Statistics II - Ankit Pangeni

Unit 1 : Sampling distribution & estimation

Population : A set of similar items or events that we take for study of statistics. It is a collection of data from which a statistical sample is drawn for study.

Parameter : A parameter is a value (piece of data) that tells you something about whole population. It is more reliable than sample.

Statistic : A piece of data that describes only the sample of a population.

Sampling : The process of selecting a sample from a population for the study or research is called sampling.

Sampling distribution : It is a probability distribution of statistic obtained from a larger number of samples drawn from specific population. It arrives through repeated sampling from a larger population.

Sampling distribution of mean :

The distribution of the values of the sample mean in repeated samples is called sampling dist. of mean (यसलाई population को repeated samples को लिएको sample mean हरुको probability distribution भनिन्छ sampling distribution of mean भनिन्छ.

$X \sim N(\mu, \sigma_{\bar{x}}^2)$

* If sample is drawn from finite population (replacement is not done)

σ = population S.D N = pop^n size
σ^2 = population variance. n = sample size

mean $(E(\bar{x})) = \mu$, $V(\bar{x}) = \frac{\sigma^2 (N-n)}{n(N-1)}$

* If sample is from infinite pop^n, $\frac{n}{N}$ can be neglected (replacement is possible)

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

S.E $= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

Properties:

i) Sample mean $(\bar{x})$ is unbiased estimate of population mean (μ).

ii) The variance $v(\bar{x})$ depends on the sample size (n) & is equal to $\frac{\sigma^2}{n}$.

Standard error $= \sqrt{v(\bar{x})} = \sqrt{\frac{\sigma^2}{n} \frac{(N-n)}{N-1}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ (finite pop^n)

Sampling distribution of proportion

A proportion is the number of elements with a given characteristics divided by the total number of elements in that group.

population proportion $P = \frac{X}{N}$, $Q = 1 - P$

X = no. of characteristics obtained in pop^n

N = population size.

sample proportion $p = \frac{x}{n}$ x = " " in sample

n = sample size.

To study any qualitative study, we need to estimate population proportion. To estimate population proportion we use sample proportion. The distribution followed by sample proportion is called sampling distribution of Prop?

* If population is infinite, (with replacement)

$E(p) = P$, $Var(p) = \frac{PQ}{n}$, Standard error $SE(p) = \sqrt{\frac{PQ}{n}}$

* If Population is finite. (without replacement)

$E(p) = P$, $Var(p) = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$

Standard error $SE(p) = \sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}$

~~properties: same~~

or $\sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}$

Central limit theorem

It states that, as the sample size gets large enough, the sampling distribution of mean is approximately normally distributed. It is true regardless of the shape of distribution of the individual values in the population.

If $x_1, x_2, x_3, \ldots, x_n$ is a random sample of size n of any population,

Sample mean $(\bar{x}) = \frac{x_1 + x_2 + \ldots + x_n}{n}$

such that it is normally distributed, then

$E(\bar{x}) = \mu$

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

Q.5

σ = population S.D (Standard deviation)
s = sample S.D

Date
Page

Important Formula.

Statistic	Standard Error
Mean(σ known, pop^n size infinite)	$S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$
Mean(σ known, pop^n size finite)	$S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Mean(σ unknown, pop^n infinite)	$S.E(\bar{x}) = \frac{s}{\sqrt{n}}$
Mean(σ unknown, pop^n finite)	$S.E(\bar{x}) = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
Difference of means (σ's are known)	$S.E(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Difference of means (σ's are unknown)	$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$
Proportion (pop^n size infinite)	$S.E(P) = \sqrt{\frac{PQ}{n}}$
Proportion (pop^n size finite)	$S.E(P) = \sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}$
Difference of prop^{n}s	$SE(P_1 - P_2) = \sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}} = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

where, N = Population size.
n = sample size
σ = pop^n standard deviation
s = Sample standard deviation
P = proportion of success, Q = " " failure.

Inferential statistics:

Descriptive statistics describes data (for example a chart or graph) and inferential statistics allows you to make predictions from that data. With inferential statistics, you take data from samples & make generalization about a population

Estimation : The process of estimating the unknown population parameters like mean, variance, proportion from the sample statistics.

Estimator : The sample statistic which is used to estimate the unknown population parameter.
Ex: Sample mean

Estimate: The particular value taken by the estimator

- **Point estimation**: A sample statistic (numerical value) is used to provide an estimate of population parameter
- **Interval estimation**: Probable range is specified within which the value of parameter might be expected to lie.

→ good estimator criteria

1) Unbiasedness
2) Consistency
3) Efficiency
4) Sufficiency

Q.5 A sample of size 25 is drawn [illegible]
[illegible]

Date
Page

* Confidence Interval (C.I) estimation of population mean(μ)

Large sample ($n \geq 30$)

$C.I = \bar{x} \pm z_\alpha \, S.E(\bar{x})$

$$C.I = \bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$C.I = \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

Small sample ($n < 30$)

$C.I = \bar{x} \pm t_{\alpha, n-1} \, S.E(\bar{x})$

where,
z_α = Signif. value of z
n = sample size
$\bar{x}$ = sample mean
σ = Standard deviation

* Confidence Interval (C.I) estimation of population proportion (π)

Large sample ($n \geq 30$) $(1-\alpha) \times 100\%$ confidence interval

$C.I = p \pm z_\alpha \, S.E(p)$

$$C.I = p \pm z_\alpha \sqrt{\frac{pq}{n}}$$

If N is given

$$p \pm z_\alpha \sqrt{\frac{pq}{n}\left(\frac{N-n}{N-1}\right)}$$

where, z_α = Significance value of z
p = Sample proportion of success
q = Sample proportion of failure
n = Sample size
N = Population size

$p = \frac{x}{n}$, $q = 1 - p$

(x characteristics taken from a sample)

DATE

* Sample size determination for the estimate of popⁿ mean (μ)

Minimum reqᵈ sample size in estimating the population mean μ is

$$n = \frac{Z_\alpha^2 \sigma^2}{e^2}$$

where, Z_α = significance value of z
ex: 1.96 for 95% C.I
σ = population S.D
e = margin of error

* Sample size determination for the estimate of population proportion (π)

Minimum required sample size in estimating the population proportion is

$$n = \frac{Z_\alpha^2 \pi(1-\pi)}{e^2} \quad / \quad n = \frac{Z_\alpha^2 P.Q}{e^2}$$

where, n = sample size
Z_α = Significance / critical value of z
π = population proportion, e = error margin

Note: * confidence level increases, sample size increases
CI $\propto$ n

* Sampling error decreases, sample size increases
$e \propto \frac{1}{n}$

* sample size increases, width of C.I decrease
It means for achieving higher no. of sample size, confidence level should increase & sampling error should decrease

DATE

Q.5

Unit-2: Testing of Hypothesis

Date
Page

Parametric Hypothesis: A statistical hypothesis is a tentative statement or supposition about the estimated value of one or more population parameter

Hypothesis: A hypothesis is a tentative theory or supposition provisionally adopted to explain certain facts & to guide in investigation

Non-parametric hypothesis:- Statistical hypothesis about attributes is called non-parametric hypothesis.

Note: If a hypothesis completely defines the population, it is a simple hypothesis, otherwise composite hypothesis.

Types of hypothesis.

1) (Null hypothesis): The supposition about the population parameter is called null hypothesis. It is a hypothesis which is tested for possible rejection under assumption that it is true.
   - → No difference betn sample statistics & parameter
   - → denoted by H_0. & set up as $H_0: \mu = \mu_0$

DATE

Date
Page

2) (Alternate hypothesis): Complementary to null hypothesis.
→ difference between sample statistic & parameter.
→ denoted by H_1 & set up as $H_1: \mu \neq \mu_0$ → two tailed

one tailed right $\mu > \mu_0$
one tailed left $\mu < \mu_0$
] one tailed

Errors in hypothesis testing.

On the basis of accepting or rejecting the hypothesis two types of errors are produced.

1) Type I error (null hyp. true हुई तभी reject जाएगा)
⇒ It is the error of rejecting null hypothesis when it is true. It's probability is α (Level of significance)

2) Type II error (null hyp. false हुई accept जाएगा)
⇒ It is the error of accepting null hypothesis when it is false (ie when alternative hypothesis is true) It's probability is β. [β = prob(accepting H_0 / H_1 is true)

Power of test $(1-\beta)$ [first explain about β)

The probability of rejecting H_0 when H_0 is false is probability of correct decision which is called ~~prob~~ power of test.

ie It is probability of not making type II error / ie. prob. of accepting H_1 when H_1 is true.

Test statistic

It is the statistic based upon approximate probability distribution. It is used to decide whether to accept or reject null hypothesis. Commonly used test statistic are-

1) Z-test: we use z-distribution under the normal curve for large sample ($n > 30$)
The z-test statistic is $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$ as $n \to \infty$

2) t-test: we use t-distribution for small sample ($n < 30$).
The t-test statistic is $t = \frac{\bar{x} - \mu}{S.E(\bar{x})} \sim t$ dist. with $n-1$ degree of freedom

Level of significance (L.O.S)

The two errors (type-I & type-II) are inversely related. Both errors cannot be minimized at the same time. Usually we fix type I & minimize type II). The maximum size of type I error prepared in testing of hypothesis is called LOS. Denoted by α.

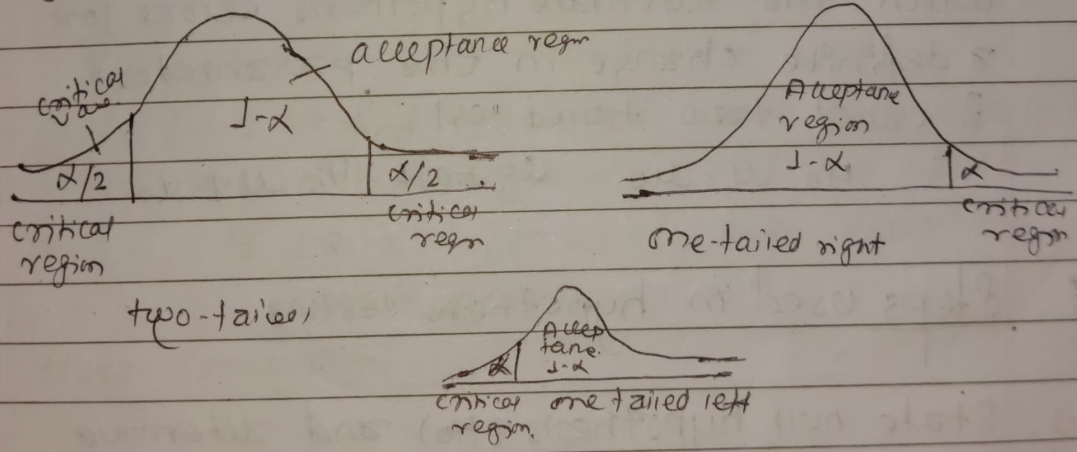
→ Commonly used L.O.S are 5%, 1%. The L.O.S should be chosen on basis of power of test. If P.O.T is too low then high L.O.S should be chosen. ie- 10%, 20%.

DATE

Critical region

Also called the rejection region. The set of all possible values of statistic is divided into two regions, one leading to rejection & other leading to acceptance of null hypothesis (H_0). The division is based on L.O.S and alternate hypo (H_1).

→ region which leads to rejection of H_0 = rejection region (ω)
→ region which " " acceptance " H_0 = acceptance region ($\bar{\omega}$)
→ If test statistic falls on rejection region, then H_0 is rejected, otherwise accepted.



critical value

The value of statistic which separates critical region and acceptance region is called critical value.

Degree of freedom (No. of independent choice / choice selection)

The no. of independent variates which make up statistic is called d.f. If sample size = n & restriction = 1

Then, d.f = n - 1

* One-tailed test (for directional hypothesis)

A test of statistical hypothesis in which the alternative hypothesis H_1 looks for a definite increase (right tail) or definite decrease (left tail) in parameter is called

Null		Alternate	
$H_0: \mu = \mu_0$	against	$H_1: \mu > \mu_0$	(right tail)
$H_0: \mu = \mu_0$	against	$H_1: \mu < \mu_0$	(left tail)

* Two-tailed test (for ~~non~~ non directional hy-

A test of statistical hypothesis in which the alternate hypothesis looks for a definite change in the parameter is called two tailed test.

$H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

* Steps used in hypothesis testing

Step-1: State null hypothesis (H_0) and alternate hypothesis (H_1)

Step-2: Select the appropriate test statistics. and calculate the required value for the test under H_0.

~~Step-3: State level of significance.~~

Step-3: State the Level of significance.

Step-4: Find the critical value (tabulated value)

Step-5: Make decision by compairing the calculated value & tabulated value.

DATE

Step-5: Make conclusion; ie. Either accept H_0 or reject H_0.

* Procedure of testing significance of mean for large sample ($n \geq 30$)

Step-1: State H_0 & H_1 ($H_0: \mu = \bar{x}_{yz}$ / $\mu = \mu_0$, $H_1: \mu > \bar{x}_{yz}$ / $\mu < \bar{x}_{yz}$ / $\mu \neq \bar{x}_{yz}$)

Step-2: Select Z-test & find test statistic under H_0

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = sample mean
μ = population/hypothesised mean
σ = population standard deviation
n = sample size

find $|cal\ z|$

Step-3: State L.O.S & find tabulated value of z $|tab\ z|$ (critical value)

Step-4: Make decision by compairing $|cal\ z|$ & $|tab\ z|$

If $|cal\ z| < |tab\ z|$ then accept H_0

reject otherwise.

Step-5: Make conclusion.

* Procedure of testing Significance of mean for small Sample ($n < 30$)

$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n-1}}}$ use this when sample sd is already given

other- use $\frac{S}{\sqrt{n}}$, coz $S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

Step1: State H_0 & H_1.

Step2: Select t-test and find test-statistic under H_0

$$t = \frac{\bar{x} - \mu}{\frac{\sigma_s}{\sqrt{n-1}}} \quad \text{or} \quad t = \frac{\bar{x} - \mu}{\frac{\sigma_s}{\sqrt{n}}} \quad \text{or} \quad t = \frac{\bar{x} - \mu}{\frac{\sigma_s}{\sqrt{n}}\sqrt{\frac{n-1}{n-1}}}$$

find $|cal\ t|$ if σ_s is already given. ($\sigma_s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$) If n is given

where, $\bar{x}$ = Sample mean

μ = population mean

s / σ_s = sample S.D = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

σ = population S.D

Step-3: State L.O.S & find tabulated value of t

Step-4: Find critical value, tab|t| and compare with cal|t|

Step-5: Make decision: If cal|t| < table|t| accept H_0 reject otherwise.

Step-6: Make conclusion.

* Procedure for testing Significance of difference between two means for small sample

Step-1: State H_0 and H_1 [$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2, \mu_1 < \mu_2, \mu_1 > \mu_2$]

Step-2: Select t-test and find test statistic under H_0

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

or $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where, s_p^2 = pooled / combined variation

s_1 = 1st sample SD

s_2 = 2nd sample SD

n_1 = 1st sample size

n_2 = 2nd sample size

Note: confidence level = (1-α)
95% = (1-α)
0.95 = 1-α
α = 0.05

Here,

$$Sp^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad \text{or} \quad Sp^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

use If s_1 & s_2 are not given.

Step-3: State Level of significance (α)
step-4: Find critical value / table value of t
Step-5: Make decision by compairing cal|t| and table|t|
If cal|t| < table|t|, accept Ho, reject otherwise.
Step-6: Make Conclusion

* Procedure for testing significance of proportion for large sample (n>30)

Step-1: " " (Ho: P = xyz, H1: P ≠, P >, P < xyz)
Step-2: Select Z-test & find teststatistics under Ho.

$$Z = \frac{Ps - P}{\sqrt{\frac{PQ}{n}}}$$

p/Ps: sample proportion = $\frac{x}{n}$
P: population proportion
Q = 1 - P
n = sample size

Step-3: " "
Step-4: " "
Step-5: If cal|z| < table|z|, accept Ho otw reject Ho
Step-6: " "

= 0.4

DATE

Date
Page

* Procedure of testing significance of difference between two proportions for large samples ($n > 30$)

Step1: State H_0 and H_1 ($H_0: P_1 = P_2$, $H_1: P_1 \neq P_2$ or $P_1 > P_2$, P_1

Step2: select z-test x " "

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

p_1 & p_2 are sample prop

$$p_1 = \frac{x_1}{n_1}, \quad p_2 = \frac{x_2}{n_2}$$

or

$$P = \frac{x_1 + x_2}{n_1 + n_2}, \quad Q = 1 - P$$

P, Q are pop^n prop's

$$z = \frac{p_1 - p_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

when P_1 and P_2 are given.

Step-3: " " " "

Step-4: " "

Step-5: Make conclusion: If (cal|z| < tabel|z|) accept H_0 else reject H_0:

95% CI, For z test, If $\alpha = 0.05$, for two tailed $z_\alpha = \frac{1-0.05}{2}$

95% CI, for t-test, If $\alpha = 0.05$, for two tailed, $t_\alpha = \frac{0.05}{2}$

Note: for z-table (not standard normal table,

If $\alpha = 0.05$

for one tailed ⇒ see $(1-\alpha)$ for Z_α (critical value)

for two tailed ⇒ see $(1-\frac{\alpha}{2})$ for Z_α (critical va

If $|z_{cal}| > |z_{table}|$ reject H_0, otw accept H_0

for t-test (table in Laptop)

for one tailed ⇒ see α and $(n-1)$ df

for two tailed ⇒ see α and $(n-1)$ df.

If $|t_{cal}| > |t_{table}|$, then reject H_0, otw accept H_0

for p-value test

for one tailed ⇒ P-value = $prob\{z > |z_{cal}|\}$

for two-tailed ⇒ P-value = $2prob\{z > |z_{cal}|\}$

If p-value $< \alpha$ ⇒ reject H_0

otherwise accept H_0

for pear t-test

X = before, Y = after

H_0: $\mu_x = \mu_y$

H_1: $\mu_x \neq \mu_y$, $\mu_x < \mu_y$, $\mu_x > \mu_y$.

Test statistic: $t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$ where $\bar{d} = \frac{\Sigma d}{n}$.

where $d = X - Y$

standard dev = $S_d = \sqrt{\frac{1}{n-1}\{\Sigma d^2 - n\bar{d}^2\}}$

$= \sqrt{\frac{1}{n-1}\{\Sigma d^2 - \frac{(\Sigma d)^2}{n}\}}$

Or, $S_d = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$

-0.4