Lec 21 Math 241 11/29/17

How to use CLT. If $X_1, \dots, X_n \overset{iid}{\sim}$ w/ mean μ, w/ var. σ^2, n is "large"

(2) $\bar{X} \overset{d}{\approx} N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$

(3) $T \overset{d}{\approx} N\left(n\mu, (\sigma\sqrt{n})^2\right)$

Random walk.. take 100 steps, what's the prob we are more than 10 steps from where we started?

$X_1, \dots, X_{100} \stackrel{iid}{\sim}$ Rademacher

$T = X_1 + \dots + X_{100}$

↑

total distance away

$$P(|T| \geq 10) = P(T > 10) + P(T < -10)$$

$$\approx P\left(\frac{T-0}{10} > \frac{10-0}{10}\right) + P\left(\frac{T-0}{10} < \frac{-10-0}{10}\right) = P(Z > 1) + P(Z < -1) = 2P(Z \geq 1)$$

symmetric bell curve

$$= 2 \cdot 16\% = \boxed{32\%}$$

$$T \stackrel{d}{\approx} N(n\mu, (\sqrt{n}\sigma)^2) = N(0, 10^2)$$

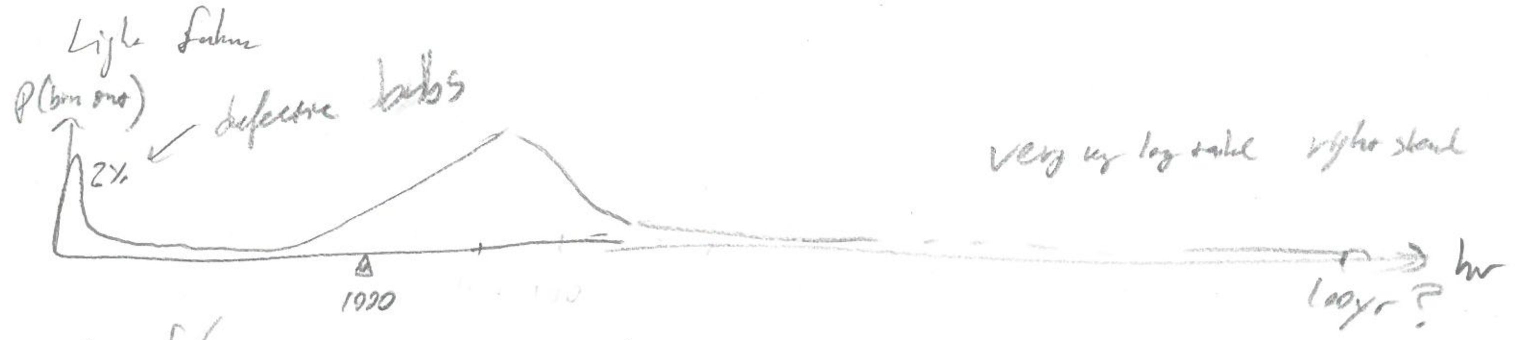
$\mu = 0, \quad \sigma = 1$

$E(X) = \quad \sigma = \sqrt{Var(X)}$

previous lectures

$\Rightarrow n\mu = 100 \cdot 0 = 0$

$\sqrt{n}\sigma = \sqrt{100} \cdot 1 = 10$



$X \sim f_{💡}(\mu \approx 1000, \sigma \approx 500)$

You get 50 bulbs. what's the prob the avg burnout is > 1300hr?

$X_1, \dots, X_{50} \stackrel{iid}{\sim} f_{💡}$

$$P(\bar{X} > 1300\text{hr}) \approx P\left(\frac{\bar{X} - 1000}{70.7} > \frac{1300 - 1000}{70.7}\right) \approx P(Z > 4.24) \approx 0$$

↑

$$\bar{X} \approx N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = N(1000, 70.7^2)$$

$$\frac{500}{\sqrt{50}} = 70.7$$

Shippers are lose on 2% of orders. In 10,000 orders, what's prob more than 3% lose?

$$P(\bar X > 3\%) = P\left(\frac{\bar X - .02}{.0014} > \frac{.03 - .02}{.0014}\right) \overset{\text{by CLT}}{\approx} P(Z > 7.14) \approx 0$$

$$X_1, \ldots, X_{10000} \overset{iid}{\sim} \text{Bern}(0.02) \overset{\text{by CLT}}{\Rightarrow} \bar X \approx N\left(\mu, \left(\frac{\sigma}{\sqrt n}\right)^2\right) = N(.02, .0014^2)$$

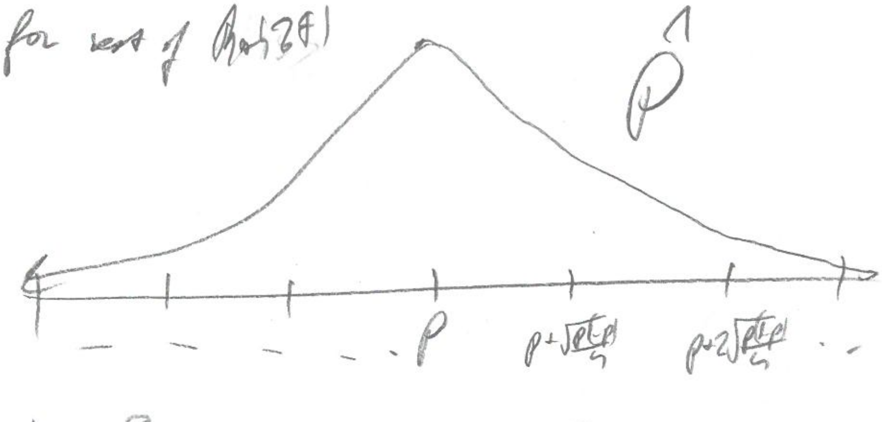
$$\Rightarrow \mu = p = 0.02, \quad \sigma = \sqrt{p(1-p)} = \sqrt{.02 \cdot .98} \Rightarrow \frac{\sigma}{\sqrt n} = \sqrt{\frac{.02 \cdot .98}{10000}} = .0014$$

$\bar X$ gets a special name if $X_1, \ldots, X_n \overset{iid}{\sim}$ Bern, $\bar X$ denoted

$\underbrace{\hat p}_{\text{sample proportion r.v.}} = \bar X$ $\qquad$ $\underbrace{\hat p}_{\text{sample proportion}} := \bar x = \frac{\sum x_i}{n} = \frac{\# 1\text{'s}}{n}$

$$\bar X \approx N\left(\mu, \left(\frac{\sigma}{\sqrt n}\right)^2\right)$$

$$\Rightarrow \underbrace{\hat p \approx N\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right)}_{\text{focus for rest of Math 241}} \quad \text{for Bernoullis}$$



Who likes mushrooms? $\qquad \hat p = \ldots$ what is p? INVERSE problem!!

$\frac{\text{PROB} \uparrow}{\text{STAT} \downarrow}$

$X_1, \ldots, X_N \overset{iid}{\sim} \text{Bern}(p)$

$|N| \approx \infty$

p is unknown as it is

$p := \frac{\sum x_i}{N}$

"Statistical Inference" has 3 goals:

Goals:

1. Estimate p as a single pt
2. Estimate a range of p's which make sense
3. Test theories about what p is.

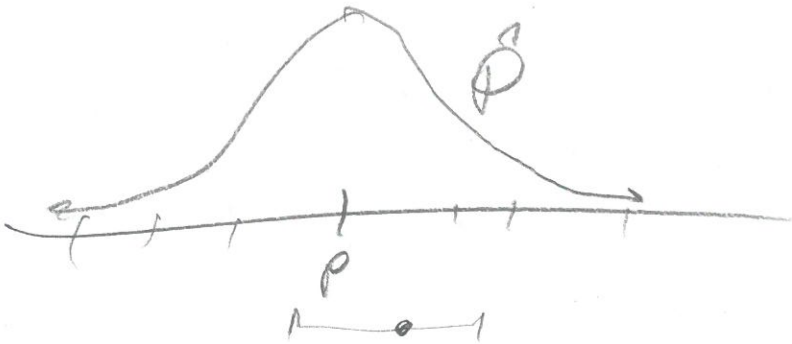
How? Take a "finite sample" or "small sample" of size $\underbrace{n << N}_{\text{many less than}}$ but is large enough for CLT to kick in

Sample

Sample must be "representative" which means it preserves iid property. How? Simple random sample. All males? All college students? No... completely at random using random # generator.

How to get best guess of p? $\hat{p} := \frac{\sum x_i}{n} = \frac{\# 1\text{'s}}{n}$ sample proportion.

Where does $\hat{p}$ come from? $\underbrace{\hat{p}}_{\text{lowercase}}$ is a value from $\underbrace{\hat{P}}_{\text{uppercase}}$



Interval procedure:

What if I take $\left[\hat{p} \pm \sqrt{\frac{p(1-p)}{n}}\right] = \left[\hat{p} - \quad , \hat{p} + \quad \right]$

What is the probability of this covers regardless of this interval procedure ~~of~~ cont. p?

$$P\left(p \in \left[\hat{P} \pm \sqrt{\frac{p(1-p)}{n}}\right]\right)$$

Lots of P's: $P()$, p, $\hat{p}$, $\hat{P}$

$$= P\left(\hat{P} - \sqrt{\frac{p(1-p)}{n}} \le p \le \hat{P} + \sqrt{\frac{p(1-p)}{n}}\right)$$

$$= P\left(-\sqrt{\frac{p(1-p)}{n}} \le p - \hat{P} \le + \sqrt{\frac{p(1-p)}{n}}\right)$$

$$= P\left(-1 \le \frac{p - \hat{P}}{\sqrt{\frac{p(1-p)}{n}}} \le 1\right)$$

$$= P\left(-1 \le -Z \le 1\right)$$

$$= P\left(1 \ge Z \ge -1\right)$$

$$= P\left(-1 \le Z \le 1\right)$$

$$= P\left(Z \in [-1, 1]\right) = .68$$

What if I do $\left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

s.t. $z_{\frac{\alpha}{2}} := F_Z^{-1}\left(1 - \frac{\alpha}{2}\right) \Leftrightarrow 1 - \frac{\alpha}{2} = \int_{-\infty}^{z_{\frac{\alpha}{2}}} f_Z(x)\,dx$

what does this mean?

$\alpha = 5\% \Rightarrow \frac{\alpha}{2} = 2.5\% \Rightarrow 1 - \frac{\alpha}{2} = 97.5\%$

what is $F_Z^{-1}(97.5\%) = 2 = z_{2.5\%}$

$\Rightarrow \frac{\alpha}{2} = \int_{z_{\frac{\alpha}{2}}}^{\infty} f_Z(x)\,dx$

via symmetry

$= \int_{-\infty}^{-z_{\frac{\alpha}{2}}} f_Z(x)\,dx$

$P\left(p \in \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right)$

$= P\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

$= P\left(\hat{p} \in \left[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}\right]\right) = F_Z\left(z_{\frac{\alpha}{2}}\right) - F_Z\left(-z_{\frac{\alpha}{2}}\right)$

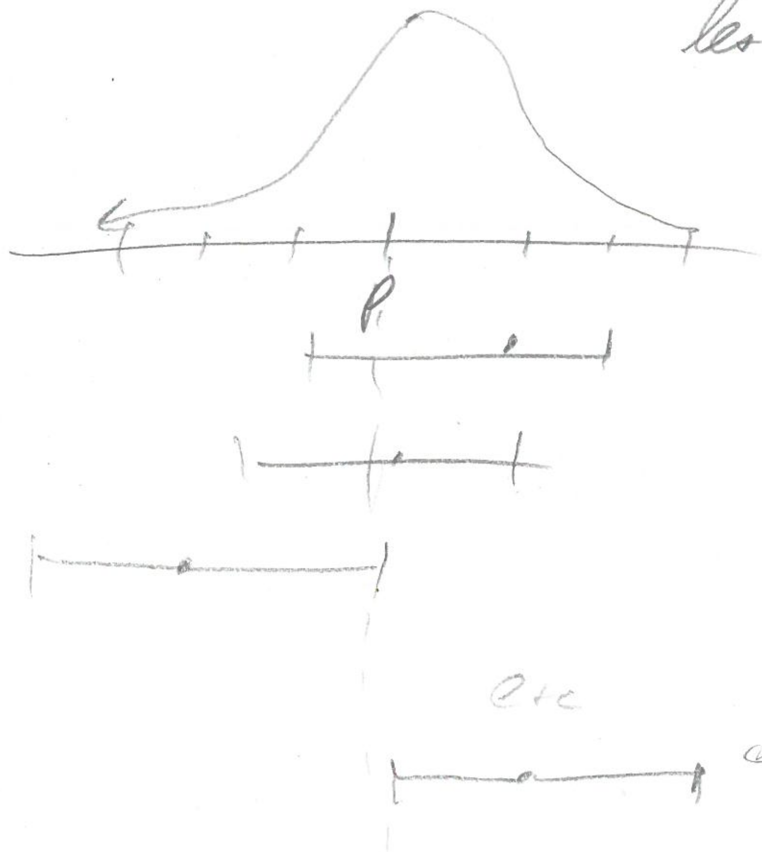
$= \left(1 - \frac{\alpha}{2}\right) - \left(\frac{\alpha}{2}\right)$

$= 1 - \alpha$

So the multiple $z_{\frac{\alpha}{2}}$ allows me to pick my prob of the interval.

So $\left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ is a procedure, when repeated, gives $1-\alpha$ prob that p is contained within it.

let $\alpha = 5\% \Rightarrow z_{\frac{\alpha}{2}} = 2$



5% of the time this happens

Big prob: p unknown! But..

$$\left[\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right] \approx \left[\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

as long as $p \not\approx 0$ and $p \not\approx 1$

Hotly debated for ≈ 100 yr

$$CI_{p,1-\alpha} := \left[\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$\underbrace{CI}_{\text{Confidence Interval}}$

Said to have $1-\alpha$ coverage of p. What does this mean?