STAT 315: Central Limit Theorem

Luc Rey-Bellet

University of Massachusetts Amherst

luc@math.umass.edu

April 29, 2025

Central Limit Theorem

Very useful $\longrightarrow$ when n is large everything looks like a normal RV! So it can be computed using the z-score

Central Limit Theorem

Suppose that $Y_1, Y_2, \cdots, Y_n$ are IID random variables with $E[Y_i] = \mu$ and $V[Y_i] = \sigma^2$. Then

$$P\left(a \le \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \le b\right) \longrightarrow P(a \le Z \le b) \text{ as } n \to \infty.$$

where Z is standard normal random variable.

$$\text{To compute } P(a \le Z \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad \text{use z-score table}$$

Note that since $\overline{Y} = \frac{1}{n}(Y_1 + \cdots, Y_n)$ we have

$$P\left(a \le \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \le b\right) = P\left(a \le \frac{Y_1 + \cdots + Y_n - n\mu}{\sigma\sqrt{n}} \le b\right)$$

so use either the sum $Y_1 + \cdots + Y_n$ or the average $\frac{1}{n}(Y_1 + \cdots + Y_n)$.

Important to remember the scaling in n.

$$\left.\begin{array}{l} E[\overline{Y}] = \mu \\ V[\overline{Y}] = \frac{\sigma^2}{n} \end{array}\right\} \Longrightarrow \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \text{ has mean } 0 \text{ and variance } 1.$$

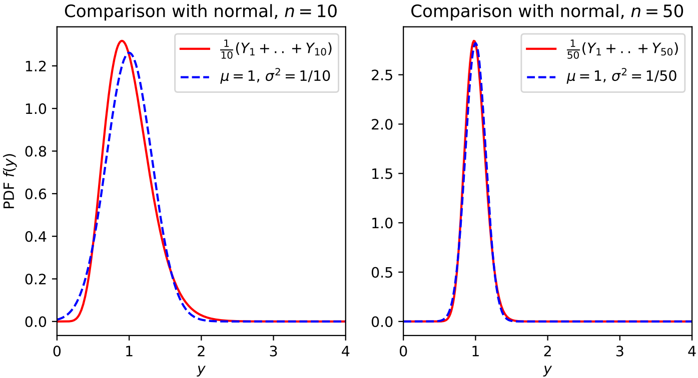

Figure: Comparing the sample mean of n exponential with $\beta = 1$ (that is a gamma with $\alpha = n$ and $\beta = 1/n$) and the corresponding normal with same mean and variance that is $\mu = 1$ and $\sigma^2 = \frac{1}{n}$

Proof of the CLT

First let us rewrite

$$\begin{aligned} U_n \equiv \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} &= \frac{Y_1 + Y_2 + \cdots + Y_n - n\mu}{\sigma\sqrt{n}} \\ &= \frac{1}{\sqrt{n}}\left(\frac{Y_1 - \mu}{\sigma} + \frac{Y_2 - \mu}{\sigma} + \cdots + \frac{Y_n - \mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{n}}\left(Z_1 + Z_2 + \cdots + Z_n\right) \end{aligned}$$

where Z_i are IID with

$$E[Z_i] = 1 \qquad V[Z_i] = 1\,.$$

We now use MGF and show that

$$m_{U_n}(t) \longrightarrow m_Z(t)$$

where Z is standard normal.

$$\begin{aligned}
m_{U_n}(t) &= E[e^{tU_n}] = E\left[e^{t\frac{1}{\sqrt{n}}(Z_1 + \cdots Z_n)}\right] \\
&= E\left[e^{t\frac{1}{\sqrt{n}}Z_1} \cdots e^{t\frac{1}{\sqrt{n}}Z_n}\right] \\
&= E\left[e^{t\frac{1}{\sqrt{n}}Z_1}\right] \cdots E\left[e^{t\frac{1}{\sqrt{n}}Z_n}\right] \text{ by independence} \\
&= E\left[e^{t\frac{1}{\sqrt{n}}Z_1}\right]^n \text{ by IID property} \\
&= m_{Z_1}\left(\frac{t}{\sqrt{n}}\right)^n
\end{aligned}$$

We use then the Taylor series of order 2:

$$f(t) = f(0) + tf'(0) + \frac{t^2}{2} f''(0) + \text{ small error}$$

applied to $m_{Z_1}(t)$.

$$m_{Z_1}(t) = m_{Z_1}(0) + t m'_{Z_1}(0) + \frac{t^2}{2} m''_{Z_1}(0) + \cdots = 1 + 0 + \frac{t^2}{2} + \cdots$$

since $m_{Z_1}(0) = 1$, $m'_{Z_1}(0) = E[Z_1] = 0$ and $m''_{Z_1}(0) = E[Z_1^2] = V[Z_1] = 1$.

Then we can conclude

$$m_{U_n}(t) = m_{Z_1}\left(\frac{t}{\sqrt{n}}\right) = \left(1 + \frac{t^2}{n}\right)^n \longrightarrow e^{t^2/2}$$

since $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$.

This is the MGF of a standard normal, so we are done.

Example

An astronomer is measuring the distance in light-years to a certain star. The measurement has mean d but is noisy due to measurement error and the variance is $\sigma^2 = 4$.

How many measurement should the astronomer perform to measure d with a precision of .5 light year and 95% confidence?.

Denote $X_1, X_2, \cdots, X_n$ the n measurement. For n large, using CLT we get

$$
\begin{aligned}
P\left(-.5 \le \overline{X} - d \le .5\right) &= P\left(\frac{-.5}{2/\sqrt{n}} \le \frac{\overline{X} - d}{2/\sqrt{n}} \le \frac{.5}{2/\sqrt{n}}\right) \\
&= P\left(\frac{-\sqrt{n}}{4} \le \frac{\overline{X} - d}{2/\sqrt{n}} \le \frac{\sqrt{n}}{4}\right) \\
&\approx P\left(\frac{-\sqrt{n}}{4} \le Z \le \frac{\sqrt{n}}{4}\right) = .95
\end{aligned}
$$

And thus

$$1.96 = \frac{\sqrt{n}}{4} \Longleftrightarrow n = (7.84)^2 = 61.47$$

Example: Poisson

The number of student enrolling in a class has a Poisson distribution with mean 100. If there are more that 120 students then we will need an extra section. What is the probability that an extra section is needed?

Exact solution: $P(X \geq 120) = 1 - \sum_{n=0}^{119} e^{-100} \frac{(100)^n}{n!} = 0.02823$ (using technology).

Careless approximation: We have $E[X] = 100$ and $V[X] = 100$. Since X takes only integer discrete values we have

$$P(X \geq 120) = P(X \geq 119.5) \quad \text{continuity correction} .$$

Let us pretend that X is normal with $\mu = 100$ and $\sigma^2 = 100$. Then we have

$$P(X \geq 119.5) = P\left(\frac{X - 100}{\sqrt{100}} \geq \frac{119.5 - 100}{\sqrt{100}}\right) \approx P(Z \geq 1.95) = 0.0256$$

Why so close to the correct answer?

Example: Poisson, continued

Recall that if X is Poisson with parameter λ then the MGF is $m(t) = E[e^{tX}] = e^{\lambda(e^t - 1)}$.

If X_1 and X_2 are independent Poisson random variables with parameters λ_1 and λ_2 then

$$m_{X_1+X_2}(t) = m_{X_1}(t) m_{X_2}(t) = e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

and thus $X_1 + X_2$ is Poisson with parameter $\lambda_1 + \lambda_2$.

Therefore if X is Poison with parameter 100 we can write

$$X = X_1 + X_2 + \cdots + X_{100}$$

where X_i are IID Poisson with $\lambda = 1$.

Normal approximation is totally reasonable by the CLT.

Example: test scores and difference of sample mean

Test scores in a standardized High School test has a statewide mean of 60 with a standard deviation of 8?

Question 1: In NHS 100 students take the tests and obtain an average score of 62. The principal congratulates their students for such an excellent score. Is this justified?

Sample size $n = 100$, Y_i=score of student i. Sample average $\overline{Y} = 62$. Using the CLT we have

$$\begin{aligned} P(\overline{Y} \geq 62) &= P\left(\frac{\overline{Y} - 60}{8/\sqrt{100}} \geq \frac{62 - 60}{8/\sqrt{100}}\right) \\ &\approx P(Z \geq 2.5) = 0.0062 \end{aligned} \qquad (1)$$

Very unlikely! The sample of NHS is not representative from the statewide population. The principal was correct!

Question 2: 100 students take the exam in NHS with an average of $\overline{X} = \frac{1}{100}(X_1 + \cdots X_{100})$ and 50 students take the tests in AHS with an average $\overline{Y} = \frac{1}{50}(Y_1 + \cdots Y_{50})$. What is the probability that the differecne between the average scores $|\overline{X} - \overline{Y}|$ is at least equal to 1?

Look at the difference $\overline{X} - \overline{Y}$. We have

$$\begin{aligned} E[\overline{X} - \overline{Y}] &= E[\overline{X}] - E[\overline{Y}] = \mu - \mu = 0 \\ V[\overline{X} - \overline{Y}] &= V[\overline{X}] + V[\overline{Y}] = \frac{\sigma^2}{100} + \frac{\sigma^2}{50} = \frac{3\sigma^2}{100} \end{aligned}$$

By the CLT we have

$$\begin{aligned} & P(|\overline{X} - \overline{Y}| > 1) = P(-1 \le \overline{X} - \overline{Y} \le 1) \\ & = P\left(\frac{-1}{8\sqrt{3/100}} \le \frac{\overline{X} - \overline{Y} - 0}{8\sqrt{3/100}} \le \frac{1}{8\sqrt{3/100}} \right) \\ & \approx P(-.721 \le Z \le .721) = .529 \quad \text{so not unlikely} \end{aligned}$$

Normal Approximation to the binomial distribution

Suppose X_i are IID Bernoulli RV (= Binomial with $n = 1$ and p) with PDF

$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

Then

$$\begin{aligned} X = X_1 + \cdots + X_n &= \text{number of successes in } n \text{ independent trials} \\ &= \text{Binomial with parameters } n \text{ and } p \end{aligned}$$

So we can use the normal approximation which is very good, even for small n!

Continuity correction: Here X take integer discrete value but the the normal RV is continuous so we can and should adjust the interval

$$P(X \leq 7) = P(X \leq 7.5) \quad \text{or} \quad P(5 \leq X \leq 12) = P(4.5 \leq X \leq 12.5)$$

Leads to much better results when using CLT where you replace a discrete RV with a continuous RV.

Example: X binomial with parameters $n = 25$ and $p = .4$ so $E[X] = np = 10$ and $V[X] = np(1-p) = 6$.

- Exact: $P(X \leq 8) = .274$
- CLT with continuity correction

$$\begin{aligned} P(X \leq 8) = P(X \leq 8.5) &= P\left(\frac{X-10}{\sqrt{6}} \leq \frac{8.5-10}{\sqrt{6}}\right) \\ &\approx P(Z \leq -0.61) = .2709 \quad \text{very good} \end{aligned}$$

- CLT witout continuity correction

$$P(X \leq 8) = P\left(\frac{X-10}{\sqrt{6}} \leq \frac{8-10}{\sqrt{6}}\right) \approx P(Z \leq -.81) = .2089$$

Not so good: always use the continuity correction!

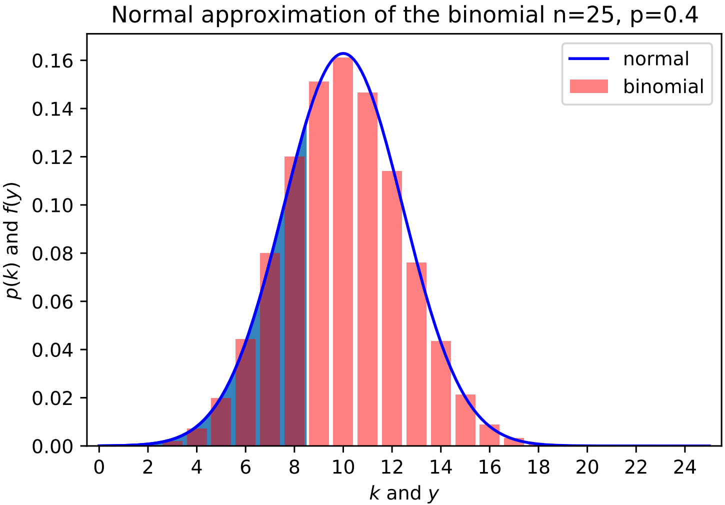

Figure: PDF of the binomial with $n = 25$ and $p = 4$ and PDF of the normal with $\mu = np = 10$ and $\sigma^2 = np(1-p) = 6$. The shaded area is $P(X \leq 8.5)$ with continuity correction.

This works well even for a single value of X

$$
\begin{aligned}
P(X=8) &= P(7.5 \le X \le 8.5) \\
&= P\left(\frac{7.5-10}{\sqrt{6}} \le \frac{X-10}{\sqrt{6}} \le \frac{8.5-10}{\sqrt{6}}\right) \\
&\approx P(-1.02 \le Z \le .61) = .1170
\end{aligned}
$$

Compare with the exact value

$$P(X=8) = .1198 \quad \text{awesome!}$$

$n = 25$ is not a big number.....

Always use the continuity correction!

Empirical rule for the normal approximation to the binomial

You can use normal approximation to the binomial if n moderately large and p not too close to 0 and 1.

$$\text{empirical rule:} \quad n > 9 \, \frac{\max(p, 1-p)}{\min(p, 1-p)}$$

For example if $p = 1/4 \leq 1/2$ then $1/4 = p \leq 1 - p = 3/4$ and the empirical rules means $n > 9\frac{1-p}{p} = 27$.

Recall that if n is large and p is very small we have the Poisson approximation to the binomial.

$$X \text{ is approximately Poisson with } \lambda = np$$

Poisson vs Normal

1 in 410 American is a lawyer (a true fact)and your town has 1,500 inhabitants. What is the probability that no lawyer lives in your town. The number of lawyers X is a binomial with $n = 1,500$ and $p = 1/410$

Exact: $P(X = 0) = \left(\frac{409}{410}\right)^{1500} = 0.02565$

Poisson approximation: n is large and p is small so X is approximately Poisson with $\lambda = np = \frac{1500}{410}$ so $P(X = 0) = e^{-\lambda} = e^{-\frac{1500}{410}} = 0.02577$

Normal approximation:

$$\begin{aligned} P(X = 0) &= P(X \le \frac{1}{2}) = P\left(\frac{X - \frac{1500}{410}}{1500\frac{1}{410}\frac{409}{410}} \le \frac{\frac{1}{2} - \frac{1500}{410}}{1500\frac{1}{410}\frac{409}{410}}\right) \\ &\approx P(Z \le -1.653) = 0.04913 \end{aligned}$$

The relative error is 100%.
The rule of thumb is violated: $n = 1500$ and $9\frac{1-p}{p} = 9 \times 409 = 3681 \ldots$