STA 250: Statistics

HW 9

Due Wed Nov 13 2013

1. A researchers surveys n college students and counts how many support, how many oppose and how many are undecided about a recently introduced federal policy. Letting X_1, X_2, X_3 denote these counts, she models $X = (X_1, X_2, X_3)$ as $X \sim \mathsf{Multinomial}(n, p)$, $p \in \Delta_3$.

   (a) Give the p-value for testing $H_0 : p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ against $p_0 \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ based on Pearson's chi-square tests for observed counts $X_1 = 140$, $X_2 = 165$, $X_3 = 195$.

   (b) The researcher wants to test whether the actual proportions of supporters and opposers in the entire college are equal. Give a mathematical formulation of this null hypothesis.

   (c) Find the restricted MLE of p under the null hypothesis in part (b) [i..e., maximize the likelihood function only over the null set].

   (d) Give the p-value for testing the null hypothesis in part (b) based on Pearson's chi-square tests, with same observed counts as in part (a).

2. A total of 309 wafer defects were recorded and the defects were classified as being one of four types, A, B, C, or D. At the same time each wafer was identified according to the production shift in which it was manufactured, 1, 2, or 3. These counts are presented in the following table.

	Type of Defect				
Shift	A	B	C	D	Total
1	15	21	45	13	94
2	26	31	34	5	96
3	33	17	49	20	119
Total	74	69	128	38	309

   Give the p-value (based on Pearson's chi-squared tests) for testing independence between type of defect and production shift.

3. The table below summarizes how the daily rain forecasts of a TV channel compared to the actual outcomes over a period of 365 days (the table entries are counts in days, total equals 365).

		Forecast	
		Rain	No-rain
Actual	Rain	95	36
	No-rain	4	230

Test the hypothesis (at 5% level) that the channel's odds of incorrect prediction is not affected by whether it's forecasting rain or no-rain.

4. DNA samples of individuals from two ethnic groups were sequenced and the genetic codes at 12 locations on chromosome 10 were analyzed for association with ethnicity. At each location, an individual may exhibit the "wild type" code (i.e., what most humans have) or have a common "alteration" (on one or both strands of the chromosome). The following tables give the two-way cross category counts of ethnicity ("E1" or "E2") and code type ("Wild" or "Alt") for the 12 locations. The sample had 506 individuals of ethnicity E1 and 494 individuals of ethnicity E2. But the column totals of the tables below may not match these numbers due to removal of some records affected by sequencing errors.

Location 1

	E1	E2
Wild	287	308
Alt	212	182

Location 2

	E1	E2
Wild	374	339
Alt	129	152

Location 3

	E1	E2
Wild	364	385
Alt	136	105

Location 4

	E1	E2
Wild	281	231
Alt	222	255

Location 5

	E1	E2
Wild	330	337
Alt	171	153

Location 6

	E1	E2
Wild	503	488
Alt	0	2

Location 7

	E1	E2
Wild	275	323
Alt	227	167

Location 8

	E1	E2
Wild	359	317
Alt	143	174

Location 9

	E1	E2
Wild	314	335
Alt	182	156

Location 10

	E1	E2
Wild	280	222
Alt	226	263

Location 11

	E1	E2
Wild	407	417
Alt	93	69

Location 12

	E1	E2
Wild	410	427
Alt	94	65

(a) For each location, calculate the p-value for H_0 : "genetic code and ethnicity are independent". Report the locations for which you will reject H_0 a 5% significance level.

(b) Is there any location for which you will be unsure of the p-value reported in part (a)? Explain.

(c) The goal of the study is to flag locations that are deemed associated with ethnicity. For this task what kind(s) of error will you be most concerned about? Fix a control level for this error rate and report which locations will be flagged under a flagging rule with error rate guaranteed to be below the chosen level. Give details and refer to known results.