11/29 (MGF)

(Everything you need to KNOW!) Recap → Def: Moment Generating Function for r.v's X:

$m_X(t) = E[e^{tX}]$ recall $e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$ (Taylor Series Expansion)

Properties: I. $m_X(t) = m_Y(t) \Rightarrow X \overset{d}{=} Y$

II. $m_X^{(k)}(0) = E[X^k]$

III. $Y = aX + c \Rightarrow m_Y(t) = e^{tc}\, m_X(at)$

IV. If X, Y indep

$m_{X+Y}(t) = m_X(t)m_Y(t) \overset{iid}{=} (m_X(t))^2$

Law of Large #s → V. $\lim_{n\to\infty} m_{X_n}(t) = m_Y(t) \Longleftrightarrow X \to Y$ P.DF this → PDF that (approx.)

$X_1, X_2, \ldots$ square of r.v's.

$\Rightarrow$ if n large $\Rightarrow X_n \overset{d}{\approx} Y$ X_n is approx equally distributed as Y.

- $X \sim Bern(p) \Rightarrow m_X(t) = 1 - p + pe^t$ (definition).
- $X \sim Binom(n,p) \Rightarrow m_X(t) = (1 - p + pe^t)^n$ from rule V (add indep r.v of ...)
- $X \sim Geom(p) \Rightarrow m_X(t) = \frac{pe^t}{1 - e^t(1-p)}$ (definition). if $t < h(\frac{1}{1-p})$
- $X \sim Exp(\lambda) \Rightarrow m_X(t) = \frac{\lambda}{\lambda - t}$ if $t < \lambda$
- $Z \sim N(0,1) \Rightarrow m_X(t) = e^{\frac{t^2}{2}}$
- $X \sim N(\mu, \sigma^2) \Rightarrow m_X(t) = e^{mt + \frac{1}{2}\sigma^2 t^2}$
- $X \sim Deg(t) \Rightarrow m_X(t) = e^{tc}$

* When X gets larger, it gets closer to the mean *.

A week LLN!

Proof not required on final only HW

- if $X_1, \dots, X_n$ iid same distrubution with mean $M < \infty$ ($E[X]$)

* $\Rightarrow \bar{X}_n \to M$

$M \sim Deg(M)$

$\lim_{n\to\infty} m_{\bar{X}_n}(t) = e^{tm} = m_Y(t) =$

$m_{\bar{X}_n}(t) \overset{\text{def. of } \bar{X}}{=} m_{\frac{T}{n}}(t) \overset{\text{fact III}}{=} m_T(\frac{t}{n})$

$= \prod_{i=1}^{n} m_{X_i}(\frac{t}{n}) = (m_X(\frac{t}{n}))^n = (E[e^{\frac{t}{n}X}])^n = (E[1 + \frac{tX}{n} + \frac{t^2X^2}{2!n^2} + \frac{t^3X^3}{3!n^3} + \dots$

fact IV — def. of mgf. — Taylor series $\exp(e^{tx})$

Thus $= (E[1 + \frac{tx}{n} + o(\frac{1}{n})])^2 = (1 + \frac{tm}{n} + E(o(\frac{1}{n})))^n$

$m_{\bar{X}_n}(t) = [(1 + \frac{tm}{n} + o(\frac{1}{n})]^n \rightarrow$ ignore expectation exponent as a technical pt. ~~with~~ $\lim_{n\to\infty}(1 + \frac{tm}{n} + o(\frac{1}{0}))^n = e^{tm}$

We say

$f(n) = o(g(n))$

"little o"

if

$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$ $g(n)$ goes ∞ larger faster then $f(n)$

$n^2 = o(n^3)$

$\lim_{n\to\infty} \frac{n^2}{n^3} = \lim \frac{1}{n} = 0$ ✓

$\frac{t^2X^2}{2!n^2} + \frac{t^3X^3}{3!n^3} + \dots = (o)(\frac{1}{n})$ pf. $\lim_{n\to\infty} \frac{\frac{t^2X^2}{2!n^2} + \frac{t^3X^3}{3!n^3}}{\frac{1}{n}} = \lim_{n\to\infty} \frac{t^2X^2}{2!n} + \frac{t^3X^3}{3!n^2} + \dots = 0.$

Remember: $\lim_{n\to\infty}(1 + \frac{1}{n})^n = e$

$\lim_{n\to\infty}(1 + \frac{a}{n})^n = e^a$ but goes slower.

Not affected if we put smth else in there

Does $\lim(1 + \frac{a}{n} + o(\frac{1}{n}))^n \overset{\text{small?}}{=} e^a$ Yes

$\frac{1}{n+1} = o(\frac{1}{n}) \Rightarrow \lim \frac{\frac{1}{n+1}}{\frac{1}{n}} = \lim \dots = 0$ ✓.

WHY standard normal is important?

- let $X_1, \dots, X_n \overset{iid}{\sim}$ with mean μ and s.e. σ

consider.

$*$ $\bar{X}$ standardized to have mean 0, s.e 1. $\rightarrow$

$$C_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} = \frac{\sqrt{n}\left(\frac{X_1+\dots+X_n}{n} - \mu\right)}{\sigma}$$

$$= \frac{\sqrt{n}\left(\frac{X_1+\dots+X_n}{n} - \frac{n\mu}{n}\right)}{\sigma} = \frac{\sqrt{n}\left(\frac{X_1-\mu+X_2-\mu+\dots+X_n-\mu}{n}\right)}{\sigma}$$

$$= \frac{(X_1-\mu)+(X_2-\mu)+\dots+(X_n-\mu)}{\sigma\sqrt{n}}$$

$$= \frac{1}{\sqrt{n}}\left(\frac{X_1-\mu}{\sigma} + \frac{X_2-\mu}{\sigma} + \dots + \frac{X_n-\mu}{\sigma}\right)$$

let $Z_i = \frac{X_i - \mu}{\sigma}$ Z_i is X_i standardized

$$= \frac{1}{\sqrt{n}}(Z_1 + Z_2 + \dots Z_i)$$

$E[Z_i] = 0$

$SE[Z_i] = 1$

☆ when you take a r.v and standardize it $\downarrow$ always: mean $= 0$, s.e $= 1$ ☆.

C_n When n is large, it behaviors like certain r.v.

- $$\lim_{n\to\infty} M_{C_n}(t) = M_{C_n}(t) = M_{\frac{1}{\sqrt{n}}(Z_1+\dots+Z_n)}(t) \underset{\text{by III}}{=} M_{Z_1+\dots Z_n}\left(\frac{t}{\sqrt{n}}\right) \underset{\text{by IV}}{=} M_{Z_1}\left(\frac{t}{\sqrt{n}}\right) \cdot \dots \cdot M_{Z_n}\left(\frac{t}{\sqrt{n}}\right)$$

$$= \left(M_Z\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

$$= \left(E\left[e^{\frac{t}{\sqrt{n}}Z}\right]\right)^n = \left(E\left[1 + \frac{tZ}{\sqrt{n}} + \frac{t^2Z^2}{2!\,n} + \underbrace{\frac{t^3Z^3}{3!\,n^{1.5}} + \frac{t^4Z^4}{4!\,n^2} + \dots}_{\text{goes to 0 quicker than } \frac{1}{n}}\right]\right)^n$$

can use the "little o" notation here. $\hookrightarrow o\left(\frac{1}{n}\right)$

$$= \left(E\left[1 + \frac{tZ}{\sqrt{n}} + \frac{t^2Z^2}{2n} + o\left(\frac{1}{n}\right)\right]\right)^n$$

$$= \left(1 + E\left[\frac{tZ}{\sqrt{n}}\right] + E\left[\frac{t^2Z^2}{2n}\right] + E\left[o\left(\frac{1}{n}\right)\right]\right)^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n$$

$\frac{t}{\sqrt{n}}E[Z] = 0$

$\frac{t^2}{n}E[Z^2] = \frac{t^2}{2n}$

Note: $Var[X] = E[X^2] - \mu^2$

if $\mu^2 = 0$

$Var[X] = E[X^2]$

if $Var[X] = 1$

$\Rightarrow E[X^2] = 1$

$$\lim_{n\to\infty} M_{C_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = \underbrace{e^{\frac{t^2}{2}}}_{\text{the mgf of } Z\sim N(0,1)} \Rightarrow C_n \to N(0,1)$$

adding a bunch of distributions –

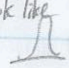
$\approx$ $e^{\frac{-x^2}{2}}$

a bell curve (cuz it looks like a bell)

— pretty much everything in the earth (seems infinite) has a mean, s.e and come out looking like this

$\bar{X} = \frac{T}{n}$

need a bunch to add up to look like



• CLT = Central Limit Theorem.

(I) If $X_1, \dots, X_n \overset{iid}{\sim}$ w/ mean(μ), Standard error (σ)

$$\Rightarrow \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow Z \sim N(0, 1).$$

$$\Rightarrow \text{If } n \text{ large} \dots \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \overset{d}{\approx} Z \sim N(0,1) \Rightarrow \bar{X} \overset{d}{\approx} N(\mu, (\tfrac{\sigma}{\sqrt{n}})^2)$$

this standardized Z bar (II) (III)

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

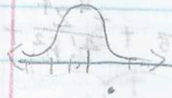
(IV)

Standard Normal r.v – $Z \sim N(0,1) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Generalized – $X = \mu + \sigma Z \sim N(\mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{\sqrt{2\pi}}(x-\mu)^2}$

Normal $\quad C_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \to Z.$

That's what the outcome is → considered normal.



defines what normal means

abnormal!

Ex : $X_1, \dots, X_{30=n} \overset{iid}{\sim}$ Geom (p) $\quad p = \frac{1}{2}$. $\quad n=30$

What is the prob. the avg realization is more than 2.75?

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

$m = \frac{1}{p} = \frac{1}{\frac{1}{2}} = 2$

$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{2} \approx 1.414$

$\frac{\sigma}{\sqrt{n}} = \frac{1.414}{\sqrt{30}} \approx .2582$

$P(\bar{X} > 2.75)$ = By the CLT $\bar{X} \overset{d}{\approx} N(2, .2502)$.

$= P(\frac{\bar{X} - 2}{.258} > \frac{2.75 - 2}{.258}) \approx P(Z > 3) = 0.15\%$ (by the graph $P(-3,3) = 99.7\%$, $(>3) = 0.15\%$)

– Random Walk 100 steps. What is the prob you are more than 10 steps away from where you start?

let $T = X_1 + X_2 + \dots + X_{100}$

$X_1, \dots, X_{100} \overset{iid}{\sim} \begin{cases} 1 & wp\ \frac{1}{2} \\ -1 & wp\ \frac{1}{2} \end{cases}$

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

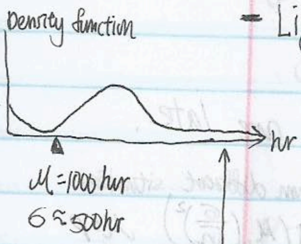
σ = expected distance from mean.

From the CLT, we know $T \overset{d}{\approx} N(n\mu, (\sqrt{n}\sigma)^2)$

$\Rightarrow T \overset{d}{\approx} N(0, 10^2)$

continue →

$$P(|T|>10) = P(T<-10) + P(T>10)$$
$$= P\left(\frac{T-0}{10} < \frac{-10-0}{10}\right) + P\left(\frac{T-0}{10} > \frac{10-0}{10}\right)$$
$$= P(Z<-1) + P(Z>1)$$
$$= P(Z \notin [-1,1]) = 32\%.$$



- Light Bulb Lifetime.

keyword! → which CLT do you take.

You buy 50 bulbs. What is the prob the avg bulb lasts more than 1300hrs?

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

Not normal b/c it has a really long right tail.

$f(t)$

$$P(\bar{X}>1300) = P\left(\frac{\bar{X}-1000}{\frac{500}{\sqrt{50}}} > \frac{1300-1000}{\frac{500}{\sqrt{50}}}\right) \overset{\text{II}}{\approx} P(Z \geq 4.24) \approx 0 \ *.$$

As n gets bigger → normal if it is CLT.

12/01 CLT if $X_1, \dots, X_n \overset{iid}{\sim}$ w/ mean μ and s.e. σ, then if n large…

(II) $\dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \overset{d}{\approx} Z \sim N(0,1)$

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

(IV) $T \approx N(n\mu, (\sigma\sqrt{n})^2)$

<u>Questions</u> • Shipments goes late 2% of the time.
In 10,000 orders, what is the prob more than 3% are late.

Assume its iid.

$X_1, \dots, X_{10{,}000} \overset{iid}{\sim} \text{Bern}(2\%)$ ⟹ Sample proportions r.v.

(Come from different situation.)
III. $\bar{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$ $\mu = p$, $\sigma = \sqrt{p(1-p)}$

$\hat{p} \sim N(p, (\sqrt{\frac{p(1-p)}{n}})^2)$

$P(\hat{p} > 3\%)$

$= P\left(\dfrac{\hat{p} - 0.02}{0.0014} > \dfrac{0.03 - 0.02}{0.0014}\right)$

by CLT ⟹ $\bar{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})^2)$

$= N(0.02, 0.0014^2)$

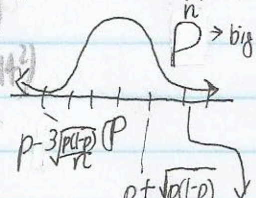
*Not gonna happen $\overset{II}{\approx} P(Z > 7.14) \approx 0$

bigger than given 2%.

$\mu = p = 0.02$

$\sigma = \sqrt{p(1-p)} = \sqrt{0.02 \cdot 0.98}$

$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{0.02 \cdot 0.98}{10000}} = 0.0014$



$\hat{p}$ → big p (a r.v.) has many little p's

$\hat{p} = \bar{X} = \frac{1}{n}\sum X_i$.

☆ μ is the real thing = The total population of the people in the whole world who like mushrooms.
↓ inaccessable.
you can't go ask all 7.5 billion ppl. but you really will not be able to know it.
$\forall n \; E[\bar{X}_n] = \mu$

- Do you like mushrooms?

$\hat{p} = \dfrac{\#\text{ Yes's}}{n} = \dfrac{19}{32} = 0.59$

<u>Goals of Inference</u>

Point estimation. ① Give me best "guess" of p.

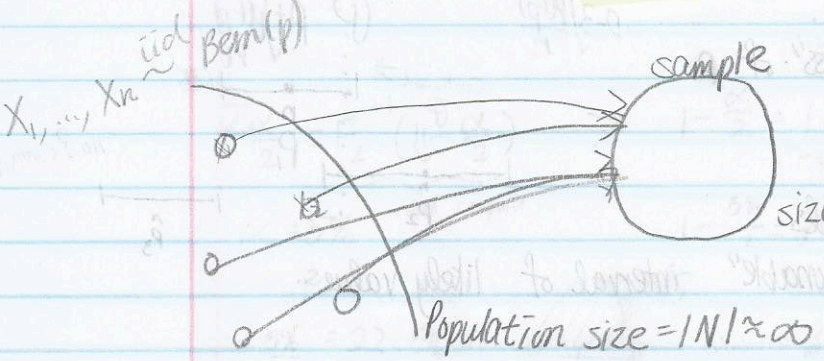
② Give me a "reasonable" interval of likely values.

③ Test "theories" about p.

<u>Statistical Inference</u>

Infer parameters from sample.

• What constimetes a sample?



$p = \frac{\sum_{i=1}^{N} X_i}{N}$ unknownable "true" parameter

$X_1, \dots, X_n \overset{iid}{\sim} Bern(p)$

We want X do sample

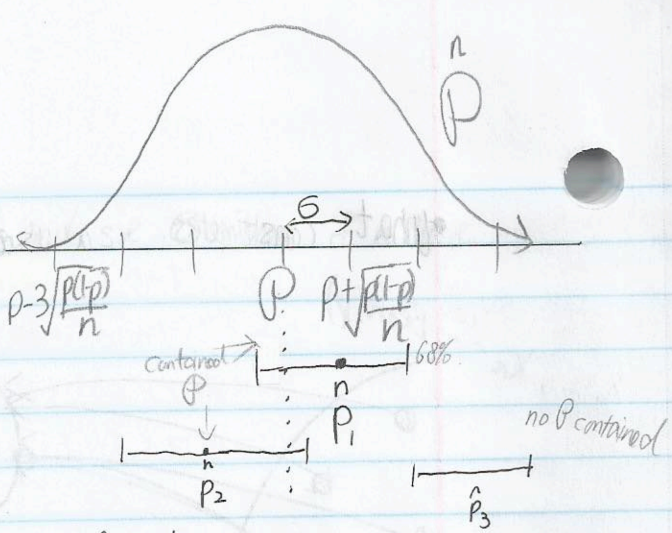
$X_1, \dots, X_n \overset{iid}{\sim} Bern(p)$
↑
sample size.

-what type of sampling procedure?

Not representitive! You want to be representitive of the population.

✳ How? → Completely random selection then you have $X_1, \dots, X_n \overset{iid}{\sim} Bern(p)$.

$\hat{p}$ is in realization from $\hat{P}$



most of your $\overset{n}{p}$ will be and p.

Goals of Inference

① Point Estimation.

On Final Question: = Give me best "guess" of p.

How? $\hat{p} = \frac{\sum X_i}{n} = \frac{\#1's}{n}$

What is you best guess $\overset{n}{p}$

② Confidence Internals.

= Give me a "reasonable" interval of likely values.

just write what $\overset{n}{p}$ is.

Consider the interval:

$[2 \pm 1] = [1,3]$

$$\left[\overset{n}{p} \pm \sqrt{\frac{p(1-p)}{n}}\right] := \left[\overset{n}{p} - \sqrt{\frac{p(1-p)}{n}},\ \overset{n}{p} + \sqrt{\frac{p(1-p)}{n}}\right].$$

not the same

What is the prob $p \in$ the interval?

4 types of p's

p, $\overset{n}{p}$, $\overset{n}{P}$, $P($ $)$

reg p. — P hood — big P hood

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

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

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

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

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

$$= P(1 \ge Z \le -1)$$

$$= P(Z \in [-1,1]) = 0.68.$$

rather have it higher than 68%

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

$$Z_{\frac{\alpha}{2}} := \underbrace{F_Z^{-1}\left(1-\frac{\alpha}{2}\right)}_{\text{need computer.}} \longrightarrow 1-\frac{\alpha}{2} = \int_{-\infty}^{Z_{\frac{\alpha}{2}}} f_Z(x)\,dx$$

$\alpha = 5\%$ $\qquad 1-\frac{5\%}{2} = 0.975 = \int_{-\infty}^{Z_{2.5\%}} f_Z(x)\,dx$

$Z_{\frac{5\%}{2}} = 22.5\%$ $\quad$ .975 Quantile of normal curve. $Z_{\frac{\alpha}{2}}$

$Z_{\frac{\alpha}{2}} = 97.5\%$

$Z_{\frac{\alpha}{3}} = 99.7\%$

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

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

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

$$= P\left(-Z_{\frac{\alpha}{2}} \le -Z \le Z_{\frac{\alpha}{2}}\right).$$

$$\frac{\alpha}{2} = \int_{-\infty}^{-Z_{\frac{\alpha}{2}}} f_Z(x)\,dx \qquad = P\left(Z_{\frac{\alpha}{2}} \ge Z \ge -Z_{\frac{\alpha}{2}}\right)$$

$$= P\left(Z \in \left[-Z_{\frac{\alpha}{2}},\ Z_{\frac{\alpha}{2}}\right]\right)$$

$$= F_Z\left(-Z_{\frac{\alpha}{2}}\right) \qquad = F_Z\left(Z_{\frac{\alpha}{2}}\right) - F_Z\left(-Z_{\frac{\alpha}{2}}\right)$$

inverse CDF

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

$$= 1-\alpha$$

Cover 100% of the time

$\alpha = 0$ $\qquad -\left[\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$ we don't know p doesn't make sense!

then plug in 0 for α $= \left[0.59 \pm Z\sqrt{\frac{p(1-p)}{32}}\right]$

and calculate.

$Z \in -\infty, 0$. ? ☆Instead consider — Definition of a $1-\alpha$ size confidence interval for parameter p.

$$\boxed{CI_{p,1-\alpha}\ \left[\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$ the whole thing fails but it's approximately true

$$\approx \left[\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$$ → ✱ subject of 100 yr of fighting

Accurate if $p \neq 0$ and $p \neq 1$.

Technically the 2 sided . . . ?