Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII (2022-2023)
OEIT6 - Data Analytics

Experiment: Statistical Analysis

Name: Pushkar Sutar **Roll No. 2019110060**

Objective : Perform statistical data analysis such as: Estimators of the main statistical measures (mean, variance, standard deviation, covariance correlation, standard error), Main distributions (Normal distribution, chi-square distribution), Hypothesis testing, pairwise association (Pearson correlation test, t-test, ANOVA), Non-parametric test.

Dataset Description:
Dataset contains the annual profit of a company till 2020. The data is normally distributed and hence hypothesis testing can be done appropriately.
Attribute Information -

- Year : Year on which profit/loss is recorded.
- Profit/Loss : Profit in thousands. If negative, then loss, else its profit made by the company for a given year.

Code and Output:

First we import the csv data.

```
proc import out=my_data
    datafile="/home/u62322946/DA2/CompanyABCProfit.csv"
    dbms=csv
    replace;
    getnames=YES;
run;
```

We print the data to verify.

```
PROC PRINT DATA=my_data;
RUN;
```

193	2013	1053
194	2014	1302
195	2015	636
196	2016	988
197	2017	895
198	2018	-178
199	2019	543
200	2020	316

Generating summary information about the contents of the dataset.

```
PROC CONTENTS DATA=my_data;
RUN;
```

Data Set Name	WORK.MY_DATA	Observations	200
Member Type	DATA	Variables	2
Engine	V9	Indexes	0
Created	11/08/2022 23:49:35	Observation Length	16
Last Modified	11/08/2022 23:49:35	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	8126
Obs in First Data Page	200
Number of Data Set Repairs	0
Filename	/saswork/SAS_workB08500014267_odaws02-apse1.oda.sas.com/SAS_work481F00014267_odaws02-apse1.oda.sas.com/my_data.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	1460444
Access Permission	rw-r--r--
Owner Name	u62322946
File Size	256KB
File Size (bytes)	262144

Hence, the dataset has 200 observations with two attributes.

We perform univariate analysis to get basic statistical information about the profit attribute.

```
PROC UNIVARIATE DATA=my_data;
  VAR "Profit(Rs '000)"n;
RUN;
```

We see that the profit is normally distributed with a mean of 1021.99Rs and standard deviation of 487.95. Other parameters can also be seen from the table

The UNIVARIATE Procedure
Variable: Profit(Rs '000)

Moments			
N	200	Sum Weights	200
Mean	1021.99	Sum Observations	204398
Std Deviation	487.955626	Variance	238100.693
Skewness	-0.1289021	Kurtosis	-0.0479365
Uncorrected SS	256274750	Corrected SS	47382038
Coeff Variation	47.7456361	Std Error Mean	34.5036732

Basic Statistical Measures			
Location		Variability	
Mean	1021.990	Std Deviation	487.95563
Median	1024.000	Variance	238101
Mode	1082.000	Range	2468
		Interquartile Range	602.00000

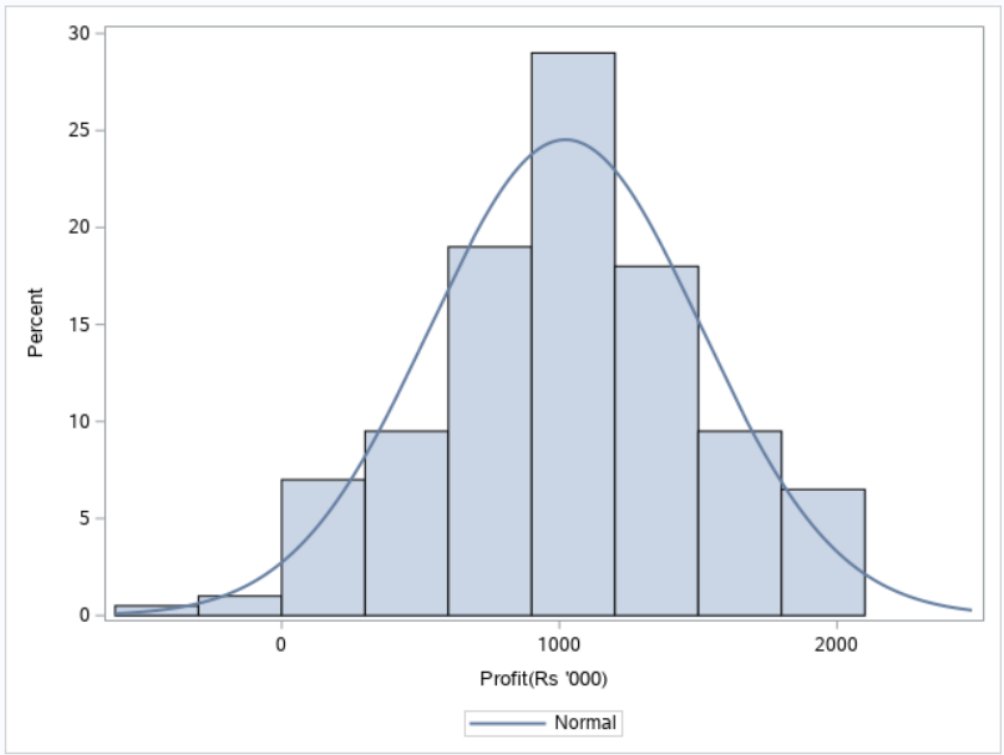
Note: The mode displayed is the smallest of 2 modes with a count of 3.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	29.61974	Pr > |t|	<.0001
Sign	M	97	Pr >= |M|	<.0001
Signed Rank	S	10007	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2055.0
99%	2036.0
95%	1900.0
90%	1683.5
75% Q3	1318.5
50% Median	1024.0
25% Q1	716.5
10%	399.5
5%	180.0
1%	-211.0
0% Min	-413.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-413	168	1973	59
-244	106	1996	169
-178	198	2027	102
26	177	2045	158
49	147	2055	43

```
proc sgplot data=my_data;
  histogram "Profit(Rs '000)"n;
  density "Profit(Rs '000)"n;
run;
```



We take a random sample of 100 points for performing hypothesis testing.

```
proc surveyselect data=my_data method=srs n=100
                  out=sample_data;
run;
```

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MY_DATA
Random Number Seed	777781619
Sample Size	100
Selection Probability	0.5
Sampling Weight	2
Output Data Set	SAMPLE_DATA

We can view the sampled data using the print procedure.

```
PROC PRINT DATA=sample_data;
RUN;
```

Hypothesis Testing -
We perform t-test on the data with 100 samples as data is numeric and the number of samples is also not very large.

We define the hypothesis as, with significance level of 0.05.

Ho : **μ** = 1000 (**μ** indicates the population mean of profit per year)
Ha : **μ** ≠ 1000

```
/* t test */
ods noproctitle;
ods graphics / imagemap=on;

/* Test for normality */
proc univariate data=WORK.SAMPLE_DATA normal mu0=1000;
  ods select TestsForNormality;
  var 'Profit(Rs ''000)'n;
run;
```

```
/* t test */
proc ttest data=WORK.SAMPLE_DATA sides=2 h0=1000 plots(showh0);
  var 'Profit(Rs ''000)'n;
run;
```

Variable: Profit(Rs '000)

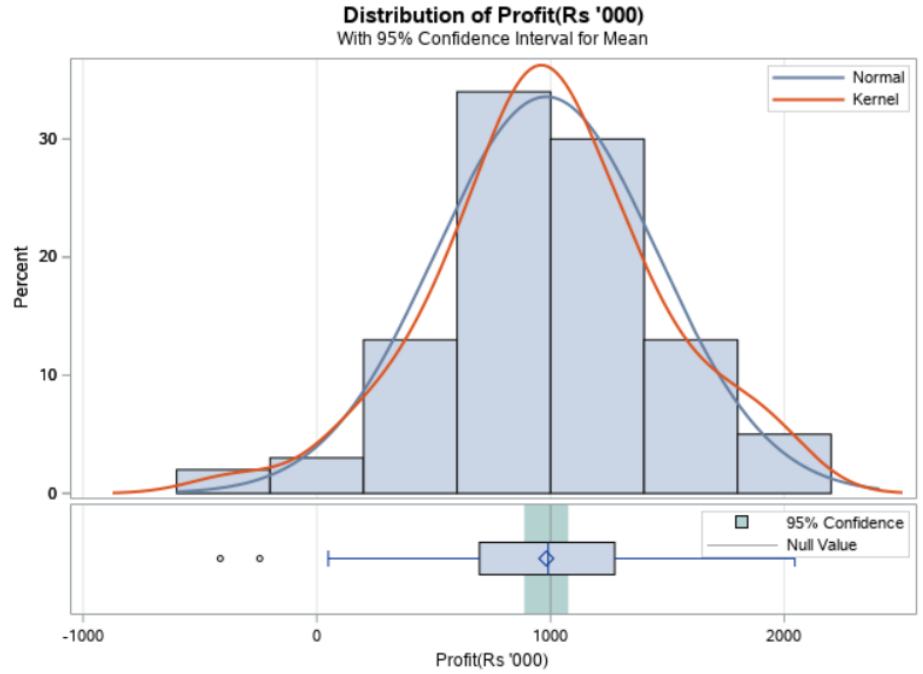
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.988539	Pr < W	0.5487
Kolmogorov-Smirnov	D	0.050464	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.057871	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.348669	Pr > A-Sq	>0.2500

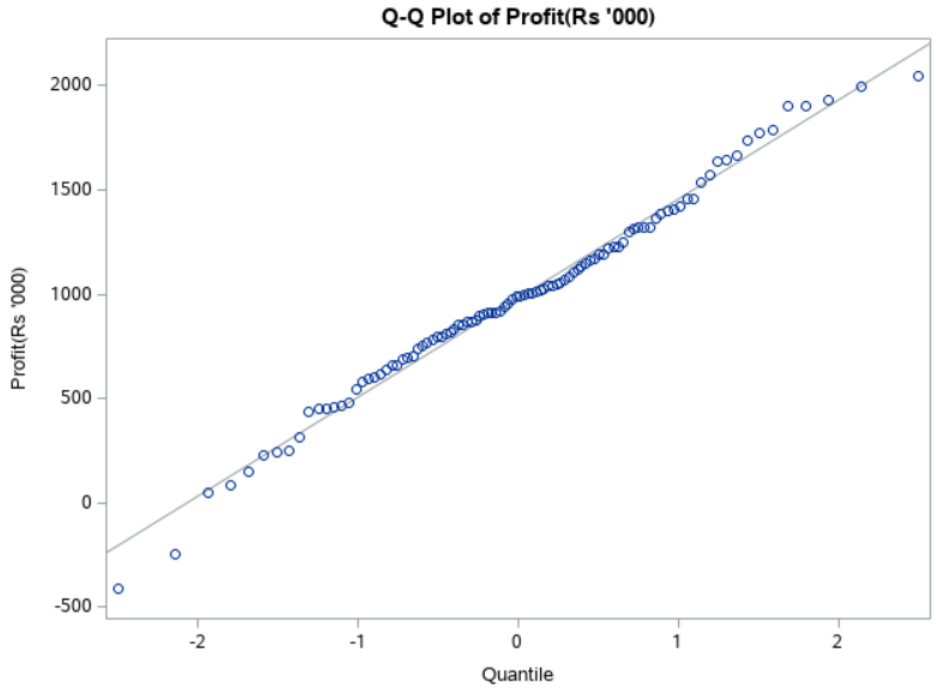
Variable: Profit(Rs '000)

N	Mean	Std Dev	Std Err	Minimum	Maximum
100	981.8	475.2	47.5191	-413.0	2045.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
981.8	887.5	1076.1	475.2	417.2	552.0

DF	t Value	Pr > |t|
99	-0.38	0.7028





Here, the p value is 0.702 which is greater than 0.05. Hence we do not reject the null hypothesis. It can be concluded that the mean of profit is 1000Rs.

Correlation Analysis:

```
/* correlation */
ods noproctitle;
```

```
ods graphics / imagemap=on;

proc corr data=WORK.SAMPLE_DATA pearson nosimple noprob plots=none;
  var 'Profit(Rs ''000)'n;
  with Year;
run;
```

1 With Variables:	Year
1 Variables:	Profit(Rs '000)

Pearson Correlation Coefficients, N = 100	
	Profit(Rs '000)
Year	0.06933

The two variables Year and Profit are not very correlated. There is no significant increase or decrease in profit as the year increases.

Conclusions :

- Statistical Analysis suggests that the population mean is 1000Rs, which means that the company had an average profit of 1000 Rs since its inception.
- Using t-test we can confirm from the randomly selected sample that indeed the profit is equal to 1000 Rs on average.
- Hypothesis testing provides a reliable framework for making any data decisions for your population of interest.
- Hypothesis testing is one of the most important processes for measuring the validity and reliability of outcomes in any systematic investigation. In this case it was the annual profit of a company.
- Profit and Year attributes are not related to each other.