Statistical Analysis on 2008 – 2012 IL & NC Alzheimer's Mortality Dataset

By: Michael Carnival

Last Updated: 12/7/2023

Contents

Introduction

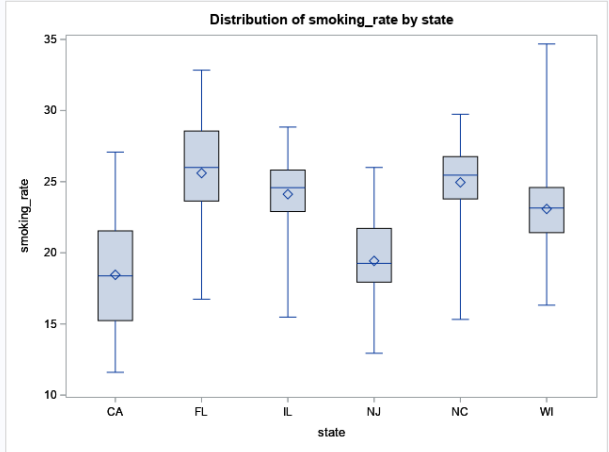
Group 2 were tasked to identify trends in the AD mortality datasets provided by Professor Amin. This project aimed to conduct statistical analysis on several states where team members picked their states and variables to analyze. I decided to conduct these analyses on states; North Carolina, and Illinois with variables; smoking_rate, Nata_cancer, and Glyphosates. The analysis included central tendency, normality, ANOVA testing, test for equal variances, multiple comparisons test, goodness of fit test, contingency tables, and linear regression analysis. The results of this analysis are described in the following sections.

Boxplot

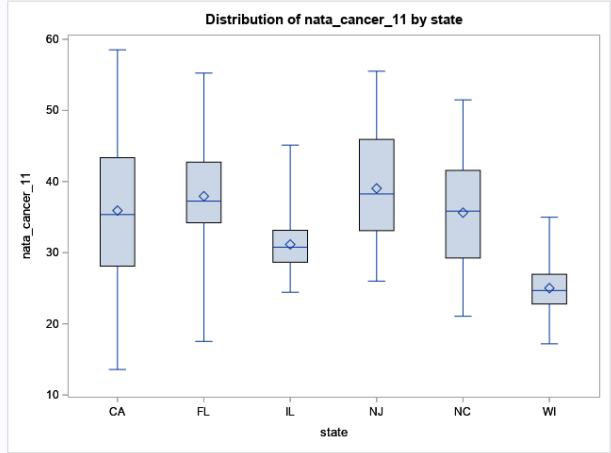
```
data ourdata;
/*Consider the data set on Alzheimer's Disease (AD) Mortality Rate in the 6 state CA FL IL NJ NC WI
input state $ smoking_rate nata_cancer_11 glyphosates deaths_5_yrs pop_5_yrs mental_distress diabetes cancer pop_dens;/*
smoking: age adjusted rate of smoking
nata_cancer: cancer caused by air pullatioon / air toxic
glyphosates: level of Glyphosates used
deaths_5yrs: death amount within 5 yr
pop_5yrs: number of
mental_distress: mental distress rate
diabetes: amount of diabetes
cancer: number of cancer
pop_dens:*/
datalines;
 CA 13.74 37.926 1.63166 2516 7570613 10 7.68 150.7400216 789.0462994
 CA 18.96 16.326 0.0006 2 5805 13 7.94 155.3348548 0.614452367
 CA 22.76 35.596 0.84414 141 189565 10 7.66 170.7211034 24.73501001
 CA 21.14 48.926 48.635 879 1100616 12 8.1 181.0384095 51.90609518
 CA 20.26 33.537 0.5808 78 227537 10 6.78 158.7001835 17.25250197
 CA 20.06 33.747 46.1488 39 106763 12 7.74 154.9369297 7.186668993
 CA 14.18 35.132 7.8804 2782 5255952 10 7.28 155.0173274 565.7351187
 CA 27.08 21.824 0.23228 34 142654 12 8 195.2045597 10.97643101
 WI 20 24.719 25.29359 286 659308 9 7.3 153.9236928 118.229675
 WI 17.82 27.851 22.3555 683 1948561 9 6.14 160.0924908 273.9173936
 WI 24.64 23.717 27.72854 180 262208 10 7.48 191.778383 27.06335639
 WI 26.3 23.11 16.28709 14 122847 10 7.48 183.6019817 15.10486205
 WI 23.04 31.33 31.55365 346 834826 10 7.76 176.7470568 148.3973959
 WI 24.08 27.402 19.64328 146 373104 10 7.62 155.7979535 36.38904627
;
ods html close;
ods html;
```

```
proc boxplot;
 plot smoking_rate*state/ haxis =state;
 plot nata_cancer_11*state/haxis =state;
 plot glyphosates*state/haxis =state;
 run;
```

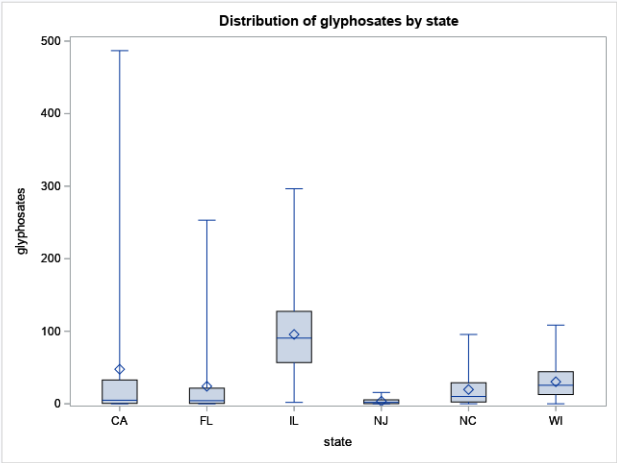
Smoking_rate: the distribution of all states: **smoking rate is the highest in Florida** on average among the states



Nata Cancer : the distribution of all states: **NJ is the highest in Nata_cancer** among the states



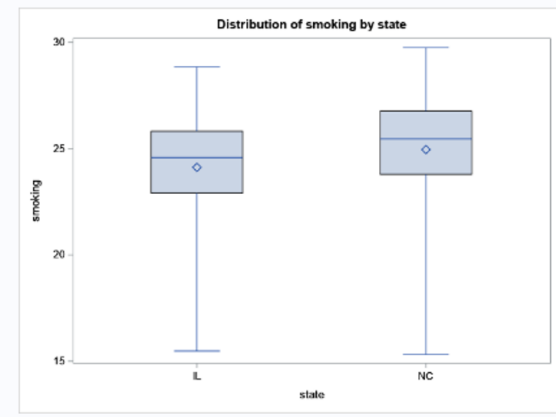
Glyphosates: the distribution of all states: **IL is the highest in glyphosates** among the states. The state with the highest **outlier is CA**



Box plots for variables for NC and IL

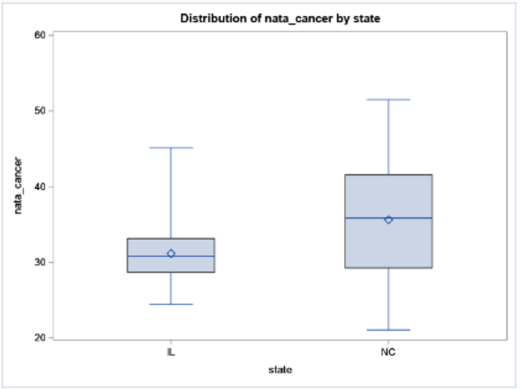
Smoking_rate: the distribution of both states is similar

```
proc boxplot data=mcdata;
plot smoking*state/ haxis =state;
```



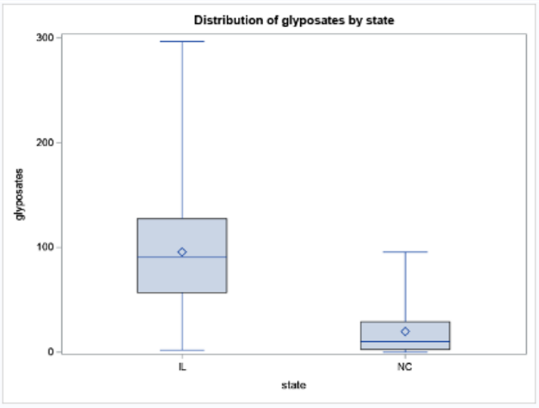
Nata_cancer: the distribution of IL is narrower than the NC.

```
proc boxplot data=mcdata;
plot nata_cancer*state/haxis =state;
```



Glyphosates: the glyphosates for NC are narrower and are less of the IL

```
proc boxplot data=mcdata;
plot glyposates*state/haxis =state;
```



Correlation coefficient

Correlation coefficient between the variables by state: $H_o: \rho = 0 \mid H_1: \rho \neq 0$

assumptions

State: IL

The correlation coefficient for smoking vs natacancer, smoking vs Glyposates and nata_cancer vs. Glyposates were 0.01928, 0.14454, -0.25356. The P_{value} associated with them were 0.8475, 0.1472, and 0.0101. **Only nata_cancer vs. glyposates showed significance.** Therefore, we **cannot conclude that the true correlation is zero for nata_cancer vs. glyposates.**

```
/* correlation of the variables across states*/
proc sort; by state;
proc corr; by state;
var smoking nata_cancer glyposates; run;
```

Pearson Correlation Coefficients, N = 102
Prob > |r| under H0: Rho=0

	smoking	nata_cancer	glyposates
smoking	1.00000	0.01928 0.8475	0.14454 0.1472
nata_cancer	0.01928 0.8475	1.00000	-0.25356 0.0101
glyposates	0.14454 0.1472	-0.25356 0.0101	1.00000

State: NC

The correlation coefficient for smoking vs nata_cancer, smoking vs Glyposates, and nata_cancer vs. Glyposates were -0.24782, 0.13100, -0.06849. The P_{value} associated with them were 0.0129, 0.1939, and 0.4983. **Only smoking vs. nata_cancer showed significance**. Therefore, we **cannot conclude that the true correlation is zero for smoking vs. nata_cancer.**

```
/* correlation of the variables across states*/
proc sort; by state;
proc corr; by state;
var smoking nata_cancer glyposates; run;
```

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	smoking	nata_cancer	glyposates
smoking	1.00000	-0.24782 0.0129	0.13100 0.1939
nata_cancer	-0.24782 0.0129	1.00000	-0.06849 0.4983
glyposates	0.13100 0.1939	-0.06849 0.4983	1.00000

States: combined

The correlation coefficient for smoking vs natacancer, smoking vs Glyposates, and nata_cancer vs. Glyposates were -0.08230, 0.00336, -0.33229. The P_{value} associated with them were 0.2443, 0.9621, and 0.0001. **Only nata_cancer vs. glyphosates showed significance.** Therefore, **we** cannot conclude that the true correlation is zero for nata_cancer vs. glyposates.

```
proc corr;
var smoking nata_cancer glyposates; run;
```

Pearson Correlation Coefficients, N = 202
Prob > |r| under H0: Rho=0

	smoking	nata_cancer	glyposates
smoking	1.00000	-0.08230 0.2443	-0.00336 0.9621
nata_cancer	-0.08230 0.2443	1.00000	-0.33229 <.0001
glyposates	-0.00336 0.9621	-0.33229 <.0001	1.00000

Shapiro Wilk's

Test for normality: $H_o: normal \mid H_1: non\ normal$

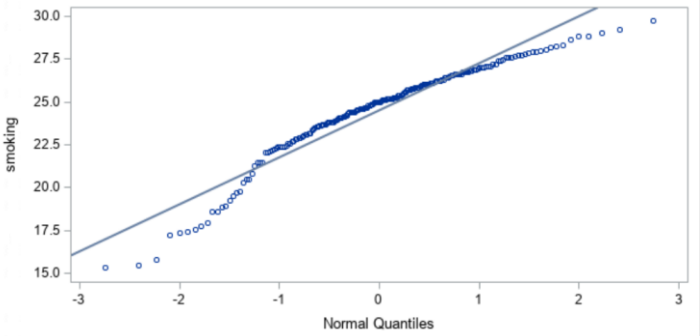
Smoking_rate:

```
proc univariate plot normal;
var smoking nata_cancer glyposates; run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.929187	Pr < W	<0.0001

At $\alpha = 0.05$ $p_{value} < \alpha$. We can reject the H_o. Therefore, **the data for smoking is non-normal**

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.929187	Pr < W	<0.0001



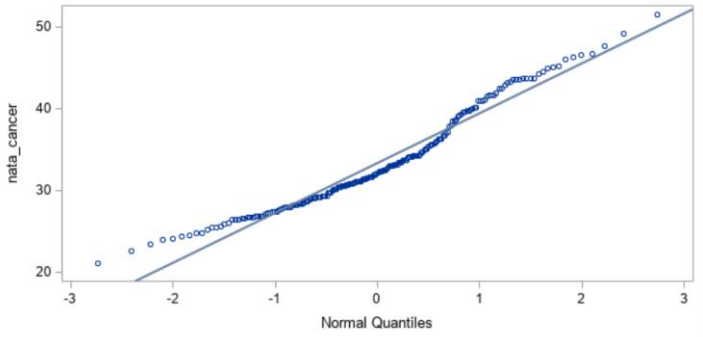
Nata_cancer:

```
proc univariate plot normal;
var smoking nata_cancer glyposates; run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.955085	Pr < W	<0.0001

At $\alpha = 0.05$ $p_{value} < \alpha$. We can reject the H_o. Therefore, **the data for NATA_cancer is non-normal**

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.955085	Pr < W	<0.0001



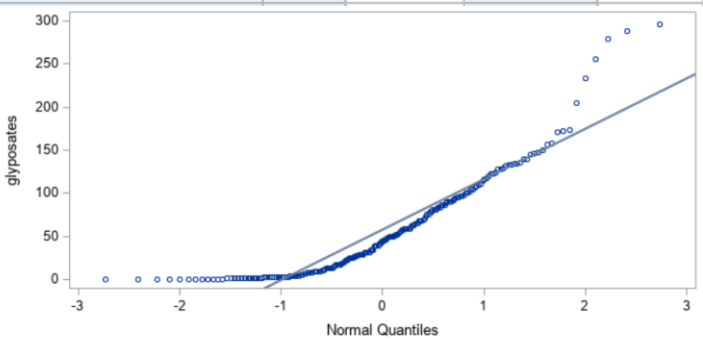
Glyphosates:

```
proc univariate plot normal;
var smoking nata_cancer glyposates; run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.854961	Pr < W	<0.0001

At $\alpha = 0.05$ $p_{value} < \alpha$. We can reject the H_o. Therefore, **the data for Glyphosates is non-normal**.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.854961	Pr < W	<0.0001



Analysis of Variance

Analysis of Variance for States: FL, NJ, WI, CA, IL, NC on Deaths_5_yrs, Pop_5_yrs, and mental_distress, Diabetes, Cancer, and Pop Density for variables, **smoking rate, NATA cancer, and Glyphosates:**

$$H_o: \mu_1 = \mu_2 = \cdots = \mu_n \mid H_1: not\ all\ means\ are\ equal$$

Assumptions: 1. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$ 2. Each population is normal 3. All samples are random

If all three criteria were met, below are the results

Smoking_rate: for state FL, NJ, WI, CA, IL, NC. At $\alpha = 0.05$, $p_{value} < \alpha$. We can **reject the H_o**. Therefore, **not all means for smoking rate are equal**.

```
/*ANOVA testing for equal means*/
proc glm;
class state;
model smoking_rate = state;
```

The GLM Procedure

Dependent Variable: smoking_rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2375.603240	475.120648	49.89	<.0001
Error	414	3942.620573	9.523238		
Corrected Total	419	6318.223813			

Nata_cancer: for state FL, NJ, WI, CA, IL, NC. IL, NC. At $\alpha = 0.05$, $p_{value} < \alpha$. We can **reject the H_o**. Therefore, **not all means for nata-cancer rate are equal**.

```
proc glm;
class state;
model nata_cancer_11 = state;
```

The GLM Procedure

Dependent Variable: nata_cancer_11

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8476.87918	1695.37584	40.05	<.0001
Error	414	17524.22429	42.32904		
Corrected Total	419	26001.10347			

Glyphosates: for state FL, NJ, WI, CA, IL, NC. At $\alpha = 0.05$, $p_{value} < \alpha$. We can **reject the H_o**. Therefore, **not all means for glyphosates rate are equal**

```
proc glm;
class state;
model glyphosates = state;
run;
```

The GLM Procedure

Dependent Variable: glyphosates

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	405672.094	81134.419	27.76	<.0001
Error	414	1209907.687	2922.482		
Corrected Total	419	1615579.782			

Kruskal Wallis

KW test for States: FL, NJ, WI, CA, IL, NC on Deaths_5_yrs, Pop_5_yrs, and mental_distress, Diabetes, Cancer, and Pop Density for variables, smoking rate, NATA cancer, and **Glyphosates:**

$$H_o: all\ distribution\ are\ identical \mid H_1: not\ all\ distribution\ are\ identical$$

Assumptions: 1. All population is normal 2. All variances are equal

```
/* KW test*/
proc npar1way wilcoxon ;
class state; var smoking_rate; run;
proc npar1way wilcoxon ;
class state; var nata_cancer_11; run;
proc npar1way wilcoxon ;
class state; var glyphosates; run;
```

<u>Smoking rate</u>: for the six states, the KW test for smoking resulted in p_{value} <0.0001, Since p_{value}< α. We **reject the H_0.** Therefore, **not all distributions are identical**

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable smoking_rate Classified by Variable state

state	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
CA	58	4387.00	12209.00	858.25349	75.637931
FL	67	18859.00	14103.50	910.90282	281.477612
IL	102	23517.50	21471.00	1066.74580	230.563725
NJ	21	1875.00	4420.50	542.17987	89.285714
NC	100	26653.50	21050.00	1059.55202	266.535000
WI	72	13118.00	15156.00	937.56884	182.194444

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
143.4507	5	<.0001

<u>Nata_cancer:</u> for the six states KW test for nata_cancer resulted in p_{value} <0.0001, Since p_{value}< α. We **reject the H_0.** Therefore, **not all distributions are identical.**

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable nata_cancer_11 Classified by Variable state

state	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
CA	58	14188.50	12209.00	858.25946	244.629310
FL	67	19512.50	14103.50	910.90917	291.231343
IL	102	18386.50	21471.00	1066.75323	180.259804
NJ	21	6153.50	4420.50	542.18365	293.023810
NC	100	24890.50	21050.00	1059.55940	248.905000
WI	72	5278.50	15156.00	937.57537	73.312500

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
152.2287	5	<.0001

Glyphosates: for the six states KW test for glyphosates resulted in p_{value} <0.0001, Since p_{value}< α. We **reject the H_o**. Therefore, **not all distributions are identical.**

The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable glyphosates Classified by Variable state

state	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
CA	58	9540.0	12209.00	858.25953	164.482759
FL	67	10222.0	14103.50	910.90924	152.567164
IL	102	33980.0	21471.00	1066.75331	333.137255
NJ	21	1797.0	4420.50	542.18369	85.571429
NC	100	17436.0	21050.00	1059.55949	174.360000
WI	72	15435.0	15156.00	937.57545	214.375000

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
158.8868	5	<.0001

Equal Variances Test

Testing for equal variances on variables smoking rate, nata_cancer, and glyphosates for the states CA, FL, IL, NJ, NC, WI: Bartlett's test, Brown & Forsythe.

$$H_o: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2 \mid H_1: at\ least\ two\ variances\ are\ different$$

Assumptions: 1) all sample are equal 2) the data is normal

If we would to proceed assuming the two criteria to be true. Below are the results

```
proc sort; by state;
proc univariate normal; by state;
 var smoking_rate nata_cancer_11 glyphosates; run;
proc glm;
 class state;
 model smoking_rate nata_cancer_11 glyphosates=state;
 means state/hovtest = bartlett;run;
```

Smoking_rate: From the bartlett's test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus, **at least two variances are different**

Nata_cancer: From the bartlett's test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus, **at least two variances are different**

Glyphosates: From the bartlett's test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus, **at least two variances are different**

The SAS System

The GLM Procedure

Bartlett's Test for Homogeneity of smoking_rate Variance

Source	DF	Chi-Square	Pr > ChiSq
state	5	17.7378	0.0033

Bartlett's Test for Homogeneity of nata_cancer_11 Variance

Source	DF	Chi-Square	Pr > ChiSq
state	5	104.7	<.0001

Bartlett's Test for Homogeneity of glyphosates Variance

Source	DF	Chi-Square	Pr > ChiSq
state	5	305.7	<.0001

```
proc sort; by state;
proc univariate normal; by state;
 var smoking_rate nata_cancer_11 glyphosates; run;
proc glm;
 class state;
 model smoking_rate nata_cancer_11 glyphosates=state;
 means state/hovtest = bf;
 run;
```

Smoking_rate: From the Brown & Forsythe test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus, **at least two variances are different**

Nata_cancer: From the Brown & Forsythe test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus **at least two variances are different**

Glyphosates: From the Brown & Forsythe test for homogeneity of variance. Since **Pvalue < α, we reject the Ho**. thus, **at least two variances are different**

The SAS System

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of smoking_rate Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
state	5	91.8945	18.3789	4.64	0.0004
Error	414	1639.6	3.9603		

Brown and Forsythe's Test for Homogeneity of nata_cancer_11 Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
state	5	1697.0	339.4	25.17	<.0001
Error	414	5583.6	13.4870		

Brown and Forsythe's Test for Homogeneity of glyphosates Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
state	5	77447.9	15489.6	6.74	<.0001
Error	414	950912	2296.9		

Multiple Comparisons test:

If we reject Ho: $H_o: \mu_1 = \mu_2 = \cdots = \mu_n \mid H_1: not\ all\ means\ are\ equal$

Which means are different?

Which error rate is controlled?

Procedures that control the comparisonwise error

$$H_o: \mu_i = \mu_j \mid H_1: \mu_i \neq \mu_j$$

Fisher's LSD method:

```
/* Fisher's LSD method */
proc glm;
 class state;
 model smoking_rate nata_cancer_11 glyphosates= state;
 means state / LSD lines;
 run;
```

Dependent Variable: smoking_rate

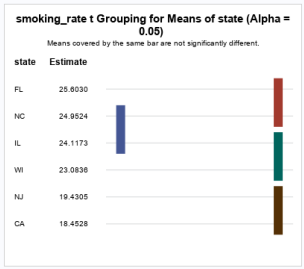
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2375.603240	475.120648	49.89	<.0001
Error	414	3942.620573	9.523238		
Corrected Total	419	6318.223813			

t Tests (LSD) for smoking_rate

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	9.523238
Critical Value of t	1.96571
Least Significant Difference	1.1798
Harmonic Mean of Cell Sizes	52.8734

Note: Cell sizes are not equal.



<u>Smoking_rate:</u> MSE = 9.5232; t_{crit} = 1.96571
All means not covered by a common bar are significantly different. That is **μ_{FL} are significantly different from μ_{IL}, μ_{WI}, μ_{NJ}, μ_{CA}.**

Dependent Variable: nata_cancer_11

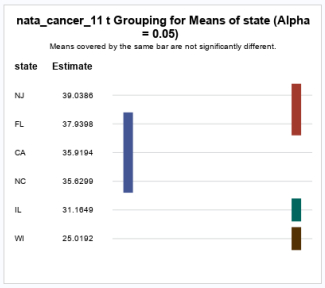
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8476.87918	1695.37584	40.05	<.0001
Error	414	17524.22429	42.32904		
Corrected Total	419	26001.10347			

t Tests (LSD) for nata_cancer_11

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	42.32904
Critical Value of t	1.96571
Least Significant Difference	2.4873
Harmonic Mean of Cell Sizes	52.8734

Note: Cell sizes are not equal.



<u>Nata_cancer:</u> MSE = 42.32904
All means not covered by a common bar are significantly different. That is **μ_{NJ} are significantly different from μ_{IL}, μ_{WI}, μ_{NJ}, μ_{CA}.**

Dependent Variable: glyphosates

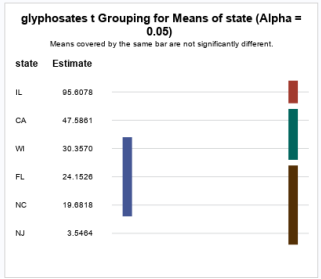
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	405672.094	81134.419	27.76	<.0001
Error	414	1209907.687	2922.482		
Corrected Total	419	1615579.782			

t Tests (LSD) for glyphosates

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	2922.482
Critical Value of t	1.96571
Least Significant Difference	20.668
Harmonic Mean of Cell Sizes	52.8734

Note: Cell sizes are not equal.



Glysphosates: MSE = 2922.482;
All means not covered by a common bar are significantly different. That is **μ_{IL} are significantly different from μ_{CA}, μ_{WI}, μ_{FL}, μ_{IL}, μ_{NC} μ_{N},.**

Procedures that control the stagewise error rate:

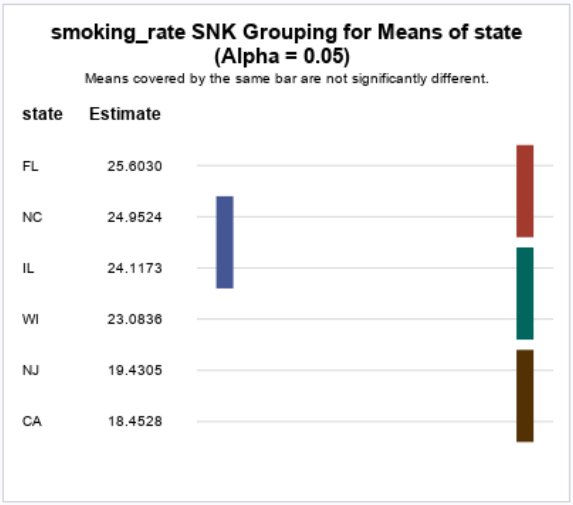
$$H_o: \mu_i = \mu_j \mid H_1: \mu_i \neq \mu_j$$

Student-Newman-Keuls Test:

```
/*SNK test */
proc glm ;
class state;
model smoking_rate nata_cancer_11 glyphosates= state;
means state / snk lines;
run;
```

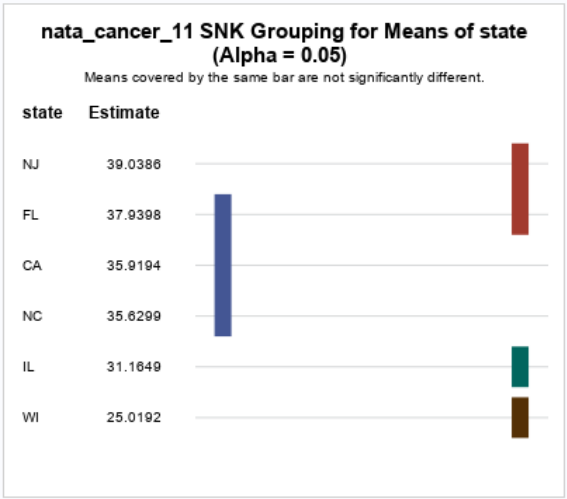
Smoking_rate: MSE = 9.5232; All means not covered by a common bar are significantly different. That is **μ_{FL} are significantly different from μ_{IL}, μ_{WI}, μ_{NJ}, μ_{CA}.**

Number of Means	2	3	4	5	6
Critical Range	1.1797998	1.4117508	1.5481872	1.6444239	1.7184155



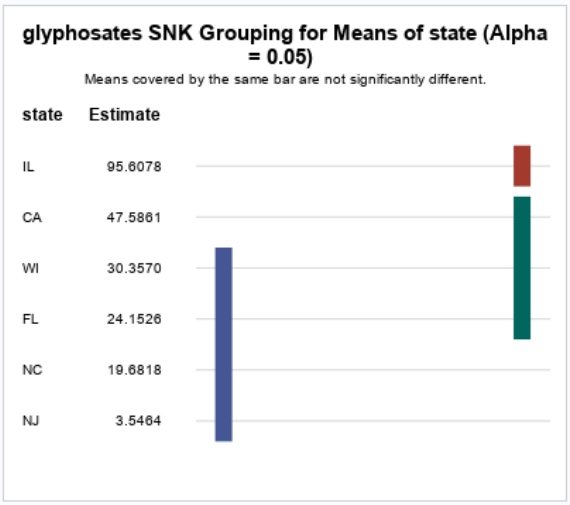
Nata_cancer: MSE = 42.32904
All means not covered by a common bar are significantly different. That is **μ_{NJ} are significantly different from μ_{CA}, μ_{NC}, μ_{IL}, μ_{WI}.**

Number of Means	2	3	4	5	6
Critical Range	2.4873405	2.9763565	3.2640018	3.4668951	3.6228895



Glysphosates: MSE = 2922.482;
All means not covered by a common bar are significantly different. That **is μ_{IL} are significantly different from μ_{CA}, μ_{WI}, μ_{FL}, μ_{NC}, μ_{NJ},**

Number of Means	2	3	4	5	6
Critical Range	20.667691	24.730999	27.121088	28.806959	30.103141

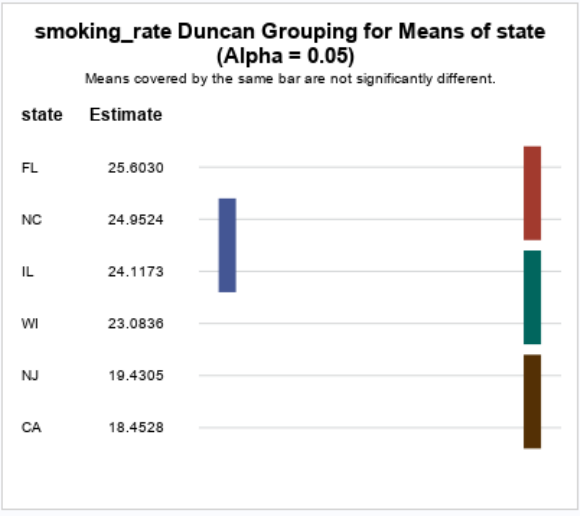


Duncan's New Multiply Range Test:

```
/*Duncan's New Multiply Range test */
proc glm ;
class state;
model smoking_rate nata_cancer_ll glyphosates= state;
means state / duncan lines;
run;
```

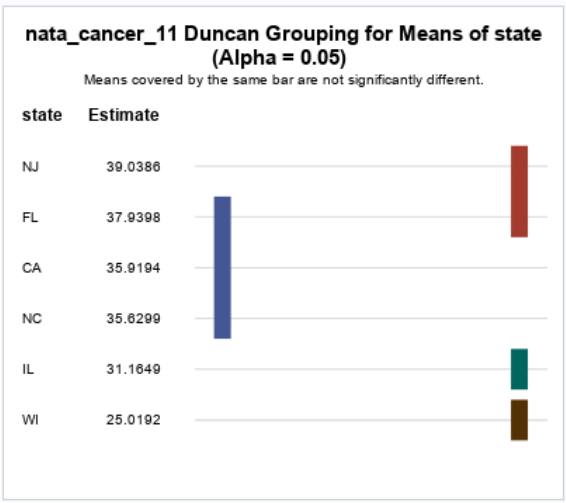
Smoking_rate: MSE = 9.5232; All means not covered by a common bar are significantly different. That is **μ_{FL} are significantly different from μ_{IL}, μ_{WI}, μ_{NJ}, μ_{CA}.**

Number of Means	2	3	4	5	6
Critical Range	1.180	1.242	1.284	1.314	1.338



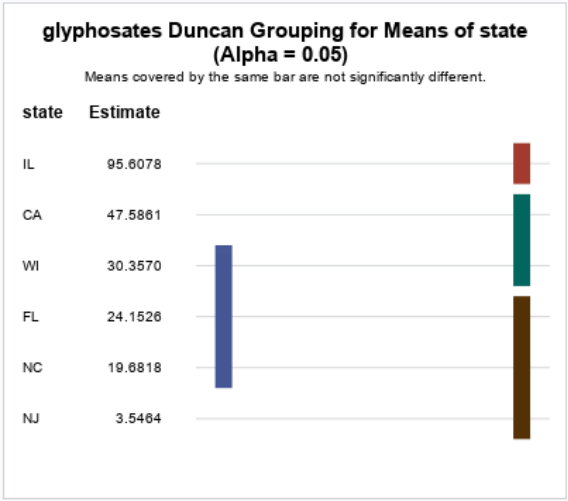
Nata_cancer: MSE = 42.32904
All means not covered by a common bar are significantly different. That is **μ_{NJ} are significantly different from μ_{CA}, μ_{NC}, μ_{IL}, μ_{WI}.**

Number of Means	2	3	4	5	6
Critical Range	2.487	2.619	2.706	2.771	2.822



Glysphosates: MSE = 2922.482;
All means not covered by a common bar are significantly different. That is **μ_{IL} are significantly different from μ_{CA}, μ_{WI}, μ_{FL}, μ_{NC}, μ_{NJ},**

Number of Means	2	3	4	5	6
Critical Range	20.67	21.76	22.49	23.03	23.45

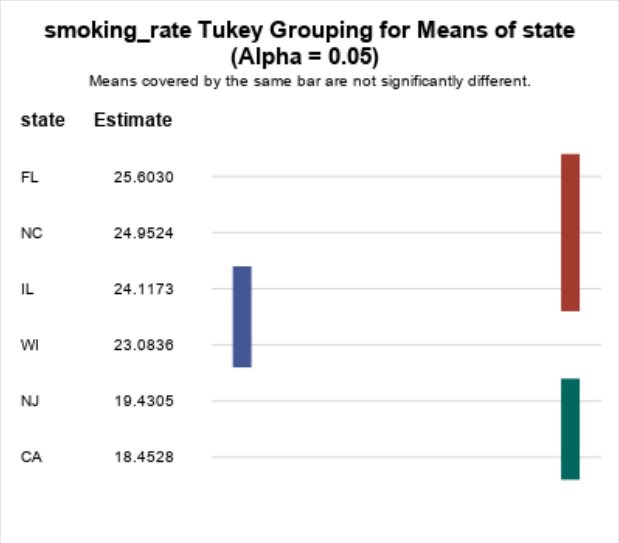


Procedures that control the experimentwise error

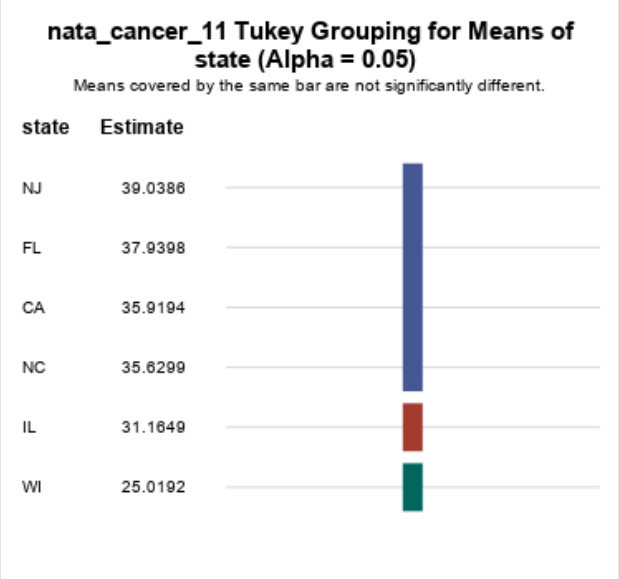
$$H_o: \mu_i = \mu_j \mid H_1: \mu_i \neq \mu_j$$

```
/*Tukey Kramer test */
proc glm ;
class state;
model smoking_rate nata_cancer_11 glyphosates= state;
means state / tukey lines;
run;
```

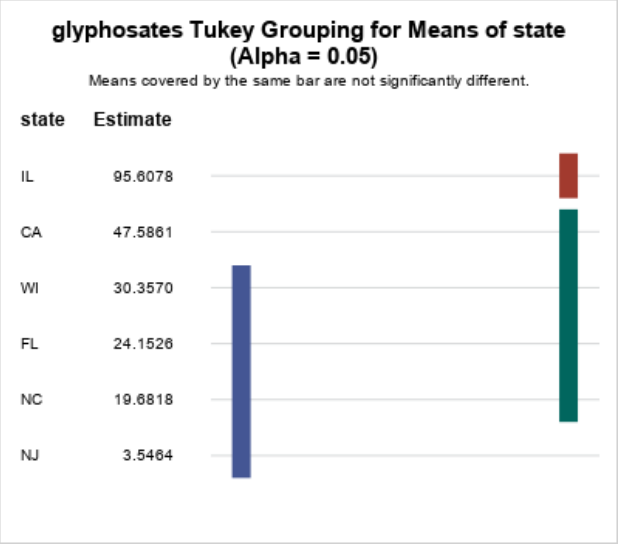
<u>Smoking_rate:</u> MSE = 9.5232; All means not covered by a common bar are significantly different. That is **μ_{FL} are significantly different from μ_{WI}, μ_{NJ}, μ_{CA}.**



<u>Nata_cancer:</u> MSE = 42.32904
All means not covered by a common bar are significantly different. That is **μ_{NJ} are significantly different from μ_{IL}, and μ_{WI}.**



<u>Glysphosates</u>: MSE = 2922.482;
All means not covered by a common bar are significantly different. That is **μ_{IL} are significantly different from μ_{CA}, μ_{WI}, μ_{FL}, μ_{NC}, μ_{NJ},**

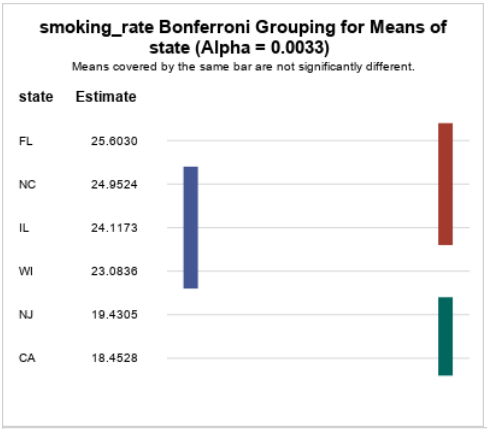


Bonferroni's method

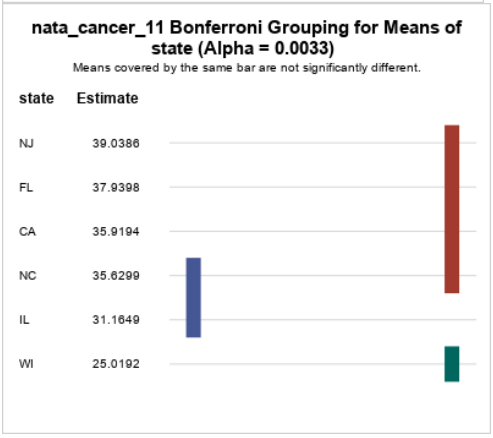
t = 6; m = t(t-1)/2 = 15; α_j = 0.05/15=0.0033 ~0.001

```
/*Bonferroni test */
proc glm alpha =0.0033;
class state;
model smoking_rate nata_cancer_11 glyphosates= state;
means state / bon lines;
run;
```

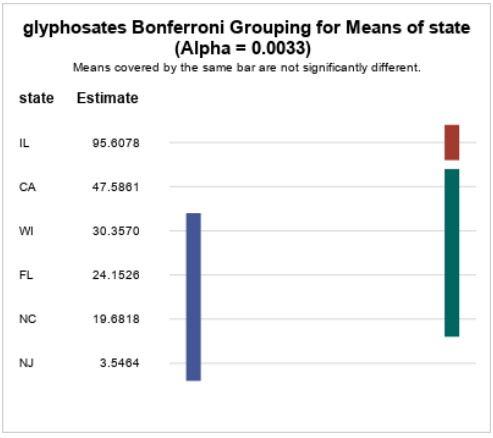
Smoking_rate: MSE = 9.5232; All means not covered by a common bar are significantly different. That is **μ_{FL} are significantly different from μ_{WI}, μ_{NJ}, μ_{CA}.**



Nata_cancer: MSE = 42.32904
All means not covered by a common bar are significantly different. That is **μ_{NJ} are significantly different from μ_{IL}, and μ_{WI}.**



Glysphosates: MSE = 2922.482;
All means not covered by a common bar are significantly different. That is μ_{IL} are significantly different from **μ_{CA}, μ_{WI}, μ_{FL}, μ_{NC}, μ_{NJ},**



Dunnetts test:

$$H_o: \mu_i = \mu_c \mid H_1: \mu_i \neq \mu_c$$

```
/*dunnett's test */
proc glm;
class state;
model smoking_rate nata_cancer_11 glyphosates= state;
means state / dunnett('NC');
run;
```

Keeping NC as control

Smoking_rate: MSE = 9.5232; All state comparisons with asterisks have significantly different means from NC. That is, **WI, NJ, and CA are significantly different from NC**

Dunnett's t Tests for smoking_rate

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	9.523238
Critical Value of Dunnett's t	2.54985

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
FL - NC	0.6506	-0.5917	1.8929	
IL - NC	-0.8351	-1.9425	0.2722	
WI - NC	-1.8688	-3.0850	-0.6526	***
NJ - NC	-5.5219	-7.4107	-3.6331	***
CA - NC	-6.4996	-7.7984	-5.2009	***

Nata_cancer: MSE = 42.32904 All state comparisons with asterisks have significantly different means from NC. That is, **IL and WI are significantly different from NC**

Dunnett's t Tests for nata_cancer_11

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	42.32904
Critical Value of Dunnett's t	2.54985

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
NJ - NC	3.4086	-0.5735	7.3908	
FL - NC	2.3098	-0.3093	4.9289	
CA - NC	0.2894	-2.4487	3.0275	
IL - NC	-4.4650	-6.7996	-2.1304	***
WI - NC	-10.6108	-13.1749	-8.0467	***

<u>Glysphosates:</u> MSE = 2922.482; All state comparisons with asterisks have significantly different means from NC. That is, **IL and CA are significantly different from NC**

Dunnett's t Tests for glyphosates

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	2922.482
Critical Value of Dunnett's t	2.54985

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
IL - NC	75.926	56.528	95.324	***
CA - NC	27.904	5.153	50.656	***
WI - NC	10.675	-10.630	31.981	
FL - NC	4.471	-17.292	26.233	
NJ - NC	-16.135	-49.224	16.953	

Keeping IL as control

<u>Smoking_rate:</u> MSE = 9.5232; All state comparisons with asterisks have significantly different means from IL. That is, **FL, NJ, and CA are significantly different from IL**

Dunnett's t Tests for smoking_rate

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	9.523238
Critical Value of Dunnett's t	2.55080

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
FL - IL	1.4857	0.2479	2.7236	***
NC - IL	0.8351	-0.2726	1.9429	
WI - IL	-1.0336	-2.2453	0.1780	
NJ - IL	-4.6868	-6.5731	-2.8005	***
CA - IL	-5.6645	-6.9590	-4.3700	***

Nata_cancer: MSE = 42.32904
All state comparisons with asterisks have significantly different means from IL. That is, **FL, NJ, and CA are significantly different from IL.**

Dunnett's t Tests for nata_cancer_11

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	42.32904
Critical Value of Dunnett's t	2.55080

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
NJ - IL	7.8736	3.8968	11.8505	***
FL - IL	6.7748	4.1651	9.3846	***
CA - IL	4.7544	2.0252	7.4837	***
NC - IL	4.4650	2.1296	6.8005	***
WI - IL	-6.1458	-8.7003	-3.5913	***

Glysphosates: MSE = 2922.482;
All state comparisons with asterisks have significantly different means from IL. **All state has different means**

Dunnett's t Tests for glyphosates

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	414
Error Mean Square	2922.482
Critical Value of Dunnett's t	2.55080

Comparisons significant at the 0.05 level are indicated by ***.

state Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
CA - IL	-48.022	-70.699	-25.344	***
WI - IL	-65.251	-86.476	-44.025	***
FL - IL	-71.455	-93.140	-49.770	***
NC - IL	-75.926	-95.332	-56.520	***
NJ - IL	-92.061	-125.106	-59.017	***

Contrast

Individual Contrast: α = 0.05

$$H_o: L = 0 \mid H_1: L \neq 0$$

North Carolina

$$H_o: \mu_{NC} = \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5} \mid H_1: \mu_{NC} \neq \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5}$$

Illinois

$$H_o: \mu_{IL} = \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5} \mid H_1: \mu_{IL} \neq \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5}$$

```
/* individual contrast: testing if the mean rate of NC is a contrast to the rest of the mean rates.
individual contrast: testing if the mean rate of IL is a contrast to the rest of the mean rates.*/
proc glm;
class state;
model  smoking_rate nata_cancer_ll glyphosates= state;
estimate 'NC vs rest' state -1 -1 -1 -1 5 -1;
estimate 'IL vs rest' state -1 -1 5 -1 -1 -1;
```

Smokng_rate: From testing a single contrast individually for NC and IL vs. the rest, we obtain a p_{value} <0.0001 for both. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude the mean rate for smoking in both NC and IL is significantly different** from the rest of the state.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs rest	-19.0566288	3.45895714	-5.51	<.0001
IL vs rest	9.0640435	1.82248340	4.97	<.0001

Nata_cancer: From testing a single contrast individually for NC and IL vs. the rest, we obtain a p_{value} <0.0001 for both. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude the mean rate for Nata_cancer in both NC and IL is significantly different** from the rest of the state.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs rest	29.5196776	7.29242700	4.05	<.0001
IL vs rest	-17.7221628	3.84229309	-4.61	<.0001

Glyphosates: From testing a single contrast individually for NC and IL vs. the rest, we obtain a p_{value} <0.0011 and p_{value} < 0.0001 respectively. Since $p_{value} < \alpha$, we **can reject the H_0 and conclude the mean rate for Glyphosates in both NC and IL is significantly different** from the rest of the state.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs rest	-199.653234	60.5938862	-3.29	0.0011
IL vs rest	352.715160	31.9261983	11.05	<.0001

Multiple Contrast

Multiple Contrast: Bonferoni correction: $\alpha = 0.05/m = 0.01$. m: number of comparisons

$$H_o: L = 0 \mid H_1: L \neq 0, F_0 - test\ or\ t_0 - test$$

North Carolina

$H_o: \mu_{NC} = \mu_{FL}$, $H_1: \mu_{NC} \neq \mu_{FL}$

$H_o: \mu_{NC} = \mu_{IL}$, $H_1: \mu_{NC} \neq \mu_{IL}$

$H_o: \mu_{NC} = \mu_{NJ}$, $H_1: \mu_{NC} \neq \mu_{NJ}$

$H_o: \mu_{NC} = \mu_{WI}$, $H_1: \mu_{NC} \neq \mu_{WI}$

$H_o: \mu_{NC} = \mu_{CA}$, $H_1: \mu_{NC} \neq \mu_{CA}$

```
/* testing  for multiple contrast of mean rate of NC vs.
CA; NC vs. FL; NC vs. IL; NC vs. NJ; NC vs. WI*/
proc glm alpha = 0.01;
class state;
model smoking_rate nata_cancer_ll glyphosates = state;
estimate 'NC vs CA' state -1 0 0 0 1 0;
estimate 'NC vs FL' state 0 -1 0 0 1 0;
estimate 'NC vs IL' state 0 0 -1 0 1 0;
estimate 'NC vs NJ' state 0 0 0 -1 1 0;
estimate 'NC vs WI' state 0 0 0 0 1 -1;
run;
```

<u>Smokng_rate</u>: at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with NC, we obtain a significance for states FL, IL, NJ, and WI. Thus, **the mean smoking rates for these states are significantly different from NC**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs CA	0.97771757	0.78592703	1.24	0.2142
NC vs FL	-6.17250888	0.77176772	-8.00	<.0001
NC vs IL	-4.68677871	0.73949483	-6.34	<.0001
NC vs NJ	-5.52192381	0.74075631	-7.45	<.0001
NC vs WI	-3.65313492	0.76534615	-4.77	<.0001

<u>Nata_cancer:</u> at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with NC, we obtain a significance for states IL, NJ, and WI. Thus, the **mean Nata_cancer rate for these states are significantly different from NC**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs CA	3.1191921	1.65694897	1.88	0.0605
NC vs FL	1.0987953	1.62709728	0.68	0.4999
NC vs IL	7.8736401	1.55905722	5.05	<.0001
NC vs NJ	3.4086314	1.56171675	2.18	0.0296
NC vs WI	14.0194187	1.61355885	8.69	<.0001

<u>Glyphosates:</u> at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with NC, we obtain a significance for states CA, and IL. Thus, the **mean glyphosates rate for these states are significantly different from NC**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
NC vs CA	-44.0396686	13.7678412	-3.20	0.0015
NC vs FL	-20.6061914	13.5197990	-1.52	0.1282
NC vs IL	-92.0613989	12.9544438	-7.11	<.0001
NC vs NJ	-16.1353560	12.9765423	-1.24	0.2144
NC vs WI	-26.8106188	13.4073061	-2.00	0.0462

Illinois

```
/* testing  for multiple contrast of mean rate of IL vs.
CA; NC vs. FL; NC vs. IL; NC vs. NJ; NC vs. WI*/
proc glm alpha = 0.01;
class state;
model smoking_rate nata_cancer_ll glyphosates = state;
estimate 'IL vs CA' state -1 0 1 0 0 0;
estimate 'IL vs FL' state 0 -1 1 0 0 0;
estimate 'IL vs NJ' state 0 0 1 -1 0 0;
estimate 'IL vs NC' state 0 0 1 0 -1 0;
estimate 'IL vs WI' state 0 0 1 0 0 -1;
run;
```

$H_o: \mu_{IL} = \mu_{FL}$ $H_1: \mu_{IL} \neq \mu_{FL}$ $H_o: \mu_{IL} = \mu_{NC}$ $H_1: \mu_{IL} \neq \mu_{NC}$ $H_o: \mu_{IL} = \mu_{NJ}$ $H_1: \mu_{IL} \neq \mu_{NJ}$ $H_o: \mu_{IL} = \mu_{WI}$ $H_1: \mu_{IL} \neq \mu_{WI}$ $H_o: \mu_{IL} = \mu_{CA}$ $H_1: \mu_{IL} \neq \mu_{CA}$

Smokng_rate: at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with IL, we obtain a significance for states CA, FL, and NC. Thus, **the mean smoking rates for these states are significantly different from IL**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs CA	5.66449628	0.50750255	11.16	<.0001
IL vs FL	-1.48573017	0.48528647	-3.06	0.0023
IL vs NJ	-0.83514510	0.43427809	-1.92	0.0552
IL vs NC	4.68677871	0.73949483	6.34	<.0001
IL vs WI	1.03364379	0.47500766	2.18	0.0301

Nata_cancer: at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with IL, we obtain a significance for all comparison. Thus, **the mean Nata_cancer rate for all states are significantly different from IL**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs CA	-4.75444794	1.06995408	-4.44	<.0001
IL vs FL	-6.77484475	1.02311652	-6.62	<.0001
IL vs NJ	-4.46500863	0.91557691	-4.88	<.0001
IL vs NC	-7.87364006	1.55905722	-5.05	<.0001
IL vs WI	6.14577859	1.00144596	6.14	<.0001

Glyphosates: at α = 0.01, The results of testing multiple contrasts on each pairwise comparison with IL, we obtain significances for all comparison. Thus, **the mean Glyphosates rate for all states are significantly different from IL**.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs CA	48.0217303	8.8904114	5.40	<.0001
IL vs FL	71.4552075	8.5012309	8.41	<.0001
IL vs NJ	75.9260429	7.6076679	9.98	<.0001
IL vs NC	92.0613989	12.9544438	7.11	<.0001
IL vs WI	65.2507800	8.3211670	7.84	<.0001

Scheffe's Multiple Contrast Test: Use only for a huge number of contrasts

α_E is controlled for all possible contrasts

```
/* scheffe's multiple contrast test for IL vs rest and NC vs rest */
proc glm;
class state;
model smoking_rate nata_cancer_ll glyphosates = state;
means state/scheffe cldiff lines;
estimate 'IL vs rest' state -1 -1 5 -1 -1;
estimate 'NC vs rest' state -1 -1 -1 -1 5 -1;
run;
```

North Carolina

$$H_o: \mu_{NC} = \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5} \mid H_1: \mu_{NC} \neq \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5}$$

Illinois

$$H_o: \mu_{IL} = \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5} \mid H_1: \mu_{IL} \neq \frac{\mu_{CA} + \mu_{FL} + \mu_{IL} + \mu_{NJ} + \mu_{WI}}{5}$$

Smokng_rate: The Scheffe's multiple contrast for for ILvs rest and NC vs. rest resulted **significantly different means for smoking rate across states.**

Dependent Variable: smoking_rate

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs rest	9.0640435	1.82248340	4.97	<.0001
NC vs rest	-19.0566288	3.45895714	-5.51	<.0001

Nata_cancer: The Scheffe's multiple contrast for for ILvs rest and NC vs. rest resulted **significantly different means for Nata_cancer across states.**

Dependent Variable: nata_cancer_11

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs rest	-17.7221628	3.84229309	-4.61	<.0001
NC vs rest	29.5196776	7.29242700	4.05	<.0001

Glyphosates: The Scheffe's multiple contrast for ILvs. rest and NC vs rest resulted **significantly different means for Glyphosates across states.**

Dependent Variable: glyphosates

Parameter	Estimate	Standard Error	t Value	Pr > |t|
IL vs rest	352.715160	31.9261983	11.05	<.0001
NC vs rest	-199.653234	60.5938862	-3.29	0.0011

Test for Chi squared goodness of fit.

$$H_o: F(x) = N(\mu, \sigma^2) \mid H_1: F(x) \neq N(\mu, \sigma^2)$$

```
/* test for categorical proportion for each variable*/
proc freq;
tables smoking_ratec/ chisq(df=1);
weight smoking_rate;
run;

proc freq;
tables natac/ chisq(df=1);
weight nata;
run;

proc freq;
tables glyphosatesc/ chisq(df=1);
weight glyphosates;
run;
```

North Carolina

```
/* making categorical variable for north carolina*/
if smoking_rate le 23.79 then smoking_ratec = '1'; /*creating categorical variable*/
if smoking_rate gt 23.79 and smoking_rate le 25.46 then smoking_ratec= '2';
if smoking_rate gt 25.46 and smoking_rate le 26.76  then smoking_ratec = '3';
if smoking_rate gt 26.76 then smoking_ratec ='4';

if nata le 29.2495 then natac = '1'; /*creating categorical variable*/
if nata gt 29.2495 and nata le 35.8435 then natac= '2';
if nata gt 35.8435 and nata le 41.5720 then natac = '3';
if nata gt 41.5720 then natac ='4';

if glyphosates le 2.541915 then glyphosatesc = '1'; /*creating categorical variable*/
if glyphosates gt 2.541915 and glyphosates le 10.054425 then glyphosatesc= '2';
if glyphosates gt 10.054425 and glyphosates le 28.953130 then glyphosatesc = '3';
if glyphosates gt 28.953130 then glyphosatesc ='4';
```

Smoking_rate: The Chi squared goodness of fit test resulted in p_{value} <0.0001. Thus, **we can reject H_0 and conclude the proportion of each category of smoking_rate are not the same**.

The FREQ Procedure

smoking_ratec	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	530.52	21.26	530.52	21.26
2	617.02	24.73	1147.54	45.99
3	681.74	27.32	1829.28	73.31
4	665.96	26.69	2495.24	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	22.2530
DF	1
Pr > ChiSq	<.0001

Nata_cancer: the Chi squared goodness of fit resulted in p_{value} <0.0001. Thus, we **can reject H_0 and conclude the proportion of each category of Nata_cancer are not the same**.

The FREQ Procedure

natac	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	657.757	18.46	657.757	18.46
2	819.399	23.00	1477.156	41.46
3	972.446	27.29	2449.602	68.75
4	1113.392	31.25	3562.994	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	129.8014
DF	1
Pr > ChiSq	<.0001

Glyphosates: the Chi squared goodness of fit resulted in p_{value} <0.0001. Thus, we can **reject the H_0, and conclude that the proportion of each category of Glyphosates are not the same**

glyphosatesc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	23.15567	1.18	23.15567	1.18
2	139.0744	7.07	162.2301	8.24
3	486.1595	24.70	648.3896	32.94
4	1319.788	67.06	1968.178	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	2092.5727
DF	1
Pr > ChiSq	<.0001

Illinois

```
/* making categorical variable for Illinois */
if smoking_rate le 22.90 then smoking_ratec = '1';
if smoking_rate gt 22.90 and smoking_rate le 24.58 then smoking_ratec= '2';
if smoking_rate gt 24.58 and smoking_rate le 25.82 then smoking_ratec = '3';
if smoking_rate gt 25.82 then smoking_ratec ='4';

if nata le 28.655 then natac = '1';
if nata gt 28.655 and nata le 30.7765 then natac= '2';
if nata gt 30.7765 and nata le 33.14 then natac = '3';
if nata gt 33.14 then natac ='4';

if glyphosates le 56.90997 then glyphosatesc = '1';
if glyphosates gt 56.90997 and glyphosates le 90.57663 then glyphosatesc= '2';
if glyphosates gt 90.57663 and glyphosates le 127.43457 then glyphosatesc = '3';
if glyphosates gt 127.43457 then glyphosatesc ='4';
```

Smoking_rate: The Chi squared goodness of fit test resulted in p_{value} <0.0001. Thus, we can **reject H_0 that the proportion of each category of smoking_rates are similar**

The FREQ Procedure

smoking_ratec	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	533.68	21.69	533.68	21.69
2	620.2	25.21	1153.88	46.91
3	631.2	25.66	1785.08	72.57
4	674.88	27.43	2459.96	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	17.0540
DF	1
Pr > ChiSq	<.0001

Nata_cancer: the Chi squared goodness of fit resulted in p_{value} <0.0001. Thus, we can **reject H_0 and conclude that the proportion of each category of Nata_cancers are not the same**

The FREQ Procedure

natac	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	702.074	22.09	702.074	22.09
2	744.38	23.42	1446.454	45.50
3	798.689	25.13	2245.143	70.63
4	933.68	29.37	3178.823	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	38.3073
DF	1
Pr > ChiSq	<.0001

Glyphosates: the Chi squared goodness of fit resulted in p_{value} <0.0001. Thus we can **reject the H_0, and conclude that the proportion of each category of Glyphosates are not the same**

The FREQ Procedure

glyphosatesc	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	826.8675	8.48	826.8675	8.48
2	1846.315	18.93	2673.183	27.41
3	2762.341	28.33	5435.524	55.74
4	4316.474	44.26	9751.998	100.00

Chi-Square Test for Equal Proportions	
Chi-Square	2698.8113
DF	1
Pr > ChiSq	<.0001

Contingency Table

Chi-Squared Test for Independence:

H_0: *row factor and column factor are independent of each other*
H_1: *the row and column factor are dependent (associated) of each other*

Assumptions:

(a) One random sample is taken
(b) Each observation may be classified into one of r different categories for row factor and one of c categories for column factor

Chi-squared test for independence on variables for all states (CA, FL, IL, NJ, NC, WI).

```
if smoking_rate le 21.29 then smoking_ratec = '1'; /*creating categorical variable for smoking_rate*/
if smoking_rate gt 21.29 and smoking_rate le 24.06 then smoking_ratec= '2';
if smoking_rate gt 24.06 and smoking_rate le 26.04 then smoking_ratec = '3';
if smoking_rate gt 26.04 then smoking_ratec ='4';

if nata le 27.4595 then natac = '1'; /*creating categorical variable for nata*/
if nata gt 27.4595 and nata le 32.4345 then natac = '2';
if nata gt 32.4345 and nata le 39.0790 then natac = '3';
if nata gt 39.0790 then natac='4';

 if glyphosates le 2.836460 then glyphosatesc = '1'; /*creating categorical variable for glyphosates*/
if glyphosates gt 2.836460 and glyphosates le 21.750745 then glyphosatesc = '2';
if glyphosates gt 21.750745 and glyphosates le 61.827565 then glyphosatesc = '3';
if glyphosates gt 61.827565 then glyphosatesc='4';


 /*test for independence between different variables smoking_rate vs. nata_cancer for all states */
proc freq;
 table smoking_ratec*natac/chisq cellchi2 expected;
 run;

 /*test for independence between different variables smoking_rate vs. glyphosates for all states */
proc freq;
 table smoking_ratec*glyphosatesc/chisq cellchi2 expected;
 run;

 /*test for independence between different variables nata_cancer vs. glyphosatesc for all states */
proc freq;
 table natac*glyphosatesc/chisq cellchi2 expected;
 run;
```

Smoking_rate vs. Nata_cancer (all states):

Statistics for Table of smoking_ratec by natac

Statistic	DF	Value	Prob
Chi-Square	9	42.3174	<.0001
Likelihood Ratio Chi-Square	9	40.7184	<.0001
Mantel-Haenszel Chi-Square	1	0.0095	0.9222
Phi Coefficient		0.3174	
Contingency Coefficient		0.3025	
Cramer's V		0.1833	

The FREQ Procedure

Table of smoking_ratec by natac

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

smoking_ratec	natac 1	2	3	4	Total
1	18 26.25 2.5929 4.29 17.14 17.14	24 26.25 0.1929 5.71 22.86 22.86	22 26.25 0.6881 5.24 20.95 20.95	41 26.25 8.2881 9.76 39.05 39.05	105 25.00
2	43 26.25 10.688 10.24 40.95 40.95	24 26.25 0.1929 5.71 22.86 22.86	22 26.25 0.6881 5.24 20.95 20.95	16 26.25 4.0024 3.81 15.24 15.24	105 25.00
3	27 26.5 0.0094 6.43 25.47 25.71	36 26.5 3.4057 8.57 33.96 34.29	26 26.5 0.0094 6.19 24.53 24.76	17 26.5 3.4057 4.05 16.04 16.19	106 25.24
4	17 26 3.1154 4.05 16.35 16.19	21 26 0.9615 5.00 20.19 20.00	35 26 3.1154 8.33 33.65 33.33	31 26 0.9615 7.38 29.81 29.52	104 24.76
Total	105 25.00	105 25.00	105 25.00	105 25.00	420 100.00

When testing for the independence of smoking_rate against nata_cancer rate, we obtain a p_{value} <0.0001. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude that the smoking_rates and nata_cancers are associated with each other.**

Smoking_rate vs. Glyphosates (all states):

Statistics for Table of smoking_ratec by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	14.5729	0.1034
Likelihood Ratio Chi-Square	9	14.5111	0.1053
Mantel-Haenszel Chi-Square	1	0.0034	0.9533
Phi Coefficient		0.1863	
Contingency Coefficient		0.1831	
Cramer's V		0.1075	

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

Table of smoking_ratec by glyphosatesc

smoking_ratec	glyphosatesc 1	2	3	4	Total
1	31 26.25 0.8595 7.38 29.52 29.52	25 26.25 0.0595 5.95 23.81 23.81	27 26.25 0.0214 6.43 25.71 25.71	22 26.25 0.6881 5.24 20.95 20.95	105 25.00
2	26 26.25 0.0024 6.19 24.76 24.76	23 26.25 0.4024 5.48 21.90 21.90	28 26.25 0.1167 6.67 26.67 26.67	28 26.25 0.1167 6.67 26.67 26.67	105 25.00
3	19 26.5 2.1226 4.52 17.92 18.10	22 26.5 0.7642 5.24 20.75 20.95	30 26.5 0.4623 7.14 28.30 28.57	35 26.5 2.7264 8.33 33.02 33.33	106 25.24
4	29 26 0.3462 6.90 27.88 27.62	35 26 3.1154 8.33 33.65 33.33	20 26 1.3846 4.76 19.23 19.05	20 26 1.3846 4.76 19.23 19.05	104 24.76
Total	105 25.00	105 25.00	105 25.00	105 25.00	420 100.00

When testing for the independence of smoking_rate against glyphosates, we obtain a p_{value} <0.1034. Since $p_{value} > \alpha$, **we cannot reject the H_0 and conclude that the smoking_rates and nata_cancers are independent from each other**

Nata_cancer vs. Glyphosates (all states):

Statistics for Table of natac by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	65.1048	<.0001
Likelihood Ratio Chi-Square	9	65.8271	<.0001
Mantel-Haenszel Chi-Square	1	8.7806	0.0030
Phi Coefficient		0.3937	
Contingency Coefficient		0.3663	
Cramer's V		0.2273	

Sample Size = 420

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

Table of natac by glyphosatesc

natac	glyphosatesc 1	2	3	4	Total
1	32 26.25 1.2595 7.62 30.48 30.48	19 26.25 2.0024 4.52 18.10 18.10	32 26.25 1.2595 7.62 30.48 30.48	22 26.25 0.6881 5.24 20.95 20.95	105 25.00
2	16 26.25 4.0024 3.81 15.24 15.24	11 26.25 8.8595 2.62 10.48 10.48	36 26.25 3.6214 8.57 34.29 34.29	42 26.25 9.45 10.00 40.00 40.00	105 25.00
3	27 26.25 0.0214 6.43 25.71 25.71	26 26.25 0.0024 6.19 24.76 24.76	26 26.25 0.0024 6.19 24.76 24.76	26 26.25 0.0024 6.19 24.76 24.76	105 25.00
4	30 26.25 0.5357 7.14 28.57 28.57	49 26.25 19.717 11.67 46.67 46.67	11 26.25 8.8595 2.62 10.48 10.48	15 26.25 4.8214 3.57 14.29 14.29	105 25.00
Total	105 25.00	105 25.00	105 25.00	105 25.00	420 100.00

When testing for the independence of nata_cancer against glyphosates, we obtain a p_{value} <0.0001. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude that the nata_cancer and glyphosates are associated with each other**

Chi-squared test for independence on variables for individual state (NC, IL)

```
/* making categorical variable for north carolina*/
if smoking_rate le 23.79 then smoking_ratec = '1';
if smoking_rate gt 23.79 and smoking_rate le 25.46 then smoking_ratec= '2';
if smoking_rate gt 25.46 and smoking_rate le 26.76  then smoking_ratec = '3';
if smoking_rate gt 26.76 then smoking_ratec ='4';

if nata le 29.2495 then natac = '1';
if nata gt 29.2495 and nata le 35.8435 then natac= '2';
if nata gt 35.8435 and nata le 41.5720 then natac = '3';
if nata gt 41.5720 then natac ='4';

if glyphosates le 2.541915 then glyphosatesc = '1';
if glyphosates gt 2.541915 and glyphosates le 10.054425 then glyphosatesc= '2';
if glyphosates gt 10.054425 and glyphosates le 28.953130 then glyphosatesc = '3';
if glyphosates gt 28.953130 then glyphosatesc ='4';

if state ne 'NC' then delete;

 /*test for independence between different variables smoking_rate vs nata_cancer  for North Carolina  */
proc freq;
 table smoking_ratec*natac/chisq cellchi2 expected;
 run;

 /*test for independence between different variables smoking_rate vs. glyphosates for North Carolina */
proc freq;
 table smoking_ratec*glyphosatesc/chisq cellchi2 expected;
 run;

 /*test for independence between different variables nata_cancer vs. glyphosatesc for North Carolina */
proc freq;
 table natac*glyphosatesc/chisq cellchi2 expected;
 run;
```

Smoking_rate vs. Nata_cancer (NC):

Statistics for Table of smoking_ratec by natac

Statistic	DF	Value	Prob
Chi-Square	9	17.6544	0.0394
Likelihood Ratio Chi-Square	9	17.3017	0.0442
Mantel-Haenszel Chi-Square	1	1.1736	0.2787
Phi Coefficient		0.4202	
Contingency Coefficient		0.3874	
Cramer's V		0.2426	

Sample Size = 100

The FREQ Procedure

Table of smoking_ratec by natac

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct	natac 1	natac 2	natac 3	natac 4	Total
smoking_ratec 1	5 6.25 0.25 5.00 20.00 20.00	5 6.25 0.25 5.00 20.00 20.00	4 6.25 0.81 4.00 16.00 16.00	11 6.25 3.61 11.00 44.00 44.00	25 25.00
2	6 6.25 0.01 6.00 24.00 24.00	6 6.25 0.01 6.00 24.00 24.00	8 6.25 0.49 8.00 32.00 32.00	5 6.25 0.25 5.00 20.00 20.00	25 25.00
3	10 6.5 1.8846 10.00 38.46 40.00	9 6.5 0.9615 9.00 34.62 36.00	2 6.5 3.1154 2.00 7.69 8.00	5 6.5 0.3462 5.00 19.23 20.00	26 26.00
4	4 6 0.6667 4.00 16.67 16.00	5 6 0.1667 5.00 20.83 20.00	11 6 4.1667 11.00 45.83 44.00	4 6 0.6667 4.00 16.67 16.00	24 24.00
Total	25 25.00	25 25.00	25 25.00	25 25.00	100 100.00

When testing for the independence of smoking_rate against nata_cancer rate, we obtain a p_{value} <0.0394. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude that the smoking_rates and nata_cancers are associated with each other.**

Smoking_rate vs. Glyphosates (NC):

Statistics for Table of smoking_ratec by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	6.1446	0.7254
Likelihood Ratio Chi-Square	9	6.2943	0.7101
Mantel-Haenszel Chi-Square	1	0.2721	0.6019
Phi Coefficient		0.2479	
Contingency Coefficient		0.2406	
Cramer's V		0.1431	

When testing for the independence of smoking_rate against glyphosates, we obtain a p_{value} <0.7254. Since $p_{value} > \alpha$, **we cannot reject the H_0. Thus, we conclude that the smoking_rates and nata_cancers are independent from each other**

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

Table of smoking_ratec by glyphosatesc

smoking_ratec	glyphosatesc 1	2	3	4	Total
1	7 6.25 0.09 7.00 28.00 28.00	9 6.25 1.21 9.00 36.00 36.00	5 6.25 0.25 5.00 20.00 20.00	4 6.25 0.81 4.00 16.00 16.00	25 25.00
2	6 6.25 0.01 6.00 24.00 24.00	4 6.25 0.81 4.00 16.00 16.00	8 6.25 0.49 8.00 32.00 32.00	7 6.25 0.09 7.00 28.00 28.00	25 25.00
3	5 6.5 0.3462 5.00 19.23 20.00	5 6.5 0.3462 5.00 19.23 20.00	8 6.5 0.3462 8.00 30.77 32.00	8 6.5 0.3462 8.00 30.77 32.00	26 26.00
4	7 6 0.1667 7.00 29.17 28.00	7 6 0.1667 7.00 29.17 28.00	4 6 0.6667 4.00 16.67 16.00	6 6 0 6.00 25.00 24.00	24 24.00
Total	25 25.00	25 25.00	25 25.00	25 25.00	100 100.00

Nata_cancer vs. Glyphosates (NC):

Statistics for Table of natac by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	33.7600	<.0001
Likelihood Ratio Chi-Square	9	33.6113	0.0001
Mantel-Haenszel Chi-Square	1	0.1014	0.7502
Phi Coefficient		0.5810	
Contingency Coefficient		0.5024	
Cramer's V		0.3355	

When testing for the independence of nata_cancer against glyphosates, we obtain a p_{value} <0.0001. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude that the nata_cancer and glyphosates are associated with each other**

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

Table of natac by glyphosatesc

natac	glyphosatesc 1	2	3	4	Total
1	13 6.25 7.29 13.00 52.00 52.00	2 6.25 2.89 2.00 8.00 8.00	6 6.25 0.01 6.00 24.00 24.00	4 6.25 0.81 4.00 16.00 16.00	25 25.00
2	4 6.25 0.81 4.00 16.00 16.00	3 6.25 1.69 3.00 12.00 12.00	6 6.25 0.01 6.00 24.00 24.00	12 6.25 5.29 12.00 48.00 48.00	25 25.00
3	5 6.25 0.25 5.00 20.00 20.00	7 6.25 0.09 7.00 28.00 28.00	5 6.25 0.25 5.00 20.00 20.00	8 6.25 0.49 8.00 32.00 32.00	25 25.00
4	3 6.25 1.69 3.00 12.00 12.00	13 6.25 7.29 13.00 52.00 52.00	8 6.25 0.49 8.00 32.00 32.00	1 6.25 4.41 1.00 4.00 4.00	25 25.00
Total	25 25.00	25 25.00	25 25.00	25 25.00	100 100.00

```
/* making categorical variable for Illinois */
if smoking_rate le 22.90 then smoking_ratec = '1';
if smoking_rate gt 22.90 and smoking_rate le 24.58 then smoking_ratec= '2';
if smoking_rate gt 24.58 and smoking_rate le 25.82 then smoking_ratec = '3';
if smoking_rate gt 25.82 then smoking_ratec ='4';

if nata le 28.655 then natac = '1';
if nata gt 28.655 and nata le 30.7765 then natac= '2';
if nata gt 30.7765 and nata le 33.14 then natac = '3';
if nata gt 33.14 then natac ='4';

if glyphosates le 56.90997 then glyphosatesc = '1';
if glyphosates gt 56.90997 and glyphosates le 90.57663 then glyphosatesc= '2';
if glyphosates gt 90.57663 and glyphosates le 127.43457 then glyphosatesc = '3';
if glyphosates gt 127.43457 then glyphosatesc ='4';
 /*test for independence between different variables smoking_rate vs nata_cancer  for Illinois */
proc freq;
 table smoking_ratec*natac/chisq cellchi2 expected;
 run;

 /*test for independence between different variables smoking_rate vs. glyphosates for Illinois */
proc freq;
 table smoking_ratec*glyphosatesc/chisq cellchi2 expected;
 run;

 /*test for independence between different variables nata_cancer vs. glyphosatesc for Illinois */
proc freq;
 table natac*glyphosatesc/chisq cellchi2 expected;
 run;
```

Smoking_rate vs. Nata_cancer (IL):

Statistics for Table of smoking_ratec by natac

Statistic	DF	Value	Prob
Chi-Square	9	5.2214	0.8146
Likelihood Ratio Chi-Square	9	5.6713	0.7723
Mantel-Haenszel Chi-Square	1	1.1152	0.2910
Phi Coefficient		0.2263	
Contingency Coefficient		0.2207	
Cramer's V		0.1306	

Sample Size = 102

The FREQ Procedure

Table of smoking_ratec by natac

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

smoking_ratec	natac 1	2	3	4	Total
1	7 6.6275 0.0209 6.86 26.92 26.92	7 6.3725 0.0618 6.86 26.92 28.00	5 6.3725 0.2956 4.90 19.23 20.00	7 6.6275 0.0209 6.86 26.92 26.92	26 25.49
2	9 6.6275 0.8493 8.82 34.62 34.62	6 6.3725 0.0218 5.88 23.08 24.00	6 6.3725 0.0218 5.88 23.08 24.00	5 6.6275 0.3996 4.90 19.23 19.23	26 25.49
3	7 6.3725 0.0618 6.86 28.00 26.92	4 6.1275 0.7387 3.92 16.00 16.00	7 6.1275 0.1243 6.86 28.00 28.00	7 6.3725 0.0618 6.86 28.00 26.92	25 24.51
4	3 6.3725 1.7849 2.94 12.00 11.54	8 6.1275 0.5723 7.84 32.00 32.00	7 6.1275 0.1243 6.86 28.00 28.00	7 6.3725 0.0618 6.86 28.00 26.92	25 24.51
Total	26 25.49	25 24.51	25 24.51	26 25.49	102 100.00

When testing for the independence of smoking_rate against nata_cancer rate, we obtain a p_{value} <0.8146. Since $p_{value} > \alpha$, we **cannot reject the H_0. Thus, smoking_rate and nata_cancer is independent from one another.**

Smoking_rate vs. Glyphosates (IL):

Statistics for Table of smoking_ratec by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	3.9991	0.9115
Likelihood Ratio Chi-Square	9	4.1777	0.8993
Mantel-Haenszel Chi-Square	1	1.2148	0.2704
Phi Coefficient		0.1980	
Contingency Coefficient		0.1942	
Cramer's V		0.1143	

Sample Size = 102

The FREQ Procedure

Table of smoking_ratec by glyphosatesc

Frequency / Expected / Cell Chi-Square / Percent / Row Pct / Col Pct	glyphosatesc 1	2	3	4	Total
smoking_ratec 1	7 6.6275 0.0209 6.86 26.92 26.92	8 6.3725 0.4156 7.84 30.77 32.00	6 6.6275 0.0594 5.88 23.08 23.08	5 6.3725 0.2956 4.90 19.23 20.00	26 25.49
2	8 6.6275 0.2843 7.84 30.77 30.77	7 6.3725 0.0618 6.86 26.92 28.00	6 6.6275 0.0594 5.88 23.08 23.08	5 6.3725 0.2956 4.90 19.23 20.00	26 25.49
3	4 6.3725 0.8833 3.92 16.00 15.38	6 6.1275 0.0027 5.88 24.00 24.00	7 6.3725 0.0618 6.86 28.00 26.92	8 6.1275 0.5723 7.84 32.00 32.00	25 24.51
4	7 6.3725 0.0618 6.86 28.00 26.92	4 6.1275 0.7387 3.92 16.00 16.00	7 6.3725 0.0618 6.86 28.00 26.92	7 6.1275 0.1243 6.86 28.00 28.00	25 24.51
Total	26 25.49	25 24.51	26 25.49	25 24.51	102 100.00

When testing for the independence of smoking_rate against glyphosates, we obtain a p_{value} <0.9115. Since $p_{value} > \alpha$, **we cannot reject the H_0. Thus, smoking_rates and nata_cancers are independent from one another**

Nata_cancer vs. Glyphosates (IL):

Statistics for Table of natac by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	9	20.0690	0.0175
Likelihood Ratio Chi-Square	9	20.7197	0.0140
Mantel-Haenszel Chi-Square	1	8.8337	0.0030
Phi Coefficient		0.4436	
Contingency Coefficient		0.4055	
Cramer's V		0.2561	

Sample Size = 102

The FREQ Procedure

Table of natac by glyphosatesc

Frequency / Expected / Cell Chi-Square / Percent / Row Pct / Col Pct	glyphosatesc 1	2	3	4	Total
natac 1	2 6.6275 3.231 1.96 7.69 7.69	9 6.3725 1.0833 8.82 34.62 36.00	5 6.6275 0.3996 4.90 19.23 19.23	10 6.3725 2.0649 9.80 38.46 40.00	26 25.49
2	6 6.3725 0.0218 5.88 24.00 23.08	2 6.1275 2.7803 1.96 8.00 8.00	11 6.3725 3.3602 10.78 44.00 42.31	6 6.1275 0.0027 5.88 24.00 24.00	25 24.51
3	6 6.3725 0.0218 5.88 24.00 23.08	8 6.1275 0.5723 7.84 32.00 32.00	6 6.3725 0.0218 5.88 24.00 23.08	5 6.1275 0.2075 4.90 20.00 20.00	25 24.51
4	12 6.6275 4.3553 11.76 46.15 46.15	6 6.3725 0.0218 5.88 23.08 24.00	4 6.6275 1.0417 3.92 15.38 15.38	4 6.3725 0.8833 3.92 15.38 16.00	26 25.49
Total	26 25.49	25 24.51	26 25.49	25 24.51	102 100.00

When testing for the independence of nata_cancer against glyphosates, we obtain a p_{value} <0.0175. Since $p_{value} < \alpha$, we can **reject the H_0 and conclude that the nata_cancer and glyphosates are associated with each other**

Chi-squared test for homogeneity:

Chi-Squared Test for homogeneity:

$$H_0: the\ column\ distributions\ are\ homogeneous$$
$$H_1: the\ column\ distributions\ are\ not\ homogeneous$$

```
/*making categoriccal variable for all states*/
if smoking_rate le 21.29 then smoking_ratec = '1';
if smoking_rate gt 21.29 and smoking_rate le 24.06 then smoking_ratec= '2';
if smoking_rate gt 24.06 and smoking_rate le 26.04 then smoking_ratec = '3';
if smoking_rate gt 26.04 then smoking_ratec ='4';

if nata le 27.4595 then natac = '1';
if nata gt 27.4595 and nata le 32.4345 then natac = '2';
if nata gt 32.4345 and nata le 39.0790 then natac = '3';
if nata gt 39.0790 then natac='4';

if glyphosates le 2.836460 then glyphosatesc = '1';
if glyphosates gt 2.836460 and glyphosates le 21.750745 then glyphosatesc = '2';
if glyphosates gt 21.750745 and glyphosates le 61.827565 then glyphosatesc = '3';
if glyphosates gt 61.827565 then glyphosatesc='4';
```

All states (CA, FL, IL, NJ, NC, WI)

Smoking_rate:

Statistics for Table of state by smoking_ratec

Statistic	DF	Value	Prob
Chi-Square	15	165.5281	<.0001
Likelihood Ratio Chi-Square	15	166.1151	<.0001
Mantel-Haenszel Chi-Square	1	1.2064	0.2720
Phi Coefficient		0.6278	
Contingency Coefficient		0.5317	
Cramer's V		0.3625	

Sample Size = 420

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Percent
Row Pct
Col Pct

Table of state by smoking_ratec

state	smoking_ratec 1	2	3	4	Total
CA	42 14.5 52.155 10.00 72.41 40.00	12 14.5 0.431 2.86 20.69 11.43	3 14.638 9.2529 0.71 5.17 2.83	1 14.362 12.432 0.24 1.72 0.96	58 13.81
FL	10 16.75 2.7201 2.38 14.93 9.52	8 16.75 4.5709 1.90 11.94 7.62	16 16.91 0.0489 3.81 23.88 15.09	33 16.59 16.231 7.86 49.25 31.73	67 15.95
IL	14 25.5 5.1863 3.33 13.73 13.33	29 25.5 0.4804 6.90 28.43 27.62	37 25.743 4.9227 8.81 36.27 34.91	22 25.257 0.42 5.24 21.57 21.15	102 24.29
NC	8 25 11.56 1.90 8.00 7.62	20 25 1 4.76 20.00 19.05	32 25.238 1.8117 7.62 32.00 30.19	40 24.762 9.3773 9.52 40.00 38.46	100 23.81
NJ	13 5.25 11.44 3.10 61.90 12.38	7 5.25 0.5833 1.67 33.33 6.67	1 5.3 3.4887 0.24 4.76 0.94	0 5.2 5.2 0.00 0.00 0.00	21 5.00
WI	18 18 0 4.29 25.00 17.14	29 18 6.7222 6.90 40.28 27.62	17 18.171 0.0755 4.05 23.61 16.04	8 17.829 5.4183 1.90 11.11 7.69	72 17.14
Total	105 25.00	105 25.00	106 25.24	104 24.76	420 100.00

When testing for the homogeneity of all states having the same ratio of smoking_rate, we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, the smoking rate distribution is not homogenous across states**

When testing for independence. we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, smoking_rate and states are associated**

Nata_cancer:

Statistics for Table of state by natac

Statistic	DF	Value	Prob
Chi-Square	15	227.8974	<.0001
Likelihood Ratio Chi-Square	15	228.6835	<.0001
Mantel-Haenszel Chi-Square	1	63.6542	<.0001
Phi Coefficient		0.7366	
Contingency Coefficient		0.5931	
Cramer's V		0.4253	

Sample Size = 420

The FREQ Procedure

Table of state by natac

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct	natac 1	2	3	4	Total
CA	11 14.5 0.8448 2.62 18.97 10.48	12 14.5 0.431 2.86 20.69 11.43	11 14.5 0.8448 2.62 18.97 10.48	24 14.5 6.2241 5.71 41.38 22.86	58 13.81
FL	3 16.75 11.287 0.71 4.48 2.86	7 16.75 5.6754 1.67 10.45 6.67	28 16.75 7.556 6.67 41.79 26.67	29 16.75 8.959 6.90 43.28 27.62	67 15.95
IL	16 25.5 3.5392 3.81 15.69 15.24	52 25.5 27.539 12.38 50.98 49.52	30 25.5 0.7941 7.14 29.41 28.57	4 25.5 18.127 0.95 3.92 3.81	102 24.29
NC	17 25 2.56 4.05 17.00 16.19	20 25 1 4.76 20.00 19.05	24 25 0.04 5.71 24.00 22.86	39 25 7.84 9.29 39.00 37.14	100 23.81
NJ	1 5.25 3.4405 0.24 4.76 0.95	3 5.25 0.9643 0.71 14.29 2.86	8 5.25 1.4405 1.90 38.10 7.62	9 5.25 2.6786 2.14 42.86 8.57	21 5.00
WI	57 18 84.5 13.57 79.17 54.29	11 18 2.7222 2.62 15.28 10.48	4 18 10.889 0.95 5.56 3.81	0 18 18 0.00 0.00 0.00	72 17.14
Total	105 25.00	105 25.00	105 25.00	105 25.00	420 100.00

When testing for the homogeneity of all states having the same ratio of nata_cancer, we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, the nata_cancer distribution is not homogenous across states**

When testing for independence. we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, nata_cancer and states are associated**

glyphosates:

Statistics for Table of state by glyphosatesc

Statistic	DF	Value	Prob
Chi-Square	15	211.1932	<.0001
Likelihood Ratio Chi-Square	15	217.4838	<.0001
Mantel-Haenszel Chi-Square	1	0.4528	0.5010
Phi Coefficient		0.7091	
Contingency Coefficient		0.5784	
Cramer's V		0.4094	

Sample Size = 420

When testing for the homogeneity of all states having the same ratio of nata_cancer, we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, the glyphosates distribution is not homogenous across states**

When testing for independence. we obtain a $p_{value} < 0.0001$. Since $p_{value} < \alpha$, we **reject the H_0. Thus, glyphosates and states are associated**

The FREQ Procedure

Table of state by glyphosatesc

Frequency / Expected / Cell Chi-Square / Percent / Row Pct / Col Pct	glyphosatesc 1	2	3	4	Total
CA	26	12	11	9	58
	14.5	14.5	14.5	14.5	
	9.1207	0.431	0.8448	2.0862	
	6.19	2.86	2.62	2.14	13.81
	44.83	20.69	18.97	15.52	
	24.76	11.43	10.48	8.57	
FL	27	24	7	9	67
	16.75	16.75	16.75	16.75	
	6.2724	3.1381	5.6754	3.5858	
	6.43	5.71	1.67	2.14	15.95
	40.30	35.82	10.45	13.43	
	25.71	22.86	6.67	8.57	
IL	2	5	25	70	102
	25.5	25.5	25.5	25.5	
	21.657	16.48	0.0098	77.657	
	0.48	1.19	5.95	16.67	24.29
	1.96	4.90	24.51	68.63	
	1.90	4.76	23.81	66.67	
NC	26	39	27	8	100
	25	25	25	25	
	0.04	7.84	0.16	11.56	
	6.19	9.29	6.43	1.90	23.81
	26.00	39.00	27.00	8.00	
	24.76	37.14	25.71	7.62	
NJ	12	9	0	0	21
	5.25	5.25	5.25	5.25	
	8.6786	2.6786	5.25	5.25	
	2.86	2.14	0.00	0.00	5.00
	57.14	42.86	0.00	0.00	
	11.43	8.57	0.00	0.00	
WI	12	16	35	9	72
	18	18	18	18	
	2	0.2222	16.056	4.5	
	2.86	3.81	8.33	2.14	17.14
	16.67	22.22	48.61	12.50	
	11.43	15.24	33.33	8.57	
Total	105	105	105	105	420
	25.00	25.00	25.00	25.00	100.00

Regression

Fitting linear model for all states:

Assumptions:

1) x_i are non-random, observed without error:
2) ϵ_i *are random variable with* $E(\epsilon_i) = 0$ *and* $var(\epsilon_i) = \sigma^2$
3) ϵ_i *are uncorrelated from observation to observation*

$$\widehat{y_{i_i}} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

Interval estimation of slope and assumption

Slope: $\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}}$

Intercept: $\hat{B}_0 \pm t_{\alpha/2\,,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$

Test for slope: $H_0: \beta_1 = 0 \mid H_1: \beta_1 \neq 0, F_0 = \frac{MS_{reg}}{MSE}$

Assumptions:

1) $\epsilon_i \; are \; uncorrelated \sim N(0, \sigma^2)$
2) $y \sim N(\beta_0 + \beta_1 x, \sigma^2) \; for \; all \; x$

Test for lack of fit:

$$H_0: y \; and \; x \; are \; associated \; (no \; LOF) \mid H_1: y \; and \; x \; are \; not \; associated \; (LOF)$$

Assumptions:

1) Are the data normal? Are residual normal?
2) No measurement errors
3) Should the model used be a linear model?

For all states <u>(CA, FL, IL, NJ, NC, WI)</u>

```
/* 95% confidence interval on the betas, test for slope, and lack of fit for
the response variable (smoking_rate)and the regressor variable (nata_cancer) */
proc reg; model smoking=nata/ CLB lackfit  alpha = 0.05; TEST nata = 0; run;
```

<u>Smoking_rate vs Nata Cancer:</u>

Dependent Variable: smoking

Number of Observations Read	420
Number of Observations Used	420

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	110.81177	110.81177	7.46	0.0066
Error	418	6207.41204	14.85027		
Lack of Fit	412	6126.63604	14.87048	1.10	0.5091
Pure Error	6	80.77600	13.46267		
Corrected Total	419	6318.22381			

Root MSE	3.85360	R-Square	0.0175
Dependent Mean	23.35933	Adj R-Sq	0.0152
Coeff Var	16.49707		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	25.53360	0.81786	31.22	<.0001	23.92597	27.14123
nata	1	-0.06528	0.02390	-2.73	0.0066	-0.11226	-0.01831

$$\widehat{y_i} = 25.53 - 0.065x_i$$

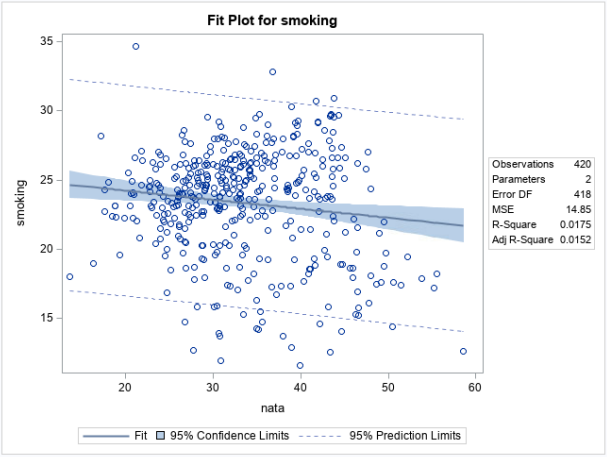
At 95 Confidence interval for true intercept & slope:

$$23.926 < \beta_0 < 27.141$$
$$-0.112 < \beta_1 < -0.018$$

Test for slope:

We get a $P_{value} = 0.0066$, since **Pvalue < α, we reject the Ho**. We can conclude slope of the linear model is not entirely 0

Test for lack of fit:



Since the R^2 (coefficient of determination) is ~ 0.0175. **It is a bad fit to use a linear model to represent the model**

We get a $P_{value} = 0.5091$, **since Pvalue > α, we fail to reject the Ho**. We can conclude smoking rate and nata_cancer is associated in a linear fashion

```
/* 95% confidence interval on the betas, test for slope, and lack of fit for
the response variable (smoking_rate)and the regressor variable (glyphosates) */
proc reg; model smoking=glyphosates/ CLB lackfit  alpha = 0.05; run;
```

Smoking_rate vs. Glyphosates:

Dependent Variable: smoking

Number of Observations Read	420
Number of Observations Used	420

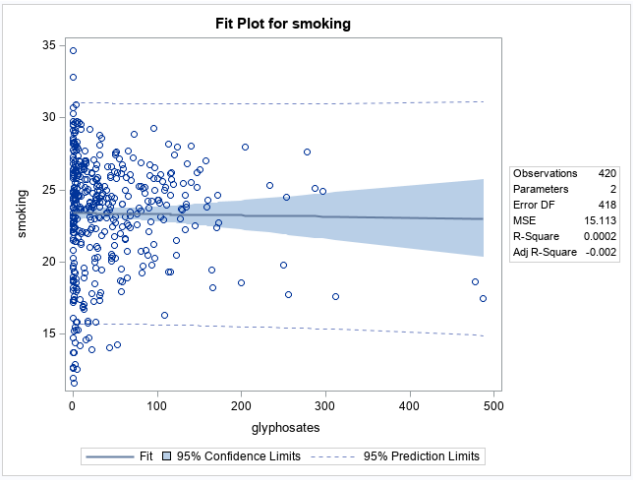
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.95425	0.95425	0.06	0.8017
Error	418	6317.26956	15.11309		
Lack of Fit	416	6256.10716	15.03872	0.49	0.8678
Pure Error	2	61.16240	30.58120		
Corrected Total	419	6318.22381			

Root MSE	3.88756	R-Square	0.0002
Dependent Mean	23.35933	Adj R-Sq	-0.0022
Coeff Var	16.64241		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	23.39293	0.23207	100.80	<.0001	22.93676	23.84910
glyphosates	1	-0.00076854	0.00306	-0.25	0.8017	-0.00678	0.00524



Since the R^2 (coefficient of determination) is ~ 0.0002, **It is a bad fit to use a linear model to represent the model**

$$\hat{y_i} = 23.393 - 0.001x_i$$

At 95 Confidence interval for true intercept & slope:

$$22.938 < \beta_0 < 23.849$$
$$-0.007 < \beta_1 < 0.005$$

Test for slope:

We get a $P_{value} = 0.80$, **since Pvalue > α, we fail to reject the Ho**. We can conclude the slope of the linear model for smoking and glyphosates is 0.

Test for lack of fit:

We get a $P_{value} = 0.8678$, **since Pvalue > α, we fail to reject the Ho**. We can conclude smoking rate and glyphosates is associated in a linear fashion.

```
/* 95% confidence interval on the betas, test for slope, and test lack of fit
for where the response variable (nata) and the regressor variable (glyphosates)*/
proc reg; model nata=glyphosates/ CLB lackfit alpha = 0.05;run;
```

Nata cancer vs. Glyphosates:

Dependent Variable: nata

Number of Observations Read	420
Number of Observations Used	420

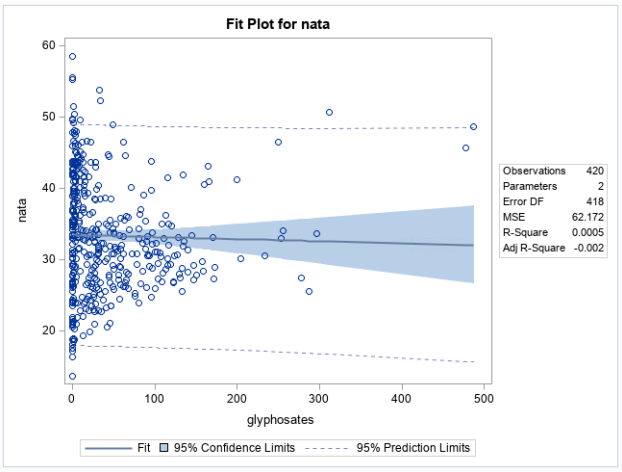
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.29812	13.29812	0.21	0.6440
Error	418	25988	62.17178		
Lack of Fit	416	25413	61.08795	0.21	0.9905
Pure Error	2	575.21709	287.60855		
Corrected Total	419	26001			

Root MSE	7.88491	R-Square	0.0005
Dependent Mean	33.30546	Adj R-Sq	-0.0019
Coeff Var	23.67452		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	33.43087	0.47070	71.02	<.0001	32.50564	34.35610
glyphosates	1	-0.00287	0.00620	-0.46	0.6440	-0.01506	0.00932



Since the R^2 (coefficient of determination) is ~ 0.0005, **It is a bad fit to use a linear model to represent the model**.

$$\hat{y}_i = 33.431 - 0.003x_i$$

At 95 Confidence interval for true intercept & slope:

$$32.506 < \beta_0 < 34.356$$
$$-0.015 < \beta_1 < 0.009$$

Test for slope:

We get a $P_{value} = 0.64$, **since Pvalue > α, we fail to reject the Ho**. We can conclude the slope of the linear model for smoking and glyphosates is 0.

Test for lack of fit:

We get a $P_{value} = 0.99$, **since Pvalue > α, we fail to reject the Ho**. We can conclude nata_cancer and glyphosates is associated in a linear fashion.

For North Carolina

```
/* 95% confidence interval on the betas and test for slope for where
the response variable (nata) and the regressor variable (glyphosates)*/
proc reg; model smoking=nata/ CLB alpha = 0.05;run;
```

Smoking_rate vs Nata Cancer:

Dependent Variable: smoking

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	46.67273	46.67273	6.41	0.0129
Error	98	713.30069	7.27858		
Corrected Total	99	759.97342			

Root MSE	2.69788	R-Square	0.0614
Dependent Mean	24.95240	Adj R-Sq	0.0518
Coeff Var	10.81212		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	28.38563	1.38238	20.53	<.0001	25.64234	31.12891
nata	1	-0.09636	0.03805	-2.53	0.0129	-0.17187	-0.02084

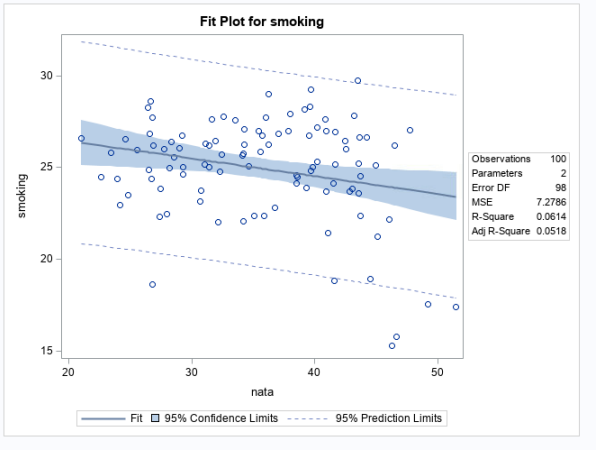
$$\hat{y_i} = 28.38 - 0.10x_i$$

At 95 Confidence interval for true intercept & slope:

$$25.64 < \beta_0 < 31.129$$
$$-0.172 < \beta_1 < -0.021$$

Test for slope:

We get a $P_{value} = 0.01$, since **Pvalue < α, we reject the Ho**. We can conclude slope of the linear model is not entirely 0



Since the R^2 (coefficient of determination) is ~ 0.06. **It is a bad fit to use a linear model to represent the model**

Smoking_rate vs. Glyphosates:

Dependent Variable: smoking

Number of Observations Read	100
Number of Observations Used	100

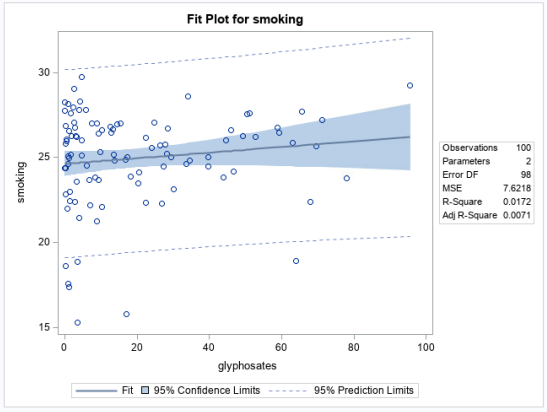
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13.04110	13.04110	1.71	0.1939
Error	98	746.93232	7.62176		
Corrected Total	99	759.97342			

Root MSE	2.76075	R-Square	0.0172
Dependent Mean	24.95240	Adj R-Sq	0.0071
Coeff Var	11.06408		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	24.63037	0.36990	66.59	<.0001	23.89631	25.36442
glyphosates	1	0.01636	0.01251	1.31	0.1939	-0.00846	0.04118



Since the R^2 (coefficient of determination) is ~ 0.0172, **It is a bad fit to use a linear model to represent the model**

$$\hat{y}_i = 24.630 - 0.016x_i$$

At 95 Confidence interval for true intercept & slope:

$$23.896 < \beta_0 < 25.364$$
$$-0.008 < \beta_1 < 0.04$$

Test for slope:

We get a $P_{value} = 0.19$, **since Pvalue > α, we fail to reject the Ho**. We can conclude the slope of the linear model for smoking and glyphosates is 0.

Nata cancer vs. Glyphosates:

Dependent Variable: nata

Number of Observations Read	100
Number of Observations Used	100

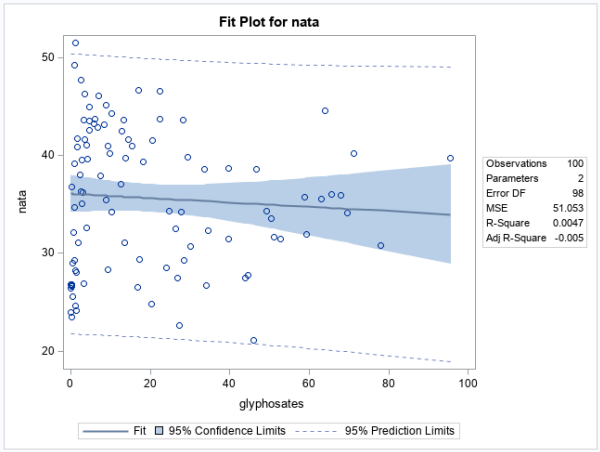
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	23.58069	23.58069	0.46	0.4983
Error	98	5003.17435	51.05280		
Lack of Fit	98	5003.17435	51.05280	.	.
Pure Error	0	0	.		
Corrected Total	99	5026.75504			

Root MSE	7.14512	R-Square	0.0047
Dependent Mean	35.62994	Adj R-Sq	-0.0055
Coeff Var	20.05371		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	36.06297	0.95734	37.67	<.0001	34.16315	37.96279
glyphosates	1	-0.02200	0.03237	-0.68	0.4983	-0.08625	0.04224



Since the R^2 (coefficient of determination) is ~ 0.0047, **It is a bad fit to use a linear model to represent the model**

$$\hat{y}_i = 36.06 - 0.022x_i$$

At 95 Confidence interval for true intercept & slope:

$$34.163 < \beta_0 < 37.963$$
$$-0.086 < \beta_1 < 0.0422$$

Test for slope:

We get a $P_{value} = 0.50$, **since Pvalue > α, we fail to reject the Ho**. We can conclude the slope of the linear model for nata_cancer and glyphosate is 0.

For Illinois

```
/* 95% confidence interval on the betas and test for slope for where
the response variable (nata) and the regressor variable (glyphosates)*/
proc reg; model smoking=nata/ CLB alpha = 0.05;run;
```

Smoking_rate vs Nata Cancer:

Dependent Variable: smoking

Number of Observations Read	102
Number of Observations Used	102

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.26603	0.26603	0.04	0.8475
Error	100	715.37641	7.15376		
Corrected Total	101	715.64243			

Root MSE	2.67465	R-Square	0.0004
Dependent Mean	24.11725	Adj R-Sq	-0.0096
Coeff Var	11.09020		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	23.69346	2.21354	10.70	<.0001	19.30186	28.08507
nata	1	0.01360	0.07052	0.19	0.8475	-0.12630	0.15350

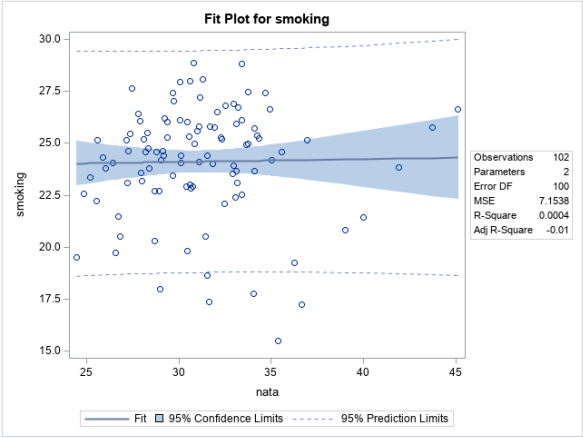
$$\hat{y}_i = 23.69 - 0.01x_i$$

At 95 Confidence interval for true intercept & slope:

$$19.302 < \beta_0 < 28.085$$
$$-0.126 < \beta_1 < 0.153$$

Test for slope:

We get a $P_{value} = 0.85$, since **Pvalue > α, we reject the Ho**. We can conclude slope of the linear model is 0



Since the R^2 (coefficient of determination) is ~ 0.0004. **It is a bad fit to use a linear model to represent the model**

Smoking_rate vs. Glyphosates:

Dependent Variable: smoking

Number of Observations Read	102
Number of Observations Used	102

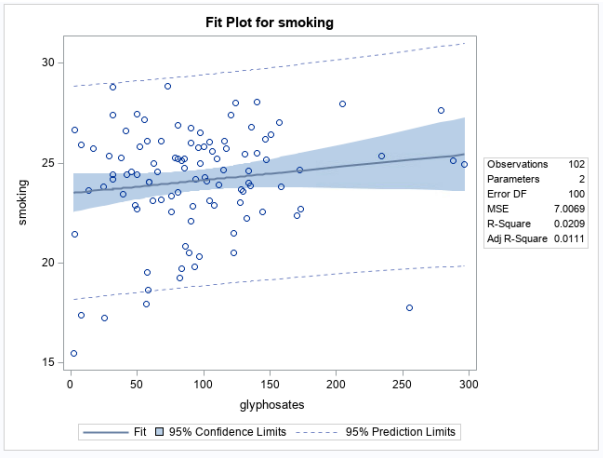
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	14.95119	14.95119	2.13	0.1472
Error	100	700.69125	7.00691		
Corrected Total	101	715.64243			

Root MSE	2.64706	R-Square	0.0209
Dependent Mean	24.11725	Adj R-Sq	0.0111
Coeff Var	10.97578		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	23.49467	0.50035	46.96	<.0001	22.50199	24.48735
glyphosates	1	0.00651	0.00446	1.46	0.1472	-0.00233	0.01536



$$\hat{y}_i = 23.495 - 0.007x_i$$

At 95 Confidence interval for true intercept & slope:

$$22.501 < \beta_0 < 24.487$$

$$-0.002 < \beta_1 < 0.015$$

Test for slope:

We get a $P_{value} = 0.15$, **since Pvalue > α, we fail to reject the Ho**. We can conclude the slope of the linear model for smoking and glyphosates is 0.

Since the R^2 (coefficient of determination) is ~ 0.0209, **It is a bad fit to use a linear model to represent the model**

Nata cancer vs. Glyphosates:

Dependent Variable: nata

Number of Observations Read	102
Number of Observations Used	102

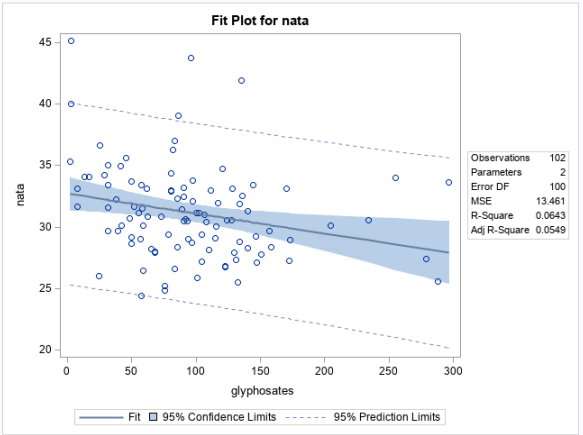
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	92.49144	92.49144	6.87	0.0101
Error	100	1346.14824	13.46148		
Corrected Total	101	1438.63967			

Root MSE	3.66899	R-Square	0.0643
Dependent Mean	31.16493	Adj R-Sq	0.0549
Coeff Var	11.77281		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	32.71343	0.69352	47.17	<.0001	31.33751	34.08935
glyphosates	1	-0.01620	0.00618	-2.62	0.0101	-0.02846	-0.00394



$$\hat{y}_i = 32.71 - 0.016x_i$$

At 95 Confidence interval for true intercept & slope:

Since the R^2 (coefficient of determination) is ~ 0.064, **It is a bad fit to use a linear model to represent the model**

$$31.338 < \beta_0 < 34.089$$
$$-0.028 < \beta_1 < 0.004$$

Test for slope:

We get a $P_{value} = 0.01$, since **Pvalue < α, we reject the Ho**. We can conclude **the slope of the linear model for nata_cancer and glyphosate not 0**.

If we wanted to make prediction interval of a mean y value when x is given

$$\hat{y}_i \pm t_{\frac{\alpha}{2},n-2}\, S_{\hat{y}_0}$$
$$S_{\hat{y}_0} = \sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

If we wanted to make prediction interval of an individual value of y value when x is given

$$\hat{y}_i \pm t_{\frac{\alpha}{2},n-2}\, S_{\hat{y}_0}$$
$$S_{\hat{y}_0} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

Conclusion

In conclusion, our investigation of the central tendency for the variables smoking_rate, nata_cancer, and glyphosates revealed the following: Florida has the highest average smoking rate, New Jersey has the highest Nata Cancer rate, and Illinois has the highest glyphosate mortality. Our normality tests for these variables indicated that they are not normally distributed. The ANOVA tests for all states across all variables were significant, as were the tests for equal variances.In the multiple comparisons test, North Carolina and Illinois showed significant differences from the rest in all variables. The Chi-squared tests for goodness of fit were significant for our states on smoking, nata, and glyphosates. The contingency table tests of independence and homogeneity were significant for smoking_rate vs nata_cancer and nata_cancer vs glyphosates, but not for smoking_rate vs glyphosates. Through regression analysis, we found that it is not recommended to use a linear model to predict the results of one variable against another for our variables. The coefficient of determination for all model results in poor R^2. This applies to all comparisons, such as smoking_rate vs Nata_cancer, smoking_rate vs glyphosates, and nata_cancer vs glyphosates.