Michael Shoemate

Professor Choudhary

Advanced Statistical Methods Project I

**Problem 1**

Please consider that the number of samples is not 50 from each sector, there are instead 35, 70, 49 and 46 samples. This does not have a significant effect the analysis.

(a) Does public opinion regarding the choice of technology options for generating electricity differ among the four groups?

This question may be tested via a Chi-Square test of homogeneity, with 6 degrees of freedom. The test is staged:

H₀: The opinion for each group is homogenous.

H₁: The opinion for each group is not homogenous.

PROC FREQ DATA=home.coalnuc;

TABLES opinion \* group / CHISQ;

WEIGHT number;

The computed test statistic is 28.18, with a p-value of <.0001. Since the p-value is less than the significance level of .05, we reject H₀. It is concluded that public opinion regarding the choice of technology options differ among the four groups.

(b) Does there appear to be more overall support for the coal option rather than the nuclear option?

The test is staged:

H₀: *p₀* = *p₁*

H₁: *p₀* > *p₁*

Where p₀ is the probability of preferring coal, and p₁ is the probability of supporting nuclear.

Use the test statistic:

Since each sample is not independent, a correction must be applied.

Consider each sample from a multinomial distribution, with 1 trial per sample.

Compute the standard deviation by taking the sum of the variances, less the covariance to compensate for dependence.

The probability of preferring nuclear or coal is .175 and .46 respectively. From SAS, the critical point of a two sided interval at significance level of .05 for coal is .26, which is less than .46. Therefore reject H₀, there is more overall support for the coal option than the nuclear option.

(c) Construct a 95% confidence interval for the percentage of environmentalists and conservationists who support the nuclear option and interpret the result.

First, transform the data into a count for the number of environmentalists and conservationists who support the nuclear option, against everyone else. The transformation is included in the code appendix. Then construct a confidence interval via binomial proportions:

PROC FREQ DATA=coalnuc\_1c;

TABLES group / BINOMIAL(ac wilson EXACT) ALPHA=.05;

WEIGHT number;

Using the Clopper-Pearson exact interval, the percentage of environmentalists and conservationists who support nuclear is within the interval [2.34%, 8.66%] with 95% confidence.

**Problem 2**

(a) Find the p-value for an appropriate test of hypothesis to determine whether the true percentages of colors produced differ from the manufacturer’s stated percentages.

To find whether the true percentages of color differ from the manufacturer’s stated percentages, apply a goodness-of-fit Chi-Square test against the provided multinomial null distribution.

The test is staged:

H₀: The null distribution is equal to the sample distribution

H₁: The null distribution is different from the sample distribution

PROC FREQ DATA=home.m\_m ORDER=DATA;

TABLES color / CHISQ testp=(.3, .2, .2, .1, .1, .1);

WEIGHT number;

The computed test statistic is 13.5405, with a p-value of 0.01821.

(b) How does the p-value from Monte Carlo simulation compare with the one obtained in (a)?

To estimate the test statistic, one could first sample from the multinomial null distribution, which will approximate a multivariate normal distribution at large sample sizes. The process of computing the test statistic standardizes the multivariate normal and squares it, approximating a chi-square distribution at 5 degrees of freedom. Therefore one can equivalently sample from

χ²(ν =5) to approximate the p-value provided by a chi-square goodness-of-fit test.

PROC IML;

samples = 100000;

CALL RANDSEED(160830);

matrix = J(samples, 1);

CALL RANDGEN(matrix, 'CHISQUARE', 5);

pvalue = sum(matrix > 13.5405) / samples;

PRINT pvalue;

The computed p-value is .0182, which is relatively close to .0188. Notice that as the estimated value gets nearer zero, a much larger number of Monte Carlo iterations is required to get an accurate result.

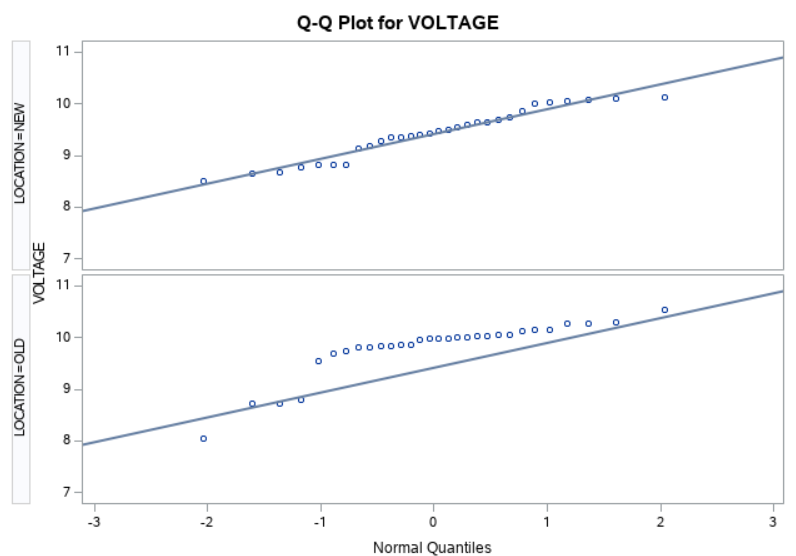
(c) Do the samples match the proportions provided by the manufacturer?

Since the p-value = .018 is smaller than α = .05, then H₀ is rejected. The samples do not match the proportions provided by the manufacturer.

**Problem 3**

Does it appear that the manufacturing process can be established locally?

Before applying a t-test, verify the assumption of normality visually from a QQ-Plot.



Notice the old location is not normal. Therefore the assumptions needed to use a t-test have been violated; a nonparametric test is required. The Wilcoxon rank sum test / Mann-Whitney U test will be used instead. We can safely assume that the samples are independent.

The test is staged:

H₀: μ₀ = μ₁

H₁: μ₀ ≠ μ₁

PROC NPAR1WAY WILCOXON CORRECT=NO DATA=home.voltage;

VAR voltage;

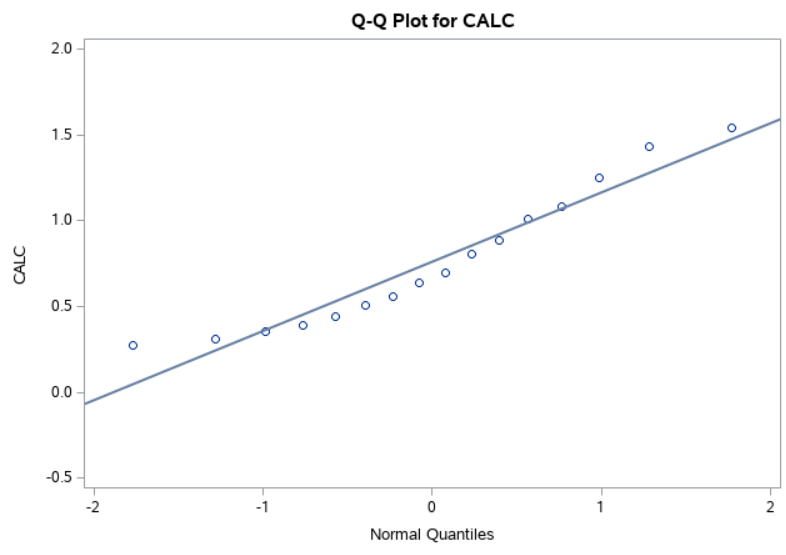
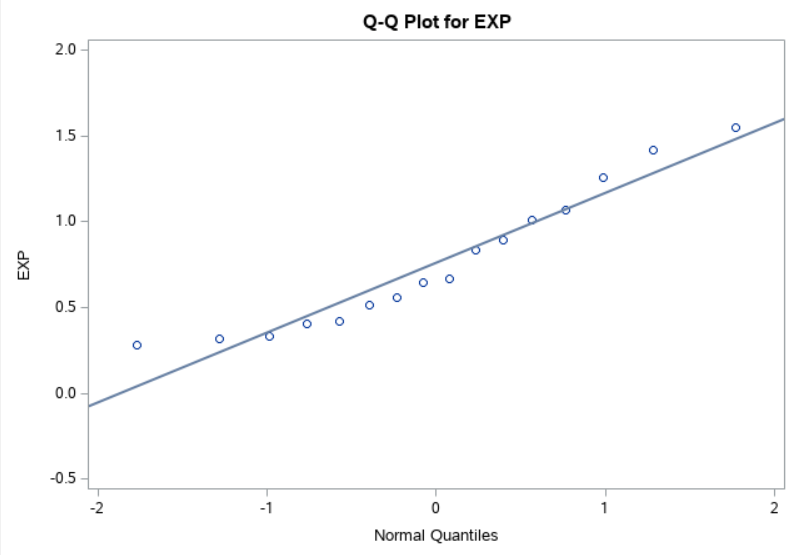
CLASS location;

The resulting p-value is .0002, which is very significant. We reject H₀ and conclude that the manufacturing process can be established locally.

**Problem 4**

Perform an appropriate analysis of the VAPOR data to see if there is sufficient evidence to indicate that the mean difference differs from zero.

First, check for normality. The data appears to be normally distributed.



Now apply the F-test for homogeneity of variances.

H₀: σ₁² = σ₂²

H₁: σ₁² ≠ σ₂²

The resulting F-ratio is .9962, which is a strong indicator for homogeneity of variances. Do not reject H₀.

PROC SQL;

CREATE TABLE vapor\_melt AS

SELECT calc AS value, 'CALC' AS label FROM home.vapor

UNION ALL SELECT exp AS value, 'EXP' AS label FROM home.vapor;

PROC GLM DATA=vapor\_melt;

CLASS label;

MODEL value = label;

Given that the samples are independent, but paired, apply the paired t-test. The assumption is made that the order of the elements in the datafile defines the pair order.

H₀: μ₀ = μ₁

H₁: μ₀ ≠ μ₁

PROC TTEST DATA=home.vapor;

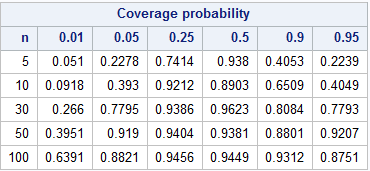
PAIRED exp\*calc;

The resulting p-value is 0.8492, which is greater than .05, so do not reject H₀. There is not sufficient evidence to indicate the mean distances differ from zero.

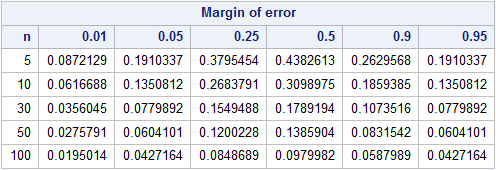
**Problem 5**

We know how to construct a large-sample confidence interval for a population proportion p. How large should n be for this interval to have acceptable accuracy?

The table of coverage probabilities is provided:



Notice as n becomes small, the coverage probability tends to decrease. Similarly, coverage probabilities decrease as p gets closer to 0 or 1. However, consider that the error margins for the population also decrease as p becomes extreme. Intuitively, of course, error margins decrease as n becomes larger.



The margin of error E is defined . To estimate the number of samples needed to satisfy an error bound, solve for n.

n =

Notice the function n(p) is maximized at .25 when p = .5. Intuitively as p approaches .5, samples become more entropic, and the confidence interval becomes wider. Larger n is needed to bound the margin of error.

Michael Shoemate

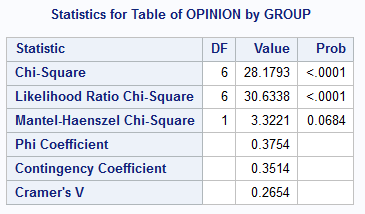
Professor Choudhary

Advanced Statistical Methods Project I – SAS Output

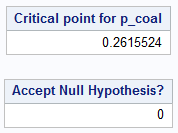
Some of the SAS output was already included in the report and has been omitted here with note.

**Problem 1**

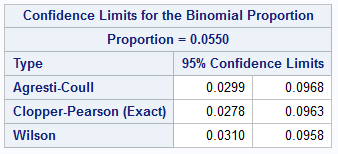
(a)



(b) Output from PROC IML;

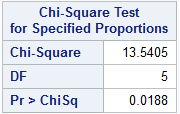


(c)



**Problem 2**

(a)

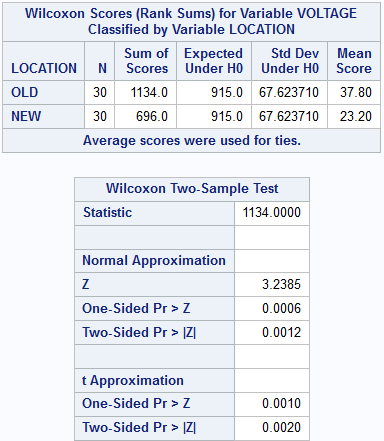


(b) Output from PROC IML;



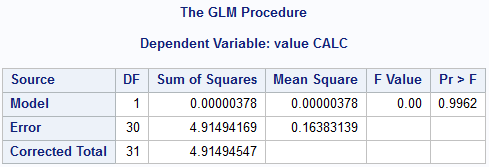
**Problem 3**

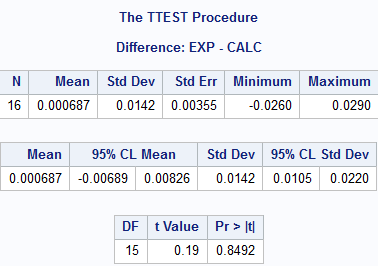
A QQ plot is included with the report.



**Problem 4**

QQ plots are included in the report.





**Problem 5**

Tables produced by PROC IML are included in the report.

Michael Shoemate

Professor Choudhary

Advanced Statistical Methods Project I – SAS Code

%LET seed=160830;

LIBNAME home '/folders/myfolders/';

TITLE '1.a.';

PROC FREQ DATA=home.coalnuc;

TABLES opinion \* group / CHISQ;

WEIGHT number;

RUN;

TITLE '1.b.';

PROC IML;

c = {92, 35, 73};

p = c/c[+];

deviation = SQRT((p[1] \* (1 - p[1]) + p[2] \* (1 - p[2]) + 2 \* p[1] \* p[2]) / c[+]);

critical\_point = p[2] + QUANTILE('NORMAL', .95) \* deviation;

PRINT critical\_point[LABEL='Critical point for p\_coal'];

result = p[1] < critical\_point;

PRINT result[LABEL='Accept Null Hypothesis?'];

TITLE '1.c. Confidence interval for percentage of environmentalists who support nuclear';

PROC SQL;

CREATE TABLE coalnuc\_1c AS

SELECT group, SUM(number) AS number FROM (

SELECT CASE WHEN group EQ 'E&C' AND OPINION EQ 'NUCLEAR' THEN 'E&C-NUCLEAR' ELSE 'NOT E&C-NUCLEAR' END AS group, \*

FROM home.coalnuc

) GROUP BY group;

PROC FREQ DATA=coalnuc\_1c;

TABLES group / BINOMIAL(ac wilson EXACT) ALPHA=.05;

WEIGHT number;

RUN;

TITLE '2.a';

PROC FREQ DATA=home.m\_m ORDER=DATA;

TABLES color / CHISQ testp=(.3, .2, .2, .1, .1, .1);

WEIGHT number;

RUN;

TITLE '2.b and 2.c';

TITLE2 'Since P(χ² > obs) ≈ .0188 < .05 = α, reject H₀.';

PROC IML;

samples = 100000;

CALL RANDSEED(&seed);

matrix = J(samples, 1);

CALL RANDGEN(matrix, 'CHISQUARE', 5);

pvalue = sum(matrix > 13.5405) / samples;

PRINT pvalue;

RUN;

TITLE '3. Data does not appear to be normal';

PROC UNIVARIATE DATA=home.voltage;

CLASS location;

QQPLOT voltage / NORMAL(MU=est SIGMA=est COLOR=red L=2);

TITLE '3. Test for equality of means';

PROC NPAR1WAY WILCOXON CORRECT=NO DATA=home.voltage;

VAR voltage;

CLASS location;

RUN;

TITLE '4. Checking normality';

PROC UNIVARIATE DATA=home.vapor;

QQPLOT exp / NORMAL(MU=est SIGMA=est COLOR=red L=2);

QQPLOT calc / NORMAL(MU=est SIGMA=est COLOR=red L=2);

TITLE '4. Checking homogeneity of variances';

PROC SQL;

CREATE TABLE vapor\_melt AS

SELECT calc AS value, 'CALC' AS label FROM home.vapor

UNION ALL SELECT exp AS value, 'EXP' AS label FROM home.vapor;

PROC GLM DATA=vapor\_melt;

CLASS label;

MODEL value = label;

run;

TITLE '4. Vapor analysis';

PROC TTEST DATA=home.vapor;

PAIRED exp\*calc;

RUN;

TITLE '5.';

PROC IML;

CALL RANDSEED(&seed);

samples = 10000;

sample\_sizes = {5, 10, 30, 50, 100};

pop\_proportions = {.01, .05, .25, .5, .9, .95};

/\* p-values stored inside \*/

results = J(NROW(sample\_sizes), NROW(pop\_proportions) + 1);

margins = J(NROW(sample\_sizes), NROW(pop\_proportions) + 1);

/\* compute p-values for every combination of sample size and probability \*/

DO i = 1 to NROW(sample\_sizes);

results[i, 1] = sample\_sizes[i];

margins[i, 1] = sample\_sizes[i];

DO j = 1 to NROW(pop\_proportions);

props = J(samples, 1);

CALL RANDGEN(props, 'BINOMIAL', pop\_proportions[j], sample\_sizes[i]);

props = props / sample\_sizes[i];

z\_stat = QUANTILE('NORMAL', .975);

interval = z\_stat \* SQRT(props # (1 - props) / sample\_sizes[i]);

results[i, j + 1] = SUM(ABS(props - pop\_proportions[j]) < interval) / samples;

margins[i, j + 1] = z\_stat \* SQRT(pop\_proportions[j] \* (1 - pop\_proportions[j]) / sample\_sizes[i]);

END;

END;

/\* pretty-print the table \*/

colnames = 'n' // CHAR(pop\_proportions);

PRINT results[COLNAME=colnames LABEL='Coverage probability'];

PRINT margins[COLNAME=colnames LABEL='Margin of error'];