# STAT 6337
# Advanced Statistical Methods I (Fall 2022)
# Project 1

**This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions if something is not clear.**

**Sign on this page below and attach with your project. You project will not be graded without it.**

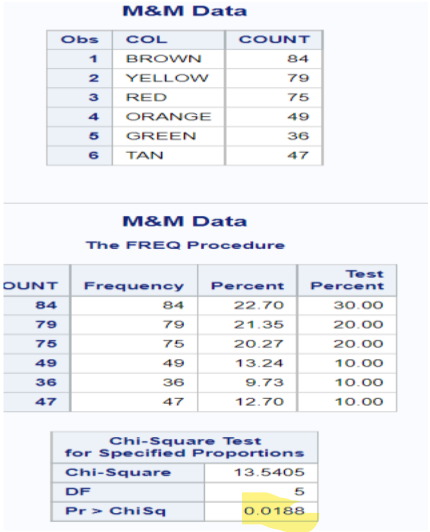
**This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* of UT Dallas https://policy.utdallas.edu/utdsp5003.**

**YOUR NAME** ASHISH MANI ACHARYA

**DATE** 10/12/2022

**YOUR SIGNATURE** ____________________

ASHISH MANI ACHARYA AXA190076 project 1



M&M Data

| Obs | COL | COUNT |
|---|---|---|
| 1 | BROWN | 84 |
| 2 | YELLOW | 79 |
| 3 | RED | 75 |
| 4 | ORANGE | 49 |
| 5 | GREEN | 36 |
| 6 | TAN | 47 |

M&M Data

The FREQ Procedure

| OUNT | Frequency | Percent | Test Percent |
|---|---|---|---|
| 84 | 84 | 22.70 | 30.00 |
| 79 | 79 | 21.35 | 20.00 |
| 75 | 75 | 20.27 | 20.00 |
| 49 | 49 | 13.24 | 10.00 |
| 36 | 36 | 9.73 | 10.00 |
| 47 | 47 | 12.70 | 10.00 |

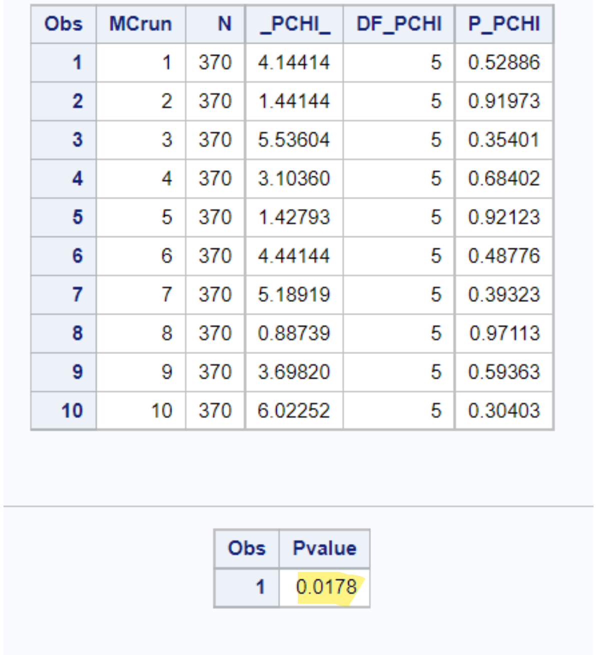
| Chi-Square Test for Specified Proportions | |
|---|---|
| Chi-Square | 13.5405 |
| DF | 5 |
| Pr > ChiSq | 0.0188 |

Q 1 A) Here H0: true percentage is different from manufacturers percentage

HA: they are not and there is no relationship between them
In other words we are trying to find out if true percentage is independent of manufacturers percentage or if there is no relationship between these two percentages.

We know that to find out it that the occurrence of outcomes for the two samples is independent we have to do chi sq test. Using given percentage as expected percentage I did chi sq test on the data and got 0.01888 p value with 5 df . If we suppose the significance value to be .05 then we have to reject this null hypothesis and accept that there is no significant difference between true percentage and manufacturers percentage. The rule being p value less than .05 rejects the null.

Q1B)

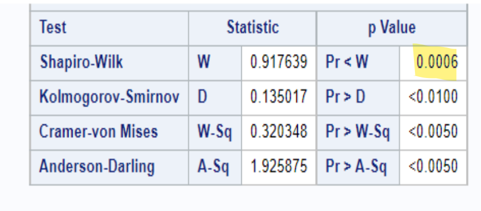


| Obs | MCrun | N | _PCHI_ | DF_PCHI | P_PCHI |
|---|---|---|---|---|---|
| 1 | 1 | 370 | 4.14414 | 5 | 0.52886 |
| 2 | 2 | 370 | 1.44144 | 5 | 0.91973 |
| 3 | 3 | 370 | 5.53604 | 5 | 0.35401 |
| 4 | 4 | 370 | 3.10360 | 5 | 0.68402 |
| 5 | 5 | 370 | 1.42793 | 5 | 0.92123 |
| 6 | 6 | 370 | 4.44144 | 5 | 0.48776 |
| 7 | 7 | 370 | 5.18919 | 5 | 0.39323 |
| 8 | 8 | 370 | 0.88739 | 5 | 0.97113 |
| 9 | 9 | 370 | 3.69820 | 5 | 0.59363 |
| 10 | 10 | 370 | 6.02252 | 5 | 0.30403 |

| Obs | Pvalue |
|---|---|
| 1 | 0.0178 |

Here we get p value as 0.01788 for chi sq test through monte Carlo simulations which is smaller than the one we got from traditional chi square test. I can say that the reason the p value might be different is because we are sampling and simulating the data again and again for hundreds of time. Hence minor differences from p value gained through usual approach is understandable. What we need to emphasis is though the p value differs, it doesn't differ significantly( differs by.001) and we still reject H0 and even when computing p value through Monte Carlo approach we reject the notion that true percentage is different from manufacturing percentage.

Q1C) At 5 % significance level through both chi square test and chi square test through Monte Carlo approach we reject H0 that true percentage is different from manufacturing percentage and we can say that these two percentages are not independent from one another and there exists a significant relationship between true percentage and manufacturers percentage.



| Test | Statistic | | p Value | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.917639 | Pr < W | 0.0006 |
| Kolmogorov-Smirnov | D | 0.135017 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 0.320348 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 1.925875 | Pr > A-Sq | <0.0050 |

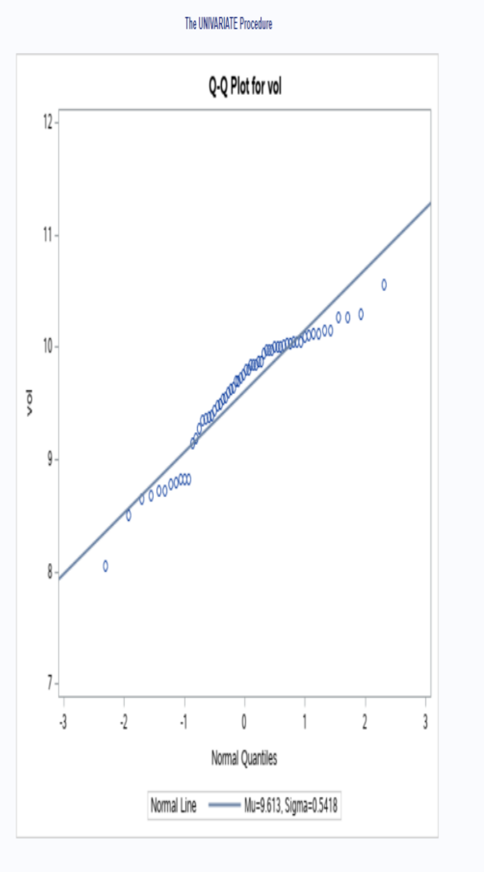
Q2) The manufacturing process can be established locally if there is no difference in the population means of voltage readings at the two locations.

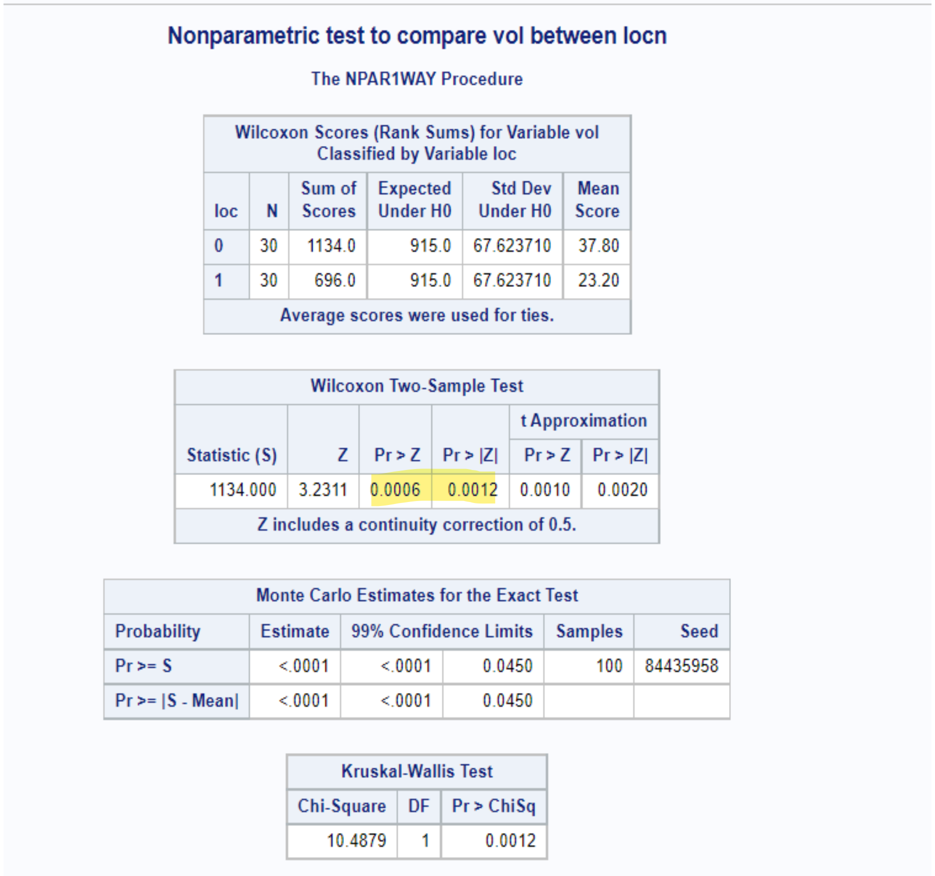
Here, H0: The difference of the mean between the two voltages of the population is equal to zero.

HA:- The difference of the mean between the two voltages of the population is not equal to zero.

To prove this hypothesis we could use a parametric test but since we found out that the voltage variable is not similar , we have to use non parametric counterpart to this parametric test which is Wilcoxon test. The proof of non normality is the Shapiro test p value which is .0006 and is less than .05 in adjoining diagram . Also QQ plot below proves non normality of the variable as the data seems disjointed and not seem to follow QQ line.

ASHISH MANI ACHARYA AXA190076 project 1





**Nonparametric test to compare vol between locn**

The NPAR1WAY Procedure

| Wilcoxon Scores (Rank Sums) for Variable vol Classified by Variable loc | | | | | |
|---|---|---|---|---|---|
| loc | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| 0 | 30 | 1134.0 | 915.0 | 67.623710 | 37.80 |
| 1 | 30 | 696.0 | 915.0 | 67.623710 | 23.20 |

Average scores were used for ties.

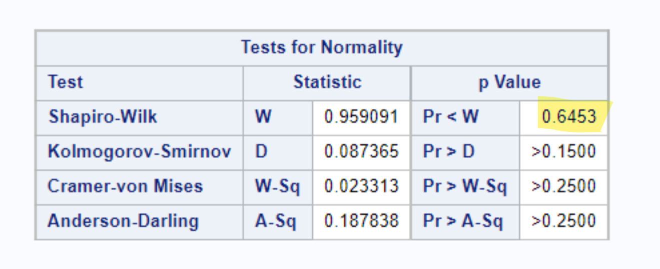
| Wilcoxon Two-Sample Test | | | | | |
|---|---|---|---|---|---|
| | | | | t Approximation | |
| Statistic (S) | Z | Pr > Z | Pr > \|Z\| | Pr > Z | Pr > \|Z\| |
| 1134.000 | 3.2311 | 0.0006 | 0.0012 | 0.0010 | 0.0020 |

Z includes a continuity correction of 0.5.

| Monte Carlo Estimates for the Exact Test | | | | | |
|---|---|---|---|---|---|
| Probability | Estimate | 99% Confidence Limits | | Samples | Seed |
| Pr >= S | <.0001 | <.0001 | 0.0450 | 100 | 84435958 |
| Pr >= \|S - Mean\| | <.0001 | <.0001 | 0.0450 | | |

| Kruskal-Wallis Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 10.4879 | 1 | 0.0012 |

We know that t if a p value less than 0.05 in Wilcoxon test means that the two data are significantly different and the p-value of 1 means that are exactly same if we take significance level 95 %. Here we have one sided p value is .0006 and two sided is .0012. since both p values specially the two sided p value is lower than .05 we reject the H0 and accept the fact that there is significant difference between voltage readings and manufacturing cant be established locally.

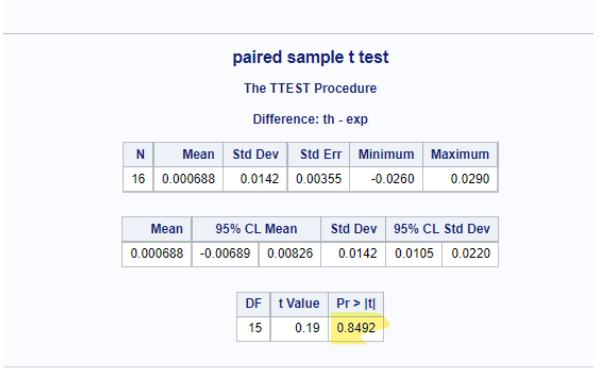


| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.959091 | Pr < W | 0.6453 |
| Kolmogorov-Smirnov | D | 0.087365 | Pr > D | >0.1500 |
| Cramer-von Mises | W-Sq | 0.023313 | Pr > W-Sq | >0.2500 |
| Anderson-Darling | A-Sq | 0.187838 | Pr > A-Sq | >0.2500 |

Q3) Here we have to prove that experimental values of vapor pressure should be equal to theoretical values of vapor pressure or in other words the difference should be almost equal to zero .We know that in such case we need to perform paired sample t test which has following hypothesis.

- $H_0$: $\mu_1 = \mu_2$ (the two population means are equal)
- $H_A$ (two-tailed): $\mu_1 \neq \mu_2$ (the two population means are not equal)

The adjoining box shows the difference has normal distribution with Shapiro Wilks p value of .64 which is way bigger than .05 . Hence we are free to use paired sample t test.



**paired sample t test**

The TTEST Procedure

Difference: th - exp

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 16 | 0.000688 | 0.0142 | 0.00355 | -0.0260 | 0.0290 |

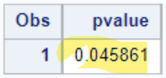
| Mean | 95% CL Mean | | Std Dev | 95% CL Std Dev | |
|---|---|---|---|---|---|
| 0.000688 | -0.00689 | 0.00826 | 0.0142 | 0.0105 | 0.0220 |

| DF | t Value | Pr > \|t\| |
|---|---|---|
| 15 | 0.19 | 0.8492 |

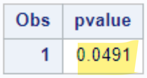
Here p value for paired t test is .8492 which is way bigger than significance level of .05. Hence, we accept the H0 that population mean are almost equal. Hence there is no sufficient evidence that the average difference differs significantly from zero.

ASHISH MANI ACHARYA AXA190076 project 1

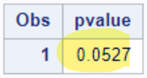
Q4) First of all we have inputed two datas separately as ethnicity and prevalence of diseases as given by the question. Then we used Monte Carlo approach to estimate the type I error in this question. For getting type I error probability based on different n sample , I got the following: for when n= 20,50,100 and 500 respectively. Here capital N is used for replication and we should not confused   with small n which is sample.



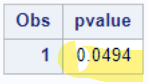
| Obs | pvalue |
|---|---|
| 1 | 0.045861 |



| Obs | pvalue |
|---|---|
| 1 | 0.0491 |



| Obs | pvalue |
|---|---|
| 1 | 0.0527 |



| Obs | pvalue |
|---|---|
| 1 | 0.0494 |

Based  on my calculation we can reject the probability of type I error when n= 20, 50 and 500 . Similarly based on my calculations there is probability of type I error when n=100.
We also have to note that since 0.0527 is pretty close to .05 , some other researcher might even get no probability of type I error when n=100 as calculations differ slightly every time we do MC simulations.

SAS CODES:-

```
/*Q1 A */
filename MNM "/home/u59316208/MNM.DAT";

DATA MNM;
INFILE MNM;
INPUT COL $ COUNT;
RUN;

/* seeing the data */
PROC PRINT DATA=MNM (obs=6); TITLE 'M&M Data';
RUN;

/* Applying the chis-square test of homogenity with the given percentages */
PROC FREQ DATA=MNM order = DATA;
        weight COUNT;
        tables COUNT / nocum chisq testp = (30 20 20 10 10 10);
OUTPUT out=B CHISQ;
run;

proc freq data= MNM;
weight COUNT;
table COL/ testp=(0.30   0.20   0.20   0.10   0.10   0.10);
exact chisq / MC;
run;

/*Q1B */

DATA B; SET B; Nruns=10000;
Run;

/*MC simulations */
DATA MC; SET B;
call streaminit(8733456);
DO MCrun=1 TO Nruns;
          DO j=1 TO N;              /* N run samples */
     x = RAND("Table", 0.30, 0.20, 0.20, 0.1, 0.1, 0.1); /* generating N( 0,1) variable*/
     OUTPUT;
     END;
END;
RUN;
```

```
PROC PRINT DATA=MC (obs=10);
RUN;

PROC FREQ DATA = MC NOPRINT;
BY MCrun;
TABLE x / CHISQ TESTP = (30, 20, 20, 10, 10, 10);
OUTPUT OUT=MC1 CHISQ;
RUN;

PROC PRINT DATA=MC1 (obs=10);
RUN;

DATA B; SET B;
DO i=1 to Nruns; OUTPUT; END;
Run;

DATA MC2; MERGE B MC1;
indicator = (13.540540541 <= _PCHI_ );
RUN;

PROC MEANS DATA = MC2 NOPRINT;
VAR indicator;
OUTPUT OUT = MC3 MEAN=Pvalue;
Run;

/* finding p value through MC simulations */
PROC PRINT DATA = MC3;
Var Pvalue;
Run;

/*Q2 */
filename c "/home/u59316208/VOLTAGE.DAT";
data c;
infile c;
input loc $ vol; /* here location and voltage been shortened to loc and vol respectively for ease of coding */
run;

/* checking normality */
proc univariate data=c normal;
qqplot vol/normal (mu=est sigma=est color=blueviolet l=1);
run;
```

```
/* wilcoxon test */
proc NPAR1WAY data=c WILCOXON;
title "Nonparametric test to compare vol between locn";
class loc;
var vol;

run;

/*Q3 */
filename vap "/home/u59316208/VAPOR .DAT";
data vap;
infile vap firstobs=2;
input temp th exp; /* here theoretical and experimental been shortened to th and exp
respectively for ease of coding */
run;



/* getting differences*/
data vap;
set vap;
diff=(th-exp);
run;

/* checking normality */
proc univariate data=vap normal;
qqplot diff/normal (mu=est sigma=est color=blueviolet l=1);
run;

/* paired sample t test */
proc ttest data=vap sides=2 alpha=.05 h0=0;
title "paired sample t test";
paired th*exp;
run;

/*Q4 */
/*when n =20 */
DATA MC;
RETAIN SEED 90076;
DO MCrun=1 TO 10000;
  DO n=1 TO 20;
    ethnicity = RAND("Table", 0.62, 0.13, 0.18, 0.07);
    disease_probabilty = RAND("Table", 0.75, 0.25);
    OUTPUT;
  END;
END;
RUN;

/*use ch-square test of independence between ethnicity and disease_probabilty */
proc freq data=MC noprint;
by MCrun;
table ethnicity*disease_probabilty / chisq;
output out=K chisq;
run;

proc print data=K (obs=100);
run;

DATA K; set K;
indicator = (0.05 >= P_PCHI);
RUN;

PROC MEANS DATA = K NOPRINT;
VAR indicator;
output out=J MEAN = pvalue;
run;

proc print data = J;
var pvalue;
run;

/*when n=50 */
DATA MC;
RETAIN SEED 90076;
DO MCrun=1 TO 10000;
  DO n=1 TO 50;
    ethnicity = RAND("Table", 0.62, 0.13, 0.18, 0.07);
    disease_probabilty = RAND("Table", 0.75, 0.25);
    OUTPUT;
  END;
END;
RUN;

/*use ch-square test of independence between ethnicity and disease_probabilty */
proc freq data=MC noprint;
```

```
by MCrun;
table ethnicity*disease_probabilty / chisq;
output out=K chisq;
run;

proc print data=K (obs=100);
run;

DATA K; set K;
indicator = (0.05 >= P_PCHI);
RUN;

PROC MEANS DATA = K NOPRINT;
VAR indicator;
output out=J MEAN = pvalue;
run;

proc print data = J;
var pvalue;
run;

/*when n=500 */
DATA MC;
RETAIN SEED 90076;
DO MCrun=1 TO 10000;
  DO n=1 TO 500;
    ethnicity = RAND("Table", 0.62, 0.13, 0.18, 0.07);
    disease_probabilty = RAND("Table", 0.75, 0.25);
    OUTPUT;
  END;
END;
RUN;

/*use ch-square test of independence between ethnicity and disease_probabilty */
proc freq data=MC noprint;
by MCrun;
table ethnicity*disease_probabilty / chisq;
output out=K chisq;
run;

proc print data=K (obs=100);
run;
```

```
DATA K; set K;
indicator = (0.05 >= P_PCHI);
RUN;

PROC MEANS DATA = K NOPRINT;
VAR indicator;
output out=J MEAN = pvalue;
run;

proc print data = J;
var pvalue;
run;

/*when n=100 */
DATA MC;
RETAIN SEED 90076;
DO MCrun=1 TO 10000;
  DO n=1 TO 100;
    ethnicity = RAND("Table", 0.62, 0.13, 0.18, 0.07);
    disease_probabilty = RAND("Table", 0.75, 0.25);
    OUTPUT;
  END;
END;
RUN;

/*use ch-square test of independence between ethnicity and disease_probabilty */
proc freq data=MC noprint;
by MCrun;
table ethnicity*disease_probabilty / chisq;
output out=K chisq;
run;

proc print data=K (obs=100);
run;

DATA K; set K;
indicator = (0.05 >= P_PCHI);
RUN;

PROC MEANS DATA = K NOPRINT;
VAR indicator;
output out=J MEAN = pvalue;
run;
```

ASHISH MANI ACHARYA AXA190076 project 1

```
proc print data = J;
var pvalue;
run;
```