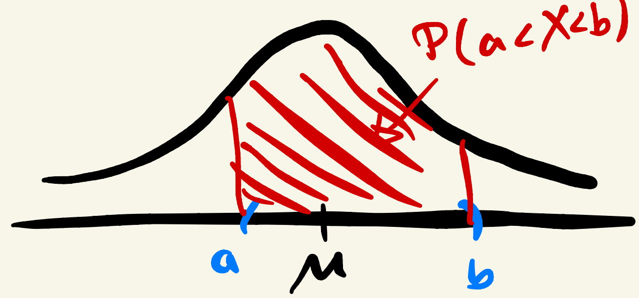
# AMAT 362 - Probability for Statistics

## Lecture 14 → The Central Limit Theorem

**Recall** The probability density function (PDF) for a normally distributed R.V. $X$ w/ $E(X)=\mu$ $SD(X)=\sigma$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} = \varphi_{\mu,\sigma}(x)$$

$$P(a < X < b) = \int_a^b f_X(x)\,dx$$

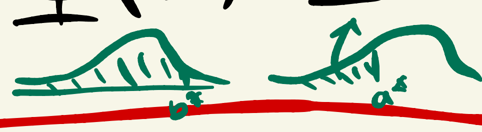



**N.B.** $\varphi_{\mu,\sigma}(x)$ HAS NO NICE ANTIDERIVATIVE!

So... we transform $X$ to $Z := \frac{X-\mu}{\sigma}$

$P(a < X < b) = P(a^* < Z < b^*)$ where $a^* = \frac{a-\mu}{\sigma}$, $b^* = \frac{b-\mu}{\sigma}$

$$= \boxed{\int_{a^*}^{b^*} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-z^2/2}}_{=\varphi} dz = \Phi(b^*) - \Phi(a^*)}$$



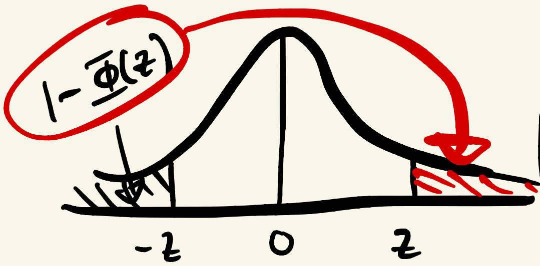
**Rmk** We say $X \sim \text{Norm}(\mu, \sigma)$ is normally distributed to mean it has PDF $\varphi_{\mu,\sigma}$

$\mu = 0 \quad \sigma = 1 \quad Z \sim \text{Norm}(0,1)$ is <u>unit normal</u>

**Rmk** When we want to compute for $Z \sim \text{Norm}(0,1)$

$$P(-z < Z < z) = \Phi(z) - \Phi(-z)$$

$$= \Phi(z) - (1 - \Phi(z))$$




$$\boxed{\Phi(-z,z) = 2\Phi(z) - 1}$$

## The BABY CLT = Normal Approx to Bin(n,p)

If $X_1, X_2, \ldots, X_n$ are IID Bernoulli/Indicator

i.e. $E(X_i) = p$ & $SD(X_i) = \sqrt{pq}$

$S_n = X_1 + \cdots + X_n \qquad S_n \sim \text{Bin}(n,p)$

then $$\boxed{P(a \le S_n \le b) \approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right)}$$

$E(S_n) = np$

$SD(S_n) = \sqrt{npq}$

OR w/ the continuity correction

$$\boxed{\approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}}\right)}$$

This is a good approximation as long as $npq > 10$

Ex Jane & Alex take a 48Q True-False exam

Jane answers correctly 75% of the time
Alex guesses ~ so 50% of the time

Passing = 30 or more correct answers

$$P(J \geq 30) = P(30 \leq J \leq 48)$$

$S_n \sim Bin(48, \frac{3}{4})$

$$= \sum_{k=30}^{48} \binom{48}{k} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{48-k}$$

$np = 48 \cdot \frac{3}{4} = 36$

$\sqrt{npq} = \sqrt{36 \cdot \frac{1}{4}} = 3$

$$\approx \Phi\left(\frac{48-np}{\sqrt{npq}}\right) - \Phi\left(\frac{30-np}{\sqrt{npq}}\right)$$

$$= \Phi\left(\frac{12}{3}\right) - \Phi\left(\frac{-6}{3}\right)$$

$$= \Phi(4) - \Phi(-2)$$

LOOK UP IN Z-TABLE

$$= 1 - .0228 \sim 98\%$$

(Do this w/ the continuity correction

$$\Phi\left(\frac{12.5}{3}\right) - \Phi\left(\frac{-6.5}{3}\right)$$

$-2.166$

$1 - .0154 \rightsquigarrow 98.46\%$ )

Probability of Alex passing

$np = 48 \cdot \frac{1}{2} = 24$

$\sqrt{npq} = \sqrt{24 \cdot \frac{1}{2}} = \sqrt{12} \approx 3.46$

$$P(30 \leq A \leq 48)$$

$$= \cancel{\Phi\left(\frac{48-24}{\sqrt{12}}\right)}^{\,1} - \Phi\left(\frac{30-24}{\sqrt{12}}\right)$$

$$= 1 - \Phi(1.73) = 1 - .9582$$

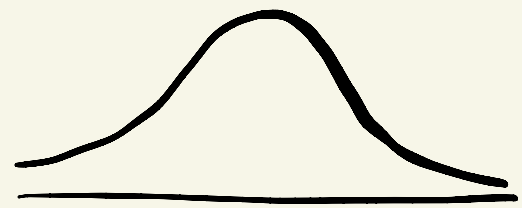
$\approx 4\%$

**Ex** Flip a fair coin 100 times

What symmetric range of heads accounts for 95% of all observations?

68-95-99.7 Rule

$\sigma$ $\quad 2\sigma \quad 3\sigma$




So $np \pm 2\sigma$

$100 \cdot \frac{1}{2} \pm 2\sqrt{npq}$

$\sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 5$

$\Rightarrow$ $50 \pm 10$ heads accounts for 95% of all observations.

# THE GROWN-UP CENTRAL LIMIT THM

If $X_1, X_2, \ldots, X_n$ are IID RVs w/

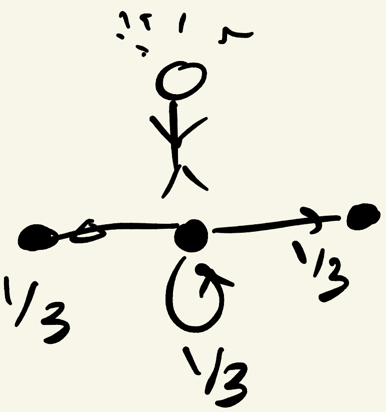
$E(X_i) = \mu \qquad SD(X_i) = \sigma \qquad S_n = X_1 + \cdots + X_n$

then
$$\boxed{P(a \le S_n \le b) \approx \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right)}$$

N.B. when $S_n \sim Bin(n,p)$ $\mu = p$ $\sigma = \sqrt{pq}$

(* You can use continuity correction when $X_i$ have possible values that are consecutive integers)

# DRUNKARD'S WALK = MODEL for DIFFUSION




At any time step we model displacement as

$$X_i = \begin{cases} -1 & \text{w/ prob } 1/3 \\ 0 & \text{"} \\ +1 & \text{"} \end{cases} \qquad X_i \sim \text{Unif}\{-1,0,1\}$$

**Q**: After 10,000 time steps what's the probability that $P(S_n > 100)$ ?

**A**: Apply CLT $\quad P(S_n > 100) = 1 - \Phi\left(\frac{100 - n\mu}{\sigma\sqrt{n}}\right)$

N.B. $\quad E(X_i) = \mu = 0 = \cancel{-1\cdot\frac{1}{3}} + 0\cdot\frac{1}{3} + \cancel{1\cdot\frac{1}{3}} = 0$

$$V(X_i) = E(X_i^2) - 0^2 = (-1)^2\cdot\frac{1}{3} + 0^2\cdot\frac{1}{3} + 1^2\cdot\frac{1}{3} = \frac{2}{3}$$

$$SD(X_i) = \sqrt{2/3}$$

Look up in z-table $\Phi\left(\frac{100 - 0}{\sqrt{2/3}\sqrt{10^4}}\right) = \Phi\left(\frac{100}{81.65}\right)$

$\Phi(1.224)$

SO $\quad 1 - .8888 \approx 11\%$

# CONFIDENCE INTERVALS

The CLT gives us better bounds on estimating $\mu = E(X_i)$ by using $\bar{X} = \frac{S_n}{n}$

Assuming we know $\sigma$ = true SD, but we can drop this and use $s_n$ = sample standard deviation.

## BASIC SET UP

If we want to determine $\mu$ up to precision $\varepsilon$ w/ 95% confidence

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \geq .95$$

($= 1 - \alpha$
$\alpha$ = confidence level
here $\alpha = .05$
$\rightarrow$ p-value!)

equiv to

$$P(\underbrace{n\mu - n\varepsilon}_{a} \leq S_n \leq \underbrace{n\mu + n\varepsilon}_{b})$$

$$\approx \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right) = \Phi\left(\frac{\cancel{n}^{\sqrt{n}}\varepsilon}{\sigma\cancel{\sqrt{n}}}\right) - \Phi\left(\frac{-\cancel{n}^{\sqrt{n}}\varepsilon}{\sigma\cancel{\sqrt{n}}}\right)$$

$$= \Phi\left(-\frac{\sqrt{n}\varepsilon}{\sigma}, \frac{\sqrt{n}\varepsilon}{\sigma}\right)$$

$$= 2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1$$

$$2\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1 \geq .95$$ ($= 1 - \alpha$ choose your $\alpha$!)

$$\Phi\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) \geq \frac{1.95}{2} = .975 \rightarrow$$ find smallest $z$

$\boxed{\text{REVERSE Z-TABLE LOOKUP}}$ $\rightarrow$ s.t. $\Phi(z) \geq .975$

$\Rightarrow z \geq 1.96$ ★

We can do 1 of 2 things

1) We know $\varepsilon$, then $\Phi\left(\frac{\sqrt{n}\,\varepsilon}{\sigma}\right) \geq .975$
to ensure

We need $\frac{\sqrt{n}\,\varepsilon}{\sigma} \geq 1.96$
Inverse Look up

$$\Rightarrow \quad n \geq \left(\frac{1.96\,\sigma}{\varepsilon}\right)^2$$

2) We know $n$, then we can give a lower bound on the spread/error in our measurements

$$\varepsilon > \frac{1.96\,\sigma}{\sqrt{n}}$$

95% Confidence interval $\approx \hat{p} \pm \frac{2\sigma}{\sqrt{n}}$

THE SQUARE ROOT LAW

Ex Roll a 6-sided die

1) Want $\varepsilon = .5 = 1/2$

$\sigma = \sqrt{\frac{35}{12}}$

$n \geq \left(\frac{1.96\,\sigma}{\varepsilon}\right)^2 \approx 44.81$ !

vs Chebyshev bound of $n \geq 243$

2) Know $n = 243$, then $\varepsilon > \frac{1.96\sqrt{\frac{35}{12}}}{\sqrt{243}} \approx .218$

$-\frac{2\sigma}{\sqrt{n}}$ $\quad$ $\frac{2\sigma}{\sqrt{n}}$

( 3.5 ) 95% confidence interval