Distribution

pdf, cdf

EX (or μ) $\operatorname{Var} X$

$\sum k\,P[X=k]$ $E[X-E(X)]^2$

Samples

$X_1, X_2, X_3 \cdots$

$$\bar{X} = \frac{X_1 + X_2 \cdots + X_n}{n}$$

$$S^2 = \frac{(X_1-\bar{X})^2 + \cdots (X_n-\bar{X})^2}{n-1}$$

$$E(S^2) = \operatorname{Var} X$$

Sum: I have samples X_1 X_2 $\cdots$ $X_n \longrightarrow (\mu, \sigma^2)$

↳ "independent"

Question: How does $Y = X_1 + X_2 + X_3 + \cdots + X_n$ look like?

→ Each time this number will be diff

$$E[Y] = n\mu$$

$$Var[Y] = n\sigma^2$$

$Var(X_1 + X_2)$

$= Var\ X_1 + Var\ X_2$

n = num_samples

Dice:

Goal is to get trial_values $= [Y_1, Y_2, Y_3 \cdots Y_{1000}]$ → hist

$\bar{Y} = \frac{Y_1 + Y_2 + \cdots Y_{1000}}{1000} = (350.5) \longrightarrow 350 = 100\ (3.5)$

$$Y = X_1 + X_2 \quad , \quad X_1 \text{ & } X_2 \text{ are indep}$$

$$\text{Var } Y = \text{Var } X_1 + \text{Var } X_2$$

$$= \sigma^2 + \sigma^2$$

$$= 2\sigma^2$$

Central Limit Theorem: $\{X_i\}$ have mean μ & variance σ^2

$Y = X_1 + X_2 + \cdots + X_n$ is approximately distributed as Normal/Gaussian with mean $= n\mu$ and variance $= n\sigma^2$

$$\frac{Y - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1)$$

Standardisation: $Y \sim N(n\mu, n\sigma^2)$

$$Z = \frac{Y - n\mu}{\sqrt{n}\,\sigma} \sim N(0,1)$$

"Standard Normal"

$E[Y] = n\mu$

$Var[Y] = n\sigma^2$

$$Z = \frac{Y - E[Y]}{\sqrt{Var\,Y}} \text{ is } N(0,1)$$

$$Y = X_1 + X_2 + \cdots + X_n$$

$$Y - n\mu = (X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)$$

Person1 : [6, 3, 1, 3, 4, 5, 6 (100 times), 2, 1 . . .]

6+3+1+3+4 ⟶ (27)

Person2 : [1, 5, 1, 3, 4, 4, 5, 2, 1 . . .] ⟶

1+5 ⟶ (35)

[27, 35, ...] → what distribution

Ex: Insurance company → 25,000 policy holders

Yearly claim → mean 320, S.D = 540

Approximately, Prob. claim is greater than 8.3 million

Y : Total Yearly claim $Y = X_1 + X_2 + \cdots + X_{25,000}$

$E(Y) = 25,000 \times 320 = 8\text{-mil}$

$EY = 8\text{-mil}$

$\text{Var}\,Y = n\,540^2$

$\sqrt{\text{Var}\,Y} = \sqrt{n}\,540$

$$P[Y > 8.3\text{ m}] = P\left[\frac{Y - 8\text{ mil}}{540\sqrt{25000}} > \frac{8.3\text{ mil} - 8\text{ mil}}{540\sqrt{25000}}\right]$$

$$= P[Z > 3.51]$$

$$= 1 - P(Z \le 3.51) = 1 - \text{CDF}(3.51)$$

$$= 1 - \text{norm.cdf}(3.51)$$

Binomial: Y: n tosses . k heads $k \leq n$ $P\{H\} = p$

$P[Y = k] \rightarrow$ "Binomial Random variable"

$$Y = X_1 + X_2 + X_3 + \cdots + X_n$$

X_1 → 0 if first toss is tails
1 if first toss is heads

X_i → 0 i^{th} toss is tails
1 heads

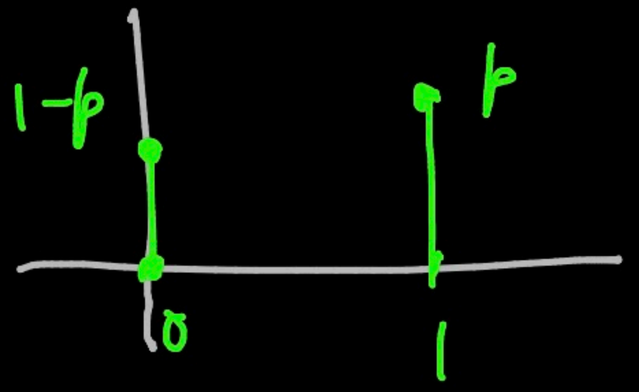
What is $\operatorname{Var} Y = n\,p\,(1-p)$.

$E\,Y = n\,p$

Bernoulli

X_i $\quad E X_i = p$

$\operatorname{Var} X_i = p(1-p)$



Gaussian Approximation to Binomial :

X_i : Bernoulli p

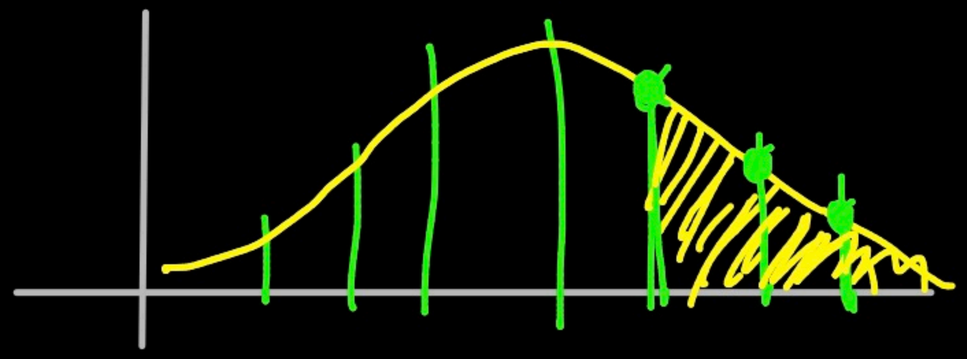
$$Y = X_1 + X_2 + \cdots + X_n$$

$$EY = np$$

$$\mathrm{Var}\, Y = np(1-p)$$

$$\frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

$np(1-p) \geq 40$



np. random. binomial (n, p)

Eg: 150 - capacity class.

Among those accepted, only 30% attend → others drop out

College accepts 450 student.

What is the prob of more than 150 people attending

→ $n = 450 \quad p = 0.3 \quad E[Y] = np = (450)(0.3) = 135$

$Var[Y] = 94.5 \rightarrow np(1-p) = (450)(0.3)(0.7)$

$$P[Y > 150] = P\left[\frac{Y - 135}{\sqrt{94.5}} > \frac{150 - 135}{\sqrt{94.5}}\right]$$

$$= P[Z > 1.5] \simeq 0.06$$

Sample mean: $X_1, X_2, X_3 \cdots$ mean μ, Variance σ^2

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$E[\bar{X}] = \mu \quad , \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$Z = \frac{\bar{X} - E[\bar{X}]}{\sqrt{Var[\bar{X}]}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

$$\downarrow$$

$$N(0,1)$$

Eg: Weights "population" → everyone

$\mu = 167$ pounds, $\sigma = 27$ → std dev.

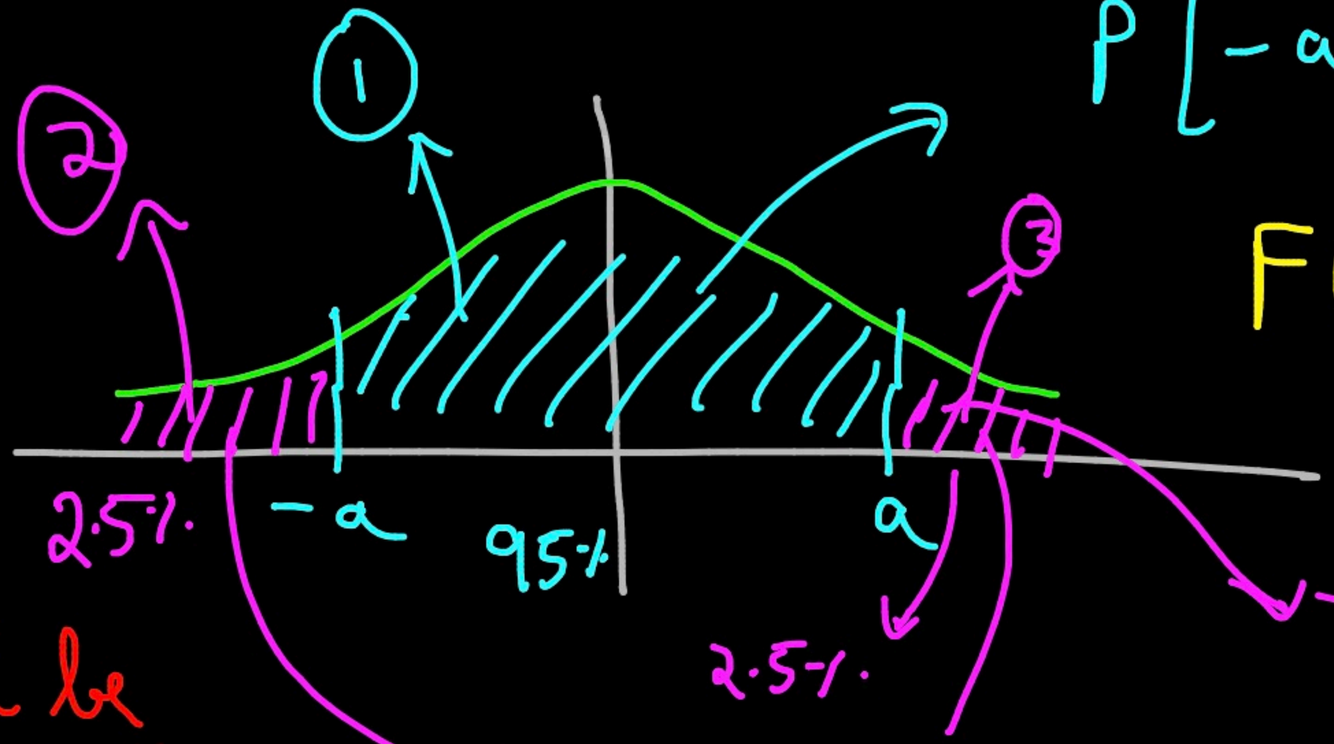
1) A sample of 36 people are taken. $\bar{X}$ → specific sample mean for those 36 people [163, 170]

$$E[\bar{X}] = 167 \qquad Var[\bar{X}] = \frac{27^2}{36} \qquad \sqrt{Var(\bar{X})} = 4.5$$

$$P[163 < \bar{X} < 170] = P\left[\frac{163-167}{4.5} < \frac{\bar{X}-167}{4.5} < \frac{170-167}{4.5}\right]$$

$$= P[-0.889 < z < 0.889]$$

→ not exact bcz discrete & cont



$P[-a < z < a]$

$F(a)$ = blue + one more
① + ②

this piece → area?

$1 - F(a)$

② = ③ = $1 - f(a)$

blue = $f(a) - (1 - F(a))$

$= 2F(a) - 1$

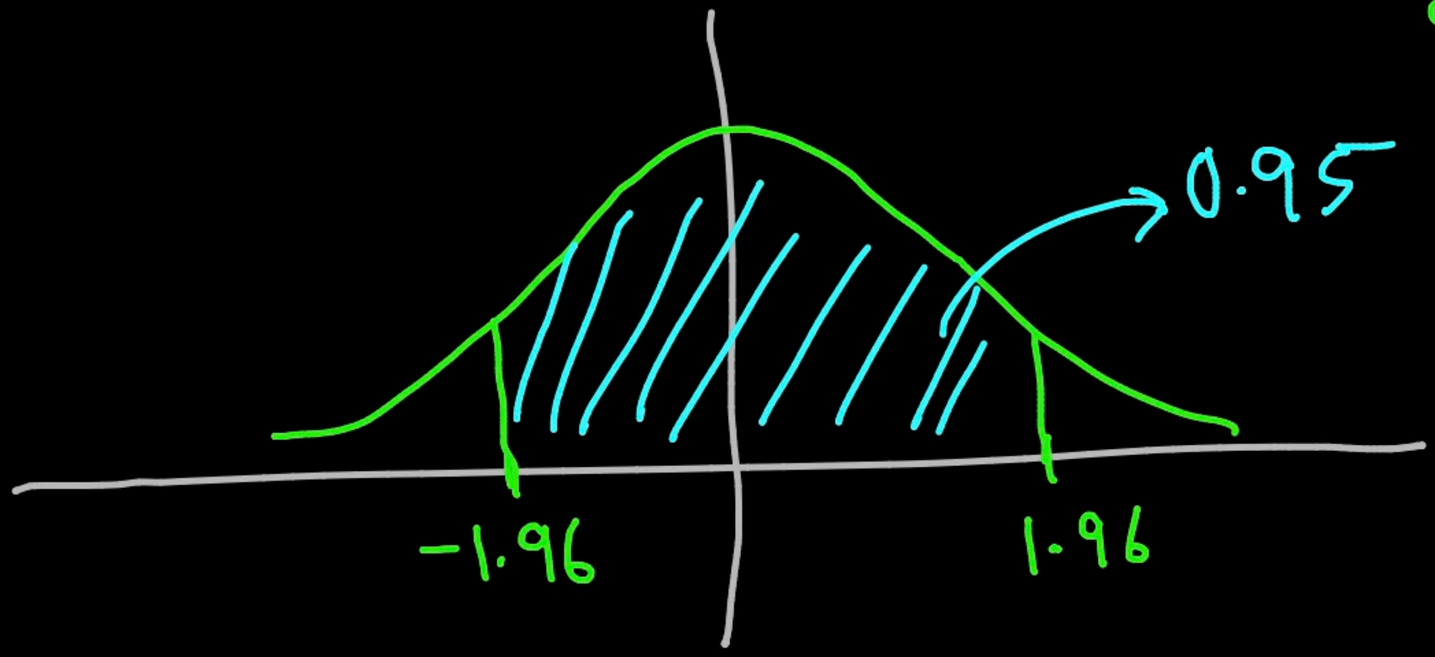
What will be the value of "a" if blue area is 95%

These two areas are equal

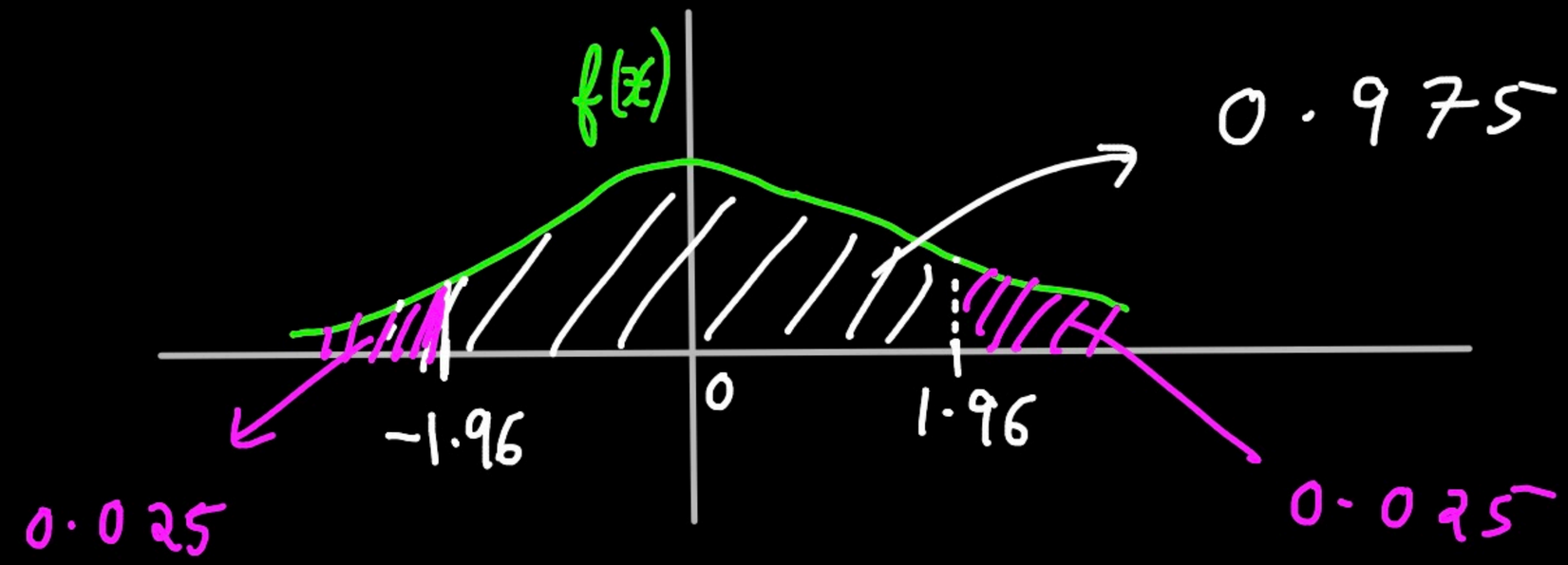
$F(a) = 0.975$

$\mu = 0$

$\sigma = 1$



$$P[-1.96 < z < 1.96] = 95$$



$$P(-1.96 < z < 1.96)$$

$$= F(1.96) - F(-1.96)$$

$$= 0.975 - 0.025$$

$$= 0.95$$

$$ppf(0.975) = 1.96$$

$$cdf(1.96) = 0.975$$

Samples $X_1\ X_2\ X_3 \ldots \to N(\mu, \sigma)$ (for now)

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$Z = \frac{\sqrt{n}\,\bar{X} - \mu}{\sigma} \to N(0,1)$$

$$P\left[-1.96 < \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} < 1.96\right] = 0.95$$

$$P\left[\frac{-1.96\,\sigma}{\sqrt{n}} < \bar{X} - \mu < \frac{1.96\,\sigma}{\sqrt{n}}\right] = 0.95 \quad -(1)$$

$$P\left[\frac{-1.96\,\sigma}{\sqrt{n}} < \mu - \bar{x} < \frac{1.96\,\sigma}{\sqrt{n}}\right] = 0.95$$

$$P\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

Summary: Population mean lies in an interval

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$ with

95% Confidence Interval

95% confidence

Sehwag

400 $\longrightarrow \mu_s$

"population mean"

50 matches randomly

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{50}}{50}$$

How far from μ_s

Dravid

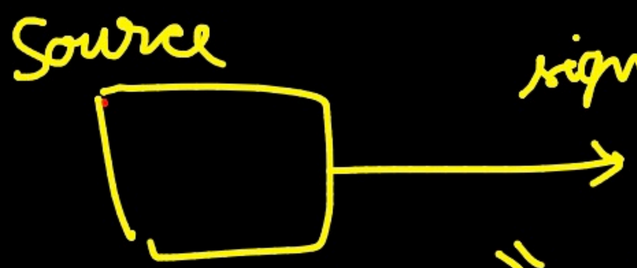
400 $\longrightarrow \mu_d$

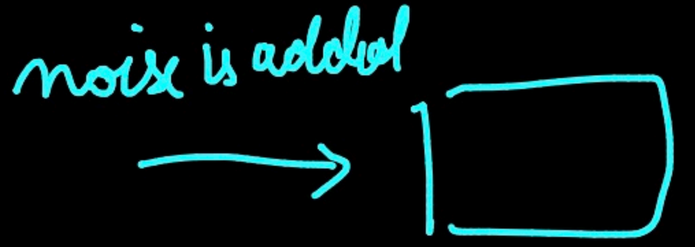
"population mean"

50 matches randomly

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{50}}{50}$$

How far from μ_d





$\mu + N \rightarrow N(0, \sigma^2 = 4)$

$\mu + N \hookrightarrow N(\mu, \sigma^2 = 4)$

$5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5 \rightarrow \frac{81}{9} = 9$

$x_1, x_2, x_3, x_4, x_5, \ldots, x_9$

$\bar{x} = 9$. Can you comment on μ?

$$\left[9 - \frac{(1.96)(2)}{\sqrt{9}}, \; 9 + \frac{(1.96)(2)}{\sqrt{9}}\right] = \underbrace{[7.69, 10.31]}_{95\% \text{ CI}}$$