09-Jun-23

DSA 503: Statistical Inference

- Estimation
- Test of Hypothesis

A population consists of all elements; individuals, items or objects - whose characteristics are being studied.

<u>Sample</u>: A $\boxed{\text{representative}}$ part of population

Ex: Blood in the body is population
The small part of it is sample. [For blood testing]

- There are some cons to work with population
  - Time consuming
  - Expensive
  - Sometimes infeasible.

<u>Parameter</u>: Population characteristics
Ex: Population mean (μ)

$\boxed{\text{Statistic:}}$ Sample characteristics
Ex: Sample mean ($\bar{x}$)

Not statistics

The population distribution is the probability distribution of the population data.

- Normal
- Binomial
- Poisson

$\sum$ Relative frequency = 1

Frequency distribution

$\sum P(x) = 1$

$\boxed{0 \leq \text{Probability} \leq 1}$

Probability range

$\boxed{\text{Probability Distribution}}$

→ ALWAYS for population

$\boxed{\text{Sampling Distribution}}$

→ probability distribution of $\boxed{\text{statistic}}$

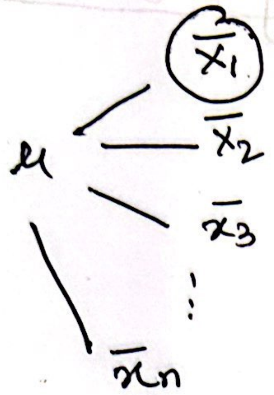
probability distribution of $\bar{x}$.

Population 5, Sample 3.

Can be taken in ${}^5C_3$ ways

$${}^nC_r = \frac{n!}{r!(n-r)!}$$

$${}^5C_3 = \frac{5!}{3!\,2!} = \frac{5 \cdot \cancel{4}^{2} \cdot \cancel{3 \cdot 2 \cdot 1}}{\cancel{3 \cdot 2 \cdot 1} \cdot \cancel{2} \cdot 1} = 10$$



← This is constant

but $\bar{x}$ altogether is variable

$\boxed{\text{If something is variable, it has a distribution.}}$
↳ a prob.

All sampling distributions are probability distribution, but all prob. dist. are not sampling distribution.

16-Jun-23

Statistical Inference

- Estimation
  - Point
  - Interval
- test
  - → MCA
  - → Proportion
  - → Chi square
  - → ANOVA F
  - → Non-parametric test

$P = \frac{X}{N}$ and $\hat{p} = \frac{x}{n}$

20% diabetic patients in a class, the proportion is 0.2.

$P \rightarrow$ population proportion (parameter)

Population mean (parameter) $\leftarrow \mu$

$\mu \rightarrow \bar{x}$

$P \rightarrow \hat{p} \rightarrow$ sample proportion (statistic)

Sample mean (statistic) rv — $\bar{x}$

যেহেতু statistic, so এটা random variable

$\bar{x} \rightarrow P(\bar{x})$

$\hat{p} \rightarrow P(\hat{p}) = ?$

$P(\bar{x})$, $P(\hat{p}) \rightarrow$ Sampling distribution

$\hat{p}$ is a point estimate.

$\hat{p}$	f	$P(\hat{p})$
0.33	3	$3/10 = 0.3$
0.67	6	$6/10 = 0.6$
1	1	$1/10 = 0.1$
		$\sum P(\hat{p}) = 1$

$$\boxed{\begin{aligned} &\text{If } X \overset{iid}{\sim} N(\mu, \sigma^2) \\ &\rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \end{aligned}}$$

	$\bar{x}$	$\hat{p}$
Mean	$\mu_{\bar{x}} = \mu$	$\mu_{\hat{p}} = p$
Variance	$\sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}$	$\sigma^2_{\hat{p}} = \dfrac{pq}{n}$
SD	$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$	$\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}$; $q = (1-p)$

$$\boxed{\begin{aligned} V(cx) &= c^2 V(x) \\ E(cx) &= cE(x) \end{aligned}}$$

Proof:

$$\text{Mean}(\bar{x}) = \mu_{\bar{x}} = E(\bar{x}) = E\left(\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}\right)$$

$$= \frac{E(x_1) + E(x_2) + E(x_3) + \dots E(x_n)}{n}$$

$$= \frac{\mu + \mu + \mu + \mu + \mu}{n}$$

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

$$= \mu$$

Proof:

$$V(\bar{x}) = \sigma_{\bar{x}}^2 = V(\bar{x}) = V\left(\frac{x_1+x_2+\dots+x_n}{n}\right)$$

$$= \frac{V(x_1)+V(x_2)+V(x_3)+\dots+V(x_n)}{n^2}$$

$$= \frac{\sigma^2+\sigma^2+\dots+\sigma^2}{n^2}$$

$$= \frac{n\sigma^2}{n^2}$$

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

$\boxed{\text{All} \rightarrow \text{population}}$

Shape needs to be understood.

Shape depends on SD.

$$z = \frac{x - \underset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

Normal

$\downarrow$

standard normal

Mean = 0

SD = 1

$$E(z) = \frac{E(x)-\mu}{\sigma}$$

$$= \frac{\mu-\mu}{\sigma} = \frac{0}{\sigma} = 0$$

$$V(z) = \frac{V(x)-0}{\sigma^2}$$

$$= \frac{\sigma^2}{\sigma^2} = 1$$

$np > 5$ and $nq > 5$

t- distribution is used when SD of population is not known.

t: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Chi-square : Sum of square of ind. std normal variables

F:

23-Jun-23

যেটা দিয়ে estimate করার হয় → Estimator

The statistic which is used to estimate a population.

The $\boxed{\text{value}}$ of estimator is estimate.

Minimum variance unbiased estimator (MVUE)

Qualities → Good estimator

- → Unbiased
- → Consistent
- → Low MSE (Mean Square Error)
- → Low MVUE

MLE $\rightarrow$ Maximum Likelihood Estimate

Poisson Distribution $\rightarrow$ For discrete variables
$\hookrightarrow$ No of accidents

$$\mu = \lambda$$
$$\bar{x} = \bar{Y}$$

$$\hat{\theta}_{MLE} = \bar{Y}$$

$$\hat{\theta}_{MOM} = \hat{\theta}_{MLE} = \hat{\theta}_{LSE} = \bar{Y}$$

When sigma is unknown

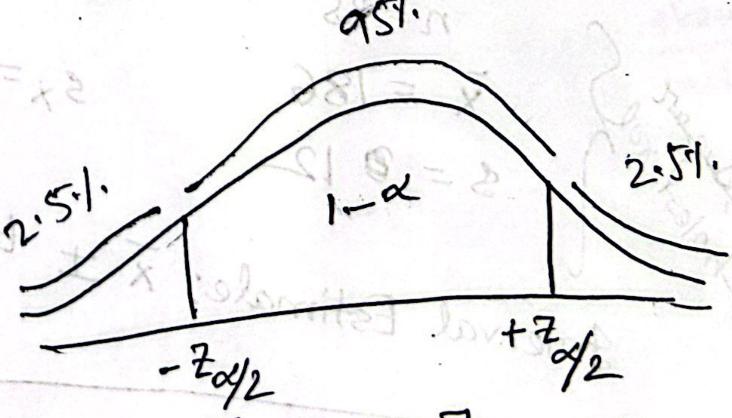
	Population Mean (μ)		~~Population Proportion~~
		Population	Sample
Point Estimate	$\bar{x}$		
Interval Estimate	$\bar{x} \pm ME \Rightarrow \bar{x} \pm z\sigma_{\bar{x}}$		$\bar{x} \pm t s_{\bar{x}} = \bar{x} \pm t\frac{s}{\sqrt{n}}$
Margin of error	$z\sigma_{\bar{x}} = z\frac{\sigma}{\sqrt{n}}$		$t s_{\bar{x}} = t\frac{s}{\sqrt{n}}$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$z\sigma_{\bar{x}} = \bar{x} - \mu$$

$$\mu = \bar{x} - z\sigma_{\bar{x}}$$

Confidence Interval (Proof :)



$$P\left[-z \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le +z\right] = 1 - \alpha$$

Population Proportion

$\hat{p}$ → sample standard deviation of $\hat{p}$

Interval Estimate: $\hat{p} \pm z(s_{\hat{p}}) = \hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Margin of error: $zs_{\hat{p}} = z\sqrt{\frac{\hat{p}\hat{q}}{n}}$

$df = n-1$

proportion always হয় Sample হয়- হয়, তাই Sample থেকে proportion draw করলে inaccurate হয়।

$z = \frac{\hat{p} - p}{s_{\hat{p}}}$

$p = \hat{p} - zs_{\hat{p}}$ $\quad \boxed{p = \hat{p} \pm zs_{\hat{p}}}$

Textbook:

$n = 25$

$\bar{x} = 145$

$\cancel{\sigma_{\bar{x}} = 35}$

$\boxed{\sigma = 35}$

ALL SUCH TEXTBOOKS population SD হবে

Sugar / Cholesterol:

$n = 25$

$\bar{x} = 186$

$s = 12$

$s_x = \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{25}}$

Preferably, just Look at নিচের Confidence Interval

$df = 24$

For these maths, always use the two-tail table.

Interval Estimate: $\bar{x} \pm ts_x$

* If সব দেয়া থাকে, ALWAYS go for the sigma because population দেয়া থাকলে fuck sample

503   7-Jul-23

- Sir reviewed previous classes (Point & Interval Estimates if σ is known/unknown & for proportion)
- Sir did the textbook math again
- Sir " " Dr. Moore Cholesterol math again
  - If we construct a lot of samples, 95% of the parameters will follow under this interval estimate)

Confidence level বাড়লে interval টা বড় হবে।
means Margin of error বড় হবে।

$\bar{x} \pm z\sigma_{\bar{x}}$

For 95%, $z = 1.96$

For 99%, $z = 2.58$

For increasing confidence interval, the Margin of error increases.

proportion শুধু large sample এর জন্য।
কিন্তু কীভাবে বুঝব এটা large sample?

$\boxed{n\hat{p} > 5 \ \& \ n\hat{q} > 5}$

P হবে আমি শেয়ার নেই।

Example math of slide (Religion)

$n = 1000$

$\hat{p} = 0.44; \ \hat{q} = 0.56$

এখন, question change করে যদি এমন দেয়, who will say religion NOT important তখন 0.56 হয়ে যাবে $\hat{p}$.

a) Point estimate of P is $\hat{p}$ 0.44

b) CI = 99%.

Interval Estimate $= \hat{p} \pm z s_{\hat{p}} = \hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}}$ → Margin of error

Alcohol Math

$n = 2401, \ \hat{p} = 0.25, \ \hat{q} = 75\%,$

$n\hat{p} = 2401 \times 0.25$ $\quad n\hat{q} = 2401 \times 0.75$

a) $\hat{p} = 0.25$

b) $CI = \hat{p} \pm z s_{\hat{p}} = \hat{p} + z\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Quiz on 14th July – On L1, 2, 3, 4

Method of Moment (MOM)

$\mu_r' = m_r'$

rth popula

First raw moment, always population mean

$Y_i \sim$ Poisson (λ)

mean $= E(Y)$

Mean = Variance = λ

By MOM,

$\mu_1' = m_1'$

$E(Y_1) = \frac{1}{n}\sum_{i=1}^{n} Y_1$

$\lambda = \bar{Y}$ (Sample Mean)

$\hat{\lambda}_{MOM} = \bar{Y}$

poisson → count data
(variable)
For discrete data; i.e.
No of accidents

Method of Maximum Likelihood: (ML)

$L(\theta_1, \theta_2, \dots \theta_k ; y_1, \dots y_n) = \prod_{i=1}^{n} f(y_i ; \theta_1, \theta_2, \dots, \theta_k)$

Semi-colon

population parameter

$L(\underset{\sim}{\theta} ; \underset{\sim}{y})$

y যেহেতু n সংখ্যক, এজন্য n সংখ্যক functions আছে।

$f(y_1 ; \theta_1, \theta_2, \theta_3, \dots \theta_k)$

$f(y_2 ; \theta_1, \theta_2, \theta_3, \dots \theta_k)$

$f(y_3 ; \theta_1, \theta_2, \theta_3, \dots \theta_k)$

আমরা estimate করতে চাচ্ছি θ

Said as Likelihood of θ factor & y factor

Discrete → pmf
population mass function

Continuous - pdf
population density function

$\boxed{\text{Data has to be independent}}$

Likelihood is product of marginal probability function.

$$L(\theta; \underset{\sim}{y}) = \prod_{i=1}^{n} f(y_i; \theta)$$ → For single parameter

$$Y_i \overset{iid}{\sim} \text{Poisson}(\theta)$$

Y_i follows poisson distribution with θ parameter

$$f(y_i, \theta) = \frac{e^{-\theta}\,\theta^{y}}{y!}$$

Def^n $L(\theta; \underset{\sim}{y}) = \prod_{i=1}^{n} f(y_i, \theta)$

$$= \prod_{i=1}^{n} \frac{e^{-\theta} \cdot \theta^{y_i}}{y_i!}$$

এর সবগুলোর উপরে product apply হবে

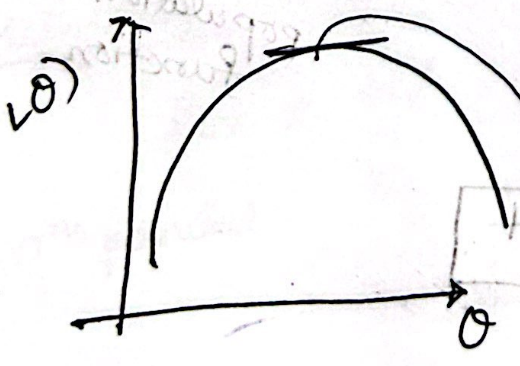
$$= \frac{e^{-n\theta}\,\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

$$\boxed{\begin{aligned} \prod_{i=1}^{2} e^{-\theta} &= e^{-\theta} \cdot e^{-\theta} \\ &= e^{-2\theta} \end{aligned}}$$

$$\boxed{\begin{aligned} \prod_{i=1}^{2} \theta^{y_i} &= \theta^{y_1} \cdot \theta^{y_2} \\ &= \theta^{y_1 + y_2} \end{aligned}}$$

$\frac{dy}{dx}$ → x এর সাপেক্ষে y এর rate of change.

$\frac{d\,L(\theta)}{d\theta}$ → θ এর সাপেক্ষে Likelihood.



→ Likelihood maximum কখন?

Likelihood is not a probability function, it just estimates which one of the parameters is more likely.

$Y_i \overset{iid}{\sim}$ Bernoulli (θ)

Bernoulli $\rightarrow$ Binomial but $n=1$

$\theta \rightarrow$ Probability of success

if $Y_i \overset{iid}{\sim}$ Binomial (θ)

$\theta \rightarrow$ Prob. of success

For Binomial, $f(y_i,\theta) = {}^{n}C_{y_i}\, \theta^{y_i} (1-\theta)^{1-y_i}$

For Bernoulli, $f(y_i,\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$

~~Case~~

$$L = \prod_{i=1}^{4} f(y_i;\theta)$$

$$= \prod_{i=1}^{4} \theta^{i} (1-\theta)^{1-y_i}$$

$$= \theta^{\sum y_i} (1-\theta)^{4-\sum y_i}$$

$\sum y_i \leftarrow$ কতবার total জিতছে

(0,0,0,0)

<u>Case-I:</u>

For $\theta = \theta_1 = 0.3 \Rightarrow L = (0.3)^0 (0.7)^{4-0} = 0.2401$

$\theta_2 = 0.8 \Rightarrow L = (0.8)^0 (0.2)^4 = 0.0016$

$\theta = 0.3$ is more likely.

Case II : (1, 1, 1, 0)

HW

Construct Likelihood for

$Y_i \overset{iid}{\sim}$ Binomidl (m, p) → known

$Y_i \overset{iid}{\sim}$ Poisson (λ)

$Y_i \overset{iid}{\sim}$ Normal (μ, σ^2)

iid = independent & identically distributed

Normal:

$$f(Y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

$$-\infty < Y < \infty$$

$$-\infty < \mu < \infty$$

$$\sigma^2 > 0$$

Mean = μ

Variance = σ^2

Binomial:

$$f(y; p) = {}^{m}C_{y}\, p^{y} (1-p)^{m-y}$$

Mean = mp

Variance = mpq