

Analysis of Frequencies

So far --> response variable (Y) - continuous

Now look at response variable that is discrete/categorical

eg presence-absence or yes-no

--> ask questions about frequencies in each category

Chi-Square Test

- compare

OBSERVED frequency of a trait

to

EXPECTED frequency of a trait

Testing for randomness

Let's say we're interested in the ratio of females to males in this class. Is the composition a RANDOM sample of the UWO population?

I will have two observed frequencies, one for each of females and males,

And

two expected frequencies, one for each of females and males.

Where do the frequencies come from?

Observed:

Count all females and males

	Count
Females	132
Males	76
Total	208

Where do the frequencies come from?

Expected:

Let's say females:males at UWO is 50:50, half are females and half are males.

So, in a RANDOM sample, expect 50% female and 50% male.

	Observed	Expected
Females	132	
Males	76	
Total	208	

	Observed	Expected
Females	132	
Males	76	
Total	208	

$$\chi^2 = \sum_{i=1}^k \frac{(f_{\text{observed}} - f_{\text{expected}})^2}{f_{\text{expected}}}$$

$$\chi^2 = \frac{(132 - 104)^2}{104} + \frac{(76 - 104)^2}{104}$$

$$= \frac{28^2}{104} + \frac{-28^2}{104} = 7.5 + 7.5 = 15$$

Expected:

Let's say females:males at UWO is 60:40.

So, in a RANDOM sample, expect 60% female and 40% male.

	Observed	Expected
Females	132	
Males	76	
Total	208	

	Observed	Expected
Females	132	124.8
Males	76	83.2
Total	208	208

$$\chi^2 = \frac{(132 - 124.8)^2}{124.8} + \frac{(76 - 83.2)^2}{83.2}$$

$$= \frac{7.2^2}{124.8} + \frac{-7.2^2}{83.2} = 0.42 + 0.62 = 1.04$$

Testing to see if your data fit a theoretical distribution

Hardy-Weinberg expectations for Mendel's Peas

Dihybrid cross testing for independent assortment of traits

smooth-yellow	9
smooth-green	3
wrinkled-yellow	3
wrinkled-green	1

Observed

	Count
smooth-yellow	152
smooth-green	53
wrinkled-yellow	39
wrinkled-green	6
TOTAL	250

Expected

	Count	Expected
smooth-yellow	152	$0.5625 \times 250 = 140.625$
smooth-green	53	$0.1875 \times 250 = 46.875$
wrinkled-yellow	39	$0.1875 \times 250 = 46.875$
wrinkled-green	6	$0.0625 \times 250 = 15.625$
TOTAL	250	

	Observed	Expected	O-E
smooth-yellow	152	140.625	11.375
smooth-green	53	46.875	6.125
wrinkled-yellow	39	46.875	7.875
wrinkled-green	6	15.625	9.625
TOTAL	250	250	

$$\chi^2 = \frac{11.375^2}{140.625} + \frac{6.125^2}{46.875} + \frac{7.875^2}{46.875} + \frac{9.625^2}{15.625} +$$

$$= 0.9201 + 0.8003 + 1.323 + 5.929$$

$$= 8.927$$

Often, collect data on more than one variable simultaneously

2 X 2 Contingency Tables

	Variable A			Total
	Column 1	Column 2	Column 3	
Row 1	O ₁₁	O ₂₁	O ₃₁	R1
Row 2	O ₁₂	O ₂₂	O ₃₂	R2
	C1	C2	C3	Total

R1 = sum of observed in Row 1
 R2 = sum of observed in Row 2
 C1 = sum of observed in Column 1
 C2 = sum of observed in Column 2
 C3 = sum of observed in Column 3
 Total = sum of all observed

	Variable A			Total
	Column 1	Column 2	Column 3	
Row 1	E ₁₁	E ₂₁	E ₃₁	R1
Row 2	E ₁₂	E ₂₂	E ₃₂	R2
	C1	C2	C3	Total

Expected calculation

For example,

A rare tree species can

be rooted in serpentine or non-serpentine soil

have pubescent or smooth leaves

		Soil Type	
		Serpentine	Non-serpentine
Leaf Morphology	Pubescent		
	Smooth		

Is leaf morphology independent of soil type?

Again, want to compare OBSERVED to EXPECTED

Observed

		Soil Type		Total
		Serpentine	Non-serpentine	
Leaf Morphology	Pubescent	12	16	28
	Smooth	22	50	72
Totals		34	66	100

Expected

		Soil Type		Total
		Serpentine	Non-serpentine	
Leaf Morphology	Pubescent	9.52	18.48	28
	Smooth	24.48	47.52	72
Totals		34	66	100

Expected

		Soil Type		Total
		Serpentine	Non-serpentine	
Leaf Morphology	Pubescent	9.52	18.48	28
	Smooth	24.48	47.52	72
Totals		34	66	100

Expected

		Soil Type		Total
		Serpentine	Non-serpentine	
Leaf Morphology	Pubescent	9.52	18.48	28
	Smooth	24.48	47.52	72
Totals		34	66	100

Expected

Leaf Morphology	Soil Type		Total
	Serpentine	Non-serpentine	
Pubescent	9.52	18.48	28
Smooth	24.48	47.52	72
Totals	34	66	100

Leaf Morphology	Soil Type		Total
	Serpentine	Non-serpentine	
Pubescent	12 (9.52)	16 (18.48)	28
Smooth	22 (24.48)	50 (47.52)	72
Totals	34	66	100

$$\chi^2 = \sum_{i=1}^k \frac{(f_{observed} - f_{expected})^2}{f_{expected}}$$

$$= \frac{6.15}{9.52} + \frac{6.15}{18.48} + \frac{6.15}{24.48} + \frac{6.15}{47.52}$$

$$= 0.6461 + 0.3328 + 0.2512 + 0.1294 = 1.3595$$

For example, varieties of tiger beetles found during four times of the year

Season	Colour Pattern		Total
	Bright Red	Not Bright Red	
Early Spring	29	11	40
Late Spring	273	191	464
Early Summer	8	31	39
Late Summer	64	64	128
Totals	374	297	671

H₀: The occurrence of tiger beetle colour types is not dependent upon time of year

H_A: The occurrence of tiger beetle colour types is dependent upon time of year

Season	Colour Pattern		Total
	Bright Red	Not Bright Red	
Early Spring	29 (22.3)	11 (17.7)	40
Late Spring	273 (258.6)	191 (205.38)	464
Early Summer	8 (21.74)	31 (17.26)	39
Late Summer	64 (71.34)	64 (56.66)	128
Totals	374	297	671

$$\chi^2 = \sum_{i=1}^k \frac{(f_{observed} - f_{expected})^2}{f_{expected}}$$

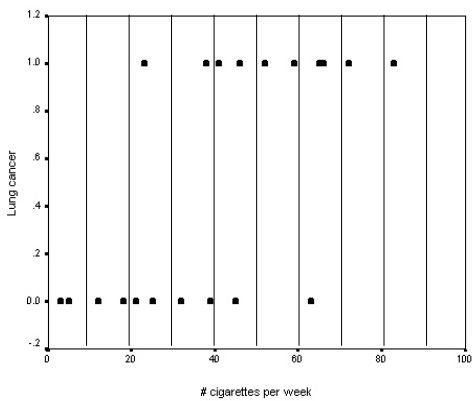
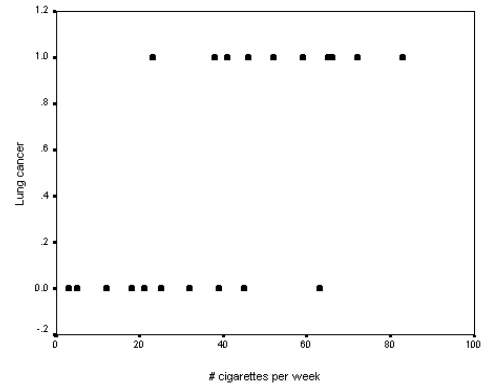
$$= \frac{(29 - 22.3)^2}{22.3} + \frac{(273 - 258.6)^2}{258.6} + \frac{(8 - 21.7)^2}{21.7} + \dots + \frac{(64 - 56.66)^2}{22.3}$$

$$= 2.01 + 0.8 + 8.68 + \dots + 0.95 = 27.68$$

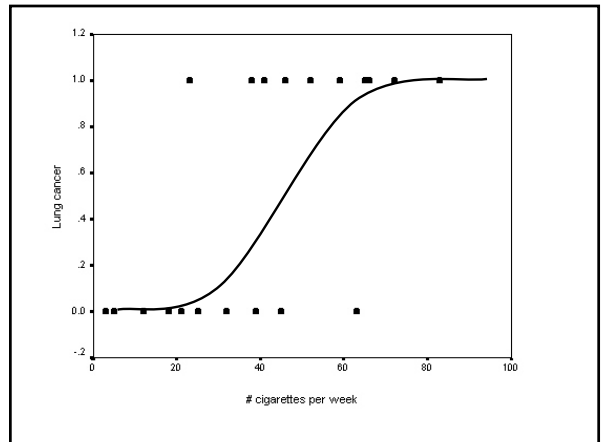
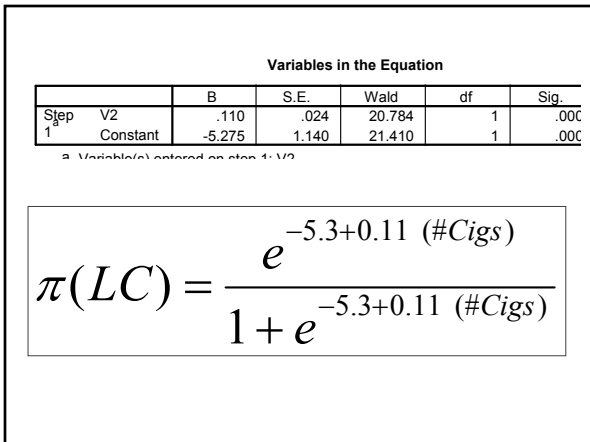
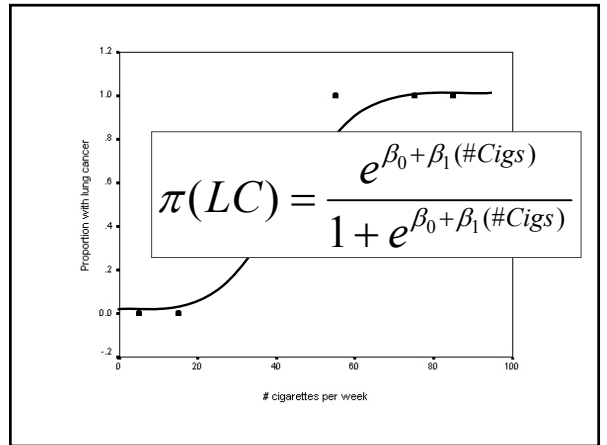
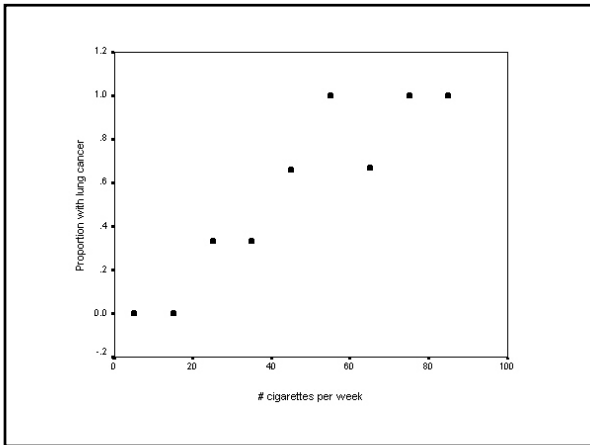
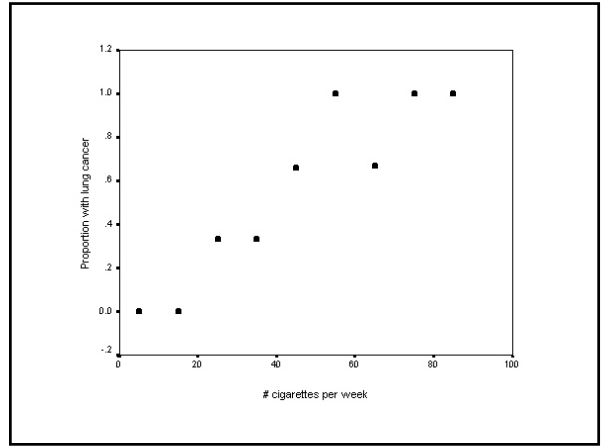
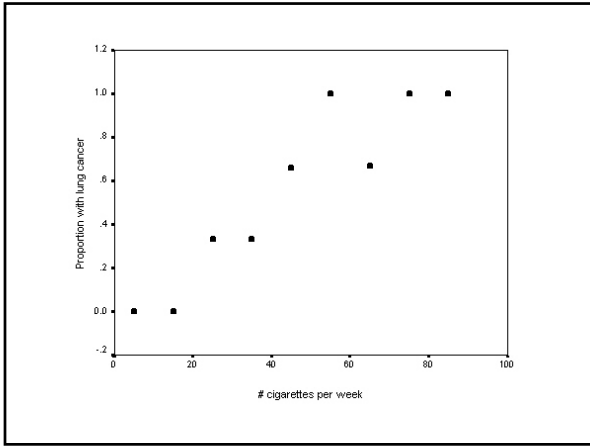
Logistic Regression

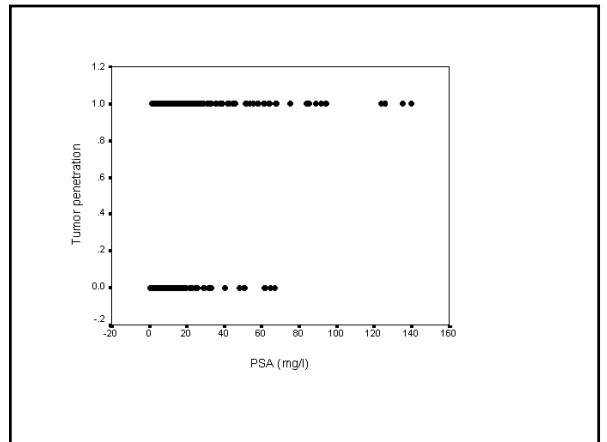
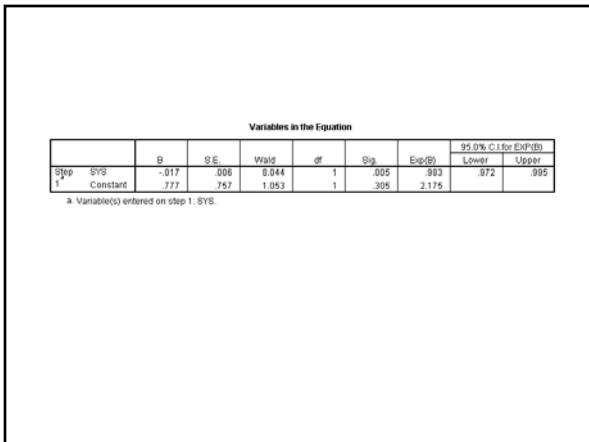
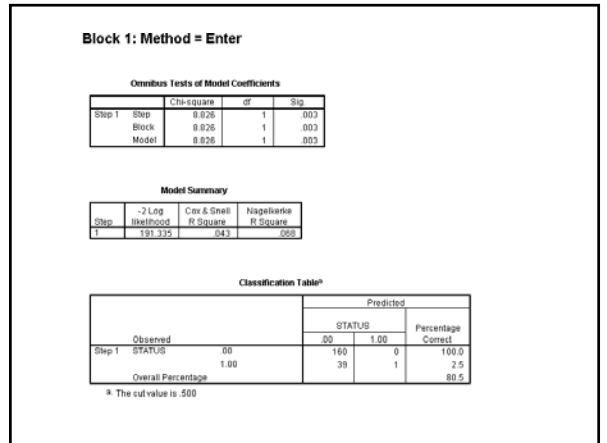
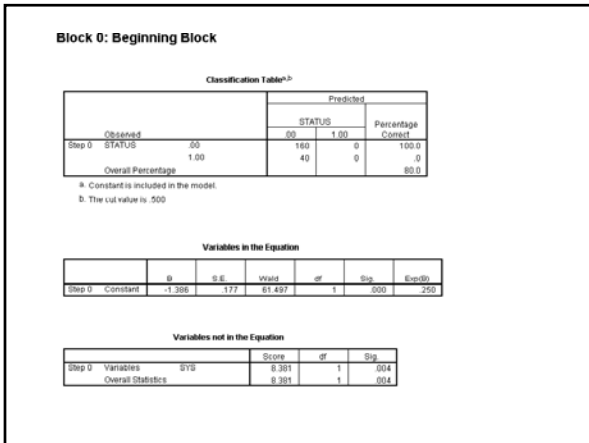
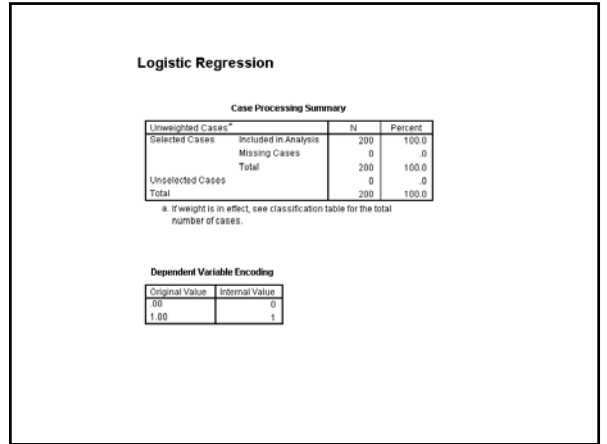
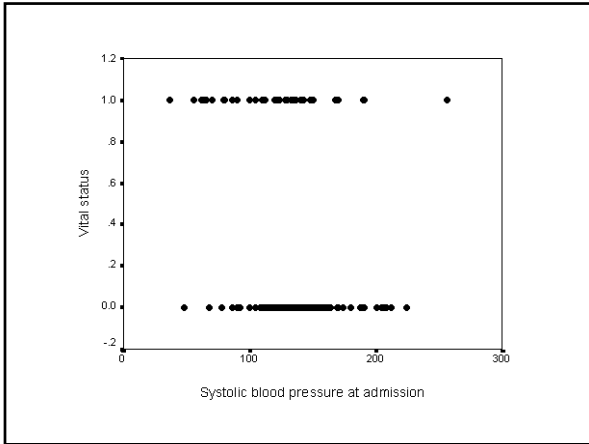
--> used to describe the relationship between

Examples of dichotomous Y variable



Smoker Group	n	Absent	Present	Proportion
0-9	2	2	0	0.0
10-19	2	2	0	0.0
20-29	3	2	1	0.33
30-39	3	2	1	0.33
40-49	3	1	2	0.66
50-59	2	0	2	1.0
60-69	3	1	2	0.67
70-79	1	0	1	1.0
80-89	1	0	1	1.0
Total				





Block 0: Beginning Block

Classification Table^a

Observed		Predicted		Percentage Correct	
		VAR00007	Constant		
Step 0	VAR00007	00	227	0	100.0
		1.00	153	0	0
Overall Percentage					59.7

a. Constant is included in the model.
b. The cut value is .500

Variables in the Equation

Step	Constant	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.395	.195	14.225	1	.000	.674

Variables not in the Equation

Step	Variables	VAR00008	Sig.	df	Sig.
Step 0	Variables	VAR00008	41.743	1	.000
Overall Statistics			41.743	1	.000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

Step	Step	Chi-Square	df	Sig.
Step 1	Block	49.120	1	.000
	Model	49.120	1	.000

Model Summary

Step	-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	463.165	.131	.164

Classification Table^a

Observed		Predicted		Percentage Correct	
		VAR00007	Constant		
Step 1	VAR00007	00	210	17	92.5
		1.00	108	45	29.4
Overall Percentage					87.1

a. The cut value is .500

Variables in the Equation

Step	Variables	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for Exp(B)	
								Lower	Upper
Step 1 ^a	Constant	-1.114	.162	47.517	1	.000	328	1.033	1.071

a. Variable(s) entered on step 1: VAR00008.