RM

__/__/____

UNIT 3

* Probability Sampling
* Stratified Sampling

* Disproportionate Sampling : When strata differ in size and variability, it is reasonable to take larger samples from variable strata

$$\frac{n_1}{N_1 \sigma_1} = \frac{n_2}{N_2 \sigma_2} = \cdots$$

$$n_i = \frac{n \cdot N_i \sigma_i}{N_1 \sigma_1 + N_2 \sigma_2 + \cdots N_k \sigma_k}$$

σ → standard deviations
N → sizes
n → sample size (total)
n_i → sample size (of i strata)

__/__/____

$N_1 = 5000$ $N_2 = 2000$ $N_3 = 3000$

$\sigma_1 = 15$ $\sigma_2 = 18$ $\sigma_3 = 5$

$n = 84$ $n_1, n_2, n_3 = ?$

$$n_i = \frac{n \cdot N_i \sigma_i}{N_1 \sigma_1 + \cdots N_k \sigma_k}$$

i) $$n_1 = \frac{n \cdot N_1 \sigma_1}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} = \frac{84\,(5000 \times 15)}{5k15 + 2k18 + 3k5}$$

$$n_1 = \frac{6300000}{126000} = 50$$

ii) $$n_2 = \frac{n \cdot N_2 \sigma_2}{126000} = \frac{84\,(2000 \times 15)}{126000} = 24$$

iii) $$n_3 = \frac{n \cdot N_3 \sigma_3}{126000} = \frac{84\,(3000 \times 5)}{126000} = 10$$

IF stratum has cost C_i

$$n_i = \frac{\frac{n \cdot N_i \sigma_i}{\sqrt{C_i}}}{\frac{N_1 \sigma_1}{\sqrt{C_1}} + \cdots + \frac{N_k \sigma_k}{\sqrt{C_k}}}$$

__/__/____

* Mean, Standard Deviation, Variance

X	f	fX	$[X-\bar{x}]^2$	$f \times [X-\bar{x}]^2$
2	40	80	4	160
4	30	120	0	0
6	20	120	4	80
8	10	80	16	160
Σ	100	400		400

* MEAN $= \bar{X} = \dfrac{\Sigma fX}{\Sigma f} = \dfrac{400}{100} = 4$

* STANDARD DEVIATION $= S = \sqrt{\dfrac{\Sigma f[X-\bar{x}]^2}{\Sigma f}} = \sqrt{\dfrac{400}{100}}$

$S = 2$

* VARIANCE $= S^2 = 4$

* Coefficient of Variance $= \dfrac{S}{\bar{X}} = \dfrac{2}{4} = 0.5$

Standard Deviation $= SD = S = \sigma$

FORMULA

* **Confidence Interval for the Mean**

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

μ lies between $\left[\bar{x} - z \frac{\sigma}{\sqrt{n}}, \bar{x} + z \frac{\sigma}{\sqrt{n}} \right]$

* **Confidence Interval for the Average**

$$\mu = \bar{x} \pm z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Mean $= \bar{x} = 10.455$

Sample $= n = 55$

S.D $= \sigma = 7.7$

Construct 90% confidence interval (CI)

Given for 90%, $z = 1.645$

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

$$\left[10.455 - (1.64)\frac{7.7}{\sqrt{55}}, \quad 10.455 + (1.64)\frac{7.7}{\sqrt{55}}\right]$$

$$\left[8.545, \quad 12.365\right]$$

UNIT 4

DATA ANALYSIS

__/__/____

* Problem Definition

Well defined problem is 50% of the project done

* Plan selection

* Hypotheris

* Data Analysis
(hypotheris rejected or accepted)

Goal: to gain info from data

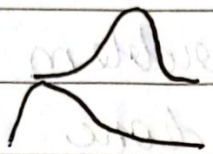
EDA – Desc stats
– Visualization

– Data Cleaning
– Correlation Analysis
– Dimensionality Reduction
– Data Transformation
(normalization, age 1-80 to 0-1 scale)
– Summary Statistics

* Data Distribution

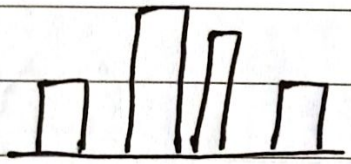
[Mean, Median, Mode] (Central Tendency)

Symmetric — mean



Asymmetric — median

Mode : which value occurs most frequently



B max times

2 x Mean + Mode = 3 x Median

- Quartile

Q1 : 25%

Q2 : 50% percentile

Q3 : 75%

IQR : Q3 - Q1

- [SD] (Variation)

$$s = \sqrt{\frac{(x - \bar{x})^2 + \cdots (y - \bar{x})^2}{n - 1}}$$

$$\sigma = s^2$$

~~Unit 3 → 3 question~~

Unit 3 → 2 question

Unit 4 → 3 question

* Statistical Estimation

– Confidence Intervals

Size (n) ↑, Confidence ↑

Problems 1, 2, 3, 4, 5

– T - Distribution

– Hypothesis Testing

__/__/____

Using t - distribution

When $n < 30$
And σ = Unknown

0) Data = {3, 1, 3, 7, 4, 2, ...}
$n = 14$
$C = 99\%$

i) Calculate mean ($\bar{x}$)

ii) Calculate SD (σ)

iii) $t_{\left(\frac{\alpha}{2}, n-1\right)} \Rightarrow \frac{\alpha}{2} = \frac{1-0.99}{2} = 0.005$
$n-1 = 14-1 = 13$

iv) Confidence Interval, (μ)

[_____]

$$\left[\bar{x} - t_{0.005} \frac{\sigma}{\sqrt{n}}, \bar{x} + t_{0.005} \frac{\sigma}{\sqrt{n}} \right]$$

__/__/____

Data = 3, 1, 3, 2, 5, 1, 2, 1, 4, 2, 1, 3, 1, 1

$n = 14$

$C = 99\%$

a) MEAN, $\bar{x} = \frac{30}{14} = 2.143$

b) SD, $\sigma = \frac{\sum(\bar{x} - x)^2}{n-1} = \frac{12.7}{13}$

$\sigma = \sqrt{0.977} = 0.988$

c) $t_{\frac{\alpha}{2}, n-1} = \frac{1 - 0.99}{2} = 0.005$

$n - 1 = 13$

$t_{(0.005, 13)} = 3.012$

d) $\bar{x} - 3.012 \frac{0.988}{\sqrt{14}} = 1.348$

$\bar{x} + 3.012 \frac{0.988}{\sqrt{14}} = 2.94$

__/__/____

* Population Proportion

\# Sample = n = 200
Proportion = x = 80
C = 95%

→ $P = \frac{x}{n} = \frac{80}{200} = 0.4$

→ $Z = 95\% = 1.96$

→ $\mu = P \pm Z \sqrt{\frac{P(1-P)}{n}}$

$= 0.4 \pm 1.96 \sqrt{\frac{0.4(1-0.4)}{200}}$

$= [\, 0.332 \;,\; 0.468 \,]$

HYPOTHESIS TESTING

__/__/____

* NULL Hypothesis $\Rightarrow$ $\mu = 20$
* ALTERNATE Hypothesis $\Rightarrow$ $\mu \neq 20$

* Significance level = α = Threshold

If $p \leq \alpha$, reject null hypothesis
If $p > \alpha$, accept null hypothesis

I. PARAMETRIC TESTS

1) t - test

i) One Sample

Compare mean of group to mean of population

→ $H_0 \Rightarrow \mu = \mu_0$ (sample mean = population mean)

→ $H_A \Rightarrow \mu \neq \mu_0$

→ $t = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ → $df = n - 1$

→ Find p value

→ $p \leq \alpha$ (or) $p \geq \alpha$

__/__/____

Data = {10.5, 9, 6.5, 8, 11, 7, 7.5, 8.5, 9.5, 12}

$n = 10$

$\bar{x} = 12$

→ Mean of sample, $\bar{x} = 8.95$

Mean of population, $\mu = 12$

→ $H_0 \Rightarrow \mu = \bar{x}$

$H_A \Rightarrow \mu \neq \bar{x}$

→ $t = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{8.95 - 12}{1.802/\sqrt{10}} = -5.35$

$\sigma = \dfrac{\sum (\bar{x} - x)^2}{n-1} = \dfrac{29.225}{9} = \sqrt{3.247} = 1.802$

→ df = 9 and Significance = 5% = 0.05

→ Tabulated Value (9, 0.05) = 2.2622

⇒ Calculated Value (t) within [−2.2, +2.2]

∵ $t < -2.2$, reject NULL hypothesis

CONCL There is significant difference in mean of sample and population

__/__/__

One Sample, Two Occasion

BP before :	180	200	-	-	-
BP after :	140	145	-	-	-
Difference :	40	55	80	85	

→ $\Sigma d = 465$ and $n = 8$

$\bar{d} = \frac{465}{8} = 58.125$

→ $\sigma = \sqrt{307.49} = 17.54$

→ $t = \frac{\bar{d}}{\sigma/\sqrt{n}} = 9.37$

→ Tabulated Value (0.05 and 7)

__/__/____

* Z - Test

- When $n > 30$ (Sample is large)

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

\# $\mu = 40$

$n = 50$ $\quad \bar{x} = 45$ $\quad \sigma = 20$

$\alpha = 0.02$

→ $H_0 : \mu \leq 40$

$H_A : \mu \neq 40$

→ $$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{45 - 40}{20 / \sqrt{50}} = 1.767$$

→ $\alpha = 0.02, \; df = 49 \Rightarrow Z_{0.02} = 2.05$

→ $p > z$

Do not reject H_0

__/__/____

$\mu = 166.17$

$\sigma = 5.89$

$n = 144$ $\quad \bar{x} = 164.65$

$\alpha = 0.05$

→ H_0 : Slum area residents are not shorter ($\mu > 166$)

H_A : Slum area residents are ~~not~~ short

→ $$Z = \frac{\bar{x} - M}{\sigma/\sqrt{n}} = \frac{166.17 - 164.65}{5.89/\sqrt{144}} = -3.097$$

→ $Z_{0.05} = [-1.65, 1.65]$

⇒ Reject null hypothesis

__/__/__

CHI SQUARE TEST (χ_2)

#		A	B	Total
	Male	30	20	50
	Female	25	25	50
	Total	55	45	100

→ H_0 : Gender and Vote are independent
H_A : Variables are dependant

→ Contigency

$E_{Male, A} = \frac{55 \times 50}{100} = 27.5$ $\quad E_{F,A} = 27.5$

$E_{Male, B} = \frac{45 \times 50}{100} = 22.5$ $\quad E_{F,B} = 22.5$

→ Chi Square

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$$\chi^2 = \frac{(30 - 27.5)^2}{27.5} + \frac{(20 - 22.5)^2}{22.5} + \dots + \frac{(25 - 22.5)^2}{22.5}$$

$\chi^2 = 1.01$

→ $df = (2-1)(2-1) = 1$

→ $\alpha = 0.05$ and $df = 1 \Rightarrow 3.841$

→ $\chi^2 < 3.841 \Rightarrow$ Do not reject H_0

__/__/____

Is Die Throw unbiased

Num	1	2	3	4	5	6	
Frequency	16	20	25	14	29	28	132

→ $E = 132 \times \frac{1}{6} = 22$

→ $X^2 = \sum \frac{(O-E)^2}{E} = \frac{(16-22)^2}{22} + \cdots$

$X^2 = 9$

→ $\alpha = 0.05$

$df = 6 - 1 = 5$

→ $P = [-11, 11]$

→ ∴ Accept H_0

∴ Die is unbiased