The central limit theorem (part 2)

Lecture 11a (STAT 24400 F24)

Addition of normal random variables (and the CLT)

Fact: if X and Y are normal and $\perp\!\!\!\perp$, then $X + Y$ is normal.

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2),\ Y \sim \mathrm{N}(\mu_2, \sigma_2^2),\ X \perp\!\!\!\perp Y \quad \Rightarrow \quad X + Y \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

This fact is often combined with the CLT:

if X is $\approx$ normal, Y is $\approx$ normal, and X and Y are independent,

then $X + Y$ is $\approx$ normal.

Example — sum of indep. variables and CLT

One coin is fair (50% chance Heads) & another is biased (25% Heads).

After flipping each coin 100 times, what is the distrib. of total # of Heads?

- Let X = # heads from the fair coin $\sim$ Binomial$(100, 0.5)$. Then

  $E(X) = np = 100 \cdot 0.5 = 50,\ \mathrm{Var}(X) = np(1-p) = 100 \cdot 0.5 \cdot (1 - 0.5) = 25$

  $$\text{by CLT, } X \ \approx\ \mathrm{N}(\ E(X),\ Var(X)\ ) = \mathrm{N}(50, 25)$$

- Let Y = # heads from the biased coin $\sim$ Binomial(100,0.25)

  $$\text{by CLT, } Y \approx \mathrm{N}(100 \cdot 0.25, 100 \cdot 0.25 \cdot (1 - 0.25)) = \mathrm{N}(25, 18.75)$$

- We know $X \perp\!\!\!\perp Y$

  $$\Rightarrow \quad X + Y \approx \mathrm{N}(75, 43.75)$$

Example — difference of indep. variables and CLT

One coin is fair (50% chance Heads) & another is biased (25% Heads).

Player A flips the fair coin 120 times.
Player B flips the biased coin 200 times.

Whoever has more Heads, wins the game. What are the odds of the game?

- Let X = # heads for Player A $\sim$ Binomial(120,0.5)

  $$\text{by CLT, } X \approx \mathrm{N}(120 \cdot 0.5, 120 \cdot 0.5 \cdot (1 - 0.5)) = \mathrm{N}(60, 30)$$

- Let Y = # heads for Player B $\sim$ Binomial(200,0.25)

  $$\text{by CLT, } Y \approx \mathrm{N}(200 \cdot 0.25, 200 \cdot 0.25 \cdot (1 - 0.25)) = \mathrm{N}(50, 37.5)$$

- We know $X \perp\!\!\!\perp Y$

Example — difference of indep. variables and CLT (cont.)

$\mathbb{P}(\text{Player A wins}) = \mathbb{P}(X > Y) = \mathbb{P}(X - Y > 0)$

- Y is $\approx \mathrm{N}(50, 37.5)$, then by symmetry, $-Y$ is also normal $\approx \mathrm{N}(-50, 37.5)$
- X and $-Y$ are still independent,
- Then $X - Y = X + (-Y)$ is $\approx$ normal, with

  $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) = 10,$

  $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2cov(X, Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 67.5.$

$$\mathbb{P}(\text{Player A wins}) = \mathbb{P}(\underbrace{X - Y}_{\approx \mathrm{N}(10,67.5)} > 0) = \mathbb{P}\left(\underbrace{\frac{(X - Y) - 10}{\sqrt{67.5}}}_{\approx \mathrm{N}(0,1)} > \frac{0 - 10}{\sqrt{67.5}}\right)$$

$$\approx 1 - \Phi\left(\frac{0 - 10}{\sqrt{67.5}}\right) = 0.8882$$

Accuracy of the sample mean

Suppose $X_1, \cdots, X_n$ are i.i.d. from a distrib. with mean μ. The sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is commonly used as an estimator of μ.

How accurately does $\bar{X}$ estimate μ, for a sample of size n?

We may ask — for a small constant $\epsilon > 0$, as $n \nearrow$:

Q1. Can we use CLT to show $\mathbb{P}\Big(|\bar{X} - \mu| > \text{(some quantity with limit 0)}\Big) \le \epsilon$?

Q2. Can we use CLT to show $\mathbb{P}(|\bar{X} - \mu| > \epsilon) \le \text{(some quantity with limit 0)}$?

Accuracy of the sample mean (by CLT for any distrib.)

Q1. Can we use CLT to show $\mathbb{P}\left(|\bar{X} - \mu| > \begin{pmatrix}\text{some quantity}\\ \text{with limit 0}\end{pmatrix}\right) \le \epsilon$?

By the CLT: for large n,

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

For a given small constant $\epsilon > 0$, we need to choose z_* s.t.

$$\mathbb{P}\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} > z_*\right) \approx \mathbb{P}(|Z| > z_*) \le \epsilon$$

By the symmetry of $N(0,1)$,

$$\mathbb{P}(|Z| > z_*) = 2[1 - \mathbb{P}(Z \le z_*)] = 2[1 - \Phi(z_*)]$$

So the desired z_* should satisfy

$$\Phi(z_*) = 1 - \frac{\epsilon}{2} \quad \Rightarrow \quad z_* = \Phi^{-1}\left(1 - \frac{\epsilon}{2}\right)$$

Fix some $\epsilon \in (0,1)$ and let $z_* = \Phi^{-1}(1 - \frac{\epsilon}{2})$.

$$\mathbb{P}\left(|\bar{X} - \mu| > z_* \cdot \frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left((\bar{X} < \mu - z_* \cdot \frac{\sigma}{\sqrt{n}}\right) + \mathbb{P}\left(\bar{X} > \mu + z_* \cdot \frac{\sigma}{\sqrt{n}}\right)$$

$$= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -z_*\right) + \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > z_*\right)$$

by CLT $\longrightarrow$ $\approx \Phi(-z_*) + [1 - \Phi(z_*)]$

by symmetry of N(0, 1) $\longrightarrow$ $= 2[1 - \Phi(z_*)]$

$= \epsilon$

Accuracy of the sample mean (and confidence intervals)

For example, for $\epsilon = 0.05$ we have $z_* = \Phi^{-1}(0.975) \approx 1.96$.

Therefore,

$$\mathbb{P}\left(|\bar{X} - \mu| \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \approx 95\%$$

This applies to i.i.d. sample means from (just about) any distributions

— a very powerful result from the CLT.

The equation can be expressed in the form of a confidence interval for μ:

$$\mathbb{P}\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \approx 95\%$$

Accuracy of the sample mean (CLT for normal sample only)

Q2. Can we use CLT to show $\mathbb{P}(|\bar{X} - \mu| > \epsilon) \leq \left(\begin{smallmatrix}\text{some quantity}\\ \text{with limit 0}\end{smallmatrix}\right)$?

Fix some $\epsilon > 0$. If the data is normal (i.e., $\bar{X}$ is exactly normal):

$$\mathbb{P}(|\bar{X} - \mu| > \epsilon) = \mathbb{P}(\bar{X} < \mu - \epsilon) + \mathbb{P}(\bar{X} > \mu + \epsilon)$$

$$= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < -\frac{\epsilon}{\sigma/\sqrt{n}}\right) + \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{\epsilon}{\sigma/\sqrt{n}}\right)$$

exact since data is normal $\longrightarrow$ $$= \Phi\left(-\frac{\epsilon\sqrt{n}}{\sigma}\right) + \left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right)$$

by symmetry of N(0, 1) $\longrightarrow$ $$= 2\left(1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\right)$$

$$\leq (\text{constant}) \cdot e^{-n \cdot (\text{constant})}$$

if data not normal, <u>cannot</u> use CLT for this step, since $\frac{\epsilon\sqrt{n}}{\sigma}$ not constant

Accuracy of the sample mean (Chebyshev's inequality for any distrib.)

For non-normal data (when the CLT result for normal sample doesn't apply),

we can bound $\mathbb{P}(|\bar{X} - \mu| > \epsilon)$ with Chebyshev's inequality (Lecture 5a):

$$\mathbb{P}(|\bar{X} - \mu| > \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

scales as $\mathcal{O}(\frac{1}{n})$
(compare to $\mathcal{O}(e^{-cn})$ for normal distrib.)

Note: The fact that $\mathbb{P}(|\bar{X} - \mu| > \epsilon) \to 0$ as $n \to \infty$ (while $\epsilon > 0$ is constant) is also known as the *Law of Large Numbers*.

Law of Large Numbers

<u>Theorem</u> — **Law of Large Numbers** (LLN)

Let $X_1, \cdots, X_n$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\bar{X} = \sum_{i=1}^n X_i / n$.

Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X} - \mu| > \epsilon) \to 0 \qquad \text{as} \quad n \to \infty$$