Sampling Distribution and the Central Limit Theorem

Goal: study the distribution of statistics defined using a random sample

Let $\{X_1, X_2, \dots, X_n\}$ be a random sample of size n ⟶ (i) X_i's have the same distribution for $i = 1, 2, 3, \dots, n$

(ii) The set $\{X_1, X_2, \dots, X_n\}$ is an independent set of random variables.

Let $X :=$ values of the random variables in the random sample, call it $X = \{X_1, X_2, X_3, \dots, X_n\}$

The joint sample space for the r.v's $X_1, X_2, \dots, X_n$ ⟶ is the n-Cartesian product of X

ie $X^n = \{(x_1, x_2, x_3, \dots, x_n) ; x_i \in X\}$

Since $X_1, X_2, \dots, X_n$ are independent r.v.s

⇓

the joint probability distribution of $X_1, X_2, X_3, \dots, X_n$

⇓

(i) $p(x_1, x_2, \dots, x_n) = p_X(x_1)\, p_X(x_2) \cdots p_X(x_n)$

if X_i's are discrete and $p_X(x)$ is the common pmf of all X_i's

(ii) $f(x_1, x_2, \dots, x_n) = f_X(x_1) \cdot f_X(x_2) \cdots f_X(x_n)$

if X_i's are continuous and $f_X(x)$ is the common pdf of all X_i's

For brevity, we will work with discrete r.v's.

⇓

continuous case is similar.

Let T be a statistic for the random sample $\{X_1, X_2, \dots, X_n\}$ ⟶ random variable calculated using r.v.s $X_1, X_2, \dots, X_n$

⇓

T is a function defined

$T : X^n \longrightarrow \mathbb{R}$, where $X^n =$ joint sample space of $X_1, X_2, \dots, X_n$

The probability distribution of the statistic T

⇓

"Sampling distribution of T"

To calculate the Sampling dist of T ⟶ use the joint distribution of the r.v's $X_1, X_2, \dots, X_n$

⇓

Let $p(x_1, x_2, \dots, x_n)$ be the joint pmf of $X_1, X_2, \dots, X_n$

(we are assuming the X_i's are discrete)

Want: the pmf of T ⟶ ie for $t \in Y$, want $P(T = t)$.

$Y :=$ Values of T

Let $E(t) = \{(x_1, x_2, \dots, x_n) \in X^n \,;\, T(x_1, x_2, x_3, \dots, x_n) = t\}$,

then

$$P(T = t) = \sum_{(x_1, x_2, \dots, x_n) \in E(t)} p(x_1, x_2, \dots, x_n)$$

$$= \sum_{(x_1, x_2, \dots, x_n) \in E(t)} p_X(x_1)\, p_X(x_2) \cdots p_X(x_n).$$

where $p_X(x)$ is the common pmf of $X_1, X_2, \dots, X_n$.

The Sample mean and sample total

Let $X = \{X_1, X_2, \dots, X_n\}$ ⟶ random sample coming from a population with mean $= \mu$, variance $= \sigma^2$

⇓

ie $E(X_i) = \mu$ for $i = 1, 2, \dots, n$

$V(X_i) = \sigma^2$ for $i = 1, 2, \dots, n$.

"The sample total" for $\{X_1, X_2, \dots, X_n\}$ ⟶ $T_0 := X_1 + X_2 + \cdots + X_n$

"The sample mean" for $\{X_1, X_2, \dots, X_n\}$ ⟶ $\overline{X} := \dfrac{X_1 + X_2 + \cdots + X_n}{n} = \dfrac{T_0}{n}$

Can Calculate:

$E(T_0) = E(X_1 + X_2 + \cdots + X_n)$

$= E(X_1) + E(X_2) + \cdots + E(X_n)$

$= \mu + \mu + \cdots + \mu$

ie $E(T_0) = n\mu$

$V(T_0) = V(X_1 + \cdots + X_n)$

$= V(X_1) + \cdots + V(X_n)$

$= \sigma^2 + \sigma^2 + \cdots + \sigma^2$

ie $V(T_0) = n\sigma^2$

Similarly,

$E(\overline{X}) = \mu$ and $V(\overline{X}) = \dfrac{\sigma^2}{n}$

Question: What is the sampling distribution of $\overline{X}$ and T_0

⇓

The Central Limit Theorem provides the answer!

Central Limit Theorem (Normal Case)

Suppose $\{X_1, X_2, \dots, X_n\}$ is a random sample coming from a population with Normal distribution $N(\mu, \sigma^2)$ ⟶ then for any $n \in \mathbb{N}$ the distributions of the sample mean $\overline{X}$ and the sample total T_0 are

$\overline{X} \sim N(\mu, \sigma^2/n)$

$T_0 \sim N(n\mu, n\sigma^2)$

Equivalently, the random variables

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad \frac{T_0 - n\mu}{\sqrt{n}\,\sigma}$$

have the standard normal Distribution

Note: for CLT to hold for all $n \in \mathbb{N}$ (especially when n is small)

⇓

need population to be normally dist.

Can drop normality assumption if n is large.

Central Limit Theorem: (General Case)

If $\{X_1, X_2, \dots, X_n\}$ is a random sample from a population with mean $= \mu$, variance $= \sigma^2$ ⟶ then if n is large enough, the distribution of $\overline{X}$ and T_0 is "approximately normal"

ie $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, $T_0 \sim N(n\mu, n\sigma^2)$

the approximation improves as $n \to \infty$.

⇓

the distributions of $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ and $\dfrac{T_0 - n\mu}{\sqrt{n}\,\sigma}$ approaches the standard normal distribution as $n \to \infty$.

Example: Suppose the population $= \{0, 1\}$

The distribution of this population is the Bernoulli dist.

ie $P(X = 1) = P(\text{success}) = p = p_X(1)$

$P(X = 0) = P(\text{failure}) = 1 - p = p_X(0)$

Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from $\{0, 1\}$

Goal: Calculate the sampling distribution of T_0

Suppose $n = 2$, then $X^2 = \{(0,0), (0,1), (1,0), (1,1)\}$

Values of $T_0 = \{0, 1, 2\}$

$p(x,y) = p_X(x)\, p_X(y)$ ⟶ be the joint pmf of X_1 and X_2

then

$P(T_0 = 0) = p(0,0) = p_X(0) \cdot p_X(0) = (1-p)^2$

$P(T_0 = 1) = p(1,0) + p(0,1) = p_X(0) p_X(1) + p_X(1) p_X(0)$

$= (1-p)p + p(1-p) = 2p(1-p)$

$P(T_0 = 2) = p(1,1) = p_X(1) p_X(1) = p^2$

Putting this into a table gives

t	0	1	2
$P(T_0 = t)$	$(1-p)^2$	$2p(1-p)$	p^2

⟶ notice any binomial structure?

Suppose $n = 3$:

Now $X^3 = \{(0,0,0), (1,0,0), (1,1,0), (1,1,1), (0,1,0), (1,0,1), (0,0,1), (0,1,1)\}$

$p(x_1, x_2, x_3) = p_X(x_1) \cdot p_X(x_2) \cdot p_X(x_3)$ and

Values of $T_0 = \{0, 1, 2, 3\}$

$\therefore P(T_0 = 0) = p(0,0,0) = p_X(0) \cdot p_X(0) \cdot p_X(0) = (1-p)^3$

$P(T_0 = 1) = p(1,0,0) + p(0,1,0) + p(0,0,1)$

$= p_X(1) p_X(0) p_X(0) + p_X(0) p_X(1) p_X(0) + p_X(0) p_X(0) p_X(1)$

$= p(1-p)(1-p) + (1-p)p(1-p) + (1-p)(1-p)p$

$= 3(1-p)^2 p$

$P(T_0 = 2) = p(1,1,0) + p(0,1,1) + p(1,0,1)$

$= p_X(1) p_X(1) p_X(0) + p_X(0) p_X(1) p_X(1) + p_X(1) p_X(0) p_X(1)$

$= 3p^2(1-p)$

$P(T_0 = 3) = p(1,1,1) = p_X(1) \cdot p_X(1) \cdot p_X(1)$

$= p^3$

The probability distribution table of T_0 is

t	0	1	2	3
$P(T_0 = t)$	$(1-p)^3$	$3(1-p)^2 p$	$3(1-p)p^2$	p^3

Suppose $n = 4$,

Now $X^4 = \{(0,0,0,0), (1,0,0,0), (1,1,0,0), (1,1,1,0), (1,1,1,1), \dots, (0,0,0,1), (0,0,1,1), (0,1,1,1)\}$

The joint dist is $p(x_1, x_2, x_3, x_4) = p_X(x_1) \cdot p_X(x_2) \cdot p_X(x_3) \cdot p_X(x_4)$

Values of $T_0 = \{0, 1, 2, 3, 4\}$

The distribution of T_0 can be calculated as

t	0	1	2	3	4
$P(T_0 = t)$	$(1-p)^4$	$4(1-p)^3 p$	$6(1-p)^2 p^2$	$4(1-p)p^3$	p^4

In general for arbitrary n

Values of $T_0 = \{0, 1, 2, 3, \dots, n\}$ and

$P(T_0 = k) = \binom{n}{k} (1-p)^{n-k} \cdot p^k \quad k = 0, 1, 2, \dots, n$

⇓

pmf of the Binomial distribution with parameters n and p

$\therefore T_0 \sim Bin(n, p)$ ⟶ makes sense because "T_0 counts the # of 1's (ie successes) for given sample of size n, where each sample entry is an outcome of independent Bernoulli trial."

⇓

The Binomial Experiment.

$\because T_0 \sim Bin(n, p) \implies E(T_0) = np$, $V(T_0) = np(1-p)$

Also the central limit theorem says,

for large enough n ⟶ $T_0 \sim N(n\mu_X, n\sigma_X^2)$, $N(np, np(1-p))$

where $\mu_X = \text{mean}(X) = p$, $\sigma_X^2 = \text{Var}(X) = p(1-p)$

⇓

Since $T_0 \sim Bin(n, p) \implies$ for large n, $Bin(n, p) \sim N(np, np(1-p))$

⇓

Normal approximation for Binomial

$\{X_1, X_2, X_3, \ldots, X_n\}$ ⟶ (i) X_i's have the same dist

↓

random sample of size n ⟶ (ii) $X_1, X_2, \ldots, X_n$ are ind r.v's.

A statistic ⟶ random variable calculated using a random sample.

↓

Can calculate the distribution of the statistic

↓

"sampling distribution".

Suppose $\mathcal{X}$ = set of values of r.v in $\{X_1, \ldots, X_n\}$

joint sample space of $X_1, X_2, \ldots, X_n$ $= \mathcal{X}^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \mathcal{X}\}$

Can calculate the "joint probability distribution"

Assume that the X_i's are discrete ⟶ want the joint pmf of $X_1, X_2, \ldots, X_n$.

∴ joint pmf: $p : \mathcal{X}^n \longrightarrow [0,1]$

$$p(x_1, x_2, x_3, \ldots, x_n) = p_X(x_1) \cdot p_X(x_2) \cdots p_X(x_n)$$

where $p_X(x_i)$ ⟶ pmf of the common dist of all X_i's

$p_X : \mathcal{X} \longrightarrow [0,1]$.

Let T be a statistic ⟶ $T : \mathcal{X}^n \longrightarrow \mathbb{R}$

to calculate the pmf T ⟶ want $P(T=t)$ for $t \in \mathcal{T}$ = values of T.

(i) Find out what the values of T are.

(ii) use the joint pmf to calculate $P(T=t)$

$t \in \mathcal{T}$, $E(t) = \{(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n \mid T(x_1, x_2, \ldots, x_n) = t\}$

$= T^{-1}(t) \subseteq \mathcal{X}^n$

Since $E(t) \subseteq \mathcal{X}^n$ ⟶ know how to calculate $P(E(t))$

$$\therefore P(T=t) = P(E(t)) = \sum_{(x_1, x_2, \ldots, x_n) \in E(t)} p(x_1, x_2, \ldots, x_n).$$

Example: Population = $\{0, 1\}$

$p_X(1) = p$, $p_X(0) = 1-p$

x	0	1
$P(X=x)$ $p_X(x)$	$(1-p)$	p

Get a random sample of size 2

↓

$\{X_1, X_2\}$ X_1 and X_2 have the Bernoulli dist defined X.

$\mathcal{X} = \{0, 1\}$

joint sample space $= \{(0,0), (0,1), (1,0), (1,1)\}$

joint pmf: $p(x,y) = p_X(x) \cdot p_X(y)$

$\Rightarrow$ $p(0,0) = p_X(0)\, p_X(0) = (1-p)(1-p) = (1-p)^2$

$p(0,1) = p_X(0) \cdot p_X(1) = (1-p) \cdot p$

$p(1,0) = p_X(1) \cdot p_X(0) = p(1-p)$

$p(1,1) = p_X(1) \cdot p_X(1) = p^2$

X \ Y	0	1
0	$(1-p)^2$	$p(1-p)$
1	$p(1-p)$	p^2

Sample mean $:= \bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n} = \dfrac{T_0}{n}$

Sample Total $:= T_0 = X_1 + X_2 + \cdots + X_n$

$\mathcal{X}^2 = \{(0,0), (0,1), (1,0), (1,1)\}$

Want: values of $\bar{X}$ and T_0

Values of $\bar{X} = \{0, \frac{1}{2}, 1\}$

$\bar{X}(0,0) = \frac{0+0}{2}$, $\bar{X}(1,0) = \bar{X}(0,1) = \frac{1}{2}$ $\bar{X}(1,1) = \frac{1+1}{2} = 1$

$T_0(0,0) = 0$ $T_0(0,1) = T_0(1,0) = 1+0 = 1$ $T_0(1,1) = 1+1 = 2$

Values of $T_0 = \{0, 1, 2\}$

To calculate the distribution of $\bar{X}$ and T_0

↓

use the joint pmf $p(x_1, x_2)$

$E(0) = \{(0,0)\}$

$P(T_0 = 0) = p(0,0) = (1-p)^2$

$E(1) = \{(0,1), (1,0)\}$

$P(T_0 = 1) \longrightarrow = p(0,1) + p(1,0) = 2 \cdot p(1-p)$

$E(2) = \{(1,1)\}$

$P(T_0 = 2) = p(1,1) = p^2$

t	0	1	2
$P(T_0 = t)$	$(1-p)^2$	$2p(1-p)$	p^2

↝ this dist table for $\text{Bin}(n=2, p=p)$

Calculating the distribution table for $\bar{X}$.

Values of $\bar{X} = \{0, \frac{1}{2}, 1\}$

$P(\bar{X} = 0) = p(0,0) = (1-p)^2$

$P(\bar{X} = \frac{1}{2}) = P(T_0 = 1) = 2(1-p)p$

$P(\bar{X} = 1) = P(T_0 = 2) = p^2$

$\bar{x}$	0	½	1
$P(\bar{X} = \bar{x})$	$(1-p)^2$	$2p(1-p)$	p^2

General formula for the sampling dist of $\bar{X}$ and T_0

↓

using the Central Limit theorem.

(I) CLT (Normal case)

If $\{X_1, X_2, \ldots X_n\}$ coming from a normal dist $N(\mu, \sigma^2)$ ↝ for all $n \in \mathbb{N}$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$T_0 \sim N(n\mu, n\sigma^2)$

Can show that

$E(\bar{X}) = \mu \quad V(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

$E(T_0) = n\mu \quad V(X) = n\sigma^2 \Rightarrow \sigma_{T_0} = \sqrt{n}\,\sigma$

CLT in normal case $\Rightarrow$ $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ and

$\frac{T_0 - n\mu}{\sqrt{n} \cdot \sigma}$

both have the standard normal dist.

(II) CLT (General Case)

Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample coming from a population with

mean = μ

variance = σ^2

↝ If n is large $\bar{X}$ and T_0 have approximately normal distributions.

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$T_0 \sim N(n\mu, n\sigma^2)$

as $n \to \infty$, the approx gets better

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ and $\frac{T_0 - n\mu}{\sqrt{n}\,\sigma}$ "approach" the standard normal dist as $n \to \infty$.

Example (Revisited).

Suppose r.s of size n coming $\{0,1\}$

$\{X_1, X_2, X_3, \ldots, X_n\} \quad \chi = \{0,1\}$

sample data $= \chi^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \{0,1\}\}$

$|\chi^n| = 2 \times 2 \times 2 \times 2 \times \ldots \times 2 = 2^n$

typical element in χ^n ↝ $(0,0,1,1,0,0,\ldots,1,\ldots,0)$

Values of T_0: $\{0,1,2,3, \ldots, n\}$

$T_0(1,1,1,\ldots,1) = 1+1+1+\cdots+1 = n$

$T_0(0,0,0,\ldots,0) = 0+0+\cdots+0 = 0$

$$\{T_0 = k\} = \{(x_1, \dots, x_n) \mid k \text{ } x_i\text{'s are equal to } 1\}$$

$$\Rightarrow |\{T_0 = k\}| = \binom{n}{k}$$

$$p(\underbrace{1,1,1,\dots,1}_{k},0,0,0) = p^k \cdot (1-p)^{n-k}$$

$$\therefore P(T_0 = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

$\Downarrow$

pmf of $Bin(n, p)$

$\therefore$ $T_0 \sim Bin(n,p)$ + CLT $\Rightarrow$ for large n

$$Bin(n, p) \sim N(np, np(1-p))$$

Values of $\bar{X}$ for $\{X_1, X_2, \dots, X_n\}$

Values of T_0: $\{0, 1, 2, \dots, n\}$

$\therefore$ Values of $\bar{X}$: $\{0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, 1\}$