chapter 14: the chi-square test

- parametric and non-parametric tests

- chi-square test of independence is non-parametric
- used for data from two categorical variables (= nominally scaled)
- tests whether observed frequencies are significantly different from the frequencies we would expect if the two variables were independent
    - null hypothesis H_0: row variable and column variable are independent

- example using table 14.1 (gender and academic major)
- make a contingency table (= observed frequencies)
- find marginal probabilities
    - find row and column totals, and overall total (table 14.2)
    - divide row and column totals by overall total
- multiply marginal probabilities
    - if independent, P(A and B) = P(A) P (B)
    - probabilities times total count = expected frequencies (table 14.3)
        - "expected" under H_0
- calculate chi-square statistic: $chi^2 = sum_i (O_i - E_i)^2 / E_i$ (table 14.6)
    - O_i = observed value, E_i = expected value
    - similar to a squared t value
    - a measure of how different the observed values are, from what we would expect according to the null hypothesis
- degrees of freedom: df = (R-1)(C-1), where R = # rows, C = # columns
- use appendix E to find critical value of chi-square statistic
    - shows probability of getting certain chi-square values, or larger, by chance if the null hypothesis is true

- worked example: gender and academic major

		academic major				
		psych	eng	bio	total	marginal prob.
gender	girl	35	50	15	100	0.5000
	p_0	0.1625	0.1875	0.1500		
	E	32.5	37.5	30		
	boy	30	25	45	100	0.5000
	p_0	0.1625	0.1875	0.1500		
	E	32.5	37.5	30		
	total	65	75	60	200	
marginal prob.		0.3250	0.3750	0.3000		

- or, shortcut version

		academic major			
		psych	eng	bio	total
gender	girl	35	50	15	100
	E	32.5	37.5	30	
		=100*65/200	=100*75/200	=100*60/200	
	boy	30	25	45	100
	E	32.5	37.5	30	
		=100*65/200	=100*75/200	=100*60/200	
	total	65	75	60	200

- $chi^2 = (32 - 32.5)^2 / 32.5 + (50 - 37.5)^2 / 37.5 + ... = 23.72$
- df = (R-1)*(C-1) = (2-1)*(3-1) = 2
- with alpha = 0.05, $t_c = 5.99$
- statistically significant
- reject null hypothesis that gender and academic major are independent
    - i.e., it seems that in this population, the probability of having each academic major is different for the two genders

- another worked example: generational status and grade level

		generational group 3+	2	1	total	marginal prob.
gender	girl	156	215	125	496	0.5433
	p_0	0.1672	0.2523	0.1238		
	E	152.65	230.35	113.03		
	boy	125	209	83	417	0.4567
	p_0	0.1406	0.2121	0.1040		
	E	128.37	193.65	94.95		
	total	281	424	208	913	
marginal prob.		0.3078	0.4644	0.2278		

- or, shortcut version

		generational group 3+	2	1	total
gender	girl	156	215	125	496
	E	152.65	230.35	113.03	
		=496*281/913	=496*424/913	=496*208/913	
	boy	125	209	83	417
	E	128.37	193.65	94.95	
		=416*281/913	=417*424/913	=417*208/913	
	total	281	424	208	913

- chi^2 = (156 - 152.65)^2 / 152.65 + (215 - 230.35)^2 / 230.25 + ... = 5.19
- df = (R-1)*(C-1) = (2-1)*(3-1) = 2
- with alpha = 0.05, t_c = 5.99
- not statistically significant
- do not reject null hypothesis that gender and generational group are independent
    - i.e., it seems that in this population, the probability of being in different generational groups is the same for both genders
    - or, other way around, probability of being male or female is the same across generational groups