The Central Limit Theorem

- Agenda: ① Prove the central limit theorem for binomially distributed random variables.
  ② Prove the central limit theorem for any RV (w/ finite variance)

- The central limit theorem will follow from using a few results on MGF's.

- **Thm:** the MGF of a Normal(μ, σ^2) R.V. is

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}$$

proof:

If $X \sim N(\mu, \sigma^2)$ then

$$f_X(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$M_X(t) = E\left[e^{tX}\right]$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\{tx\} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{tx - \frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\right\} dx$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{tx - \frac{1}{2\sigma^2}x^2 + \frac{1}{\sigma^2}x\mu - \frac{1}{2\sigma^2}\mu^2\right\} dx$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \left(\frac{\mu}{\sigma^2} + t\right)x - \frac{1}{2\sigma^2}\mu^2\right\} dx$$

Let's focus on the term in the exponent and try completing the square.

$$-\frac{1}{2\sigma^2}x^2 + \left(\frac{\mu}{\sigma^2} + t\right)x - \frac{1}{2\sigma^2}\mu^2 = -\frac{1}{2\sigma^2}\left(x^2 - 2(\mu + t\sigma^2)x + \mu^2\right)$$

$$= -\frac{1}{2\sigma^2}\left(x^2 - 2(\mu + t\sigma^2)x + (\mu + t\sigma^2)^2 - (\mu + t\sigma^2)^2 + \mu^2\right)$$

$$= -\frac{1}{2\sigma^2}\left[\left(x - (\mu + t\sigma^2)\right)^2 - \cancel{\mu^2} - 2t\mu\sigma^2 - t^2\sigma^4 + \cancel{\mu^2}\right]$$

$$= -\frac{1}{2\sigma^2}\left(x - (\mu + t\sigma^2)\right)^2 + t\mu + t^2\sigma^2/2$$

So

$$M_X(t) = \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(x - (\mu + t\sigma^2)\right)^2 + t\mu + t^2\sigma^2/2\right\} dx$$

$$= \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(x - (\mu + t\sigma^2)\right)^2\right\} \exp\left\{t\mu + t^2\sigma^2/2\right\} dx$$

$$= \exp\left\{t\mu + t^2\sigma^2/2\right\} \underbrace{\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(x - (\mu + t\sigma^2)\right)^2\right\} dx}_{\text{p.d.f. of a } N(\mu + t\sigma^2,\ \sigma^2) \text{ R.V.}}$$

Since p.d.f.'s must integrate to 1, we have that the above integral is 1 so

$$M_X(t) = \exp\left\{t\mu + t^2\sigma^2/2\right\} \quad //$$

- The other fact we will use (but not prove) is
- If two random variables have the same moment generating function, then they have the same distribution.

Subtract by np to center data

Claim: Let X be Binomial(n, p)

Let $Z_n := \frac{X - np}{\sqrt{n}}$

Then

$$M_Z(t) \to \exp\{ t^2 \sigma^2 / 2 \}$$

for some σ^2 that depends on p but not n. That is, for large n, the distribution of Z is well approximated by a normal distribution.

proof:

Recall $M_X(t) = [(1-p) + pe^t]^n$

So $M_{Z_n}(t) = E[e^{tZ_n}]$

$$= E\left[e^{t \frac{X - np}{\sqrt{n}}}\right]$$

$$= E\left[e^{\frac{t}{\sqrt{n}} X - \frac{npt}{\sqrt{n}}}\right]$$

$$= E\left[e^{-\frac{npt}{\sqrt{n}}} e^{\frac{t}{\sqrt{n}} X}\right]$$

$$= e^{-\frac{npt}{\sqrt{n}}} E\left[e^{\frac{t}{\sqrt{n}} X}\right] \quad \text{let } u := \frac{t}{\sqrt{n}}$$

$$= e^{-\frac{npt}{\sqrt{n}}} E\left[e^{uX}\right]$$

$$= e^{-\frac{npt}{\sqrt{n}}} M_X(u)$$

$$= e^{-\frac{npt}{\sqrt{n}}} \left[(1-p) + pe^u\right]^n$$

$$= e^{-\frac{npt}{\sqrt{n}}} \left[(1-p) + pe^{t/\sqrt{n}}\right]^n$$

$$= \left[e^{-\frac{pt}{\sqrt{n}}} \left(1 - p + pe^{t/\sqrt{n}}\right)\right]^n$$

<u>Note:</u> $t/\sqrt{n}$ is small no matter the value of t because n is "large"

<u>Note:</u> $e^h \approx 1 + h + h^2/2$ for small h

$\boxed{\text{See Demonstration}}$

So $M_{Z_n}(t) \approx \left[\left(1 - \frac{pt}{\sqrt{n}} + \frac{p^2t^2}{2n}\right)\left(1 - p + p\left(1 + \frac{t}{\sqrt{n}} + \frac{t^2}{2n}\right)\right)\right]^n$

$= \left[\left(1 - \frac{pt}{\sqrt{n}} + \frac{p^2t^2}{2n}\right)\left(1 - p + p + \frac{pt}{\sqrt{n}} + \frac{pt^2}{2n}\right)\right]^n$

$= \left[\left(1 - \frac{pt}{\sqrt{n}} + \frac{p^2t^2}{2n}\right)\left(1 + \frac{pt}{\sqrt{n}} + \frac{pt^2}{2n}\right)\right]^n$

$= \left[1 + \cancel{\frac{pt}{\sqrt{n}}} + \frac{pt^2}{2n} - \cancel{\frac{pt}{\sqrt{n}}} - \frac{p^2t^2}{n} - \frac{p^2t^3}{2n^{3/2}} + \frac{p^2t^2}{2n} + \frac{p^3t^3}{2n^{3/2}} + \frac{p^3t^4}{2n^2}\right]^n$

$= \left[1 + \frac{pt^2}{2n} - \frac{p^2t^2}{n} + \frac{p^2t^2}{2n} + a_n\right]^n$

a_n collects terms that have higher powers of $\frac{1}{n}$ eg. $1/n^{3/2}$, $1/n^2$

Note: for n large enough, $\frac{a}{n} + \frac{b}{n^{3/2}} + \frac{c}{n^2} \approx \frac{a}{n}$ for any a, b, c

$\approx \left[1 + (pt^2 - 2p^2t^2 + pt^2)/(2n)\right]^n$

$= \left[1 + (pt^2 - p^2t^2)/(2n)\right]^n$

$= \left[1 + t^2 p(1-p)/(2n)\right]^n$

Recall: $\left(1+\frac{u}{n}\right)^n \xrightarrow[n\to\infty]{} e^u$ for any u

$\boxed{\text{Demonstration}}$

Thus, $M_{Z_n}(t) \xrightarrow[n\to\infty]{} \exp\{t^2 p(1-p)/2\}$

- Recall that MGF of a Normal is $\exp\{t\mu + t^2\sigma^2/2\}$

Here $\mu = 0$ and $\sigma^2 = p(1-p)$

- Since same MGF $\Rightarrow$ same distribution, we have

$Z_n \approx N(0, p(1-p))$ for large n.

- So we proved that a Binomial R.V. (properly centered and scaled) converges to a Normal R.V.

- Actually, a binomial R.V. is equal to a sum of Bernoulli R.V.'s

Let $X_1, \dots, X_n$ be independent Bernoulli (p) R.V.'s

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

Claim: $X = \sum_{i=1}^{n} X_i \sim Bin(n,p)$

proof:

use MGF's!

$$E[e^{X}] = E[e^{t\sum X_i}]$$

$$= E[e^{tX_1} \cdot \ldots \cdot e^{tX_n}]$$

$$= E[e^{tX_1}] \cdot \ldots \cdot E[e^{tX_n}] \leftarrow \text{since independent}$$

$$= M_{X_1}(t) \cdot \ldots \cdot M_{X_n}(t)$$

MGF's of Bernoulli(p) are all the same

$$= [M_{X_i}(t)]^n$$

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

$$= \text{MGF of Binomial} \;//$$

- So we actually showed that

$$Z = \frac{\sum X_i - np}{\sqrt{n}}$$

$$= \frac{\sum X_i - E[\sum X_i]}{\sqrt{n}}$$

$\approx$ Normal for large n

- Turns out the X_i need not be Bernoulli but can be any random variable with finite second moments!

- Let $X_1, \dots, X_n$ be independent with the same distribution $f_X(x)$
- Let $\mu := E[X_i]$, $\sigma^2 = Var(X_i)$
- As before, let

$$Z_n := \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}}$$

$$= \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu \right)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

let $Y_i := X_i - \mu$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

Note that $E[Y_i] = 0$ since

$$E[Y_i] = E(X_i - \mu)$$

$$= E[X_i] - \mu$$

$$= \mu - \mu = 0$$

so

$$\sigma^2 = Var(Y_i) = E(Y_i^2) - E[Y_i]^2 = E(Y_i^2)$$

- The claim is that Z_n here is also well approximated by a normal distribution, even though we defined Z_n in terms of (almost) arbitrary random variables.

- The key tool we use is the Taylor Expansion of the moment generating function about 0.

- Simple version of Taylor series:
  For any function 2-times differentiable at 0, we have

$$f(x) = f(0) + f'(0)\,x + f''(0)\,x^2/2 + r(x)$$

where $\frac{r(x)}{x^2} \xrightarrow[x \to 0]{} 0$

$r(x)$ is the remainder and is negligible compared to the lower order terms.

- Taylor series expansion of MGF of Y_i

$$M_{Y_i}(t) \approx \underbrace{M_{Y_i}(0)}_{} + t\,\underbrace{M'_{Y_i}(0)}_{} + t^2\,\underbrace{M''_{Y_i}(0)}_{}/2$$

$$\begin{aligned} &= E[e^{0 Y_i}] \quad && E[Y_i] \quad && E[Y_i^2] = Var(Y_i) = \sigma^2 \\ &= E[e^0] && \;\; \| \\ &= E[1] = 1 && \;\; 0 \end{aligned}$$

so

$$M_{Y_i}(t) \approx 1 + \frac{\sigma^2}{2} t^2$$

$$\text{So} \quad M_{Z_n}(t) = E\left[e^{\frac{t}{\sqrt{n}}\sum Y_i}\right]$$

$$= E\left[\prod_{i=1}^{n} e^{\frac{t}{\sqrt{n}} Y_i}\right]$$

$$= \prod_{i=1}^{n} E\left[e^{\frac{t}{\sqrt{n}} Y_i}\right] \quad \text{(by independence)}$$

$$= \prod_{i=1}^{n} M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)$$

$$\approx \prod_{i=1}^{n} \left(1 + \frac{\sigma^2}{2}\left(\frac{t}{\sqrt{n}}\right)^2\right) \quad \text{(by approx. of } M_Y(t)\text{)}$$

$$= \prod_{i=1}^{n} \left(1 + \frac{\sigma^2}{2n} t\right)$$

$$= \left(1 + \frac{\sigma^2}{2n} t\right)^n$$

$$\xrightarrow[n\to\infty]{} e^{\sigma^2 t/2}$$

$$= \text{MGF of a normal}(0, \sigma^2) \text{ R.V.!}$$

version 1

- Conclusion: As n gets larger, Z_n (the centered sum, scaled by $\sqrt{n}$) of n iid. R.V.'s with variance σ^2 gets closer and closer to a $N(0,\sigma^2)$ R.V.

version 2

- A more common way is to say

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{} N(0,1)$$

$$= \frac{\frac{1}{n}\sum X_i - \mu}{\sigma/\sqrt{n}} = \frac{n}{n}\,\frac{\frac{1}{n}\sum X_i - \mu}{\sigma/\sqrt{n}} = \frac{\sum X_i - n\mu}{\sqrt{n}\,\sigma} = \frac{Z_n}{\sigma}$$

- Recall: $\text{var}(Z_n) = \sigma^2$

$$\text{var}\left(\frac{Z_n}{\sigma}\right) = \frac{1}{\sigma^2}\text{var}(Z_n) = 1$$

- Version 3: $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

$$E[\bar{X}] = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu = \frac{1}{n} n \mu = \mu$$

$$\text{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \text{var}\, X_i = \frac{1}{n^2}\sum \sigma^2 = \frac{1}{n^2} n \sigma^2 = \frac{1}{n}\sigma^2$$

- For most RV's, we can approximate their average with a normal, but if $X_1, \ldots, X_n$ are all $N(\mu, \sigma^2)$, this relationship is exact.

$$M_{X_i}(t) = e^{t\mu + t^2\sigma^2/2}$$

$$M_{\bar{X}}(t) = E\left[e^{\frac{t}{n}\sum X_i}\right]$$

$$= E\left(\prod e^{\frac{t}{n}X_i}\right)$$

$$= \prod E\left(e^{\frac{t}{n}X_i}\right)$$

$$= \prod M_{X_i}\left(\frac{t}{n}\right)$$

$$= \prod e^{\frac{t}{n}\mu + \frac{t^2}{n^2}\sigma^2/2}$$

$$= \exp\left\{\sum\left(\frac{t}{n}\mu + \frac{t^2}{n^2}\sigma^2/2\right)\right\}$$

$$= \exp\left\{n\frac{t}{n}\mu + n\frac{t^2}{n^2}\sigma^2/2\right\}$$

$$= \exp\left\{t\mu + t^2\left(\frac{\sigma^2}{n}\right)/2\right\} = N\left(\mu, \frac{\sigma^2}{n}\right)$$