Stat 480B – First Lecture (Spring 2022)

- Previously, we assumed we knew all the ~~variables'~~ distributions' parameters...
- We used CDF Technique, MGF Method, Joint stuff, etc., to derive fcns of R.V.s.
- Order-statistics – we learned about
- Approx. distributions used for n (sample size) sufficiently large...

First lecture

§7.2 – Sequences of R.V.s.

<u>Scenario:</u> We have a seq. of R.V.s $Y_1, Y_2, Y_3, \dots$, with respective CDFs $G_1(y), G_2(y), G_3(y), \dots$; that is, $G_n(y) = P[Y_n \le y]$, $n = 1, 2, 3, \dots$. Goal is to find the limiting distribution of Y_n.

<u>Def:</u> Let $Y_n \sim G_n(y)$, $n = 1, 2, 3, \dots$, and let $G(y)$ be some CDF. If $\lim_{n\to\infty} G_n(y) = G(y)$ for all y at which $G(y)$ is continuous, then the seq. $Y_1, Y_2, \dots$ <u>converges in distribution</u> to $Y \sim G(y)$; that is, $Y_n \xrightarrow{d} Y$.

- first, need a seq. Then, ~~state~~ want CDF of n^{th} item (arbitrary n) of seq.; take limit as $n \to \infty$. If conv. to $G(y)$ satisfying the properties of a CDF, then definition applies.

<u>Handout #1)</u> $F_X(x) = \frac{1+x^3}{2}$, $-1 \le x \le 1$; random sample of size n. $\Rightarrow$ iid

$X_{n:m}$ — the n^{th} entry in the list, size of sample

<u>Want:</u> limiting dist. of $X_{n:n} = \max\{X_1, X_2, \dots, X_n\}$. Seq. of Interest: $Y_1 = X_{1:1}$ ← max, if have a sample size of $n=1$

$Y_n =$ ← idea

$Y_2 = X_{2:2}$ ← max " " $n=2$

$Y_3 = X_{3:3}$, etc.

$Y_n = X_{n:n}$

cont'd →

1st step: Need $G_n(y) = P[Y_n \leq y]$

out of a sample of up to size n

$= P[X_{n:n} \leq y]$ (prob. that the maximum is $\leq$ some #

$\Rightarrow$ $= P[\text{all } X_i\text{'s} \leq y]$

cuz iid

$= [F_X(y)]^n$

$G_n(y) = \left[\frac{1+y^3}{2}\right]^n, \quad -1 \leq y \leq 1$ support 2

$\leftarrow$ cuz each $|X_i| \leq 1$, and $|X_i| < y$, so intuitive.

CDF of $X_{n:n} = Y_n$

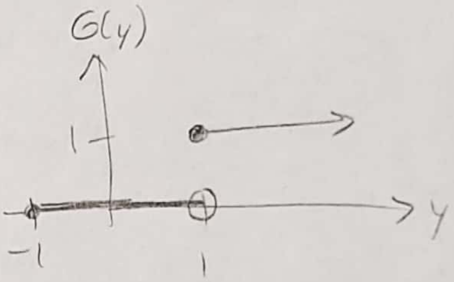
Want: $G(y) = \lim_{n\to\infty} G_n(y)$

$1+y^3 < 2$ for $-1 < y < 0$

$1+y^3 < 2$ for $0 < y < 1$, also.

smaller than denom, so as $n \to \infty$, $\lim \to 0$.

$$= \begin{cases} 0 & \text{if } -1 \leq y < 1 \\ 1 & \text{if } y \geq 1 \end{cases}$$



This is called a degenerative distribution at $y=1$

$\Rightarrow$ y can only equal 1 (cuz have a CDF)

jumps up once at some y value, equals that value forever.

has mass ~~=~~ 1 at $y=1$, discrete CDF.

applies just to the CDF (not the pdf, etc.)

$\longrightarrow$

STAT 480B

Examples for Section 7.2

1. Consider a random sample of size n from a distribution with cdf

$$F_X(x) = \frac{1+x^3}{2}, \quad -1 \leq x \leq 1.$$

Derive the limiting distribution of $X_{n:n}$.

2. Consider a random sample of size n from a distribution with cdf

$$F_X(x) = 1 - \frac{1}{x}, \quad x \geq 1.$$

Derive the limiting distribution of $X_{1:n}$.

3. Consider a random sample of size n from a distribution with cdf

$$F_X(x) = 1 - \frac{1}{x^2}, \quad x > 1.$$

   (a) Derive the limiting distribution of $Y_n = \frac{X_{n:n}}{\sqrt{n}}$.

   (b) Derive the limiting distribution of $X_{n:n}$.

- **Alternate term:** A sequence of R.V's $Y_1, Y_2, \ldots$ is said to **converge stochastically** to a constant c if its limiting distribution is degenerate at $y=c$.

means your seq. conv. in dist. to a degenerate distribution. Is a special case of convergence in distribution.

- So, in Handout #2, the seq. of the smallest order statistic converges stochastically to $y=1$.

- **Recall:** $\lim_{n\to\infty}\left(1+\frac{c}{n}\right)^n = e^c$,

when have limit of something raised to a power, provided the inside-something's limit exists, you can take that limit and raise the result to the original power.

$\lim_{n\to\infty}\left(1+\frac{c}{n}\right)^{nb} = e^{cb}$,

limit trick can be used here. ask/check

d a fn of n

$\lim_{n\to\infty}\left[1+\frac{c}{n}+\frac{d(n)}{n}\right]^{nb} = e^{cb}$ if $\lim_{n\to\infty} d(n) = 0$.

as n

- **Handout #3)** Random sample of size n w/ CDF $F_X(x) = 1-\frac{1}{x^2}, x>1$.

(a) **Want:** $Y_n = \frac{X_{n:n}}{\sqrt{n}}$

- need $G_n(y) = P[Y_n \le y]$ (CDF technique)

$= P\left[\frac{X_{n:n}}{\sqrt{n}} \le y\right]$

solve for our R.V. $(X_{n:n})$ $\to$ $= P[X_{n:n} \le y\sqrt{n}]$, $y\sqrt{n} > 1$.

$= P[\text{all } X_i\text{'s} \le y\sqrt{n}]$

if the maximum is $\le$ a #, then all the X_i's must also be.

$= [F_X(y\sqrt{n})]^n = \left[1-\frac{1}{ny^2}\right]^n, \; y > \frac{1}{\sqrt{n}}$

$= G_n(y)$

take limit of support to get support of $G(y)$

- So, $\lim_{n\to\infty} G_n(y) = \lim_{n\to\infty}\left[1-\frac{1}{ny^2}\right]^n$

$= \lim_{n\to\infty}\left[1+\frac{-\frac{1}{y^2}}{n}\right]^n = e^{-\frac{1}{y^2}}, \; y>0$

$= G(y)$

To verify legitimate CDF, $\lim_{y\to\infty} G(y) = 1$ ✓, ~~so yes~~, our CDF we found is the limiting distribution.

Handout #2) Given $F_X(x) = 1 - \frac{1}{x}$, $x \geq 1$. Have random sample of size n. w/ this CDF. Derive limiting dist. of $X_{1:n}$.

Want: $Y_n = X_{1:n}$

Need: CDF of Y_n, which is $G_n(y)$.

$$G_n(y) = P[Y_n \leq y]$$

$$= P[X_{1:n} \leq y] \quad \text{(prob. that the min is} \leq \text{some } y)$$

$$= 1 - P[X_{1:n} > y] \quad \text{(prob. that the min is} > \text{some } y)$$

$$\Rightarrow G_n(y) = 1 - P[\text{all } X_i\text{'s} > y]$$

w/ iid

$$= 1 - [1 - F_X(y)]^n, \quad y \geq 1$$

$$= 1 - [\cancel{1} - (\cancel{1} - \tfrac{1}{y})]^n$$

$$G_n(y) = 1 - \left(\frac{1}{y}\right)^n, \quad y \geq 1$$

i.e., $1 - \frac{1}{y^n}$, $y \geq 1$

So, $G(y) = \lim_{n \to \infty} G_n(y)$

$$= \lim_{n\to\infty} \left[1 - \frac{1}{y^n}\right], \quad y \geq 1$$

$$\boxed{= \begin{cases} 0 & y \leq 1 \\ 1 & y > 1 \end{cases}}$$

$y \geq 1 \Rightarrow y^n \geq 1^n = 1$

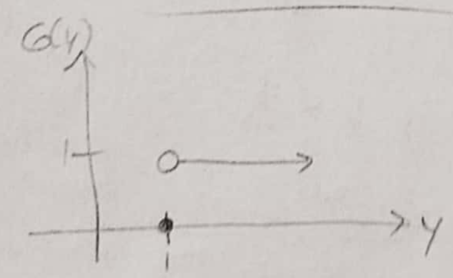
$\Rightarrow \frac{1}{y^n} \leq 1$

$\Rightarrow -\frac{1}{y^n} \geq -1$

$\Rightarrow 1 - \frac{1}{y^n} \geq 0$

$\Rightarrow -\frac{1}{y^n} \geq -1$

$\Rightarrow$



degenerate at $y = 1$

Cont'd →

Handout #3 (b)) Want $Y_n = X_{n:n}$

~~Need~~

$G_n(y) = P[Y_n \le y] = [F_X(y)]^n = [1 - \frac{1}{y^2}]^n, \; y > 1$

$\lim_{n\to\infty} G_n(y) = \lim_{n\to\infty} [\underbrace{1 - \frac{1}{y^2}}_{<1}]^n$

$\Rightarrow \lim_{n\to\infty} G_n(y) = G(y) = 0, \; y > 1,$

$\boxed{\text{but this is not a valid CDF, so the limiting dist. DNE}}$

$y > 1$

$\Rightarrow y^2 > 1 \Rightarrow 0 < y^2 < 1$

$\Rightarrow \frac{1}{y^2} < 1$

$\Rightarrow -\frac{1}{y^2} > -1$

i.e., $0 > -\frac{1}{y^2} > -1,$

$\Rightarrow 1 > 1 - \frac{1}{y^2} > 0$

$\Rightarrow 1 - \frac{1}{y}^2 < 1,$

so $(1 - \frac{1}{y^2})^n < (1)^n$
$= 1.$

- ~~All~~ each Handout example features a different type of outcome.
  1) limiting dist. is a degenerative dist. (ex #1 and #2)
  2) limiting dist. is some valid CDF (have to test to confirm) (ex #3a)
  3) limiting dist. DNE (ex #3b)

∘ use CDF to get limiting distributions.
  - what if the distribution doesn't have a CDF w/ a closed form? (eg., normal, gamma). Use MGF.

Thm 1: (See Handout) We use $M_n(t)$, the MGF of Y_n. Then instead of the CDF, we take the limit of $M_n(t)$ to hopefully get $M(t)$, which we must recognize as the MGF of a known dist.

~~AKA Central Limit Theorem~~

- if can't identify the limiting distribution is via $M(t)$, then $M(t)$ is useless.
  - then need CLT.

Example 1 (Handout):

I.e., $x = \begin{cases} 0 & \text{if "failure"} \\ 1 & \text{if "success"} \end{cases}$

$X \sim Bern(p)$, x = # of "successes" in 1 trial,

p = prob. of success. ($p = P(\text{"success"})$)

Want: $Y_n = \sum_{i=1}^{n} X_i$

= # of "successes" in n trials (cuz have n X_i's, and thus n trials).

($\Rightarrow Y_n \sim BIN(n, p)$)

CDF has no closed form, but the MGF does.

$\Rightarrow M_n(t) = (pe^t + q)^n$

$= [pe^t + 1 - p]^n$

$= [1 + p(e^t - 1)]^n$; want limiting dist.; conditions needed:

By the problem statement: For large n, $np \to \mu$, so,

$M_n(t) = \left[1 + \frac{\mu}{n}(e^t - 1)\right]^n$.

Now we take

$\lim_{n\to\infty}\left[1 + \frac{c}{n}\right]^n = e^c$

$$\lim_{n\to\infty} M_n(t) = \lim_{n\to\infty}\left[1 + \frac{\mu(e^t - 1)}{n}\right]^n$$

$$= e^{\mu(e^t - 1)}$$

at this point, need recognize which dist. has an MGF looking like this.

$M(t)$

$\Rightarrow$ MGF of a $POI(\mu)$. (poisson dist. w/ param μ).

If couldn't recognize MGF, need MGF technique

What happened:
From a $BIN(n,p)$ for Y_n, where as $n \to \infty$, $np \to \mu$, we got a limiting fcn ~~which~~ that was $POI(\mu)$ (as MGF)

needs to be a nonzero, small, finite #.

Theorem 1: *Let $Y_1, Y_2, Y_3, \ldots$ be a sequence of random variables with corresponding cdfs $G_1(y), G_2(y), G_3(y), \ldots$ and mgfs $M_1(t), M_2(t), M_3(t), \ldots$.*

If $M(t)$ is the mgf of a cdf $G(y)$ and if $\lim_{n\to\infty} M_n(t) = M(t)$ for all t in an open interval containing 0, $-h < t < h$, then

$$\lim_{n\to\infty} G_n(y) = G(y)$$

for all continuity points of $G(y)$.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. If $p \to 0$ as $n \to \infty$ such that $np = \mu$ for some fixed $\mu > 0$, determine the limiting distribution of $Y_n = \sum_{i=1}^{n} X_i$.

if n

- If $p \to 0$ (as $n \to \infty$), then we approx. BIN w/ a POI, which we obtain w/ the above method. The POI has one parameter μ, which is equal to np.
- If p is fixed and n is sufficiently large, we approx a BIN w/ a Normal dist. via the CLT.
  - Just in the BIN case, we need check $np > 5$ and $nq > 5$.

O.W., skewed dist: to apply CLT, need $n \geq 30$ (roughly)
symm. need $n > 5$.

Only use for approximations

Thm: Central Limit Theorem (CLT)

If $X_1, X_2, \ldots, X_n$ is a random sample from a dist. with mean μ and variance $\sigma^2 < \infty$, then

i.e., X_i's are iid. Don't need know their dists, tho. Just need meet these conditions.

(both μ is finite and σ^2 is finite)

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \xrightarrow{d} Z \sim N(0,1).$$

this is why the (stan.) Normal dist. is so special.

- Let $Y_n = \sum_{i=1}^{n} X_i$.

cuz linearity property of exp. value. indep. unneeded.

$$E(Y_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

all have same μ

$$= \sum_{i=1}^{n} \mu$$

$$= n\mu.$$

cuz X_i's are indep (cuz r. samp.)

- This is the only time you really get to say this result.

Also,

$$Var(Y_n) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

$$= \sum_{i=1}^{n} \sigma^2$$

$$= n\sigma^2.$$

30AX

- Std. dev $(Y_n) = \sqrt{n}\,\sigma$.

how we get stand. normal dist.

So $$Z_n = \frac{Y_n - E(Y_n)}{\sqrt{Var(Y_n)}} \xrightarrow{d} Z \sim N(0,1).$$

I.e., $Y_n \xrightarrow{d} Y \sim N(n\mu, n\sigma^2)$

$X_1, X_2, \ldots, X_n$ r.s. $N(\mu, \sigma^2)$

if start w/ a normal dist. w/ any n, the sum will be normal. If don't start w/ normal dist., then limiting distribution

$\sum_{i=1}^{n} X_i = Y_n$, as $n \to \infty$, limiting dist becomes $\sim N(n\mu, n\sigma^2)$

(as $n \to \infty$, we approach a (roughly) normal dist w/ params $n\mu, n\sigma^2$)

General guidelines IF symmetric dist!

how large n needs to be for the CLT to give a good approximation depends on the shape of the starting distribution.

- If the dist of X_i's is symm., then $n \geq 5$ is adequate.
- If the dist of X_i's is skewed, then n must be at least 30. (eg, exponential dist)

sample mean

Question: What if the limiting dist. of $\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$?

- If $X_1, X_2, \dots, X_n$ is a random sample (r.s.) from a dist. with mean μ & variance $\sigma^2 < \infty$, then

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

note; $Y_n = \overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$, then

$$E(Y_n) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)$$

n ↑ just a constant

$$= \frac{1}{n} \sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{\cancel{n}} \cdot \cancel{n}\mu$$

$$= \mu$$

$$Var(Y_n) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i)$$ cuz X_i's are indep

$$= \frac{1}{n^2} \cdot n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

want stand. dev., so take $\sqrt{Var(Y_n)} = \sqrt{\frac{\sigma^2}{n}} = \sigma/\sqrt{n}$

So $Y_n = \overline{X}_n \xrightarrow{d} Y \sim N(\mu, \frac{\sigma^2}{n})$.

This is general formula.

Ex using (CLT) Let $X_1, X_2, \ldots, X_{20}$ be a rand. sample from UNIF(0,1).

(a) Find $P(\sum_{i=1}^{20} X_i \leq 12)$.

- Do we know the dist. of $\sum_{i=1}^{20} X_i$? (A result from Stat 480A?)

  -No!

  (Unif dists are symm.)

- But, since $n = 20 \geq 5$, we use the CLT.

$$\sum_{i=1}^{20} X_i \xrightarrow{d} N(n\mu, n\sigma^2) \text{ where } \mu = E(X_i) = \frac{1}{2},$$

$$\sigma^2 = Var(X_i) = \frac{1}{12}.$$

I.e., $20(\frac{1}{2}) = 10$, $20(\frac{1}{12})^2 = \frac{5}{3}$

$$\sum_{i=1}^{20} X_i \xrightarrow{d} N(10, \tfrac{5}{3}), \text{ hence}$$

cuz CLT gives an approximation

stand normal dist, do

$$P(\sum_{i=1}^{20} X_i \leq 12) \approx P(Z \leq \frac{12-10}{\sqrt{5/3}})$$

$= P(Z \leq 1.55) \to$ go to Table...,

$= 0.9394.$

I.e., have $\approx 93\%$ chance that if take sum of 20 Unif R.V.s, the sum will be ≤ 12.

If asked what is $P(\sum^{20} X_i \leq 25)$,

$= 1$ b/c ~~Unif~~ each $X_i \sim UNIF(0,1)$ means they can each only be as large as 1, so

STAT 480B

Problem Set 1

Due Date: Wednesday, January 19, 2022

Solve the following problems completely and neatly. You are expected to work **independently** of each other. Use the appropriate notation and if applicable, encircle your final answer. **No solution, no credit.**

1. Let X and Y be independent EXP(1) random variables. Determine the joint probability density function of $U = X + Y$ and $V = e^X$.

2. Let X and Y have joint pdf $f(x, y) = e^{-y}$, $0 < x < y < \infty$. Use the convolution method to derive the probability density function of $S = X + Y$.

3. Let $X_1, X_2, \ldots, X_k$ form a random sample from a standard normal distribution. Use the MGF technique to derive the distribution of $W = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} X_i$.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x) = \theta B^{-\theta} x^{\theta - 1}, \quad 0 \leq x \leq B.$$

(a) Derive the limiting distribution of $X_{1:n}$.

(b) Derive the limiting distribution of $X_{n:n}$.

c) Derive the limiting distribution of $\left(\frac{X_{n:n}}{B}\right)^n$.

5. **(Graduate Students Only:)** You are given a random sample of size n from a distribution with cdf $F(x) = \frac{1}{1+e^{-x}}$, $x \in \mathbb{R}$. Derive the limiting distribution of

$$\frac{e^{X_{n:n}}}{n}.$$

ask see page.

ask Ptwise vs uniform conv. of sequences of fns? Thms applicable? - irrelevant. Talking about a diff. context.

ask When is CLThm needed? Can it save time on any of these?

ask/check discrete case versions?

Matthew Miller - Problem Set 1, due Wedn. Jan 19th $\frac{69}{70}$ very good!

1) Let X and Y be indep. EXP(1) R.V.s. Determine the joint pdf of $U = X+Y$ and $V = e^X$.

- Given that $X, Y \overset{iid}{\sim} EXP(1)$, so $f_X(x) = \frac{1}{(1)} e^{-x/(1)} = e^{-x}, \ x > 0$;
  similarly, $f_Y(y) = \frac{1}{(1)} e^{-y/(1)} = e^{-y}, \ y > 0$.
  Since X, Y indep., $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) = e^{-x} \cdot e^{-y} = e^{-(x+y)}, \ x > 0, \ y > 0.$

needed cuz want joint pdf

- Applying the ~~convolution formula~~ transformation...,

Need check new bounds at this point $\Rightarrow \ln v > 0 \Rightarrow v > 1$

$$\begin{matrix} U = X+Y \\ V = e^X \end{matrix} \xrightarrow[\text{of eqns for } X, Y]{\text{solve system}} \begin{matrix} V = e^X \Rightarrow X = \ln(V) = w_2(u,v), \\ U = X+Y \Rightarrow Y = U - X = U - \ln(V). \end{matrix}$$

I.e., $Y = U - \ln(V) = w_1(u,v)$. $\Rightarrow u - \ln v > 0$

$$J = \begin{vmatrix} \frac{dw_1}{du} & \frac{dw_1}{dv} \\ \frac{dw_2}{du} & \frac{dw_2}{dv} \end{vmatrix} = \begin{vmatrix} 1 & -\frac{1}{v} \\ 0 & \frac{1}{v} \end{vmatrix} = (1)\left(\tfrac{1}{v}\right) - \cancel{\left(-\tfrac{1}{v}\right)(0)}^{0} = \frac{1}{v}.$$

Thus, $V = e^X \Rightarrow V > 0$, so $\left|\frac{1}{v}\right| = \frac{1}{v}$.

$$f_{U,V}(u,v) = \cancel{f_{X,Y}(\ln v, y) \cdot |J|}$$
$$= f_{X,Y}(\ln(v), u - \ln(v)) \cdot \left|\tfrac{1}{v}\right|$$
$$= e^{-[\cancel{\ln(v)} + u - \cancel{\ln(v)}]} \cdot \frac{1}{v}$$
$$= e^{-u} \cdot \frac{1}{v},$$ where $x > 0 \Rightarrow$ ~~$V = e^x > e^0 = 1$~~
and $x, y > 0 \Rightarrow$ ~~$U = x+y > 0+0 = 0$.~~

9/10

$$\boxed{\text{Hence, } f_{U,V}(u,v) = \frac{1}{v} e^{-u}, \text{ where } u > 0, \ e^u > v > 1.}$$

draw support

$u - \ln v > 0$
$u > \ln v > 0$
$e^u > v \Rightarrow 1 < v < e^u$
$\ln v > 0$
$v > 1 \Rightarrow u > 0$

How to check if your bds are correct:
do $f_{U,V}(u,v) = \iint \text{joint pdf } d\, d \overset{?}{=} 1$

2) Let X and Y have joint pdf $f(x,y) = e^{-y}$, $\underbrace{0 < x < y < \infty}_{\Rightarrow x \neq y}$. Use the convolution method to derive the ~~joint~~ pdf of $S = X + Y$.

Not needed

are want marginal [dist] [of a sum]

- $S = X + Y$ is not one-to-one; so, let $T = X$ be our dummy variable. Then,

$$S = X + Y,\quad T = X \xrightarrow[S,T]{\text{solve for}} T = X \Rightarrow X = T = w_2(s,t),$$

$$S = X + Y \Rightarrow Y = S - X = S - T, \text{ so } Y = S - T = w_1(s,t).$$

$$J = \begin{vmatrix} \frac{\partial w_1}{\partial s} & \frac{\partial w_1}{\partial t} \\ \frac{\partial w_2}{\partial s} & \frac{\partial w_2}{\partial t} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = (1)(1) - \cancel{(-1)(0)}^{0} = 1.$$

Thus, the joint pdf of S and T is

$$f_{S,T}(s,t) = f_{X,Y}(x,y) \cdot |J|$$

$$= f_{X,Y}(t, s-t) \cdot |(1)|$$

$$= e^{-(s-t)};\ \text{where } 0 < x < y < \infty \Rightarrow 0 < t < s-t < \infty$$

$0 < t < s-t < \infty$

$-s < t-s < -t < \infty$

Original spt/bds in this form

$\Rightarrow 0 < t < 2t < s < \infty$

$\Rightarrow 0 < 2t < s < \infty.$

So $f_{S,T}(s,t) = e^{-(s-t)}$, $0 < 2t < s < \infty$

I.e., $0 < t < \frac{s}{2} < \infty$

~~Thus,~~ Convolution method (p. 209)

10/10

$$f_S(s) = \int_0^{\frac{1}{2}s} f_{S,T}(s,t)\,dt$$

$$= \int_0^{\frac{1}{2}s} e^{-(s-t)}\,dt = \int_0^{\frac{1}{2}s} e^{-s} e^{t}\,dt = e^{-s}\int_0^{\frac{1}{2}s} e^{t}\,dt$$

$$= e^{-s}\left[e^{t}\Big|_{t=0}^{t=\frac{1}{2}s}\right] = e^{-s}\left(e^{\frac{1}{2}s} - e^{0}\right) = e^{-s}\left(e^{\frac{1}{2}s} - 1\right)$$

$$= \boxed{e^{-\frac{1}{2}s} - e^{-s},\quad 0 < s < \infty}$$

Convolution is used to find a dist. of sums. Gives u a direct path to find a sum:

$$\int_{-\infty}^{\infty} f(\overset{x}{t}, \overset{y}{s-t})\,dt$$

she gave us the transformation already, so we could go straight to conv. method.

$$\frac{e^{\frac{1}{2}s} - 1}{e^{s}} = \frac{e^{\frac{1}{2}s}}{e^{s}} - \frac{1}{e^{s}}$$

$$= e^{-\frac{1}{2}s} - e^{-s}$$

$$= \frac{1}{e^{\frac{1}{2}s}} - \frac{1}{e^{s}}$$

3) Let $X_1, X_2, \ldots, X_k$ form a r. sample from a standard normal dist.
Use the MGF technique to derive the dist. of $W = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} X_i$.

- Let $L = \sum_{i=1}^{k} X_i = X_1 + X_2 + \ldots + X_k$ such that $W = \frac{1}{\sqrt{k}} L$.
- Since L is a sum of iid R.V.s, we can apply the MGF technique to find the mgf $m_L(t)$ of L. Hence,

10/10

$$m_L(t) = m_{X_1}(t) \cdot m_{X_2}(t) \cdot \ldots \cdot m_{X_k}(t) \quad \left[\text{where } m_{X_i}(t) = e^{(0)t + \frac{t^2(1)}{2}} = e^{t^2/2}\right]$$
$$= (e^{t^2/2}) \cdot (e^{t^2/2}) \ldots (e^{t^2/2})$$
$$= \left[e^{t^2/2}\right]^k$$
$$= e^{\frac{t^2 k}{2}},$$ which implies that $L \sim N(0, k)$.

This is not the efficient way!

- Now, since $W = \frac{1}{\sqrt{k}} L$ ~~and L is a cont. R.V.~~, we can apply the transformation method to obtain the pdf of W. Here, $W = \frac{1}{\sqrt{k}} L$ is one-to-one and the pdf of L is $f_L(l) = \frac{1}{(\sqrt{k})\sqrt{2\pi}} e^{\left[\frac{-(l-0)^2}{2(k)}\right]} = \frac{1}{\sqrt{k}\sqrt{2\pi}} e^{-l^2/2k}$ where $-\infty \le l \le \infty$. Thus, $W = \frac{1}{\sqrt{k}} L \Rightarrow L = \sqrt{k} W$, and

$$f_W(w) = f_L(\sqrt{k}\, w) \cdot \left| \frac{d}{dw}(\sqrt{k}\, w) \right|$$ ← [this term needed b/c L is a cont. R.V.]

$$= \frac{1}{\cancel{\sqrt{k}}\sqrt{2\pi}} e^{\frac{-(\sqrt{k}\,w)^2}{2k}} \cdot |\cancel{\sqrt{k}}|$$
$$= \frac{1}{\sqrt{2\pi}} e^{\frac{-k w^2}{2k}}$$
$$= \frac{1}{\sqrt{2\pi}} e^{\frac{-w^2}{2}}, \quad -\infty \le w \le \infty$$

$N(\mu, \sigma^2) \rightarrow f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[\frac{-(y-\mu)^2}{2\sigma^2}\right]}, \ y \in \mathbb{R},$

$N(0, k) \rightarrow f_L(l) = \frac{1}{\sqrt{k}\sqrt{2\pi}} e^{\left[\frac{-(l-0)^2}{2k}\right]}, \ l \in \mathbb{R},$

$N(0,1) \rightarrow f_W(w) = \frac{1}{(1)\sqrt{2\pi}} e^{\left[\frac{-(w-0)^2}{2(1)}\right]}, \ w \in \mathbb{R}.$

This, finally, shows that $W \sim N(0,1)$.

I.e., W adheres to the standard normal distribution.

use property of mgf for linear combination, i.e., if $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$ or
$M_W(t) = E[e^{tW}] = E[e^{\frac{tL}{\sqrt{k}}}] = M_L\left(\frac{t}{\sqrt{k}}\right) = e^{\frac{k}{2}\left(\frac{t}{\sqrt{k}}\right)^2} = e^{\frac{t^2}{2}}$ which is the mgf of $N(0,1)$.

[Want to show $Y_n \sim G_n(y) \xrightarrow{d} Y \sim G(y)$]

4) Let $X_1, \dots, X_n$ be a random sample from a distribution w/ pdf $f(x) = \theta B^{-\theta} x^{\theta-1}$, where $0 \le x \le B$.

(a) Derive the limiting distribution of $X_{1:n}$.

- First, we need the CDF of $Y_n = X_{1:n}$, the CDF of which we denote by $G_n(y)$. Using the given pdf, we can integrate to get the CDF of each X_i in the random sample:

$$F_{X_i}(x) = \int_0^x f(t)\,dt = \int_0^x \theta B^{-\theta} t^{\theta-1}\,dt = \frac{\theta}{B^\theta}\int_0^x t^{\theta-1}\,dt$$

$$= \frac{\theta}{B^\theta}\left[\frac{t^{\theta-\cancel{1}+\cancel{1}}}{\theta-\cancel{1}+\cancel{1}}\Big|_{t=0}^{t=x}\right] = \frac{\cancel{\theta}}{B^\theta}\left[\frac{t^\theta}{\cancel{\theta}}\Big|_{t=0}^{t=x}\right]$$

$$= \frac{x^\theta - \cancel{0^\theta}}{B^\theta} = \left(\frac{x}{B}\right)^\theta, \quad 0 \le x \le B.$$

CDF of each of the X_i's.

- Thus, we have that

$$G_n(y) = P(Y_n \le y) = P(X_{1:n} \le y) = 1 - P(X_{1:n} > y)$$

$$= 1 - P(\text{all } X_i\text{'s} > y) = 1 - P(X_1 > y, X_2 > y, \dots, X_n > y)$$

cuz indep. $= 1 - [P(X_1 > y)\cdot P(X_2 > y)\cdots P(X_n > y)]$

cuz identical $= 1 - [P(X_i > y)]^n = 1 - [1 - P(X_i \le y)]^n$

$$= 1 - [1 - F_{X_i}(y)]^n = 1 - [1 - (\tfrac{y}{B})^\theta]^n, \quad 0 \le y \le B.$$

For each X_i, $0 \le x \le B$. $\therefore$, so is the smallest X_i, or $X_{1:n}$. So, $0 \le X_{1:n} \le B$ $\Rightarrow 0 \le Y_n \le B$, so the support of all the CDFs $G_i(y)$ is...

10/10

If $y = 0$, then we have

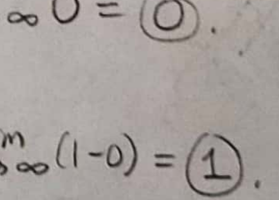
$$\lim_{n\to\infty}(1 - [1 - (\cancel{0})^{\cancel{\theta}}]^n) = \lim_{n\to\infty}(1 - (1)^n) = \lim_{n\to\infty}(1-1) = \lim_{n\to\infty} 0 = 0.$$

- If $y = B$, then we have

$$\lim_{n\to\infty}(1 - [1 - (\tfrac{B}{B})^\theta]^n) = \lim_{n\to\infty}(1 - [1-1]^n) = \lim_{n\to\infty}(1 - 0^n) = \lim_{n\to\infty}(1-0) = 1.$$

- If $0 < y < B$, then $0 < \frac{y}{B} < 1$, so $0 < (\frac{y}{B})^\theta < 1$, which means $\lim_{n\to\infty}(1 - \underbrace{[1 - (\tfrac{y}{B})^\theta]^n}_{\text{goes to } 0}) = 1$. Thus, $\lim_{n\to\infty} G_n(y) = G(y) = \begin{cases} 0 & \text{if } y \le 0 \\ 1 & \text{if } y > 0 \end{cases}$

$$\boxed{\text{So, } Y_n = X_{1:n} \xrightarrow{d} Y \sim G(y) = \begin{cases} 0 & \text{if } y \le 0 \\ 1 & \text{if } y > 0 \end{cases}}$$



AKA Y has a degenerate dist at $y = 0$.

(ADX) CDF tech. plug $\left(\frac{X_{n:n}}{B}\right)^n$ into $0 \le x \le B$ to get support of Y_n.

(b) Derive the limiting distribution of $X_{n:n}$.

- From part (a), already know that $F_{X_i}(x) = \left(\frac{x}{B}\right)^\theta$, $0 \le x \le B$.

Thus the CDF of $Y_n = X_{n:n}$ can be found to be

$$G_n(y) = P(Y_n \le y) = P(X_{n:n} \le y) = P(\text{all } X_i\text{'s} \le y)$$

$$= P(X_1 \le y, X_2 \le y, \ldots, X_n \le y)$$

cuz indep. $\rightarrow = P(X_1 \le y) \cdot P(X_2 \le y) \cdots P(X_n \le y)$

cuz identical $\rightarrow = [P(X_i \le y)]^n = [F_{X_i}(y)]^n = \left[\left(\frac{y}{B}\right)^\theta\right]^n, \ 0 \le y \le B$

$\Rightarrow 0 \le X_{n:n} \le B$
$\Rightarrow 0 \le Y_n \le B$

8/8

- If $y = 0$, $G_n(0) = \left[\left(\frac{0}{B}\right)^\theta\right]^n = [0]^n = 0$, so $\lim_{n\to\infty} G_n(y) = 0$ for $y = 0$.
- If $y = B$, $G_n(B) = \left[\left(\frac{B}{B}\right)^\theta\right]^n = [(1)^\theta]^n = [1]^n = 1$, so $\lim_{n\to\infty} G_n(y) = 1$ for $y = B$.
- If $0 < y < B$, then $0 < \frac{y}{B} < 1 \Rightarrow 0 < \left(\frac{y}{B}\right)^\theta < 1$,

so $\lim_{n\to\infty} G_n(y) = \lim_{n\to\infty} \left[\left(\frac{y}{B}\right)^\theta\right]^n = 0$ for $0 < y < B$.

Thus $\lim_{n\to\infty} G_n(y) = G(y) = \begin{cases} 0 & \text{if } y < B \\ 1 & \text{if } y \ge B \end{cases}$

so $\boxed{Y_n = X_{n:n} \xrightarrow{d} Y \sim G(y) = \begin{cases} 0 & \text{if } y < B \\ 1 & \text{if } y \ge B \end{cases}}$

could say either as final answer.

AKA ~~Y has a degenerate dist at y = B~~

ask: Can the result from (b) be used in some way? NO, lol. cuz need CDF tech.

(c) Derive the limiting distribution of $\left(\frac{X_{n:n}}{B}\right)^n$.

all X_i's between 0 and B, so the largest X_i, $X_{n:n}$, must also be. So have $0 \le x \le B \Rightarrow 0 \le X_{n:n} \le B$
$\Rightarrow 0 \le \frac{X_{n:n}}{B} \le 1$
$\Rightarrow 0 \le \left(\frac{X_{n:n}}{B}\right)^n \le 1$
$\Rightarrow 0 \le Y_n \le 1$, hence

12/12

$Y_n = \left(\frac{X_{n:n}}{B}\right)^n$, $F_{X_i}(x) = \left(\frac{x}{B}\right)^\theta$, $0 \le x \le B$.

$$\Rightarrow G_n(y) = P(Y_n \le y) = P\left(\left(\frac{X_{n:n}}{B}\right)^n \le y\right) = P\left(\frac{X_{n:n}}{B} \le \sqrt[n]{y}\right)$$

$$= P(X_{n:n} \le B\sqrt[n]{y}) = P(\text{all } X_i\text{'s} \le B\sqrt[n]{y}) = [P(X_i \le B\sqrt[n]{y})]^n$$

$$= [F_{X_i}(B\sqrt[n]{y})]^n = \left[\left(\frac{B\sqrt[n]{y}}{B}\right)^\theta\right]^n = [(\sqrt[n]{y})^\theta]^n = [y^{\theta/n}]^n = y^\theta, \ 0 \le y \le 1.$$

Thus $G(y) = \lim_{n\to\infty} G_n(y) = \lim_{n\to\infty} y^\theta = y^\theta$, $0 \le y \le 1$

- Check if valid CDF: $\lim_{y\to-\infty} y^\theta = 0$ ✓, $\lim_{y\to\infty} y^\theta = 1$ ✓

(the smallest y^θ can be is when $y = 0$: $(0)^\theta = 0$.)
(the largest y^θ can be is when $y = 1$: $(1)^\theta = 1$.)

$$\boxed{\therefore Y_n = \left(\frac{X_{n:n}}{B}\right)^n \xrightarrow{d} Y \sim G(y) = y^\theta, \ 0 \le y \le 1}$$

5) Are given a random sample of size n from a distribution with CDF $F(x) = \frac{1}{1+e^{-x}}$, $x \in \mathbb{R}$. Derive the limiting distribution of $\frac{e^{X_{n:n}}}{n}$.

- Let $X_1, X_2, \ldots, X_n$ represent the above-mentioned random sample; then we are given that $F_{X_i}(x) = \frac{1}{1+e^{-x}}$, $x \in \mathbb{R}$.

$e^{X_{n:n}} > 0$, $n \geq 1$, $\frac{e^{X_{n:n}}}{n} > \frac{0}{1} = 0$ $\Rightarrow y > 0$

- $Y_n = \frac{e^{X_{n:n}}}{n}$, so

$$G_n(y) = P(Y_n \leq y) = P\left(\frac{e^{X_{n:n}}}{n} \leq y\right) = P(e^{X_{n:n}} \leq ny)$$
$$= P(X_{n:n} \leq \ln(ny)) = P(\text{all } X_i\text{'s} \leq \ln(ny))$$
$$= [P(X_i \leq \ln(ny))]^n = [F_{X_i}(\ln(ny))]^n$$
$$= \left[\frac{1}{1+e^{-(\ln(ny))}}\right]^n = \left[\frac{1}{1+e^{\ln(ny)^{-1}}}\right]^n = \left[\frac{1}{1+(ny)^{-1}}\right]^n$$
$$= \left[\frac{1}{1+\frac{1}{ny}}\right]^n = \frac{1}{\left(1+\frac{1}{ny}\right)^n} = \left(1+\frac{1}{ny}\right)^{-n} = \left(1+\frac{1/y}{n}\right)^{n(-1)}$$

$x \in \mathbb{R} \Rightarrow X_{n:n} \in \mathbb{R}$, but $e^{X_{n:n}} \in (0,\infty)$, $\Rightarrow \frac{e^{X_{n:n}}}{n} \in (0,\infty)$ (n = sample size), $n \in \mathbb{N}$ $\Rightarrow Y_n \in (0,\infty)$. I.e., $Y_n > 0$. Hence support for $G_n(y)$ is $y > 0$.

where $y > 0$.

Hence, $G_n(y) = \left(1+\frac{1/y}{n}\right)^{n(-1)}$, $y > 0$

10/10

- Now,

$$\lim_{n\to\infty} G_n(y) = \lim_{n\to\infty}\left(1+\frac{1/y}{n}\right)^{n(-1)}$$

$= \lim_{n\to\infty}\left(1+\frac{c}{n}\right)^{nb}$, $c = \frac{1}{y}$, $b=-1$, $= e^{cb}$

$$= e^{\left(\frac{1}{y}\right)(-1)}$$
$$= e^{-1/y}, \quad y > 0.$$

So $G(y) = e^{-1/y}$, $y > 0$.

- Check if valid CDF:

$\lim_{y\to\infty} G(y) = \lim_{y\to\infty} e^{-1/y} = 1$ ✓

$\lim_{y\to-\infty} G(y) = \lim_{y\to-\infty} e^{-1/y} = 0$ ✓

(Also, $G(y) > 0$ $\forall y > 0$ since e raised to any power will always be positive-valued)

- Is a valid CDF; thus,

$$\boxed{Y_n = \frac{e^{X_{n:n}}}{n} \xrightarrow{d} Y \sim G(y) = e^{-1/y}, \ y > 0.}$$

Start of lecture, 1/19/22

- $Y_n \xrightarrow{d} Y \sim G(y)$

$\Rightarrow \lim_{n\to\infty} G_n(y) = G(y)$.

- If $G(y)$ is degen. at $y = c$, then we say Y_n converges stochastically to c. That is, $Y_n \xrightarrow{p} c$, c some constant.

implies all its mass is at c

("$\Rightarrow$ (p)robability")

- To use CLT, need finite mean and variance first.

technically can apply w/o knowing what the dist is, so long as know μ and σ^2.

<u>CLT</u>: $X_1, \ldots, X_n$ iid from a random sample ~~??~~ from a dist. with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$.

Then

we derived these results previously

(sum) $Y_n = \sum_{i=1}^{n} X_i \xrightarrow{d} Y \sim N(n\mu, n\sigma^2)$,

$Y_n = \bar{X}$ (sample mean)

$= \dfrac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{d} Y \sim N(\mu, \frac{\sigma^2}{n})$.

<u>Example</u>: $X_1, X_2, \ldots, X_{20}$ iid $UNIF(0,1)$.

symmetric, so already approach the needed sample size to apply CLT

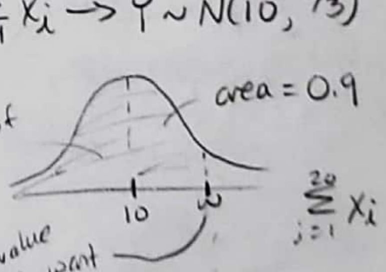
(b) Find the 90^{th} percentile of $\sum_{i=1}^{20} X_i$.

- Recall: $\mu = E(X) = \frac{1}{2}$, $\sigma^2 = Var(X) = \frac{1}{12}$.
- Find: $P(\underbrace{\sum_{i=1}^{20} X_i}_{\text{the R.V.}} \le w) = 0.90$

$\sum_{i=1}^{20} X_i \xrightarrow{d} Y \sim N(10, 5/3)$

Normal dist., so <u>we'll need to standardize</u>.

$= P\left(Z \le \frac{w-10}{\sqrt{5/3}}\right) = 0.90$.



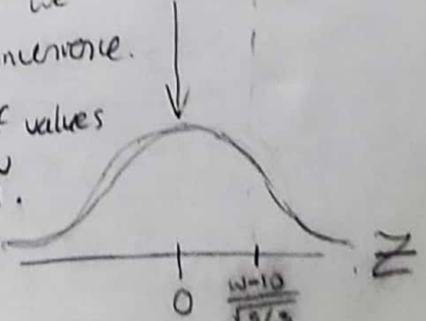
Now need table. ~~Look~~ Find # closest to that which gives you 0.90. Or use percentile for convenience.

($z_\gamma = 1.282$ for $\gamma = 0.90$.

$\Rightarrow 1.282 = \frac{w-10}{\sqrt{5/3}} \Rightarrow \boxed{w = 11.6551}$

so 90% of values are below 11.6551.

(or using table body, $w = 11.6525$ for $z = 1.28$)



(c) Find prob. that the sample mean is at most 0.35.

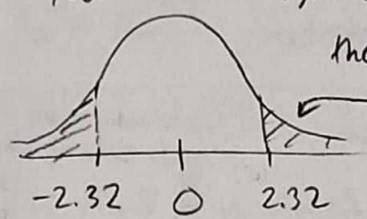
- Find $P(\bar{X} \le 0.35)$. need use μ and $\frac{\sigma^2}{n}$ for $\bar{X}$, not $n\mu$, $n\sigma^2$

  - figured out already that $\bar{X}$ is normal, so need to standardize,

$\frac{\sigma^2}{n} = \frac{\frac{1}{12}}{20} = \frac{1}{240}$

$$P(\bar{X} \le 0.35) = P\left(Z \le \frac{0.35 - 0.50}{\sqrt{1/240}}\right) = P(Z \le -2.32) \leftarrow$$ not a typo. Due to symmetry!

Normal dist is symm., so use that fact!



$$= 1 - P(Z \le 2.32)$$

$$= 1 - 0.9898$$

$$= \boxed{0.102}$$

(so roughly 1% of the time will our sample mean be $\le$ 0.35.)

(d) $P(\bar{X} > 0.5) = 0.5$

Why? Logically, $\bar{X}$ is normal and its symm. ~~mean is 0.5~~, and said mean will therefore be in the middle since Normal dist is symm.

mean is 0.5, so $P(\bar{X} > 0.5) = 0.5$ (half the values are before μ, half after μ).

$$Y_n = \sum_{i=1}^{n} X_i \xrightarrow{d} Y \sim N(n\mu, n\sigma^2) \iff \sum_{i=1}^{n} X_i \sim AN(n\mu, n\sigma^2)$$

asymptotically normal

If you read something is "asymptotically normal," it just means it converges in dist. to a normal dist. (shorthand, basically)

· Our binomial table only goes up to $n = 20$, ...

§7.4: Approximations to the Binomial Distribution

Consider the case where p is fixed and $n \to \infty$.

Let $Y_n \sim BIN(n, p)$.

Here, Y_n = # of "successes" out of n trials

(a sum of X_i's) (a $\frac{\text{sum of } X_i\text{'s}}{n}$)

· In order to use CLT, or Y_n has to be a sum or a mean

(write as sum) $= \sum_{i=1}^{n} X_i$ where $X_i \overset{iid}{\sim} BER(p)$.

Recall: $\mu = E(X_i) = p$, (cuz Bernoulli)

$\sigma^2 = Var(X_i) = pq = p(1-p)$

cuz ~~so~~ is either 0 or 1 — p finite, so pq is finite.

$\therefore$ CLT can be applied for large n.

· Well, we care about large n. so,

By CLT, (by formula for CLT)

$Y_n \xrightarrow{d} Y \sim N(np, npq)$.

(cuz CLT always → Normal, which is contin.)

· Since $BIN(n,p)$ is discrete, we need a continuity correction factor. (in order to standardize)

Let $a, b \in \mathbb{Z}$. Then, (cuz as asymptotically normal)

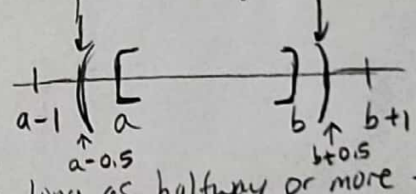
book uses "$\doteq$" to mean "$\approx$" (approximately)

$P(a \le Y_n \le b) \approx$ (equality matters)

(equality no ~~[illegible]~~ longer matters)

$P\left(\frac{a - 0.5 - np}{\sqrt{npq}} \le Z \le \frac{b + 0.5 - np}{\sqrt{npq}}\right)$.

continuity corr. factor

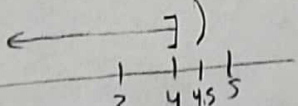


(so long as halfway or more to a, then gets rounded up to a. if less than halfway to b+1, round down to b)

Ex) $P(Y_n < 5)$

$= P(Y_n \le 4)$

$\approx P\left(Z \le \frac{4.5 - np}{\sqrt{npq}}\right)$.



anything less than 4.5 will be rounded to 4 as well

~~ask~~ why?

How big should n be to apply CLT? Here, $np > 5$ AND $nq > 5$. ✓

can't just check if $n > 5$ cuz BIN isn't a symmetric dist.

Guideline for a BIN. R.V.
Check that both np and $nq > 5$.

STAT 480B

Example for Section 7.4

1. Paulo forgot to study and decides to answer a 20-question True or False test in a completely random fashion. Determine the probability that he will get at least 70%.
   (guess) – answers independent from one another
2. Suppose the test consisted of 50 True or False problems. Determine the probability that Paulo will get at least 70%.

1) Y_{20} = # of correct answers out of n = 20.

$20 = n$

only two choices (True/False)

$Y_{20} \sim BIN(\overset{n=}{20},\ p = 0.50)$

70% of 20

Want $P(Y_{20} \geq 14)$

cuz discrete!

$= 1 - P(Y_{20} < 14) = 1 - P(Y_{20} \leq 13)$ use BIN table

$= 1 - (0.9423)$

$= \boxed{0.0577}$ exact probability cuz can use table instead of approxing w/ CLT.

Must check these things specifically, not just n, for a BIN dist, because BIN is not symmetric.

· Approxing w/ CLT:

$np = 20(0.5) = 10 > 5$ ✓

$nq = 20(0.5) = 10 > 5$ ✓

So yes, can use.

13

$P(Y_{20} \geq 14) = 1 - P(Y_{20} \leq 13) \approx 1 - P\left(Z \leq \frac{13.5 - 10}{\sqrt{20(0.5)(0.5)}}\right)$

$= 1 - P(Z \leq 1.57)$ ← round to 2 decimals

$= 1 - 0.9418 = \boxed{0.0582}$

(not as good of an answer)

2) Now, since table only goes up to n=20 and we have n=50 here, we must use CLT (which definitely applies).

70% of 50 is 35 → · Want $P(Y_{50} \geq 35) = 1 - P(Y_{50} \leq 34)$

$np = \mu = (50)(0.5) = 25$

$\approx 1 - P\left(Z \leq \frac{34.5 - 25}{\sqrt{50(0.5)(0.5)}}\right)$

$= 1 - P(Z \leq 2.69)$

$= 1 - 0.9964$

$= \boxed{0.0036}$

STAT 480B
Problem Set 2
Due Date: Wednesday, January 26, 2022

Name:

Solve the following problems completely and neatly. You are expected to work **independently** of each other. If the problem involves random variables, make sure that you define the random variables. If you are using approximations, justify why your approximation is valid. Use the appropriate notation and if applicable, encircle your final answer. **No solution, no credit.**

1. Batteries are known to deteriorate in storage. The shelf life of a battery is the length of time for which it remains usable. The shelf life, in months, of batteries produced by Supplier A can be modeled using a Weibull distribution with shape parameter 2 and scale parameter 10. A random sample of 35 batteries from Supplier A was obtained. Determine the probability that the mean shelf life of the sampled batteries is at least one year.

2. The thickness, in inches, of a metal sheet manufactured by Latem Company can be modeled by the cdf

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 1 \\ 2\left(x - 2 + \frac{1}{x}\right) & \text{if} \quad 1 \leq x \leq 2 \\ 1 & \text{if} \quad x > 2 \end{cases}$$

   A random sample of 50 metal sheets was taken from a production line.

   (a) Determine the probability that the total thickness of the metal sheets in the sample is between 78 inches and 85 inches.

   (b) Determine the 80^{th} percentile of the sample mean thickness of the metal sheets.

3. According to a report of the Nielsen Company, 76% of internet searches use the Google search engine. A random sample of 27 internet searches was taken. Determine the probability that more than 18 use the Google search engine.

4. **(Graduate Students Only:)** Suppose $X_1, X_2, \ldots, X_5$ form a random sample from a Poisson distribution with parameter θ.

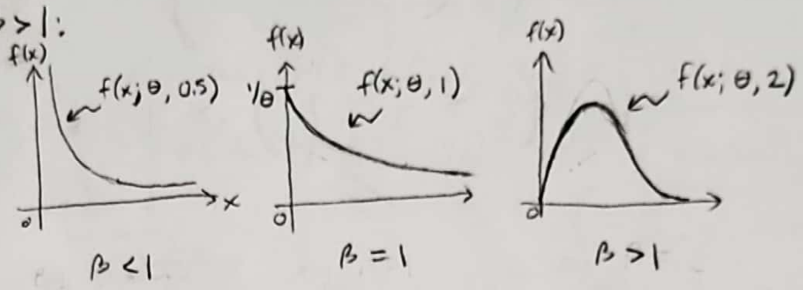
   (a) Find the exact distribution of $W = \sum_{i=1}^{5} X_i$.

   (b) If $\theta = 0.4$, determine the probability that the sample mean is between 0.6 and 1.2, inclusive.

Matthew Miller - Problem Set 2 - due Wedn., Jan 26^{th} @ 6:00pm

$\frac{50}{50}$ excellent! :)

[See page 116 for Weibull Dist. definition.] [β is shape param. For a Weibull dist., the pdfs take on a distinct shape depending on whether $\beta<1$, $\beta=1$, or $\beta>1$:



]

$\frac{10}{10}$

1) Let $X_1, X_2, \dots, X_{35} \overset{iid}{\sim} WEI(\theta=10, \beta=2)$, where each X_i represents the shelf life of a battery measured in months. Here the sample size of batteries is $n=35$.

← Continuous R.V., so if CLT used, then no continuity correction necessary.

- Let $\frac{X_1+X_2+\dots+X_n}{n} = \frac{\sum_{i=1}^{35} X_i}{35} = \overline{X}$ denote the average of the shelf lives of the 35 randomly selected batteries. Since each $X_i \sim WEI(10,2)$, we have a skewed distribution; thus, to apply the CLT we need $n>30$. Well, $n=35>30$, so the CLT applies.
- Then by the CLT, letting $Y_n=\overline{X}$, $Y_n = \overline{X} \xrightarrow{d} Y \sim N(\mu_{X_i}, \frac{\sigma^2_{X_i}}{n})$. $(\mu_Y)\ (\sigma_Y^2)$

Here, $\mu_Y = \mu_{X_i} = \theta\Gamma(1+\frac{1}{\beta})$

$$= (10)\Gamma(1+\tfrac{1}{2}) = 10\Gamma(\tfrac{3}{2}) = 10[(\tfrac{3}{2}-1)\Gamma(\tfrac{3}{2}-1)]$$

$$= 10[(\tfrac{1}{2})\cdot\Gamma(\tfrac{1}{2})] = 5\sqrt{\pi}, \text{ and}$$

$$\sigma_Y^2 = \frac{\sigma_{X_i}^2}{n} = \frac{1}{n}\left[\theta^2\left[\Gamma(1+\tfrac{2}{\beta}) - \Gamma^2(1+\tfrac{1}{\beta})\right]\right]$$

$$= \frac{1}{35}\left[(10)^2\left[\Gamma(1+\tfrac{2}{(2)}) - \Gamma^2(1+\tfrac{1}{(2)})\right]\right]$$

$$= \frac{100}{35}\left[\Gamma(2) - [\Gamma(\tfrac{3}{2})]^2\right] = \frac{20}{7}\left[(2-1)! - (\tfrac{1}{2}\sqrt{\pi})^2\right]$$

$$= \frac{20}{7}[1-\tfrac{\pi}{4}] = \frac{20}{7} - \frac{5\pi}{7}$$

- So by the CLT, $Y_{35} = \overline{X} \xrightarrow{d} Y \sim N(5\sqrt{\pi}, \frac{20-5\pi}{7})$, and we need to find $P(\overline{X} \geq 12 \text{ months})$.

cont'd →

note: $\bar{X}$ is continuous, so no continuity correction needed when apply CLT in a moment.

- Well, $P(\bar{X} \geq 12) = 1 - P(\bar{X} \leq 12)$

standardize via CLT

$$\approx 1 - P\left(Z \leq \frac{12 - \mu_Y}{\sqrt{\sigma_Y^2}}\right)$$

$$= 1 - P\left(Z \leq \frac{12 - (5\sqrt{\pi})}{\sqrt{\frac{20 - 5\pi}{7}}}\right)$$

$$= 1 - P(Z \leq 4.007\ldots) \approx 1 - P(Z \leq 4.01)$$

$$= 1 - (1)$$

$$= \boxed{0.}$$

2) Given that $X_1, X_2, \ldots, X_{50}$ are a random sample of metal sheets, where each X_i denotes the thickness in inches of the ith sheet and said thickness can be modeled by the CDF $F_{X_i}(x) = \begin{cases} 0 & \text{if } x < 1 \\ 2(x - 2 + \frac{1}{x}) & \text{if } 1 \leq x \leq 2 \\ 1 & \text{if } x > 2. \end{cases}$

(a) Need to find $P(78 \text{ inches} \leq \underbrace{\sum_{i=1}^{50} X_i}_{\text{the total thickness of all 50 metal sheets}} \leq 85 \text{ inches})$. Can apply CLT because $n = 50 > 30$ ✓.

- Let $Y_n = \sum_{i=1}^{n} X_i$; then by the CLT, $Y_n = \sum_{i=1}^{n} X_i \xrightarrow{d} Y \sim N(\underbrace{n\mu_{X_i}}_{\mu_Y}, \underbrace{n\sigma_{X_i}^2}_{\sigma_Y^2})$.
- Need to derive the pdf $f_{X_i}(x)$ using the given CDF in order to derive the values of $\mu_{X_i} = E(X_i)$ and $\sigma_{X_i}^2 = Var(X_i) = E(X_i^2) - [E(X_i)]^2$.
  - For $x < 1$, $f_{X_i}(x) = \frac{d}{dx}(0) = 0$.
  - For $1 \leq x \leq 2$, $f_{X_i}(x) = \frac{d}{dx}\left[2x - 4 + \frac{2}{x}\right] = 2 - \frac{2}{x^2}$.
  - For $x > 2$, $f_{X_i}(x) = \frac{d}{dx}(1) = 0$.

  So $f_{X_i}(x) = 2(1 - \frac{1}{x^2})$ if $1 \leq x \leq 2$, and 0 o.w.

- Thus $\mu_{X_i} = E(X_i) = \int_{-\infty}^{\infty} x f_{X_i}(x)\,dx = \int_1^2 x \cdot 2(1 - \frac{1}{x^2})\,dx = \int_1^2 (2x - \frac{2x}{x^2})\,dx$

$$= 2\int_1^2 (x - \frac{1}{x})\,dx = 2\left[\frac{1}{2}x^2 - \ln|x|\right]_1^2$$

$$= 2\left[\frac{1}{2}(2)^2 - \ln|2| - \frac{1}{2}(1)^2 + \cancel{\ln|1|}^{0}\right] = 2\left[2 - \ln(2) - \frac{1}{2}\right]$$

$$= 4 - 2\ln(2) - 1 = 3 - \ln(4) \text{ inches } (\approx 1.61 \text{ inches}).$$

cont'd →

15/15

- Also, $E(X_i^2) = \int_{-\infty}^{\infty} x^2 f_{X_i}(x)\,dx = \int_1^2 x^2 (2)(1-\frac{1}{x^2})\,dx = 2\int_1^2 (x^2-1)\,dx$

$= 2\left[\frac{1}{3}x^3 - x\right]_1^2 = 2\left[\frac{1}{3}(2)^3 - (2) - \frac{1}{3}(1)^3 + (1)\right]$

$= 2\left[\frac{8}{3} - 2 - \frac{1}{3} + 1\right] = 2\left(\frac{7}{3} - 1\right) = \frac{14}{3} - 2 = \frac{14}{3} - \frac{6}{3} = \frac{8}{3}$.

- Thus, $\sigma_{X_i}^2 = Var(X_i) = E(X_i^2) - [E(X_i)]^2 = \left(\frac{8}{3}\right) - (3 - \ln(4))^2$ same thing

$= \frac{8}{3} - [(3-\ln(4))\cdot(3-\ln(4))] = \frac{8}{3} - [9 - 3\ln(4) - 3\ln(4) + (\ln(4))^2]$

$= \frac{8}{3} - 9 + 6\ln(4) - (\ln(4))^2 = -\frac{19}{3} + 6\ln(4) - (\ln(4))^2$

$(\approx 0.063 \text{ inches}^2)$

- So $\mu_Y = n\mu_{X_i} = (50)(3-\ln(4)) = 150 - 50\ln(4)$ inches, and

$\sigma_Y^2 = n\sigma_{X_i}^2 = (50)\left(-\frac{19}{3} + 6\ln(4) - (\ln(4))^2\right) = -\frac{950}{3} + 300\ln(4) - 50(\ln(4))^2 \text{ in}^2$.

- Therefore by the CLT, $Y_{50} = \sum_{i=1}^{50} X_i \xrightarrow{d} Y \sim N\left(150 - 50\ln(4), -\frac{950}{3} + 300\ln(4) - 50(\ln 4)^2\right)$

and $P(78 \le \sum_{i=1}^{50} X_i \le 85) = P(\sum_{i=1}^{50} X_i \le 85) - P(\sum_{i=1}^{50} X_i \le 78)$

$= P\left(Z \le \frac{85 - (150 - 50\ln 4)}{\sqrt{-\frac{950}{3} + 300\ln 4 - 50(\ln 4)^2}}\right) - P\left(Z \le \frac{78 - (150 - 50\ln 4)}{\sqrt{-\frac{950}{3} + 300\ln 4 - 50(\ln 4)^2}}\right)$

$= P\left(Z \le \frac{85 - 80.685\ldots}{\sqrt{3.1310\ldots}}\right) - P\left(Z \le \frac{78 - 80.685\ldots}{\sqrt{3.1310\ldots}}\right)$

$= P(Z \le 2.44) - P(Z \le -1.52)$ by symmetry of Normal distribution

$= P(Z \le 2.44) - (1 - P(Z \le 1.52))$

$= 0.9927 - (1 - 0.9357)$

$= 0.9927 - 0.0643$

$= \boxed{0.9284.}$

~~Ask~~ Easier/quicker way?
- Nope!

(b) This time let $Y_n = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ represent the average thickness of the $n=50$ metal sheets. Then by the CLT, $Y_n = \bar{X} \xrightarrow{d} Y \sim N\left(\mu_{X_i}, \frac{\sigma_{X_i}^2}{n}\right)$. (μ_Y), σ_Y^2

Here, from the work done in part (a), $\mu_Y = \mu_{X_i} = 3 - \ln 4$ inches and

$\sigma_Y^2 = \frac{\sigma_{X_i}^2}{n} = \frac{1}{50}\left(\frac{8}{3} - (3 - \ln 4)^2\right)$ inches^2.

I.e., $Y_{50} = \bar{X} \xrightarrow{d} Y \sim N\left(3 - \ln 4, \frac{1}{50}\left(\frac{8}{3} - (3-\ln 4)^2\right)\right)$. cont'd →

- We need to find $P(\overline{X} \leq w) = 0.80$ (i.e., the value w which represents the 80^{th} percentile of the sample mean thickness of the 50 metal sheets).

So, using the CLT,

$$P(\overline{X} \leq w) = 0.80 \iff P\left(Z \leq \frac{w - \mu_Y}{\sqrt{\sigma_Y^2}}\right) = 0.80$$

$$\Rightarrow P\left(Z \leq \frac{w-(3-\ln 4)}{\sqrt{\frac{1}{50}\left(\frac{8}{3}-(3-\ln 4)^2\right)}}\right) = 0.80$$

[the closest value to 0.8 in the table is 0.7995 for $z = 0.84$.]

5/5

$$\Rightarrow (0.84) = \frac{w-(3-\ln 4)}{\sqrt{\frac{1}{50}\left(\frac{8}{3}-(3-\ln 4)^2\right)}}$$

$$\Rightarrow w-(3-\ln 4) = 0.84\sqrt{\frac{1}{50}\left(\frac{8}{3}-(3-\ln 4)^2\right)}$$

$$\Rightarrow w = 0.84\sqrt{\frac{1}{50}\left(\frac{8}{3}-(3-\ln 4)^2\right)} + (3-\ln 4)$$

$$\Rightarrow w = 1.643432806\ldots \text{ inches}$$

$$\Rightarrow \boxed{w \approx 1.64 \text{ inches.}}$$

3) We can represent each observed internet search as a Bernoulli trial where there are $n = 27$ iid trials (internet searches) and $p = 0.76$ is the probability of "success" (probability that the internet search was done using Google as the search engine). To do so, let $X_1, X_2, \ldots, X_{27} \overset{iid}{\sim} BER(0.76)$.

Then $Y_{27} = \sum_{i=1}^{27} X_i$ represents the total # of times Google was the search engine used among the 27 observed internet searches. It follows that $Y_{27} \sim BIN(n=27, p=0.76)$ and we can consider applying the CLT.

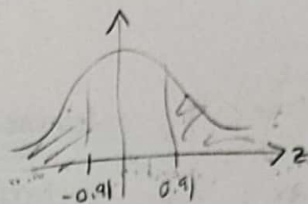
10/10

For a Binomial distribution, we need both that $np > 5$ and $nq > 5$. Here, we have that $(27)(0.76) = 20.52 > 5$ and $(27)(1-0.76) = 6.48 > 5$, so the CLT applies.

- Therefore, to find $P(\sum_{i=1}^{27} X_i > 18)$, we use the fact that

$Y_{27} = \sum_{i=1}^{27} X_i \xrightarrow{d} Y \sim N(np, npq)$ by the CLT, where $np = 20.52$ and $npq = (27)(0.76)(0.24) = 4.9248$, and we implement a continuity correction to take into account the fact that Y_{27} is discrete originally.

cont'd →



∴ In other words,

$$P(Y_{27} > 18) = P(Y_{27} \geq 19)$$

$$= 1 - P(Y_{27} < 19)$$

can go straight here

$$= 1 - P(Y_{27} \leq 18)$$

anything below 18.5 will get rounded down to 18; we use $\leq$ here but since Z is continuous the equality is irrelevant

$$\approx 1 - P\left(Z \leq \frac{18 + 0.5 - np}{\sqrt{npq}}\right)$$

$$= 1 - P\left(Z \leq \frac{18.5 - (20.52)}{\sqrt{(4.9248)}}\right) = 1 - P\left(Z \leq \frac{-2.02}{2.219...}\right)$$

$$= 1 - P(Z \leq -0.91024...) \approx 1 - P(Z \leq -0.91)$$

by symmetry of Normal dist.

$$\stackrel{\vee}{=} \cancel{1} - (\cancel{1} - P(Z \leq 0.91)) = P(Z \leq 0.91)$$

$$= \boxed{0.8186}$$

(Q:) Couldn't we just derive the CDF of Y_{27} using the known formula for its pdf since it's just a regular BIN dist? Wouldn't that be easier?

- If $n < 20$, can use BIN table, O.W., do sum of individual pdfs. cuz BIN doesn't have a closed-form CDF.

---

4) Let $X_1, X_2, \ldots, X_5 \stackrel{iid}{\sim} POI(\mu = \theta)$.

(a) Need to find the exact distribution of $W = \sum_{i=1}^{5} X_i$. [I.e., don't use the CLT, which yields an approximation distribution for sufficiently large n cases.]

10/10

- W is a sum of the X_i's and the mgf of each X_i is given by $m_{X_i}(t) = e^{(\theta)(e^t - 1)}$. So, using the mgf technique (since the X_i's are iid),

$$m_W(t) = m_{X_1}(t) \cdot m_{X_2}(t) \cdots m_{X_5}(t) = \left[e^{\theta(e^t-1)}\right]\left[e^{\theta(e^t-1)}\right]\cdots\left[e^{\theta(e^t-1)}\right]$$

$$= \left[e^{\theta(e^t-1)}\right]^5 = e^{5\theta(e^t-1)} \text{ (where } t > 0\text{)}$$

$$\Rightarrow \boxed{W \sim POI(\mu = 5\theta).}$$

cont'd →

(b) If $\theta = 0.4$, determine the probability that the sample mean is between 0.6 and 1.2, inclusive.

$5(0.4) = 5 \cdot \frac{2}{5}$

- In part (a) we found that $W = \sum_{i=1}^{5} X_i \sim POI(\mu = 5\theta)$.
- Let $\bar{X} = \frac{W}{5}$ and let $\theta = 0.4$ such that $W \sim POI(\mu = 2)$.
- Need to find $P(0.6 \le \bar{X} \le 1.2)$

Here, $P(0.6 \le \bar{X} \le 1.2) = P(0.6 \le \frac{W}{5} \le 1.2) = P(3 \le W \le 6)$.

[Use Poisson distribution table.]

So, $P(0.6 \le \bar{X} \le 1.2) = P(3 \le W \le 6)$

$$= P(W \le 6) - P(W < 3)$$

$$= P(W \le 6) - P(W \le 2)$$

$$= 0.9955 - 0.6767$$

$$= \boxed{0.3188.}$$

$$P(3 \le W \le 6) = P(W=3) + P(W=4) + P(W=5) + P(W=6)$$

$$= \frac{2^3 e^{-2}}{3!} + \frac{2^4 e^{-2}}{4!} + \frac{2^5 e^{-2}}{5!} + \frac{2^6 e^{-2}}{6!}$$

$$P(W \le 6) - P(W \le 3) = P(W=0) + P(W=1) + P(W=2) + P(W=3) + P(W=4) + P(W=5) + P(W=6)$$
$$- P(W=0) - P(W=1) - P(W=2) - P(W=3)$$

$$\ne P(W \le 6) - P(W < 3) = P(3 \le W \le 6) = P(W \le 6) - P(W \le 2).$$

Next lecture - Ch 8

"Parameter" = a value that describes a population

$\Rightarrow$ this is a constant value, but is typically unknown in practice.

e.g., population mean

"unbiased" used when expected value of statistic is equal to value of parameter.

Ch. 8: Statistics and Sampling Distribution

§8.2 - Statistics

Def: A statistic is a fcn of observable random variables that is free of unknown parameters.

$T = t(\underbrace{X_1, \ldots, X_n}_{\text{a R. sample}})$ — function of R.V.s

I.e., T is a fcn of your random sample.

$\Rightarrow$ T is a R variable.

"Statistic" = a value that describes a sample.

$\Rightarrow$ is a random variable.

- values can change from sample to sample.

common examples of statistics are sample mean, sample variance

Ex. $T = t(X_1, \ldots, X_n)$

$$= \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} \text{ is a r.v.}$$

little $\to$ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ = a single value (of $\bar{X}$), not a R. variable.

- Properties of Sample mean, $\bar{X}$, and Sample Variance, S^2

size of n irrelevant

I) If $X_1, X_2, \ldots, X_n$ form a r.s. from a population w/ pdf $f(x)$ with mean μ and variance σ^2, then $\underset{=\text{statistic}}{E(\bar{X})} = \underset{\substack{\text{population mean}\\=\text{parameter}}}{\mu}$ and $\underset{=\text{statistic}}{Var(\bar{X})} = \frac{\sigma^2}{n}$ = parameter (a constant)

mean of all the sample means

Note: Since $E(\bar{X}) = \mu$, $\bar{X}$ is unbiased for μ.

no tendency to over-/under-estimate pop. mean μ