14) Central Limit Theorem

<u>Theorem 4.1</u>: Let $X \sim N(\mu_X, \sigma_X^2)$ $Y \sim N(\mu_Y, \sigma_Y^2)$ be two independant random variables with $\mu_X, \sigma_X \in \mathbb{R}$ and $\mu_Y, \sigma_Y \in \mathbb{R}$. Then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

<u>proof</u> We know that for normal distribution

$$E[X] = \mu_X \qquad Var(X) = \sigma_X^2$$

$$E[Y] = \mu_Y \qquad Var(X) = \sigma_Y^2$$

$$E[X+Y] = E[X] + E[Y]$$

linearity of expectations Thm 10.2

$$= \mu_X + \mu_Y$$

$$Var(X+Y) = Var(X) + Var(Y) + cov(X,Y)$$

$X \perp\!\!\!\perp Y \Rightarrow Cov(X,Y) = 0$. So

$$Var(X+Y) = Var(X) + Var(Y)$$

$$= \sigma_X^2 + \sigma_Y^2.$$

Hence $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

↳ why normal, see year 2 $\square$

Now consider an iid sample $X_1, \ldots, X_n$ from a random variable X

Sample mean

$$\bar{X}_n = \frac{(X_1 + X_2 + \ldots + X_n)}{n}$$

From chapter 13:

$$E[\bar{X}_n] = E[X]$$

$$Var(\bar{X}_n) = \frac{Var(X)}{n}$$

As seen in R labs, for large sample size, the sample mean is normally distributed (approximately)

↳ even if X is not normally distributed.

Rule of thumb 14.2: For sufficiently large sample size n, mean is approximately distributed,

$$\boxed{\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n \mathrel{\dot\sim} N\left(E[X], \frac{Var(X)}{n}\right)}$$

Equivalently multiplying sample mean $\bar{X}_n$ by n gives:

$$n\bar{X}_n = X_1 + X_2 + X_3 + \cdots + X_n$$

$$\begin{aligned} E[X_1 + X_2 + \cdots + X_n] &= E[n\bar{X}_n] \\ &= nE[\bar{X}_n] \\ &= nE[X] && \text{Thm 7.16} \end{aligned}$$

$$\begin{aligned} Var(X_1 + X_2 + \cdots + X_n) &= Var(n\bar{X}_n) \\ &= n^2 Var(\bar{X}_n) && \text{Thm 7.25} \\ &= n^2 \frac{Var(X)}{n} \\ &= nVar(X) \end{aligned}$$

So

by Thm 14.1

$$\boxed{X_1 + X_2 + \cdots + X_n \mathrel{\dot\sim} N(nE[X], nVar(X))}$$

The amazing fact is that no matter what the distribution of the random variable X is, for sufficiently large sample size X, sample mean is approximately normally distributed.

14.2 is not a theorem as the term sufficiently large is vague.

To get a precise theorem, we need to take limit as $n \to \infty$.

$\bar{X}_n$ does not have a nice limit as $n \to \infty$ as its variance goes towards 0.

Therefore we consider the standardised random variable

$$\boxed{Z_n = \frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{Var(\bar{X}_n)}} = \frac{\bar{X}_n - E[X_i]}{\sqrt{\frac{Var(X_i)}{n}}} \mathrel{\dot{\sim}} N(0,1)}$$

which for all all n has variance one

proof

$$E[Z_n] = E\left[\frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{Var(\bar{X}_n)}}\right]$$

linearity of expectations

number $\in \mathbb{R}$

$$= E\left[\frac{\bar{X}_n}{\sqrt{Var(\bar{X}_n)}}\right] - E\left[\frac{E[X_n]}{\sqrt{Var(X_n)}}\right]$$

$$= \cancel{\frac{E[\bar{X}_n]}{\sqrt{Var(\bar{X}_n)}}} - \cancel{\frac{E[X_n]}{\sqrt{Var(X_n)}}}$$

$$= 0 \Rightarrow \boxed{E[Z_n] = 0}$$

$$Var(Z_n) = Var\left(\frac{\bar{X}_n - E[\bar{X}_n]}{\sqrt{Var(X_n)}}\right)$$

$\in \mathbb{R}$

$$= Var\left(\frac{\bar{X}_n}{\sqrt{Var(\bar{X}_n)}} - \frac{E[\bar{X}_n]}{\sqrt{Var(X_n)}}\right)$$

by Thm 7.25

$$= \frac{\cancel{Var(\bar{X}_n)}}{\cancel{Var(\bar{X}_n)}} = 1 \Rightarrow \boxed{Var(Z_n) = 1}$$

So

$$\boxed{Z_n \sim N(0,1)}$$

$\blacksquare$

Theorem: 14.3 (Central Limit Theorem):

For any $n \in \mathbb{N}$, let $X_1, X_2, \ldots$ be an iid sample from a distribution with finite expectation μ and finite variance σ^2.

Let

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

Then at any point $x \in \mathbb{R}$

$$\lim_{n \to \infty} F_{Z_n}(x) = \phi(x)$$

Convergence in distribution where ϕ is the distribution function of standard normal distribution.

<u>Example</u>: (Using smarties example, chapter 10):

Y: no of yellow smarties

$$Y = \sum_{i=1}^{n} Y_i \qquad Y_i = \begin{cases} 1 & \text{ith smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

$$\left[n = 40, \quad p_Y = \frac{1}{8}\right]$$

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$P(|\bar{Y}_n - p_Y| > 0.1) \rightarrow$ probability that difference of estimate $\bar{Y}_n$ and true value $p_Y = 1/8$ is greater than 0.1

estimate that

→ "probability that we are 0.1 away from our estimate

Since Y_i is an indicator random variable, it is a Bernoulli distributed so

$$E[Y_i] = \mu = p_Y \qquad \left(\begin{array}{c}\text{for } X \sim Ber(p) \\ E[X] = p\end{array}\right)$$

For bernoulli

$$\operatorname{Var}(Y_i) = p_Y(1-p_Y) > 0$$

$$P\left(|\bar{Y}_n - p_Y| > 0.1\right) = P\left(\frac{|\bar{Y}_n - p_Y|}{\sqrt{\frac{\operatorname{Var}(Y_i)}{n}}} > \frac{0.1}{\sqrt{\frac{\operatorname{Var}(Y_i)}{n}}}\right)$$

$$= P\left(\frac{|\bar{Y}_n - p_Y|}{\sqrt{\frac{(1-p)p}{n}}} > \frac{0.1}{\sqrt{\frac{(1-p)p}{n}}}\right)$$

let $Z_n = \dfrac{\bar{Y}_n - E[Y_i]}{\sqrt{\frac{\operatorname{Var}(Y_i)}{n}}} = \dfrac{\bar{Y}_n - p_Y}{\sqrt{\frac{(1-p)p}{n}}} \dot{\sim} N(0,1)$

let $\dfrac{0.1}{\sqrt{\frac{(1-p)p}{n}}} = b$

$$= P(|Z_n| > b)$$

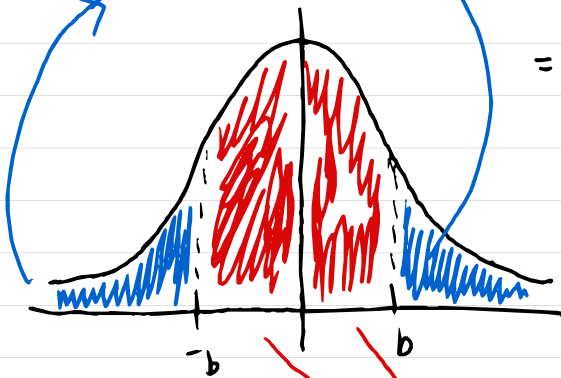
$P(Z_n > b)$ or (+) $P(-Z_n > b)$

$P(Z_n > b) + P(Z_n < -b)$

Need to calculate

$$P(|Z_n| > b) = P(|Z_n - 0| > b)$$

$$= P(Z_n > b) + P(Z_n < -b)$$



or

$$P(|Z_n| > b) = 1 - P(|Z_n| \leq b)$$

$$= 1 - P(-b \leq Z_n \leq b)$$

$$= 1 - \left(P(Z_n \leq b) - P(Z_n \leq -b)\right)$$

$$= 1 - P(Z_n \leq b) + P(Z_n \leq -b)$$

$$= P(Z_n > b) + P(Z_n < -b)$$

Due to symmetry of normal distribution,

$$P(Z_n > b) = P(Z_n < -b)$$ (check week 5 exercises for proof)

So

$$P(|Z_n| > b) = P(Z_n > b) + P(Z_n < -b)$$

$$= 2P(Z_n < -b)$$

$$= 2\,\Phi(-b)$$

↳ distribution fn of standard normal distribution

$$\approx 0.06$$

Calculating actual value, not approximation using central limit thm:

$$Y \sim Bin(n, p_Y) \Rightarrow Y \sim Bin(40, 1/8)$$

$$P(|\bar{Y}_n - p_Y| > 0.1) = P(n|\bar{Y}_n - p_Y| > 0.1n)$$

(since $n > 0$) $$= P(|n\bar{Y}_n - np_Y| > 0.1n)$$

$\left[\begin{array}{l} n\bar{Y}_n = Y \text{ as} \\ \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ Y = \sum_{i=1}^{n} Y_i \end{array}\right]$ $$= P(|Y - 5| > 4)$$

$$P(|Y-5|>4)$$

$$= 1-P(|Y-5| \leq 4)$$

$$= 1-[P(-4 \leq Y-5 \leq 4)]$$

$$= 1-(P(Y-5 \leq 4) - P(Y-5 \leq -4))$$

$$= 1-P(Y \leq 9) + P(Y \leq 1)$$

$$= 0.05487806$$

$$\approx 0.06$$

pretty accurate to approximation by central limit thm.

used pbinom(q, n, p) fn in R.