Chapter 4
Hypothesis Formulations

PART A: - Primary Data
Null hypothesis

1) Investor's Investment objectives and their demographic factors are significantly correlated.
2) There is a strong positive correlation between type of Investor and Investor demographics.
3) Young Investors have higher degree of risk tolerance.
4) Behavioral factors do not influence portfolio decisions.

PART B: - Secondary Data
Set 1 Based On Descriptive Statistics Covering;
Daily log normal returns empirical distribution analyzed :- four moments
(central tendency, dispersion, symmetry, kurtosis) - normality test et al.
1. Mean daily returns are zero. H₀:- µ = 0 v/s H₁ :- µ ≠0.
2. Density (Mass) distribution is Symmetric. H₀:- Skew = 0 v/ s H₁:- SKEW ≠0
3. Density distribution has no fat tails. H₀:- Excess-Kurtosis = 0, v/s H₁:- Excess Kurtosis ≠0.
4. H₀: Half of the data observations fall between the quartiles Q1 & Q3.
5. H₀: Median of the distribution = Mean

❖ Random walk (Unit Root Existence) (ADF) (Augmented Dickey Fuller) test and time series data are interdependent.
  • H₀: There is presence of unit root existence ( Random walk) in all formulations.(ADF)
• H0: The log returns time series exhibit strong interdependency (ACF & PACF)

Set 3 Serial/Auto Correlation And Homoscedasticity (Equal Variance) Assumptions Tested On The Empirical Data Time Series.

• H0: Log returns time series exhibit white noise (no serial/auto) correlation. (LJUNG-BOX).
• H0:-Log returns time series exhibits an ARCH EFFECT.(ARCH EFFECT)

SET 4 Residual Diagnostic (Standardized) Test Analysis
- Population Mean (H0:- µ =0)
- Population Standard Deviation (H0:- σ =1)
- Population Skew (H0:- S =0)
- Population Excess Kurtosis (H0:- K =0)
- White noise or serial (auto) correlation test (H0:- ρ1 =ρ2=…………… = ρk = 0)
- Normality test: H0:- εt ~Φ(0,1)
- White Noise & Arch Effect:- Residuals are Identical & Independently distributed.H0:-εt ~ i. i. d.
- H0:- The negative returns deviate more from normality than positive returns.

SET 5: Goodness of Fit of the Empirical Time Series Data
• H0: – Random walk model is best suited for modeling log return time series.(LLC/AIC)

SET 6: Price Range Volatility Estimators.
- H0:-Extreme value intraday volatility estimators are more efficient than traditional/classical close to close volatility estimators.
- H0:- High frequency data impacts significantly the forecasting ability of the model.

**SET 7: Linear Regression Based**

- OLS Regression Estimators
- ANOVA
- Residual Diagnostics
- Chow (regression stability) test
- Role of influencers/outliers on paradigm shift of empirical distribution towards normality a contrast to the normal rule for this time series study.

Before we consider the Regression Model, we must answer the following question.

- “Is the Regression Model Statistically Significant or a statistical data anomaly”?

- A.N.O.V.A. Test (F test statistic is computed along with its corresponding p value, which enables us to test the overall significance of the Regression Model. We regress the Daily log- Arithmetic returns on a certain stock (Y) v/s the NSE Stock Market Daily log arithmetic returns.(X). If the p value is less than 5% our chosen level of significance, to test the statistical significance of the Null Hypothesis,(False) mentioned under significance column in our reports), we reject the null hypothesis.

- \[ H_0: - \beta_k = 0 \] (Slope coefficient of the linear regression line is insignificant) v/s \[ H_1: - \exists \beta_k \neq 0 \] (Slope coefficient is statistically significant)

- If the Alternative Hypothesis is accepted (\( H_1: \) True or \( H_0: \) False), we may infer that the linear regression model is statistically valid and it does explain some of the variation in the values of the dependent variable (Y).
To test the Individual significance of the O.L.S. Regression estimators (Ordinary Least squares Principle method), we employ the student’s t test to test individually the statistical significance of the estimators, again using the same concept as explained under ANOVA, using the p values computed of the t test statistic.

Residual Diagnostics (Analysis Tableau):- Once having confirmed that the regression model does explain some variation in the values of the response variable (dependent variable/Endogenous variable), we then proceed to examine the residuals to make sure that the underlying assumptions of the model are complied with.

- $H_0$: $\mu = 0$ v/s $H_1$: $\mu \neq 0$ (Average).
- $H_0$: $\sigma = 1$ v/s $H_1$: $\sigma \neq 1$ (Standard Deviation (Volatility)).
- $H_0$: Skew =0 v/s $H_1$: Skew $\neq 0$ (Skewness)
- $H_0$: Kurtosis =0 v/s Kurtosis $\neq 0$ (Kurtosis)
- $H_0$: Residuals are Independent and identically distributed random variables v/s $H_1$: Residuals are Not Normally distributed. (Normality).
- Regression Stability Test (Chow Test):- In this we partition the data into two parts and run the regression independently, for each set, and the output displays the chow statistics with its corresponding p value. In our case we had 1246 daily returns observations hence we partitioned the data set into two halves, each half containing 623 y and x variables pair set each. We ran the usual regression and if the p value is less than 5% we reject the Null hypothesis that is $H_0$: - Regression beta is stable in the two parts v/s $H_1$: - Beta is unstable.
OUTLIERS/INFLUENCERS:- Refer CD Where we have attempted a case on how to identify whether a given data set is an outlier or not, and whether it has influencing properties of navigating the regression line better, once we drop that outlier/influencing observation. This is a tedious process which we experimented only on Reliance Industries stocks, as we need to compute the Cook’s Distance (D), for each observation and compare with the threshold value (Approximately 20% of the number of sample observations), and we drop those points one by one, if the cook’s distance exceeds the threshold limits (Highest value observation first, omit that observation, sample size changes, so also the threshold, again repeat the same step, we had to go through 90 plus iterations to weed out all influencing observations/outliers, which to quite an extent improved the regression model. Refer the $r^2$ measure before and after omitting those outliers. Also refer working in Regression Sheet in MS Excel.