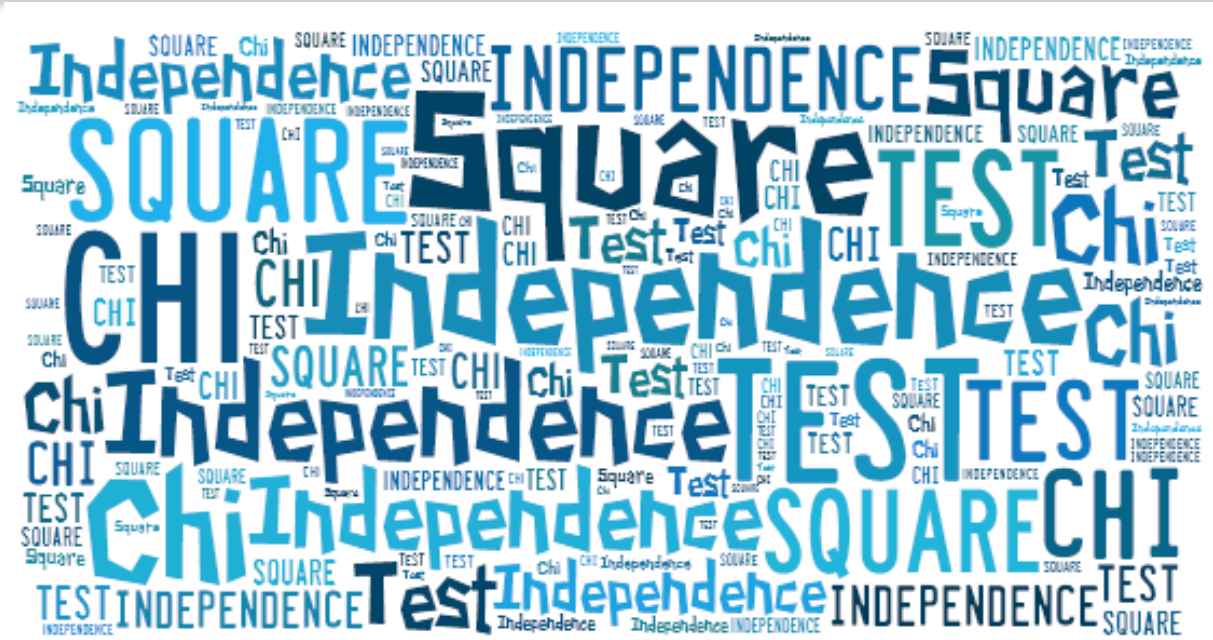
WESLEYAN
UNIVERSITY



Variable types and inferential tools

WESLEYAN UNIVERSITY

Chi Square Test of Independence

Analysis of Variance

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C → C	C → Q
	Quatitative	Q → C	Q → Q

Chi Square (X^2) Test of Independence

WESLEYAN UNIVERSITY

"Random Roadside Survey"

	Gender	Drove drunk?
Driver 1	M	Y
Driver 2	F	N
Driver 3	F	Y
•	•	•
•	•	•
•	•	•
Driver 619	M	N

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

(Craig v. Boren, 429 U.S. 190, 1976)

Calculating conditional percentages

WESLEYAN UNIVERSITY

Gender (X)	Drank Alcohol in Last 2 Hours (Y)? Yes	No	Total
Male	77/481=16.0%	404/481=84.0%	100%
Female	16/138=11.6%	122/138=88.4%	100%

H_o: There is no difference in the drunk driving rate between males and females under 20.

H_a: There is a difference in the drunk driving rate between males and females under 20.

.

Stating the Hypotheses

WESLEYAN UNIVERSITY

H_o: There is no relationship between the two categorical variables. (They are independent.)

H_a: There is a relationship between the two categorical variables. (They are not independent.)

H_o: proportion of male drunk drivers = proportion of female drunk drivers

H_a: proportion of male drunk drivers ≠ proportion of female drunk drivers

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Observed counts

Calculating the expected counts

WESLEYAN UNIVERSITY

$$P(A \text{ and } B) = P(A) * P(B)$$

Applying the rule to the first (top left) cell, if driving drunk and gender were independent then:

P(drunk and male) = P(drunk) * P(male)

By dividing the counts in our table, we see that:

P(Drunk) = 93 / 619 and

P(Male) = 481 / 619,

and so,

P(Drunk and Male) = (93 / 619) (481 / 619)

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Therefore, since there are total of 619 drivers, **if drunk driving and gender were independent**, the **count** of drunk male drivers that I would **expect** to see is:

$$619 * P(Drunk \ \ and \ \ Male) = 619\left(\frac{93}{619}\right)\left(\frac{481}{619}\right) = \frac{93*481}{619}$$

Formula for expected counts

$$Expected\ Count = \frac{Column\ \ Total * Row\ \ Total}{Table\ \ Total}$$

Expected count

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total	
Male	(93*481)/619 →		481	row total
Female	↓		138	
Total	93	526	619	

column total

table total

Expected vs. Observed Counts

WESLEYAN UNIVERSITY

Observed Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Expected Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	(93*481)/619=72.3	(526*481)/619=408.7	481
Female	(93*138)/619=20.7	(526*138)/619=117.3	138
Total	93	526	619

Chi Square Formula

WESLEYAN UNIVERSITY

$$\chi^2 = \sum_{all\ \ cells} \frac{(Observed\ \ Count - ExpectedCount)^2}{Expected\ \ Count}$$

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77 (72.3)	404 (408.7)	481
Female	16 (20.7)	122 (117.3)	138
Total	93	526	619

$$\frac{(77-72.3)^2}{72.3} + \frac{(404-408.7)^2}{408.7} + \frac{(16-20.7)^2}{20.7} + \frac{(122-117.3)^2}{117.3} = .306 + .054 + 1.067 + .188 = 1.62$$

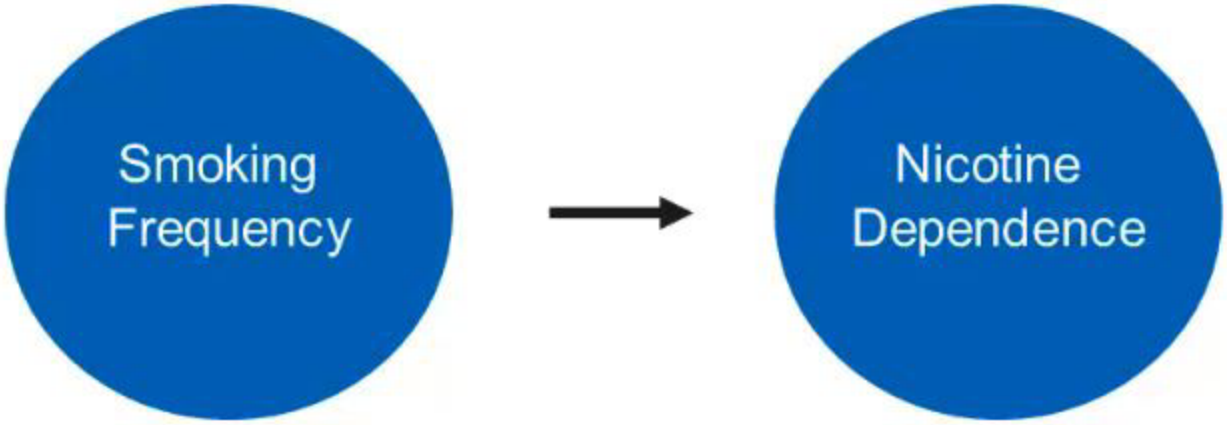
p= 0.201.

Chi Square Test of Independence using SAS

WESLEYAN UNIVERSITY

6-Level Categorical
days per month
(1, 2.5, 5, 14, 22, 30)
USFREQMO

2-Level Categorical
(1=Present/0=Absent)
TAB12MDX



H_o: There is no relationship between the two categorical variables. (They are independent.)

H_a: There is a relationship between the two categorical variables. (They are not independent.)

Percentages presented in Cross Tabs

WESLEYAN UNIVERSITY

Row %

table A

Region	Uninsured	Insured	**Total**
Northeast	12.6%	87.4%	100%
Midwest	12.0%	88.0%	100%
South	18.2%	81.8%	100%
west	17.4%	82.6%	100%

Total %

table B

Region	Uninsured	Insured	**Total**
Northeast	2.3%	16.2%	18.5%
Midwest	2.7%	19.6%	22.3%
South	6.6%	29.5%	36.1%
West	4.0%	19.1%	23.1%
Total	15.6%	84.4%	100%

Column %

table C

Region	Uninsured	Insured
Northeast	15.0%	19.2%
Midwest	17.1%	23.3%
South	42.1%	35.0%
West	25.8%	22.6%
Total	100%	100%

Chi Square Test of Independence using SAS

WESLEYAN UNIVERSITY

Table of TAB12MDX by USFREQMO

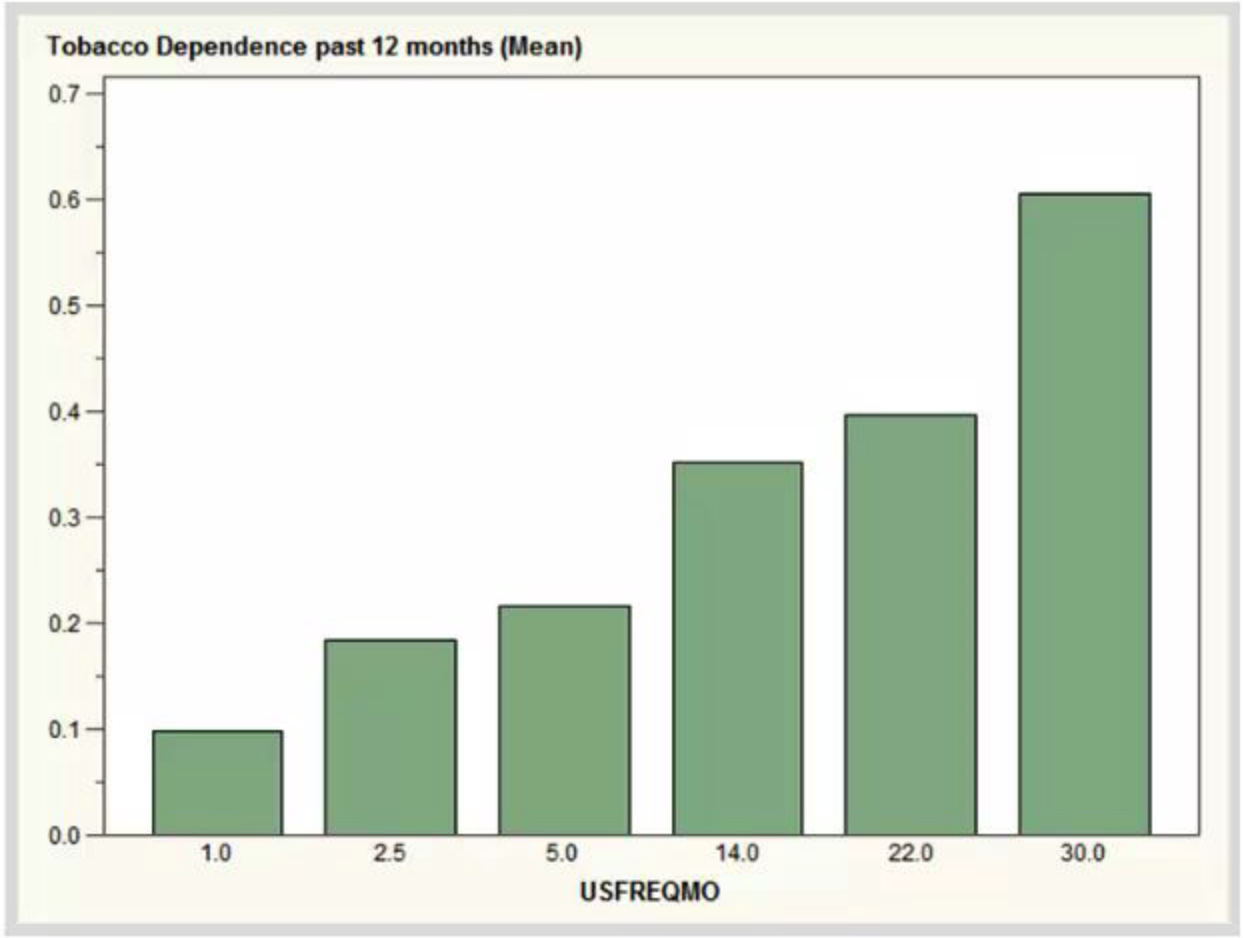
TAB12MDX(Tobacco Dependence past 12 months)		USFREQMO 1	2.5	5	14	22	30	Total
0	Frequency	64	53	69	59	41	521	807
	Percent	3.76	3.11	4.05	3.46	2.41	30.59	47.39
	Row Pct	7.93	6.57	8.55	7.31	5.08	64.56	
	Col Pct	90.14	81.54	78.41	64.84	60.29	39.47	
1	Frequency	7	12	19	32	27	799	896
	Percent	0.41	0.70	1.12	1.88	1.59	46.92	52.61
	Row Pct	0.78	1.34	2.12	3.57	3.01	89.17	
	Col Pct	9.86	18.46	21.59	35.16	39.71	60.53	
Total	Frequency	71	65	88	91	68	1320	1703
	Percent	4.17	3.82	5.17	5.34	3.99	77.51	100.00

Frequency Missing = 3

Statistic	DF	Value	Prob
Chi-Square	5	165.2732	<.0001
Likelihood Ratio Chi-Square	5	176.1834	<.0001
Mantel-Haenszel Chi-Square	1	162.8952	<.0001
Phi Coefficient		0.3115	
Contingency Coefficient		0.2974	
Cramer's V		0.3115	

Rate of ND by # Days Smoked per Month Categories

WESLEYAN UNIVERSITY



$p = .0001$

H_a: Not all ND rates are equal across smoking frequency categories.

Why not test multiple ANOVAs examining each pair?

WESLEYAN UNIVERSITY

Remember that we accept 'significance' and reject the null hypothesis at $P \leq 0.05$ (i.e. a 5% chance that we are wrong)

Performing multiple tests therefore means that our overall chance of committing a type I error is >5%.

# Tests	Comparison α	Familywise α
1	.05	.05
3	.05	.14
6	.05	.26
10	.05	.40
15	.05	.54

Which post hoc test do I use within ANOVA?

WESLEYAN UNIVERSITY

0.05 / c

(c=number of comparisons)

# comparisons	calculation	adjusted Bonferroni p value
3	.05 / 3	.017
6	.05 / 6	.008
10	.05 /10	.005
15	.05 /15	.003

p values for each X^2 paired comparison of ND rates

WESLEYAN UNIVERSITY

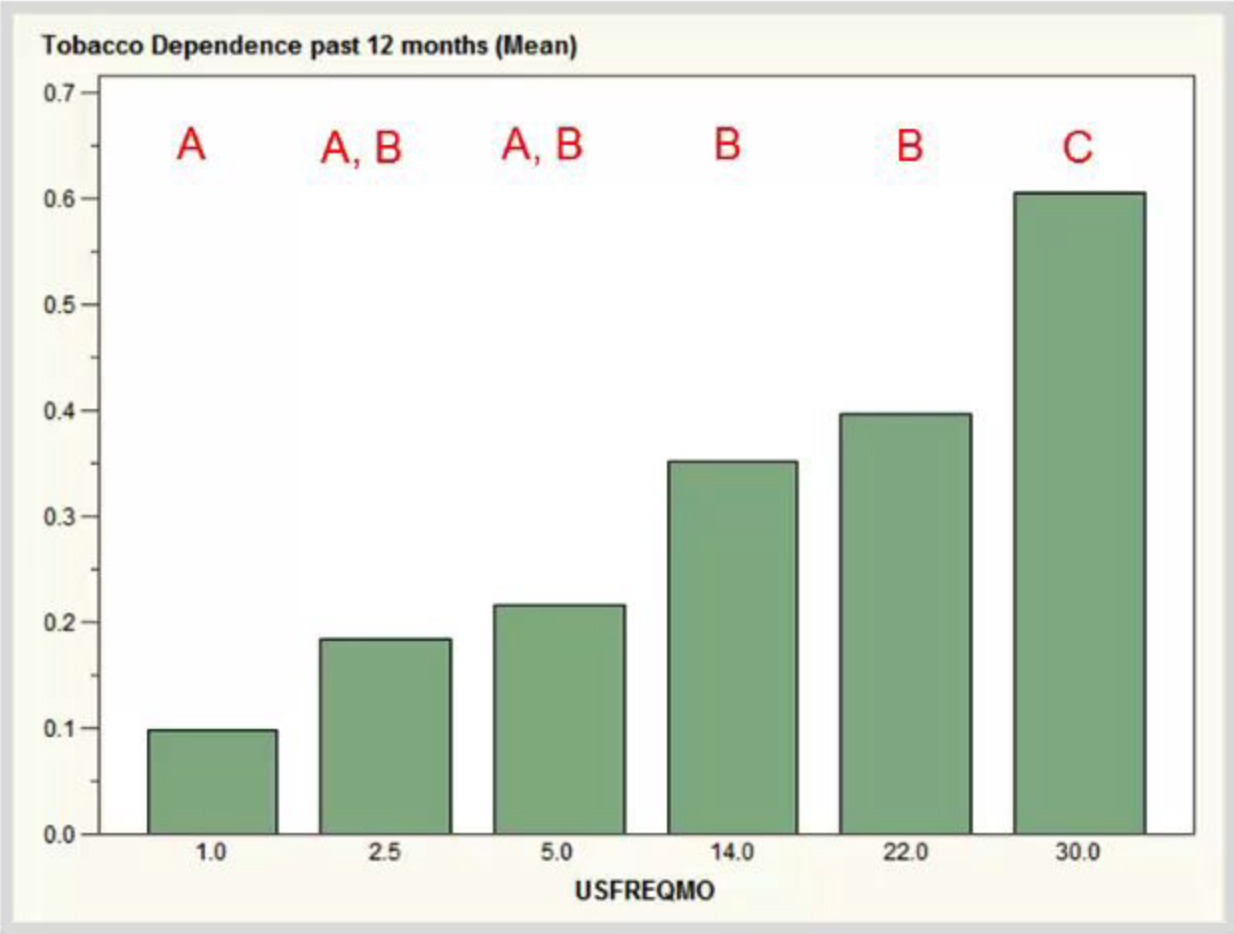
Number of cigarettes smoked per month

	1	2.5	5	14	22	30
1	*					
2.5	0.15	*				
5	0.05	0.63	*			
14	0.0002	0.02	0.04	*		
22	0.0001	0.007	0.01	0.56	*	
30	0.0001	0.0001	0.0001	0.0001	0.0006	*

Bonferroni = 0.05 / 15 = .003

Rate of ND by # Days Smoked per Month Categories

WESLEYAN UNIVERSITY



X^2 p value = .0001 (Bonferroni Adjustment=.003)

Nicotine Dependence rates with the same letter are not significantly different

Summary: Chi Square Test of Independence

WESLEYAN UNIVERSITY

		Response	
		Categorical	Quantitative
Explanatory	Categorical	C → C	C → Q
	Quatitative	Q → C	Q → Q

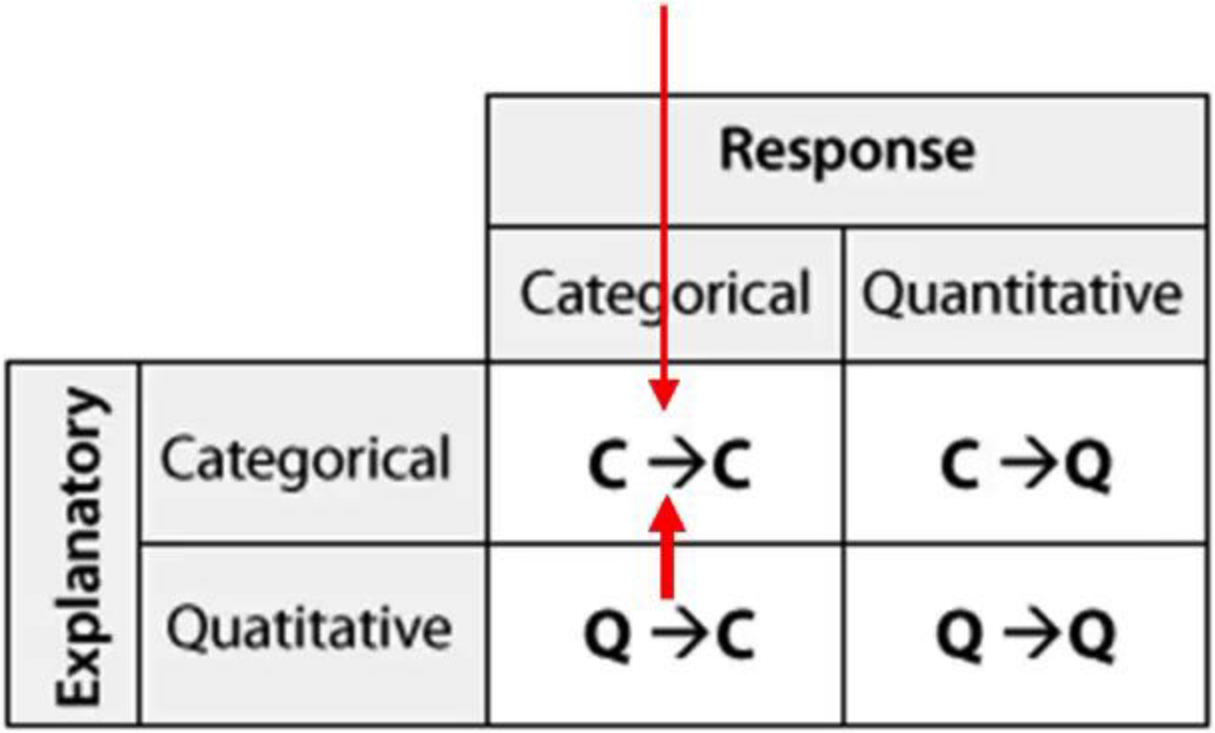
H_o: There is no relationship between the two categorical variables. (They are independent.)

H_a: There is a relationship between the two categorical variables. (They are not independent.)

$$\chi^2 = \sum_{all\ \ cells} \frac{(Observed\ \ Count - ExpectedCount)^2}{Expected\ \ Count}$$

PROC FREQ; TABLES CAT_RESPONSE*CAT_EXPLANATORY /CHISQ;

Chi Square Test of Independence



Explanatory		Response: Categorical	Response: Quantitative
	Categorical	C → C	C → Q
	Quatitative	Q → C	Q → Q

WESLEYAN

UNIVERSITY