AMS 7
1 Dec 15

this time: categorical data analysis

next time: final exam review

today: LN pp. 290 + 309

①

Y	X (G, P, I)
1	P
0	I
⋮	⋮

smoking? 1 = Y, 0 = N

n = 692

sort on X, then sort on Y

count	Y	X
59	N	G
191	Y	G
57	N	P
263	Y	P
27	N	I
95	Y	I

"by" or "times" → 3×2 contingency table

both categorical → outcome (response) variable; predictor (independent) variable

	NS	S	
G	59	191	250
I	27	95	122
P	57	263	320
	143	549	692

8.53

$$\frac{23.6\% - 17.8\%}{17.8\%} \doteq \frac{6}{18} \doteq 33\%$$

increase

large in practical terms for (G) over (P)

Karl Pearson (1895)

$(\hat{O}_{ij} - \hat{E}_{ij})$

deviations

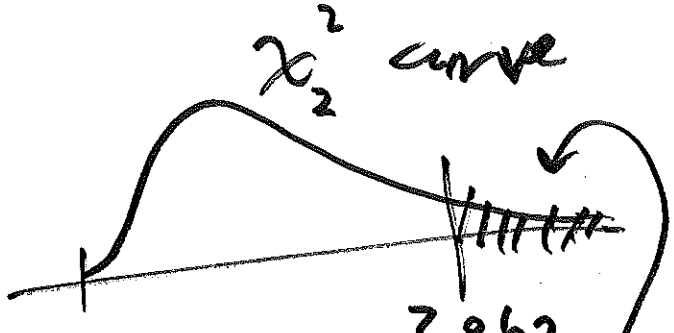
i \ j	1 ✓	2 ✗	
1	+7.3	−7.3	0
2	+1.8	−1.8	0
3	−9.1 ✗	+9.1 ✗	0
	0	0	0

$$\frac{(\hat{O}_{11} - \hat{E}_{11})^2}{\hat{E}_{11}} + \frac{(\hat{O}_{12} - \hat{E}_{12})^2}{\hat{E}_{12}} + \cdots + \frac{(\hat{O}_{32} - \hat{E}_{32})^2}{\hat{E}_{32}}$$

$$= \frac{(59 - 51.7)^2}{51.7} + \cdots + \frac{(263 - 253.9)^2}{253.9} = 3.062$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} = \chi^2$$

(# rows = I, # columns = J) 'chi-squared'



long-run hist of χ^2 if null true

p-value = 0.22

test statistic

in general: $(I-1) \cdot (J-1)$ degrees of freedom

	1	2	..	(J-1)	J
1	✓	✓	..	✓	X
2	✓	✓	..	✓	X
⋮	⋮	⋮	⋱	⋮	
I-1	✓	✓	.	✓	X
I	X	X	X	X	X