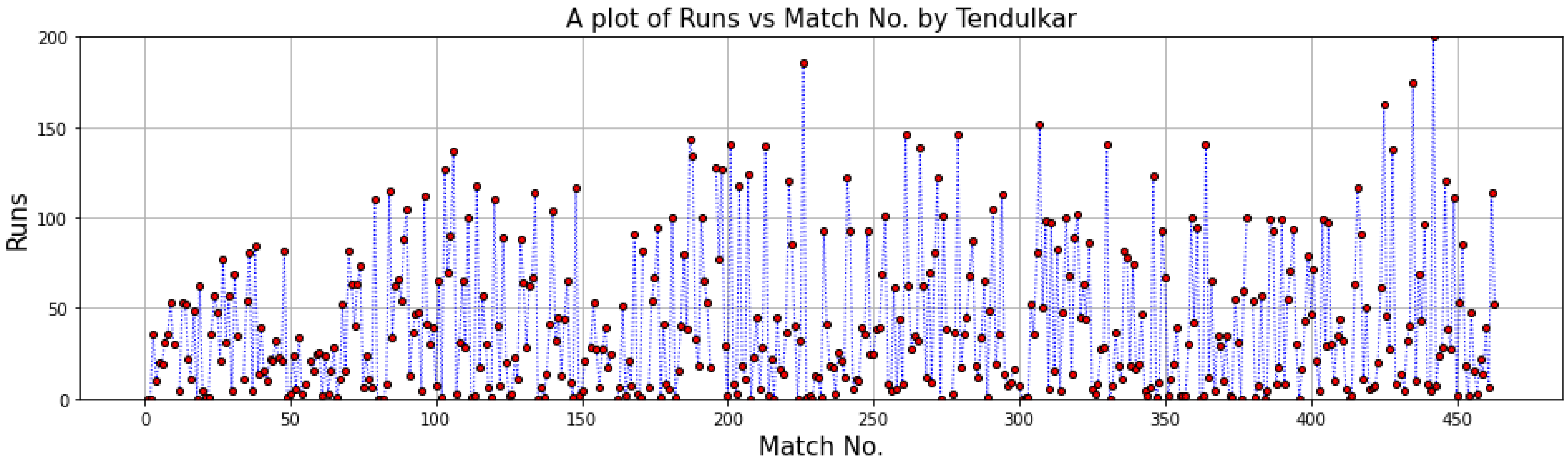
Data Characterization

Match No.	Runs	Balls
463	52	48
462	114	147
461	6	19
460	39	30
459	14	15
458	22	23
457	3	12
456	15	24
455	48	63
454	2	6
453	18	14
285	18	16
284	87	67
283	68	79
282	45	60
281	36	43
280	17	42
279	146	132
278	37	35
10	30	29
9	53	41
8	36	22
7	31	26
6	19	35
5	20	25
4	10	12
3	36	39
2	0	2
1	0	2

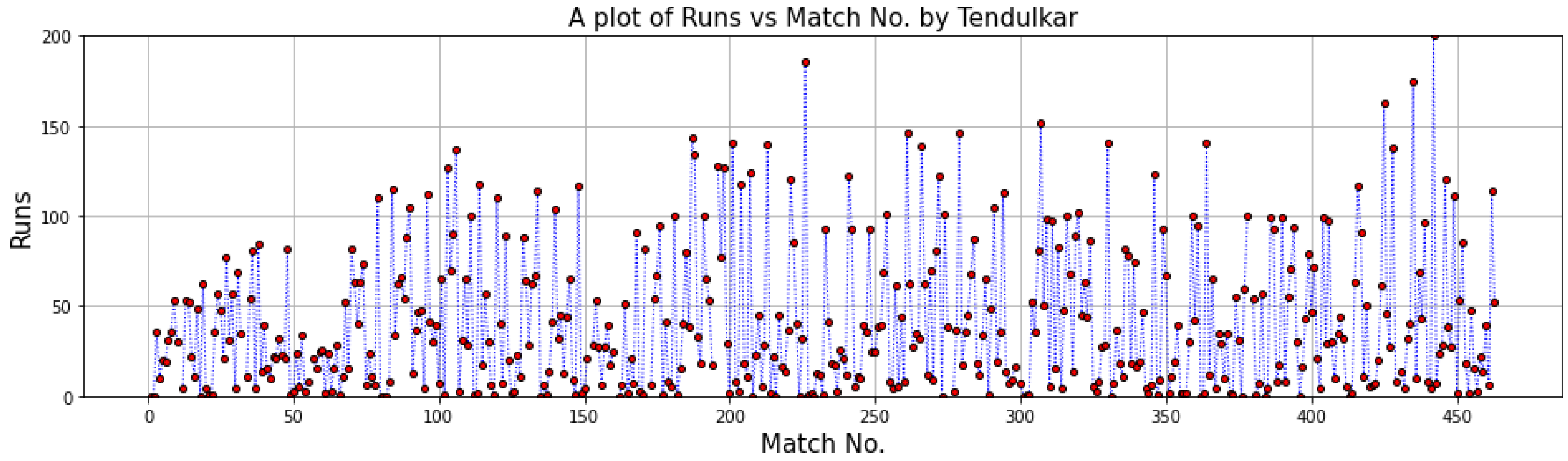
THREE Important Characteristics of Data

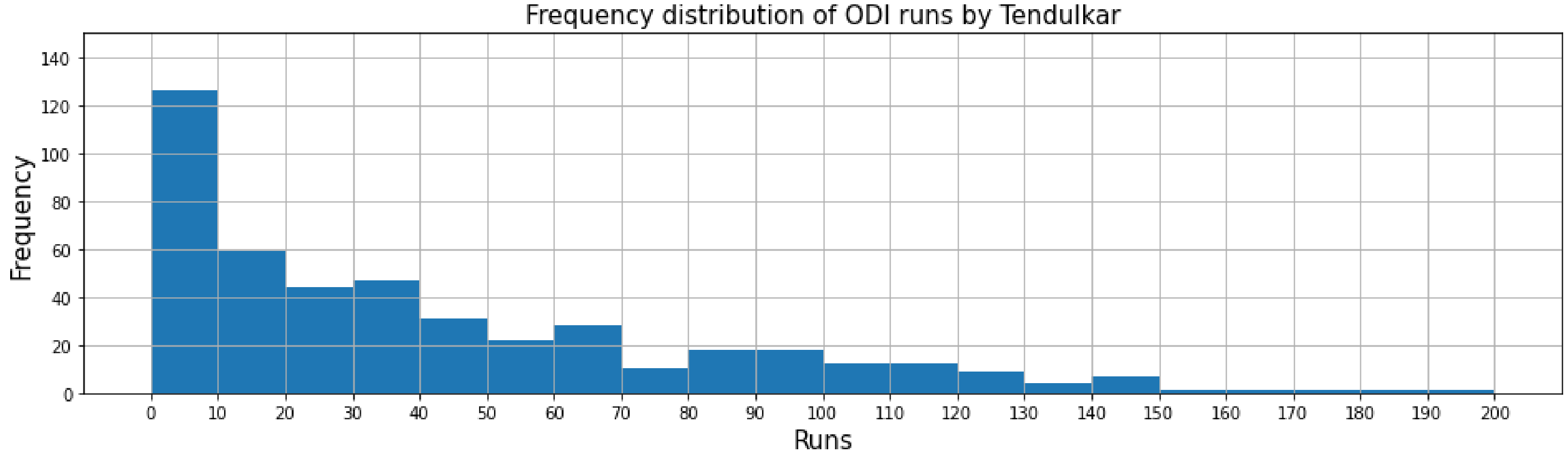
- Central Tendency
- Variability or Dispersion
- Shape of Frequency Distribution



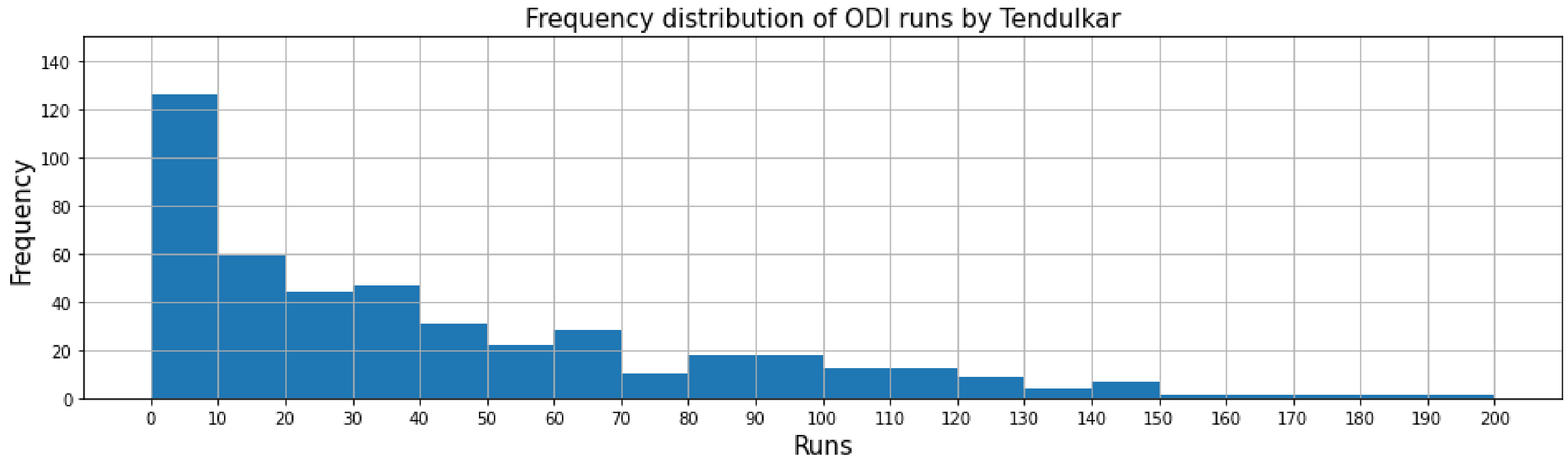
Shape of Frequency Distribution

CEP2022_Notebook (1.4)





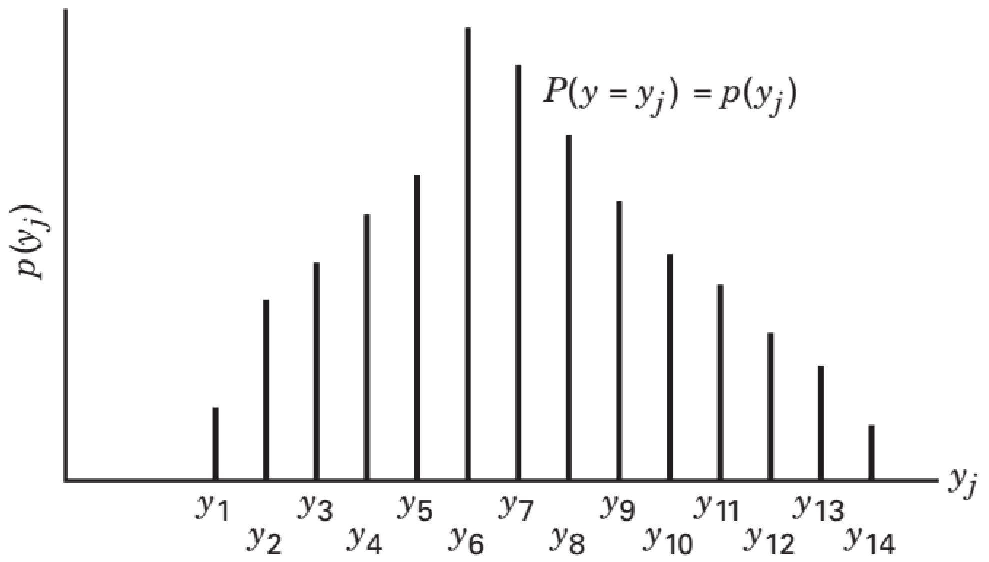
Shape of Frequency Distribution



Questions:

- What is the area under the curve?
- Given such data, how would you calculate the probability of Tendulkar scoring a given number of runs?
- How would you then convert the Y-axis to probability?
- What happens when the bin size $\rightarrow 0$

Probability Distribution



(a) A discrete distribution

y discrete: $\quad 0 \le p(y_j) \le 1 \quad$ all values of y_j

$$P(y = y_j) = p(y_j) \quad \text{all values of } y_j$$

$$\sum_{\substack{\text{all values}\\ \text{of } y_j}} p(y_j) = 1$$

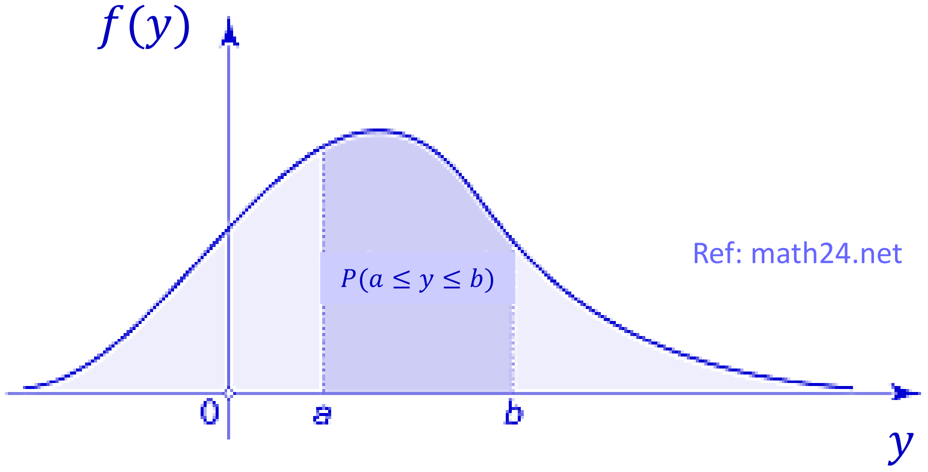
Probability Density/Distribution Function

- For a continuous random variable ‘y’, the probability behavior is described by a function called ‘probability density function’ (PDF) = $\boldsymbol{f}(y)$
- What are the properties of such PDF?

$$\boldsymbol{f(y) \geq 0}$$

$$\boldsymbol{\int_{-\infty}^{\infty} f(y)dy = 1}$$

$$\boldsymbol{Probability\,(a \leq y \leq b) = \int_a^b f(y)dy}$$



- Cumulative distribution function (CDF) for a continuous random variable x with pdf $f(X)$

$$F(y) = Probability(Y \leq y) = \int_{-\infty}^{y} f(Y)dY$$

Note: $f(y) = \dfrac{dF(y)}{dy}$

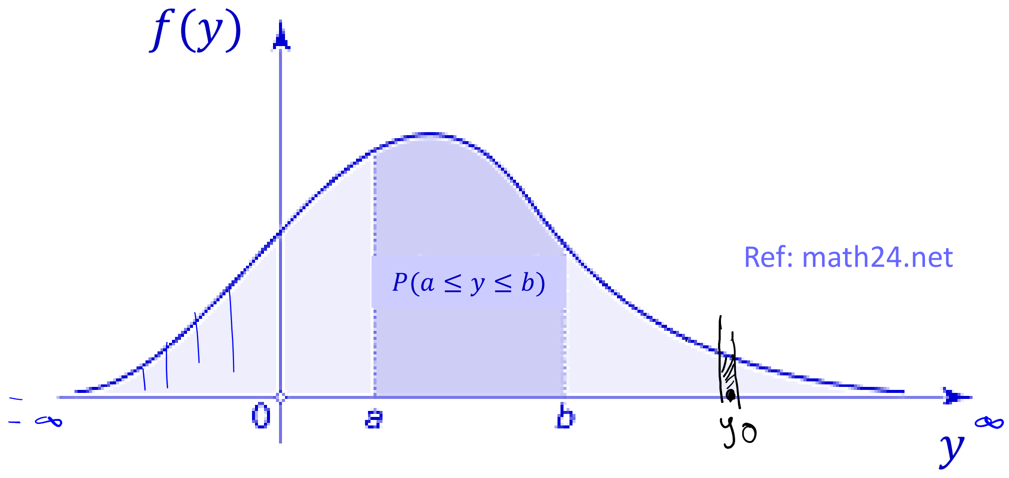
Probability Density Function

- Given $f(y)$, how would you find the *true arithmetic mean (μ)* value of 'y'?

$$\mu = \int_{-\infty}^{\infty} y\, f(y)\, dy$$

- What about *true variance (σ^2)*?

$$\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\, dy$$



- The expectation of a function g(y) of a random variable 'y' with pdf 'f(y)' is defined as,

$$\boldsymbol{E}(g(y)) = \int_{-\infty}^{\infty} g(y) f(y) dy$$

$E(y) = \mu$

$E((y-\mu)^2) = \sigma^2$

Rules for Expectation

DIY

Mean (Population)

$$\mu = E(y) = \begin{cases} \int_{-\infty}^{\infty} yf(y)\,dy & y \text{ continuous} \\ \sum_{\text{all } y} yp(y) & y \text{ discrete} \end{cases}$$

Variance (Population)

$$V(y) = E[(y-\mu)^2] = \sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy & y \text{ continuous} \\ \sum_{\text{all } y} (y-\mu)^2 p(y) & y \text{ discrete} \end{cases}$$

Identities

1. $E(c) = c$
2. $E(y) = \mu$
3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$
7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$
8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2\,\text{Cov}(y_1, y_2)$

$$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)]$$

11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$

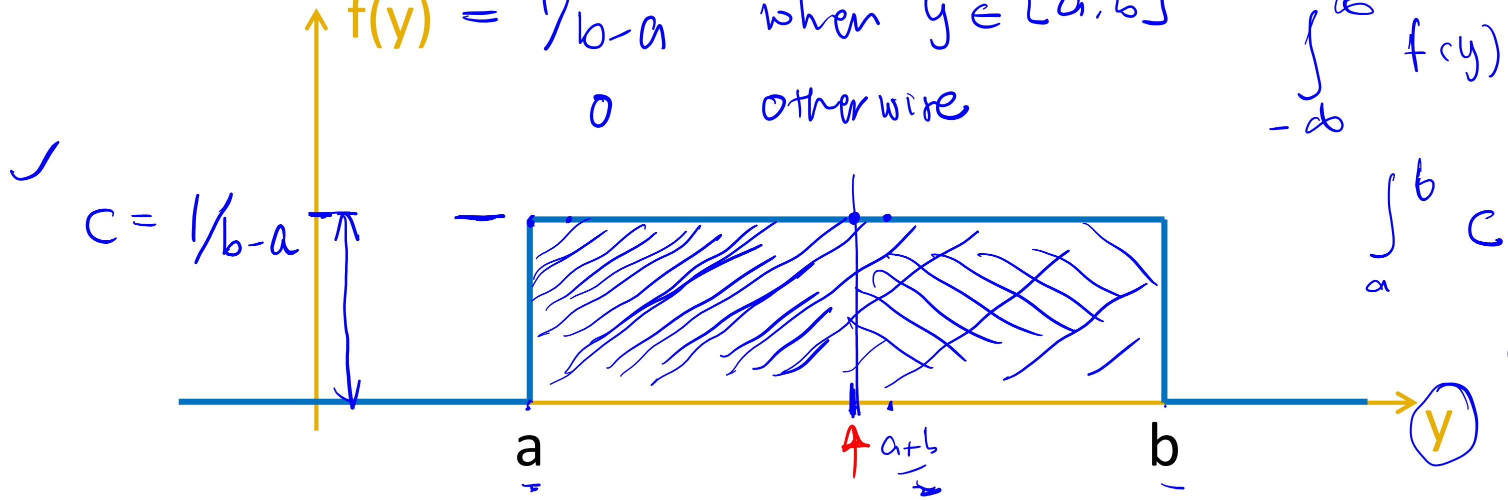
However, note that, in general

12. $E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

regardless of whether or not y_1 and y_2 are independent.

Uniform or Rectangular PDF

$f(y) = \begin{cases} 1/(b-a) & \text{when } y \in [a,b] \\ 0 & \text{otherwise} \end{cases}$



$$\int_{-\infty}^{\infty} f(y)\,dy = 1$$

$$\int_a^b C\,dy = 1$$

$$C\, y\Big|_a^b = 1$$

$$C(b-a) = 1$$

- What is mean and variance?

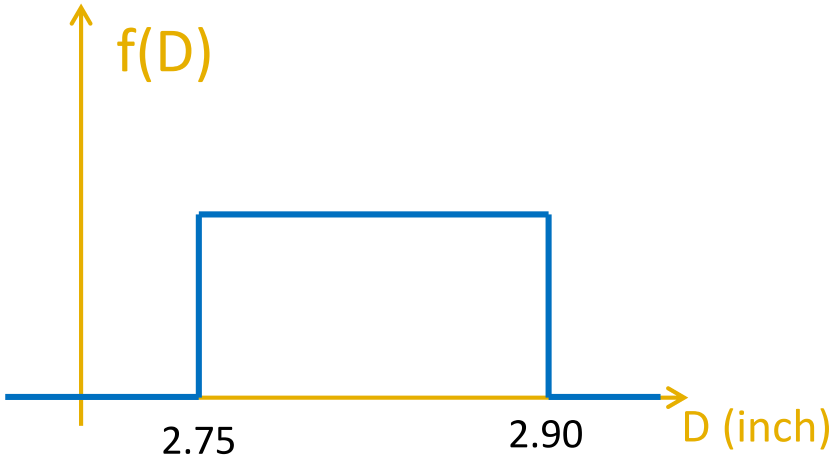
$$\mu = E(y) = \int_{-\infty}^{\infty} y\, f(y)\,dy = \int_a^b y\left(\frac{1}{b-a}\right)dy = \frac{1}{(b-a)}\,\frac{y^2}{2}\Big|_a^b = \frac{1}{2(b-a)}(b^2-a^2) = \left(\frac{a+b}{2}\right)$$

$$V(y) = E\left((y-\mu)^2\right) = \int_a^b \left(y - \frac{a+b}{2}\right)^2 \frac{1}{(b-a)}\,dy =$$

- What is median and mode?

$$\text{median} = \left(\frac{a+b}{2}\right) \qquad \text{mode} = \text{any } y \in [a,b]$$

Uniform PDF Example



Suppose a cricket ball manufacturer is making cricket balls of a **specified diameter of 2.83 inches**.

BUT due to **inaccuracies/variations** in the making process, the actual diameter of the balls made is **uniformly distributed over the range of 2.75 inches to 2.90 inches**.

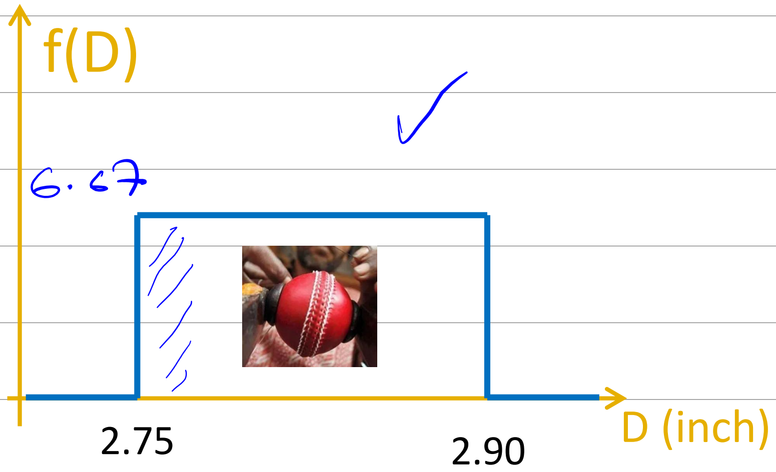
Now, the balls with diameters between **2.80-2.86 inches are still acceptable** to BCCI and can be sold for a **profit of 100 Rs/ball**.

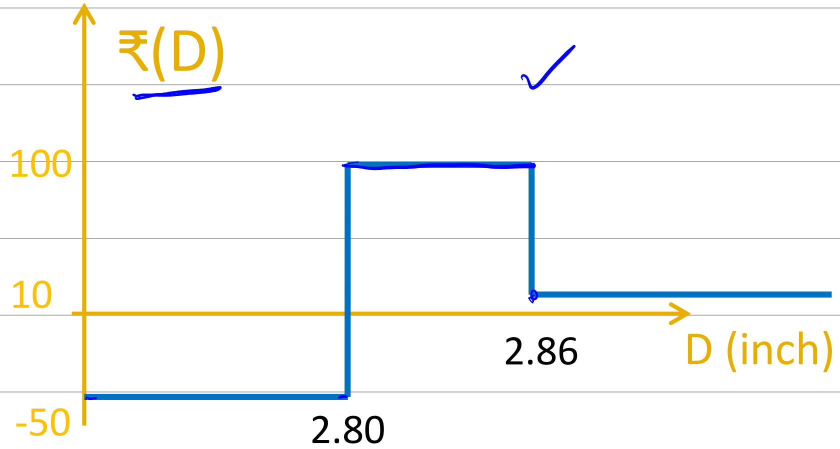
If the ball is **oversized (D > 2.86),** it can be sold, but at a **smaller profit of 10 Rs/ball**.

If the ball is **undersized (D < 2.80),** it needs to be discarded, and there is **a loss of 50 Rs/ball**.

Question: What is the expected profit (Rs/ball)?

Uniform PDF Example





$$E(\text{₹}(D)) = \int_{-\infty}^{\infty} \text{₹}(D)\, f(y)\, dy$$

$$= \int_{-\infty}^{2.75} \cancel{-50 \times 0}\, dy + \int_{2.75}^{2.80} (-50)\, 6.67\, dy + \int_{2.80}^{2.86} 100 \times 6.67\, dy$$

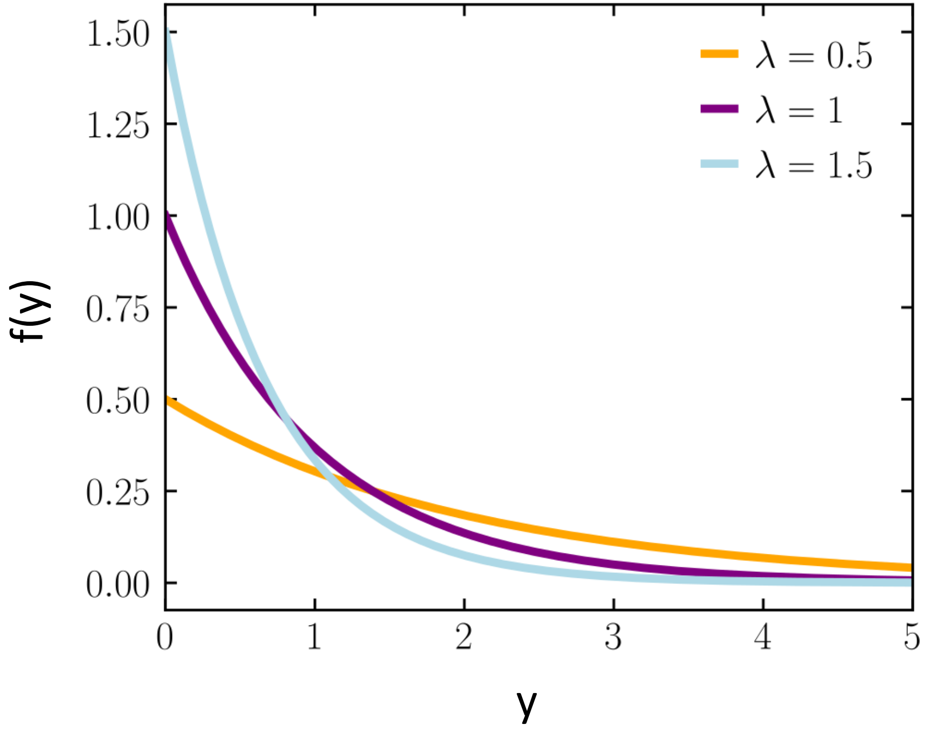
$$+ \int_{2.86}^{2.90} 10 \times 6.67\, dy + \int_{2.90}^{\infty} \cancel{10 \times 0}\, dy$$

$$= 0.05 \times 6.67 \times (-50) + 0.06 \times 6.67 \times 100 + 0.04 \times 10 \times 6.67$$

$$= 26.01 \quad \text{Rs/ball}$$

Exponential PDF

Ref: wikipedia



$$f(y) = \lambda\, e^{-\lambda y}, \qquad y \geq 0$$

$$f(y) = 0, \qquad y < 0$$

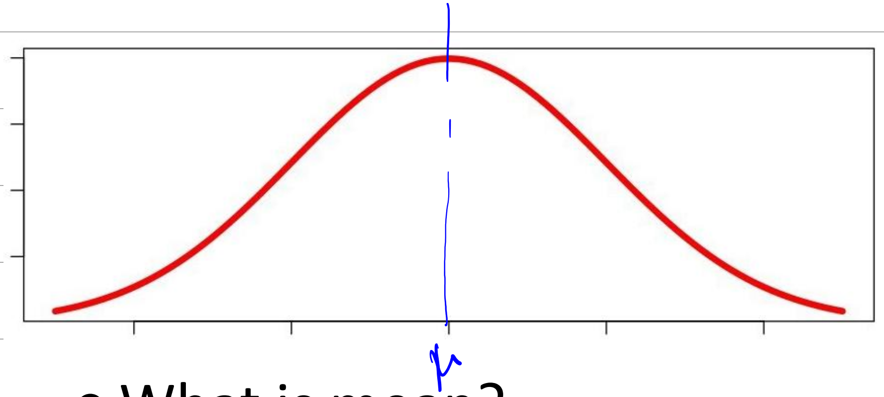
Find mean, std. deviation, median and mode **DIY**

$$\text{Mean} = \mu = \frac{1}{\lambda} \qquad \text{Std. Dev} = \sigma = \frac{1}{\lambda}$$

$$\text{Median} = \frac{\ln(2)}{\lambda} \qquad \text{Mode} = 0$$

Normal or Gaussian PDF

DIY



$$f(y) = \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-b}{a}\right)^2\right) \quad y \in [-\infty, \infty]$$

DIY

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

- What is mean?

$\mu = b$

- What is variance and std. deviation?

$\sigma^2 = a^2 \quad , \quad \sigma = a$

- What are median and mode?

mean = median = mode = μ = b