Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII

Experiment : Hypothesis Testing

Name: Sruthi Shivaramakrishnan Batch:A UID:2019110059 Branch: ETRX

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, pair-wise association (Pearson correlation test, t-test, ANOVA), Non-parametric test (Spearman rank0

Platform : SAS
Data Set : SAS help Snacks dataset
Code and ouput:
Considering a few sets of rows for hypothesis testing using different rows.

```
 data first_N_rows;
    set sashelp.snacks;
    if _N_ <= 2000
    then output;
run;
data first_N1_rows;
    set sashelp.snacks;
    if _N_ <= 4000
    then output;
run;
```

Printing the first

```
proc print data=first_N_rows;
```

Obs	QtySold	Price	Advertised	Holiday	Date	Product
1	0.00	1.99	0	0	01JAN2002	Baked potato chips
2	0.00	1.99	0	0	02JAN2002	Baked potato chips
3	0.00	1.99	0	0	03JAN2002	Baked potato chips
4	0.00	1.99	0	0	04JAN2002	Baked potato chips
5	0.00	1.99	0	0	05JAN2002	Baked potato chips
6	0.00	1.99	0	0	06JAN2002	Baked potato chips
7	0.00	1.99	0	0	07JAN2002	Baked potato chips
8	0.00	1.99	0	0	08JAN2002	Baked potato chips
9	0.00	1.99	0	0	09JAN2002	Baked potato chips
10	0.00	1.99	0	0	10JAN2002	Baked potato chips
11	0.00	1.99	0	0	11JAN2002	Baked potato chips
12	0.00	1.99	0	0	12JAN2002	Baked potato chips
13	0.00	1.99	0	0	13JAN2002	Baked potato chips
14	0.00	1.99	0	0	14JAN2002	Baked potato chips
15	0.00	1.99	0	0	15JAN2002	Baked potato chips
16	0.00	1.99	0	0	16JAN2002	Baked potato chips

Printing the mean median mode of the dataset

proc means data=sashelp.snacks mean median mode std var min max;

Variable	Label	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
QtySold	Quantity sold	5.1785795	3.0000000	0	7.5573322	57.1132704	-1.0000000	121.0000000
Price	Retail price of product	2.1018954	1.9900000	2.9900000	0.7763450	0.6027115	0.9900000	3.4900000
Advertised	Advertised (1=yes)	0.0273134	0	0	0.1629973	0.0265681	0	1.0000000
Holiday	Holiday (1=yes)	0.2866928	0	0	0.4522231	0.2045057	0	1.0000000
Date	Date of sale	15851.50	15851.50	15341.00	295.0299703	87042.68	15341.00	16362.00

Hypothesis testing:
T test
Considering the Hypothesis:

H0: The product purchased is related to presence of Holiday
Ha: The product purchased is related to presence of Holiday

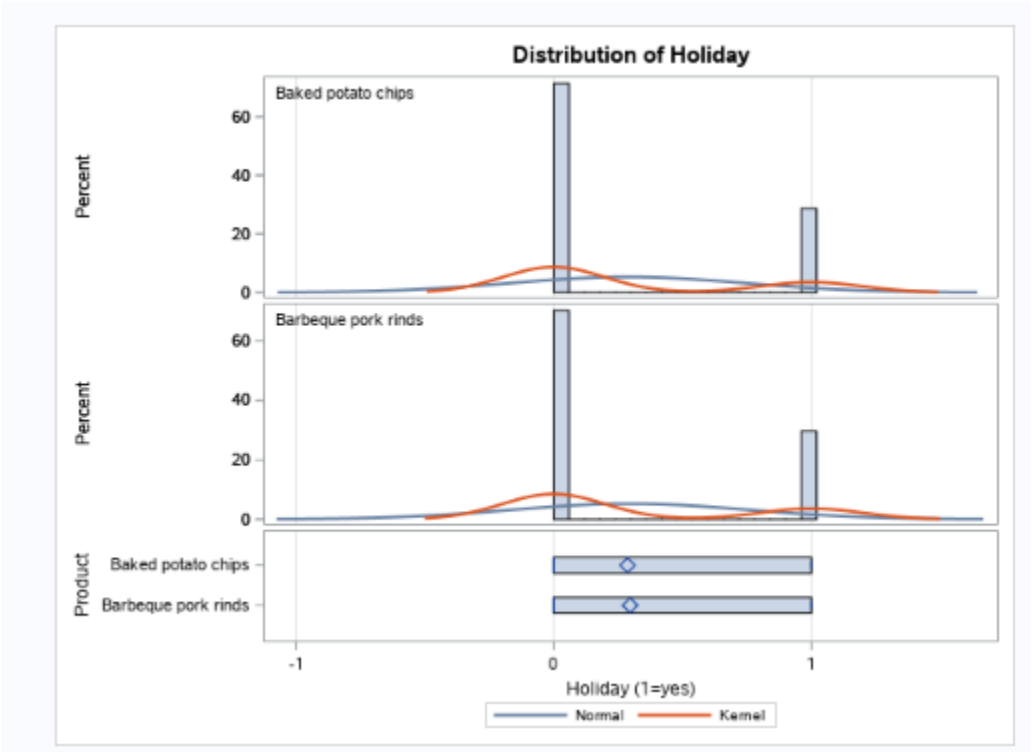
PROC TTEST DATA= first_N_rows;
CLASS Product;
VAR Holiday;
RUN;

Output:

Product	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Baked potato chips		1022	0.2867	0.4524	0.0142	0	1.0000
Barbeque pork rinds		978	0.2965	0.4570	0.0146	0	1.0000
Diff (1-2)	Pooled		-0.00983	0.4547	0.0203		
Diff (1-2)	Satterthwaite		-0.00983		0.0203		

Product	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Baked potato chips		0.2867	0.2589	0.3145	0.4524	0.4336	0.4730
Barbeque pork rinds		0.2965	0.2678	0.3252	0.4570	0.4376	0.4782
Diff (1-2)	Pooled	-0.00983	-0.0497	0.0301	0.4547	0.4410	0.4692
Diff (1-2)	Satterthwaite	-0.00983	-0.0497	0.0301			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1998	-0.48	0.6289
Satterthwaite	Unequal	1992.2	-0.48	0.6290



P value is greater than 0.05 which is 0.6289 hence we accept the null hypothesis which means the product purchased is related to presence of Holiday.

Chi square test:
H0: The presence of 4 products are 25% each of the total number of products.
Ha: The presence of 4 products are greater than 25% each of the total number of products.

```
PROC FREQ DATA =first_N1_rows;
 TABLES Product / CHISQ TESTP=(25 25 25 25);
```

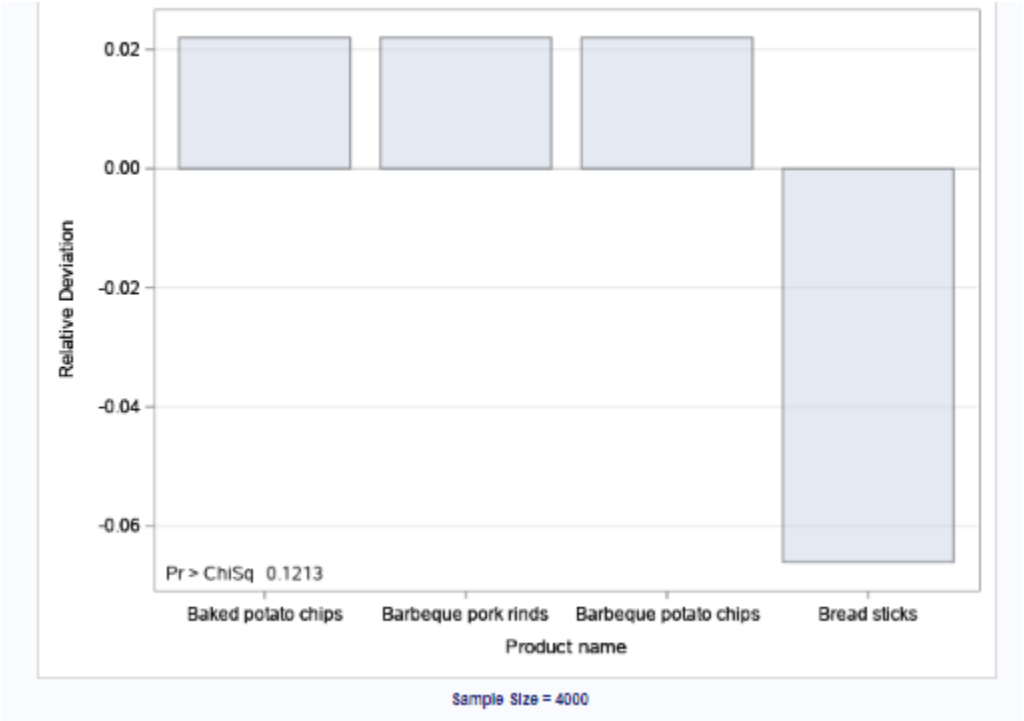
RUN;

Output:

The FREQ Procedure

Product name					
Product	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Baked potato chips	1022	25.55	25.00	1022	25.55
Barbeque pork rinds	1022	25.55	25.00	2044	51.10
Barbeque potato chips	1022	25.55	25.00	3066	76.65
Bread sticks	934	23.35	25.00	4000	100.00

Chi-Square Test for Specified Proportions	
Chi-Square	5.8080
DF	3
Pr > ChiSq	0.1213



Here p value is greater than 0.005 hence we accept the null hypothesis stating that each product has 25% frequency of the total number of products.

Regression:
H0: Holiday and Price attributes are related
Ha : Holiday and Price attributes are not related

PROC REG DATA=sashelp.snacks;

MODEL Price =Holiday;
RUN;

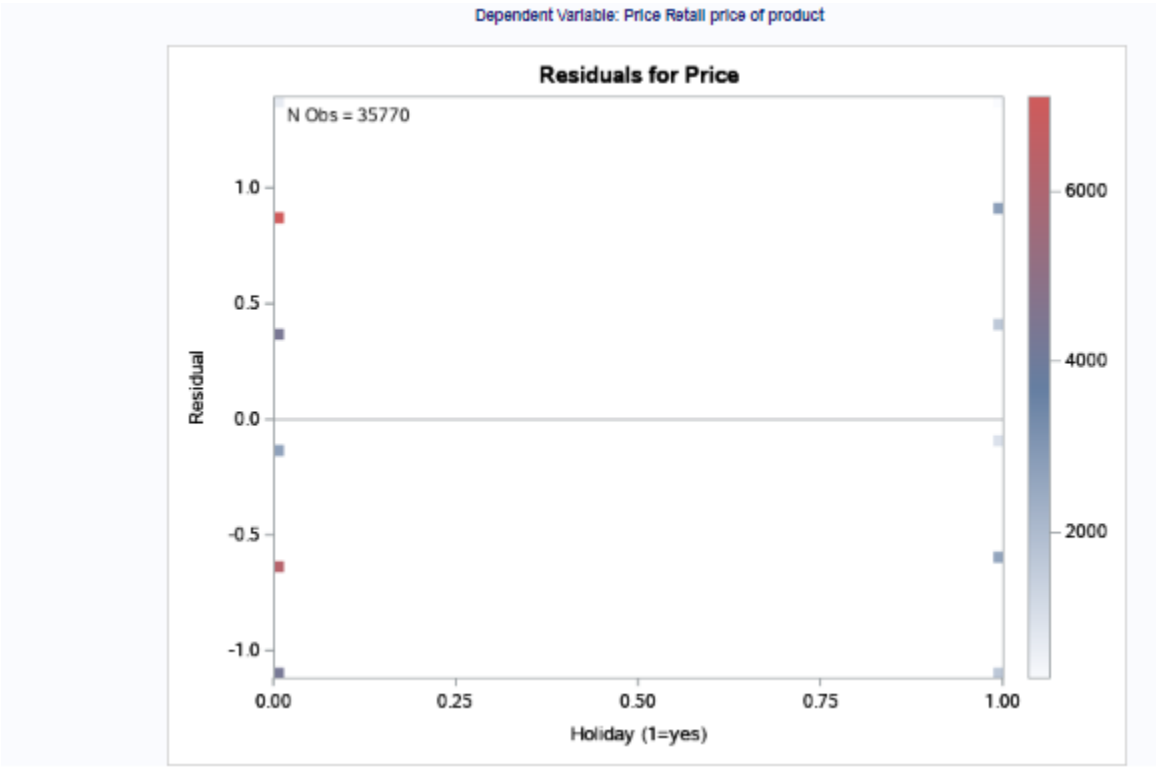
Analysis of Variance

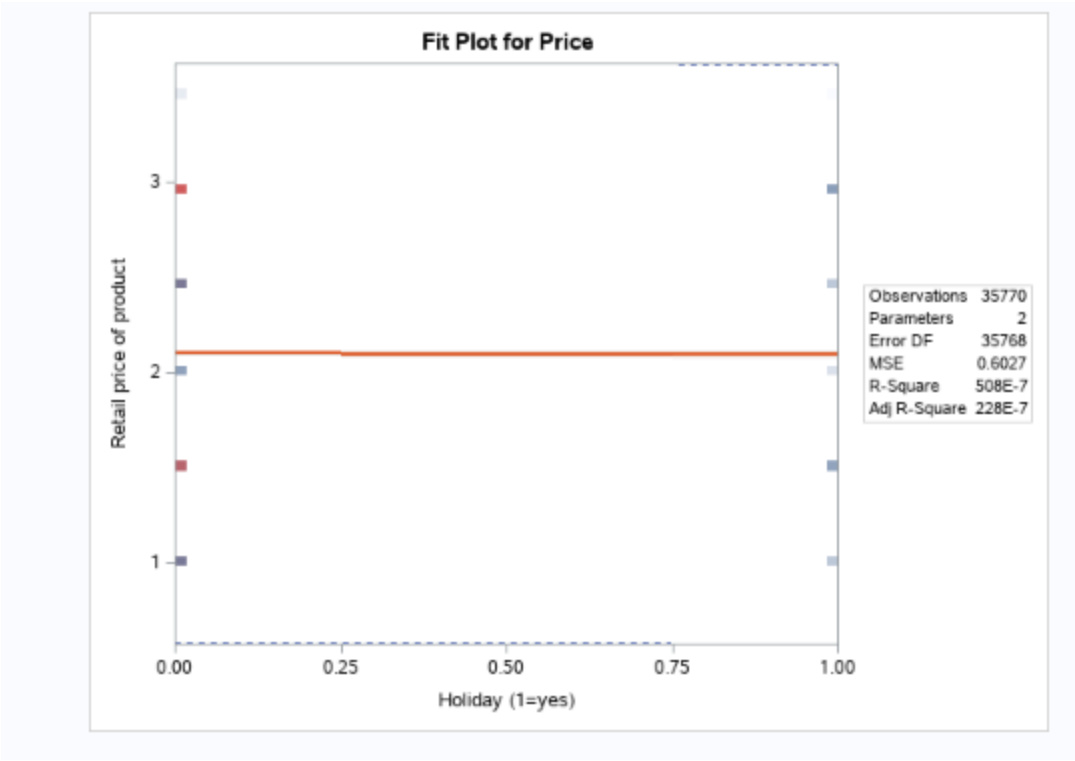
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.09475	1.09475	1.82	0.1777
Error	35768	21557	0.60270		
Corrected Total	35769	21558			

Root MSE	0.77634	R-Square	0.0001
Dependent Mean	2.10190	Adj R-Sq	0.0000
Coeff Var	36.93505		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.10540	0.00486	433.19	<.0001
Holiday	Holiday (1=yes)	1	-0.01223	0.00908	-1.35	0.1777





Here p value is greater than 0.005 hence we accept the null hypothesis stating that Holiday and price attributes are related.

Regression for multivariate analysis:

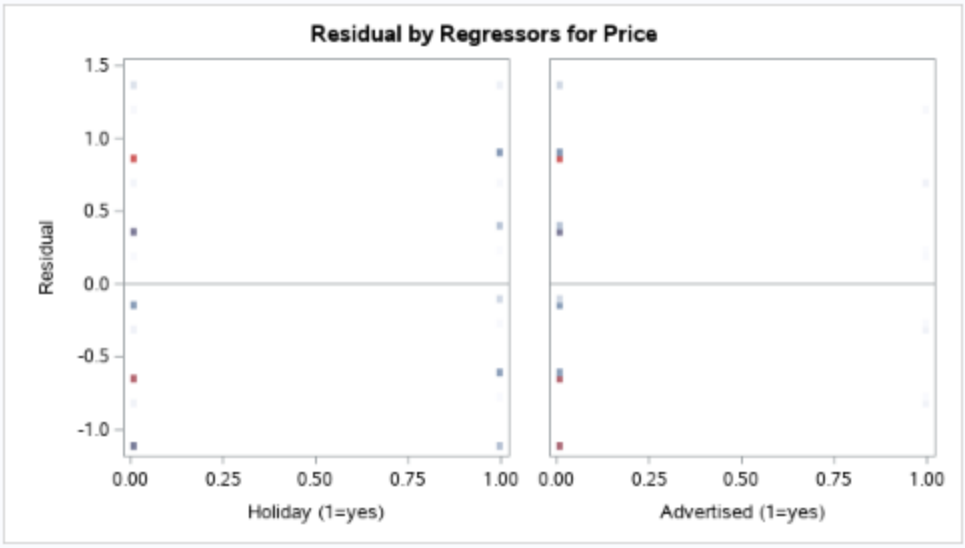
H0: Holiday, Advertised attributes and Price are related
Ha: Holiday, Advertised attributes and Price are not related

Code:

```
PROC REG DATA=sashelp.snacks;
MODEL  Price =Holiday Advertised;
RUN;
```

Output:

The REG Procedure
Model: MODEL1
Dependent Variable: Price Retail price of product



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	103.32345	51.66172	86.12	<.0001
Error	35767	21455	0.59988		
Corrected Total	35769	21558			

Root MSE	0.77450	R-Square	0.0048
Dependent Mean	2.10190	Adj R-Sq	0.0047
Coeff Var	36.84788		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.11559	0.00491	430.78	<.0001
Holiday	Holiday (1=yes)	1	-0.01650	0.00906	-1.82	0.0686
Advertised	Advertised (1=yes)	1	-0.32820	0.02514	-13.05	<.0001

Here P value is less than 0.005 hence we reject the null hypothesis stating that Holiday ,Advertised attributes and Price are not related

Inference:

1. The product purchased and and the presence of Holiday are related to each other.
2. The quantity of products has a 25% frequency in the dataset.
3. The attributes price and attribute are related to each other
4. The attributes Holiday and advertised are not related with price.
5. The attributes which are closely related can be dropped causing dimensionality reduction.