[Section-I: SPSS]

1.

a) Open JSSA.sav file. Identify three categorical variables and write the variable name, variable label, and category values on your answer sheet. Construct a frequency distribution for age (age of respondent) using the following intervals as categories: <30, 30-39, 40-49, 50-59, 60-69, 70-79, >=80. Considering your analysis what percentage of population has between 40-69 years?

b) Construct a bar chart showing age category (horizontal) and respondent's sex (sex). Add title of the chart "Number of Population by age groups and Sex". Change the color pink (for female bar), yellow (for male bar).

2.

a) Duration of stay (in month) in the U.S.A. by residents of three countries is given below. Create a data file by using text file latin.txt or create by Syntax.

i) Which of the three countries has the highest median and the lowest mean value? Does this support your idea that respondents from Latin American countries that are closer to the United States have a higher median duration of stay in the United States?

ii) For each country, identify the direction in which (if any) the distribution is skewed? Provide specific evidence using your answers from the previous questions.

iii) Calculate the variance each of three countries. Which country has the least variability? Select a country and construct the box plot.

b) Calculate average relaxation hours (HRSKELAX). Find the 90% confidence interval for Race (1=White, 2=Black and 3=others) [Use racecen1 and recode info a new variable racecat (1=White, 2=Black and 3=others) for the JSSA.sav_data].

3.

a) Using JSSA.SAV, test the null hypothesis that men and women work the same number of hours each week by using the variable HRS1. What are the assumptions underlying the test you identified? Determine whether you are able to reject the hypothesis or not. (Assume that alpha is set at .05 for a two-tailed test). What do you conclude?

b) Use NUT.SAV. Is there a relationship between mother's educational level (educat) and stunted child under 5 years (nt_ch_stunt), or are they independent?
i) What is the total sample size (both weighted number and un-weighted number)?
ii) Calculate the expected frequencies for each cell (weighted).
iii) What percentage of the primary level group children are "stunted" (weighted)?
iv) If there is a relationship, how large is this effect and what does it look like? (weighted).

4.

a) Test whether the mean level of education (educ) differs across Race
[Use *racecat* (1=White, 2=Black, and 3=others)]. State the null hypothesis in both written and symbolic form [Use *JSSA.sav* data].
i) On the basis of your analysis, can we reject the null hypothesis? If so, on what basis? Conduct a post-hoc test and give your comments.
ii) If you add respondent's sex (sex) in your model, has it any effect on education? If so, how much does it improve the model?

b) Use the **hsb2.sav** file. Conduct the following regression analyses:

- Dependent variable: **science**
- Predictor variables: **math, female, socst, read**

i) Create a table showing Multiple R, then each of the variables that significantly influence the dependent variables.
ii) Does the relationship appear to be negative, positive, or neither?
iii) Write down the full regression equation in proper notation and provide a one-sentence interpretation of the slope coefficient.

[Section-II: STATA]

[Instruction: Save your command in a do-file or save all results in a log file; if needed, copy your graphs and tables in a word document with your roll number.]

5.

Open descriptive **hsb2.dta** and do a detailed summary of the variable **write** (writing score). Interpret the mean and median values. Also, create a histogram of the variable.

i) Type the command **numlabel ses, add**, and then do a tabulation of **ses**.
ii) Find the correlation matrix of the variables **read write math science female**.
iii) Construct a scatter plot of the variables **read write** and comment on the plot.
iv) Do you expect **females to have less science score (science)** than males? If the result is statistically significant, how substantively significant is the difference?
v) How many cases do you need to have 80% power?

Answer Sheet

22.1.a) *Open JSSA.sav file. Identify three categorical variables and write the variable name, variable label, and category values on your answer sheet. Construct a frequency distribution for age (age of respondent) using the following intervals as categories: <30, 30-39, 40-49, 50-59, 60-69, 70-79, >=80. Considering your analysis what percentage of population has between 40-69 years?*

1.

- Variable Name: Sex
- Variable Label: Sex of respondent
- Category Values:
  - 1 = Male
  - 2 = Female

2.

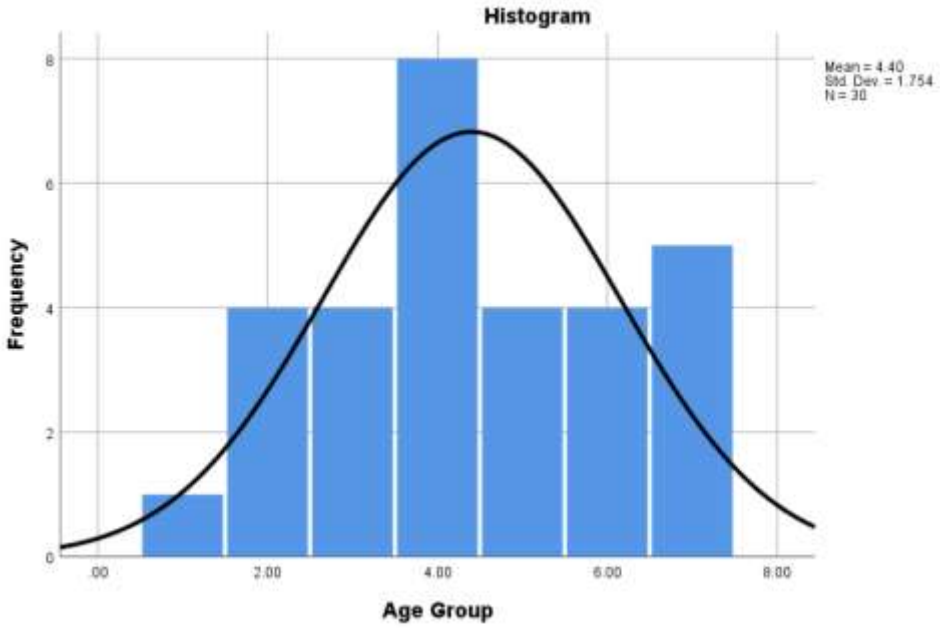
- Variable Name: Marital_Status
- Variable Label: Marital Status
- Category Values:
  - 1 = Single
  - 2 = Married
  - 3 = Others

3.

- Variable Name: Education
- Variable Label: Level of Education
- Category Values:
  - 1 = Primary
  - 2 = Secondary
  - 3 = Higher Secondar
  - 4 = Graduate
  - 5 = Postgraduate

Age Group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<30	1	3.3	3.3	3.3
	30-39	4	13.3	13.3	16.7
	40-49	4	13.3	13.3	30.0
	50-59	8	26.7	26.7	56.7
	60-69	4	13.3	13.3	70.0
	70-79	4	13.3	13.3	83.3
	>=80	5	16.7	16.7	100.0
	Total	30	100.0	100.0	

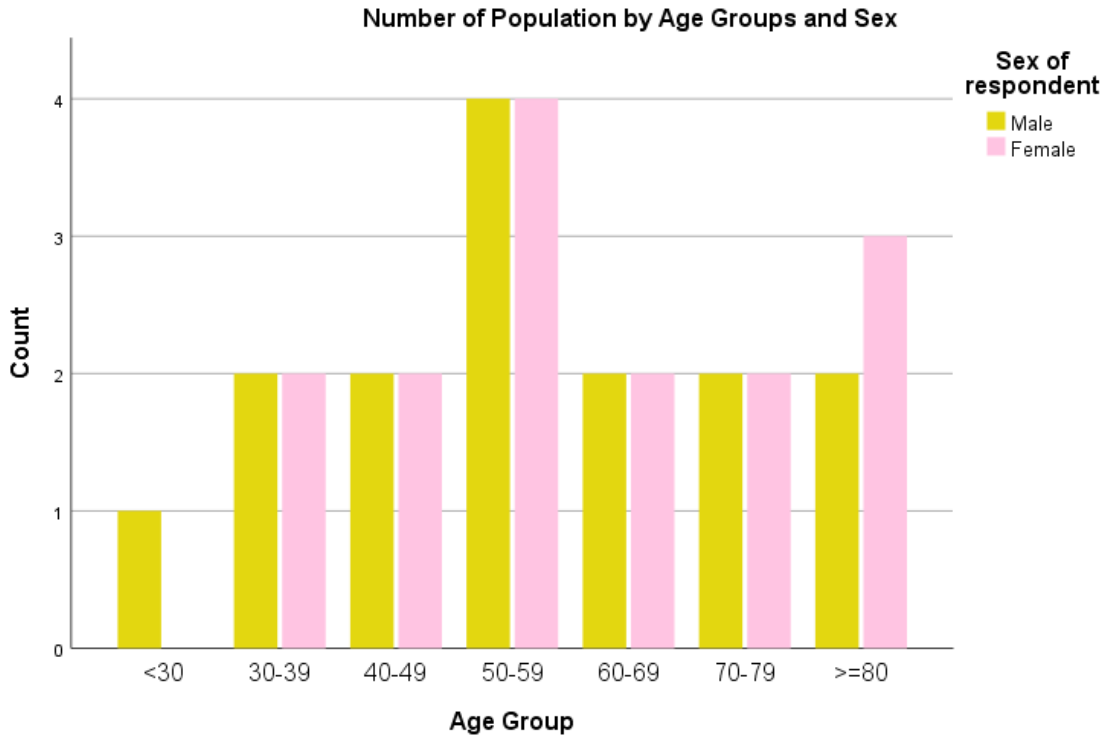


We combine age groups **40–49, 50–59, 60–69**.

Percentage=13.3%+26.7%+13.3%=53.3%

✅ Answer: 53.3% of the population is between 40–69 years.

22.1.b) *Construct a bar chart showing age category (horizontal) and respondent's sex (sex). Add title of the chart "Number of Population by age groups and Sex". Change the color pink (for femal*



22.2.a) *Duration of stay (in month) in the U.S.A. by residents of three countries is given below. Create a data file by using text file latin.txt or create by Syntax.*

i) Which of the three countries has the highest median and the lowest mean value? Does this support your idea that respondents from Latin American countries that are closer to the United States have a higher median duration of stay in the United States?

ii) For each country, identify the direction in which (if any) the distribution is skewed? Provide specific evidence using your answers from the previous questions.

iii) Calculate the variance each of three countries. Which country has the least variability? Select a country and construct the box plot.

(i)

Highest median = European 45

Lowest mean = South Asia 16.26

Interpretation: Country European has the highest median duration, and Country South Asia has the lowest mean duration. This does not support the idea that respondents from Latin American countries closer to the U.S. have a higher median duration of stay."

(ii)

Country	Mean	Median	Skewness	Skewness Direction
Latin American	24.77	24.25	0.168	Positive (slightly right-skewed) – tail is slightly toward higher values
South Asia	16.26	9	2.872	Positive (strongly right-skewed) – long tail toward higher values
European	45.42	45	-0.103	Negative (slightly left-skewed) – tail is slightly toward lower values

Interpretation

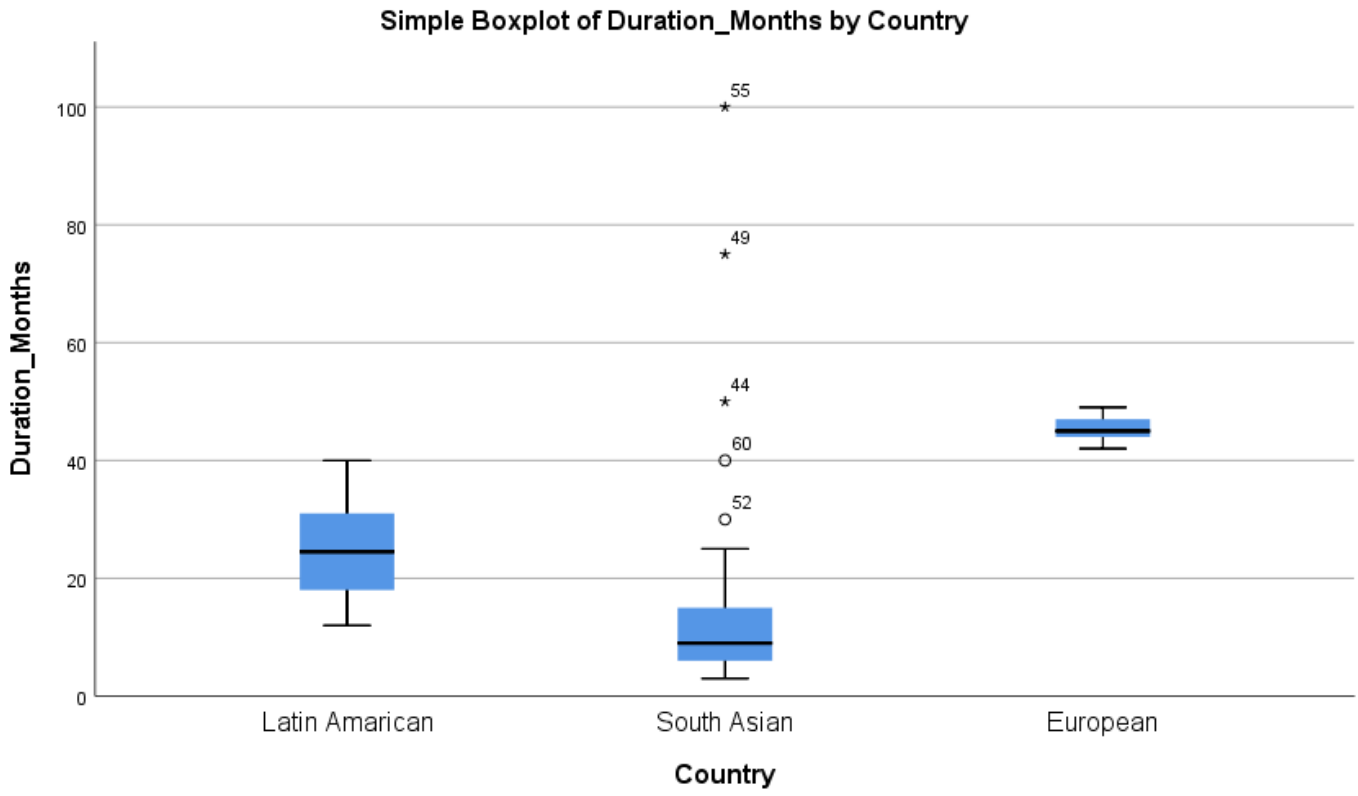
- **Latin American:** Slightly right-skewed distribution; most respondents have durations near the median, but a few stayed longer.
- **South Asia:** Strongly right-skewed; many respondents stayed for shorter durations, but a few stayed very long, pulling the mean up.
- **European:** Slightly left-skewed; most respondents have durations near the median, but a few shorter stays pull the mean slightly below the median.

(iii)

The variances for each country:

- **European** → 3.827
- **Latin American** → 65.465
- **South Asian** → 430.608

The country with the **least variability** is **European**



22.2.b) *Calculate average relaxation hours (HRSKELAX). Find the 90% confidence interval for Race (1=White, 2=Black and 3=others) [Use racecen1 and recode info a new variable racecat (1=White, 2=Black and 3=others) for the JSSA.sav_data].*

Race	Mean	CI
White	6.4156	[6.0937, 6.7375]
Black	6.4737	[6.1518, 6.7956]
Others	6..4545	[6.1309, 6.7783]

3.

a) Using JSSA.SAV, test the null hypothesis that men and women work the same number of hours each week by using the variable HRS1. What are the assumptions underlying the test you identified? Determine whether you are able to reject the hypothesis or not. (Assume that alpha is set at .05 for a two-tailed test). What do you conclude?

b) Use NUT.SAV. Is there a relationship between mother's educational level (educat) and stunted child under 5 years (nt_ch_stunt), or are they independent?
i) What is the total sample size (both weighted number and un-weighted number)?
ii) Calculate the expected frequencies for each cell (weighted).
iii) What percentage of the primary level group children are "stunted" (weighted)?
iv) If there is a relationship, how large is this effect and what does it look like? (weighted).

SPSS

→spss এর তিনটা window আছে

1. Data View window
2. Variable View window
3. Output window

1. Data file → .sav (data.sav)
2. Output file→ .spo (data.spo)
3. Syntax file → .sps (data.sps)

To open an SPSS data file

Manually: File → Open → Data → Browse file from source

Syntax: GET FILE ="path"

Example:

GET FILE='C:\SPSS Training\DayI\Obesity.sav'.

GET SAS FILE="C:\SPSS Training\Day I\ Obesity.sas7bdat".

GET STATA FILE="C:\SPSS Training\Day I\ Obesity.dta".

Open Excel files in SPSS:

Manually:

File → Open → Data → Browse Excel file → OK

Syntax:

GET DATA /TYPE=XLSX

/FILE='C:\SPSS Training\DayI\Obesity_sample2.xlsx'

/SHEET=name Obesity_sample2‘

/CELLRANGE=full /READNAMES=on

/ASSUMEDSTRWIDTH=32767.

EXECUTE.

COMPUTE Function:

Suppose we want to calculate standardized value of LIFESPAN (X). The formula is,

$$Z = \frac{X - mean(X)}{Standarddeiation(X)}$$

Suppose mean and standard deviation of LIFESPAN are 835 and 275.

Manually

Transform Compute Variable > Define target variable

(STD_LIFESPAN) > ((LIFESPAN-835)/275)” > OK

Syntax:

COMPUTE STD_LIFESPAN=(LIFESPAN-835)/275.

EXECUTE.

Recode into different variable:

Syntax:

RECODE LIFESPAN (Lowest thru 299=1) (300 thru 599=2) (600 thru 899=3) (900 thru 1199=4) (1200 thru Highest=5) INTO CAT_LIFESPAN.

VARIABLE LABELS CAT_LIFESPAN 'Distribution of Lifespan'.

VALUE LABELS CAT_LIFESPAN 1 "<300" 2 "300-600" 3 "600-900" 4 "900-1200"

5 "≥1200".

EXECUTE.

Merge files: Add Cases

Syntax:

Get FILE='G:\Data File\data1.sav'.

ADDFILES /FILE=*

/FILE='G:\Data File\data2.sav'.

SAVE OUTFILE 'G:\Data File\data3.sav'.

EXECUTE.

Merge files: Add Variables

Syntax:

GET FILE= "D:\spss\dads.sav".

SORT CASES BY famid.

SAVE OUTFILE="D:\spss\dads2.sav".

GET FILE="D:\spss\faminc.sav".

SORT CASES BY famid.

SAVE OUTFILE="D:\spss\faminc2.sav".

MATCH FILES FILE="D:\spss\dads2.sav"

/FILE="D:\spss\faminc2.sav"

/BY famid. SAVE OUTFILE="D:\spss\OnetoOneMerge.sav"

Merge files: One to Many Merge

GET FILE="D:\spss\dads.sav".

SORT CASES BY famid.

SAVE OUTFILE="D:\spss\dads2.sav".

GET FILE="D:\spss\kids.sav".

SORT CASES BY famid.

SAVE OUTFILE="D:\spss\kids2.sav".

MATCH FILES FILE="D:\spss\kids2.sav"

/Table="D:\spss\dads2.sav"

/BY famid.

SAVE OUTFILE="D:\spss\OnetoManyMerge.sav".

Aggregate Statistics:

Find mean age and weight of children for each family.

Syntax:

AGGREGATE

/OUTFILE=* MODE=ADDVARIABLES

/BREAK=famid

/age_mean_1=MEAN(age)

/wt_mean_1=MEAN(wt)

→Dialog box এর last step এ "OK" এর পরিবর্তে "Paste" চাপলে সেই operation এর syntax code syntax file এ লেখা হয়ে যায়। এরপর syntax file থেকে Run করলে output file এ output টা দেখতে পারা যাবে।

এর ফলে কোন operation এর syntax code প্রয়োজন হলে সেখান থেকে নেওয়া যাবে।

→ 1 inch = 0.0254 m

→1 kg = 2.20462 pound

Correlation Analysis:

Karl Pearson's correlation coefficient (r)

Spearman's rank correlation coefficient (ρ)

STATA

→command এ br লিখলে import করা data file দেখাবে

→**gen:** For creating new variable

gen loss = STARTWEIGHT- ENDWEIGHT

→Recode Command:

Syntax: old values = new value

Example: recode loss (0/17=0) (18/25=1) , gen(loss_cat)

The **gen** option tells recode to create a new variable (loss_cat) to store the results. If you don't include a gen option, recode will change the original variable.

→**Variable labels:** label variable loss_cat “loss of weight”

→**Value labels:**

label define los 0 “less than 17” 1 “more than 17” label values loss_cat los

→ **Rename command:** loss_cat loss_category

→ describe: dataset এর সবগুলো variable এর একটা সাধারণ বিবরণী পাওয়ার জন্য ব্যবহার হয়

→Codebook: ভেরিয়েবলগুলোর নাম, লেবেল এবং ডেটা পরীক্ষা করে ডেটাসেটের একটি বর্ণনামূলক কোডবুক তৈরি করে।

→summarize: minimum, maximum, std, mean এই সব দিবে

→by: প্রতিটা group এর জন্য আলাদা ভাবে কোন operation run করার জন্য

Syntax: bysort varname: command

Example: bysort gender: summarize income

bysort mpg: summarize price (একই সাথে sort ও করবে)

→sort: data কে sort করার জন্য

sort mpg

→if = do this only for observations that satisfy this condition

যেসব observation এই condition পূরণ করে, শুধু সেগুলোর ক্ষেত্রেই এটি প্রয়োগ করো

Summarize price if foreign == 1

→ in = do this only for these observation numbers

Syntax: command varlist in range

Example: sort price

summarize mpg in 1/10

→ tab = "show a table of counts (and sometimes percentages)"

Example: tab foreign

→help command : লিখলে এই command কি কাজ করে সেটা দেখায়

help tab

→ table = "create a summary table of statistics for groups of data"

Example: table loss_category, contents(freq mean loss)

→drop var_name: কোন column কে delete করার জন্য

→ drop if var1==1

Example: drop gender =1

→dataset থেকে n সংখ্যক data sample হিসাবে নেওয়ার জন্য

set seed 200304038

sample n, count

n সংখ্যক data রেখে বাকি সবগুলো data file থেকে delete করে দেবে

→if you want **120 observations stratified by gender**

set seed 200304038

bysort gender: sample 120, count

→ **One sample T-test:**

$$t_{cal} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

$$\text{where, } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

For two-tailed:

ttab<-qt(alpha/2,df)

pval<-2*(1-pt(abs(tcal),df))

For one-tailed:

ttab<-qt(alpha,df)

pval<-1-pt(abs(tcal),df)

→ Stata use করে one sample t-test করার জন্যঃ-

Syntax:

ttesti num_of_obs mean sd null_hypothese_value, level(1-alpha*100)

Example:

ttesti 24 44.364 0.019 44.350, level(95)

→ one-taild হোক বা two-taild formula same হবে।

→one-taild এর ক্ষেত্রেঃ-

hypothesis greater than হলে greater than চিহ্ন এর নিচের p value ব্যবহার করব।

hypothesis lower than হলে lower than চিহ্ন এর নিচের p value ব্যবহার করব।

→two-taild এর ক্ষেত্রে greater than ও lower than এর নিচের p value same হবে।

→ **Two sample T-test:**

$$t_{cal} = \frac{\overline{x_1} - \overline{x_2}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$\text{where, } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}, \text{ pooled variance}$$

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i} - \overline{x_1})^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2i} - \overline{x_2})^2$$

→ Stata use করে two sample t-test করার জন্যঃ-

Syntax:

ttesti obs1 mean1 sd1 obs2 mean2 sd2

Example:

ttesti 10 126.1 21.8 10 133.1 32.2

→Recode the Gender variable into numerical coding (0=Male, 1=Female)

encode Gender, gen(tempvar) (Gender var কে একটি temporary var এ নেওয়া হয়েছে)

recode tempvar (1=0 "Male") (2=1 "Female"), gen(Sex) label(Gender) (temporary var এর value male=0 আর Female=1 set করে সেটাকে Sex নামক নতুন var এ রাখা হয়েছে যার label হলো Gender)

drop tempvar(temporary var column কে delete করা হয়েছে)

→Paired *t*-test

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

$$\text{Given that } n = 10,\ d_i = \text{before}_i - \text{after}_i\ (i = 1,2,\dots,10)$$

The test statistic,

$$t_{\text{cal}} = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} \sim t_{\alpha/2}(n-1)$$

If $t_{cal} > t_{\alpha/2}$ (or, $p < \alpha$), we may reject the H_0.

→STAT এর মধ্যে **Paired *t*-test**

ttest before == after

→Pearson correlation in stata:

pwcorr resting_hr age weight_kg height_cm fitness_score stress_level caffeine_mg, sig star(0.05)

pairwise correlation দেখাবে

→correlate resting_hr age weight_kg height_cm fitness_score stress_level caffeine_mg

শুধু covariance matrix দেখাবে

→ simple linear regression:

reg resting_hr age weight_kg height_cm fitness_score stress_level caffeine_mg

→ Standardized regression (to compare strength)

reg resting_hr age weight_kg height_cm fitness_score stress_level caffeine_mg, beta