Lecture (2)

Sampling Distribution

First we should understand the difference between Population & sample

Population : Collection of all items of interest
[set of all elements under discussion or inverstigation]

+ Values we obtain from Population are Called **Parameters**

Sample subset of population

+ values we obtain from sample Called **statistics**

Ex to study IQ level of 700,000 student

Population
every student (in campus, at home, part time, ...)
⇒ hard to define
⇒ hard to observe in real life
⇒ hard to contact

Sample
any subset of 100 student
⇒ Much easier to gather
⇒ less time Consuming
⇒ less cost
But Choice of Sample is really a big deal

That's why most of statistics studies work on sample of data

Sample must have two characteristics:

Random (Indep)

each member of sample is chosen from population by chance (unbiased)

representative

Sample should reflect accurately the member of entire population

Ex Survey on student for the entire school
You Contact your class mates
∴ the group you study called Sample, the values you get called statistics

---

Random sample

Ex measuring IQ level of 700,000 student
Choose a sample of 100 student

Sample1 : $X_1, X_2, \dots X_n$ Random variable

$n = 100$ sample size

$x_1, x_2, \dots, x_n$ observed values of a random sample

observations readings values

Sample2 : $X_1, X_2, \dots, X_n$

$x_1, x_2, \dots, x_n$ different observed values

Conditions of a Random Sample

[1] $X_1, X_2, \dots, X_n$ are indep. R.V.

$$f(x_1, x_2, \dots x_n) = f(x_1)\, f(x_2) \dots \dots f(x_n)$$

joint pdf — marginal pdfs

this means that selection #i does not depend on selection #i

(1^{st} selection does not affect $2^{nd}, 3^{rd}, \dots$)

[2] $X_1, X_2, X_3, \dots, X_n$ have the same distribution (identically distributed)

$$E(X_1) = E(X_2) \dots = E(X_n) = \mu \rightarrow \text{Population mean}$$

$$V(X_1) = V(X_2) \dots = V(X_n) = \sigma^2 \rightarrow \text{Population variance}$$

X_1: R.V. can take any value in each R.S.

$$\mu_{\text{population mean}} = \frac{\sum (X_i)}{N \rightarrow 700{,}000}$$

$\sum (X_i)$ ← all iQ level of all 700,000 students

In most cases μ is unknown

So we use estimate for μ called $\hat{\mu}$ (ch.2, ch.3)

But $E(X_i) = \mu$ identically distributed

$V(X_i) = \sigma^2$

From 1, 2 Any random sample

S: $X_1, X_2, \cdots, X_n$ are i.i.d

independent identically distributed

Statistics It is a function of observable R.V.'s which itself an observable R.V. and doesnot contain any unknow parameter

For example, If $X_1, X_2 \cdots, X_n$ R.S.

then $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$\overline{X}$: Sample mean; n: sample size

$$S^2 = \frac{1}{n-1} \Sigma (x_i - \overline{x})^2$$

S^2: Sample Variance

$\overline{X}, S^2$ are R.V.s too

$\overline{X}, S^2$ are statistics

but μ, σ^2 are Parameters (Constants)

Note

$\mu \neq \bar{X}$

population mean $\neq$ sample mean

but if we take a good sample we can converge to μ

- $\hat{\mu}$ = value — Point estimation ch.2
- $a < \hat{\mu} < b$ — Confidence interval ch.3

→ Each sample of same size n

Each sample has $\bar{X}$, S^2

$\bar{X}$, S^2 are R.v.'s too

$$E(\bar{X}) = E\left(\frac{\sum x_i}{n}\right)$$

$$= \frac{1}{n} E(x_1 + x_2 \cdots + x_n)$$

$$= \frac{1}{n} [E(X_1) + E(X_2) \cdots + E(X_n)]$$

($X_1, X_2, \ldots$ identically distributed; $E(X_i) = \mu$)

$$= \frac{1}{n} \cdot n\mu = \mu$$

$\therefore \boxed{E(\bar{X}) = \mu}$

average Mean of Sample mean = Population mean

$$V(\bar{X}) = V\left(\frac{\sum X_i}{n}\right)$$

$$= \frac{1}{n^2} V(X_1 + X_2 + \cdots X_n)$$

($X_1, X_2, \ldots$ identically distr.)

$$= \frac{1}{n^2} [n\sigma^2]$$

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

variance of sample mean = $\frac{\text{Population var}}{\text{sample size}}$

The distribution of $\bar{X}$

If $X_1, X_2, \dots X_n$ a R.S. from a normal population with mean μ, variance σ^2

if Population $\sim$ Normal(μ, σ^2)

then $\bar{X} \sim$ Normal $\left(\mu, \frac{\sigma^2}{n}\right)$

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

this means that any R.S. from normal Population $\overset{\text{Normal}(\mu,\sigma^2)}{x}$ have sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where n is sample size

what happen if population is not normal

$\Rightarrow$ **Central Limit theorem**

If $X_1, X_2, \dots, X_n$ a R.S. from any population with mean μ and variance σ^2 then for large enough n $(n \geq 30)$

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

$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ as $n \geq 30$

Ex Certain tubes manufacture by a company have a mean lifetime of 900 hrs and standard deviation of 50 hrs

Find the probability that a random sample of 64 tubes taken from the group will have mean lifetime between 895, 910 hrs

Population parameters $\leftarrow \begin{cases} \mu = 900 \text{ hrs} \\ \sigma = 50 \text{ hr} \end{cases}$

Since $n = 64 \geq 30$ large enough

by Central Limit theorem

$$\overline{X} \sim N\left(\mu_{\overline{x}} = \mu = 900, \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{8}\right)$$

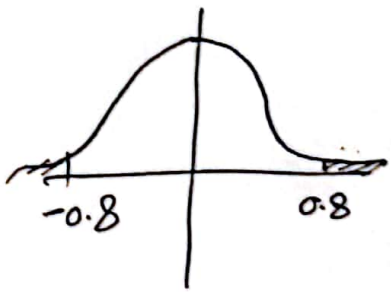
$$\overline{X} \sim N\left(900, \frac{50}{8}\right)$$

$Pr(895 < \overline{X} < 910)$ by standarization

Convert $\overline{X}$ to Z (standard normal)

$$Pr\left(\frac{895-900}{\frac{50}{8}} < \frac{\overline{X}-\mu_{\overline{x}}}{\sigma_{\overline{x}}} < \frac{910-900}{\frac{50}{8}}\right)$$

$$P(-0.8 < Z < 1.6)$$



$$= \Phi(1.6) - \Phi(-0.8)$$

$$= \Phi(1.6) - [1 - \Phi(0.8)]$$

$$= \Phi(1.6) + \Phi(0.8) - 1 = 0.733$$

using table or any package

As $n \to \infty$ $\overline{x} \to 900$