S520 Homework 11

Enrique Areyan
April 20, 2012

13.4.#2: This problem can be formulated as a test of goodness-of-fit. The null hypothesis would be that the candies are distributed as claimed by the Mars company. In mathematical terms, this can be translated as:

$$\text{Let } \Pi_0 = \{(0.13, 0.14, 0.13, 0.24, 0.20, 0.26)\} \subset R^6 \text{ and } \Pi_1 \neq \{(0.13, 0.14, 0.13, 0.24, 0.20, 0.26)\} \subset R^6$$

$$H_0 : \hat{p} \in \Pi_0 \text{ vs. } H_1 : \hat{p} \in \Pi_1$$

Where each number represent the proportion of candies, by color, found in bags of M&M's in the following order: brown, yellow, red, blue, orange and green.
The total number of M&M's is 898. The following table summarizes the most important information:

	Brown	Yellow	Red	Blue	Orange	Green
$\check{e}_i$	116.74	125.72	116.74	215.52	179.6	143.68
o_i	121	84	118	226	226	123
$o_i/\check{e}_i$	1.0365	0.6682	1.0108	1.0486	1.2584	0.8561
$o_i log(o_i/\check{e}_i)$	4.3368	-33.8722	1.2668	10.7308	51.9354	-19.1147

$$\text{Thus, } G^2 = 2\sum_{i=1}^{6} o_i log(o_i/e_i) = 2 * 15.2829 = 30.5658$$

The null hypothesis has a single point and therefore, the correct degrees of freedom is 5. Testing the null hypothesis, $1 - pchisq(30.5658, df = 5) = 1.140998e^{-05}$, at any of the standard significance levels we can conclude that there is compelling evidence against the null hypothesis.

13.4.#6:

(a)

$$\bar{x} = (0{\cdot}57+1{\cdot}203+2{\cdot}383+3{\cdot}525+4{\cdot}532+5{\cdot}408+6{\cdot}273+7{\cdot}139+8{\cdot}45+9{\cdot}27+10{\cdot}10+11{\cdot}4+12{\cdot}0+13{\cdot}1+14{\cdot}1)/2608 = 3.8715$$

(b) To compute expected counts, I used in R: $dpois(Count_i, \bar{x}) * 2608$ for all except the last for which I use: $(1 - ppois(9, \bar{x})) * 2608$

E_i	Counts	o_i	e_i	$o_i \cdot log(o_i/e_i)$	Pearson's
1	0	57	54.3144	2.7509	0.1328
2	1	203	210.2810	-7.1534	0.2521
3	2	383	407.0565	-23.3312	1.4217
4	3	525	525.3131	-0.3130	0.0002
5	4	532	508.4439	24.0936	1.0914
6	5	408	393.6931	14.5638	0.5199
7	6	273	254.0337	19.6573	1.4160
8	7	139	140.5006	-1.4925	0.0160
9	8	45	67.9943	-18.5743	7.7762
10	9	27	29.2493	-2.1605	0.1730
11	$10 >$	16	17.1202	-1.0827	0.0733
			Sum	6.9579	12.8726
			G^2	13.9159	

We need only to know the number of degrees of freedom. The unrestricted dimension is 10 while the restricted dimension is 1 (only one parameter for Poisson distribution). Thus, the correct degrees of freedom is 9.
If the counts of alpha-particle scintillations follow a Poisson distribution, we obtain a $\mathbf{p} = 1 - pchisq(13.9159, df = 9) = 0.1253$ for the likelihood ratio test statistic. At any conventional alpha level we would fail to reject the null hypothesis and conclude that there is not enough evidence to dismiss the hypothesis that the data was drawn from a Poisson distribution. Interestingly, for the Pearson's test we obtain a very similar but slighter bigger result for the significance probability and thus, also fail to reject the null hypothesis.

13.4.#8: Here is the data:

$o_{ij} * log(o_{ij}/e_{ij})$

	drink	abstain
arson	0.8993	-0.8822
rape	9.2633	-8.2223
violence	15.8501	-14.1203
stealing	21.0263	-19.7866
coining	1.1376	-1.0611
fraud	-34.7143	55.8688

$(o_{ij} - e_{ij})^2/e_{ij}$

	drink	abstain
arson	0.0162	0.0181
rape	0.9760	1.0920
violence	1.6222	1.8151
stealing	1.1668	1.3055
coining	0.0719	0.0805
fraud	19.6172	21.9491

In both cases, the two cells for which the deviation is the greatest is the fraud drink/abstain.

13.4.#9: By removing the last two rows, the total n changes to 1219 and thus, the e_i and the test statistic also changes. The new degrees of freedom are $(5-1)(2-1) = 4$ and thus,

$$G^2 = 1.1228 \Longrightarrow 1 - pchisq(1.1228, df = 4) = 0.8906381$$

We obtain a significance probability of about 89%. Interestingly, using Pearson's chi-squared, we obtain a test statistic of 1.1219, very close to the likelihood ratio chi-squared statistic.
We would fail to reject the null hypothesis (by a lot) using any of the standard significance levels. The data suggest that there is no relation between crime and drinking.

14.6.#1:

(a) i. Positive pattern of association

ii. Physical space is a scare resource and thus is expensive. *Ceteris paribus*, the more space a house has, the more value. Conversely, the smaller the space the cheaper the selling prince.

(b) i. Positive pattern of association

ii. The more young a person is, the faster it is suppose to be able to run. So, younger implies less time. Conversely, the older a person the more time he will take to run 5km.

(c) i. Positive pattern of association

ii. Often it is the case that older people have higher income than younger people. If we want to be consistent with the reason provided in (b), it should be the case that, in general, more income will be associated with older people and thus a longer running time. However, if we fix the age of the men, higher income usually comes with a better preparation (better feeding habits, for instance) and should yield a lower running time and thus, a negative patter of association. Most probably, the combination of these two factors will determine the final correlation, all subject, of course, to sampling variation.

(d) i. Positive pattern of association

ii. One would assume that the better the coach, the more money he makes. A better coach should lead the team to better performance as measured by the OHSAA.

(e) i. Positive pattern of association

ii. A heavier student should run slower than a lighter student. Thus, more weight will result in more time, in seconds, to run 50 yards. The converse should be true as well.

14.6.#6:

(a) To construct a 0.95-level confidence interval, we use $q_z = 1.96$. FIrst we compute:

$$z = \frac{1}{2} log(\frac{1-0.81}{1+0.81}) = -1.12703 \Longrightarrow z \pm \frac{q_z}{\sqrt{n-3}} = -1.12703 \pm \frac{1.96}{\sqrt{69}} \Longrightarrow (-1.3629, -0.8911)$$

(b) The test statistic is

$$\frac{r * \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.81 * \sqrt{72-2}}{\sqrt{1-(-0.81)^2}} = \frac{-0.81 * 8.3666}{0.5864} = -11.5569$$

and the significance probability is

$$\mathbf{p} = 2 * pt(-11.5569, 70) = 6.922943e^{-18}$$

This is very strong evidence that $\rho \neq 0$

14.6.#9:

(a) Looking at the diagram, it seems that the concentration ellipse is a fair summary of the data. Except for maybe one outlier at the very top of the diagram, it is plausible that this data was drawn from a bivariate normal distribution.

(b) Using binorm.estimate, we obtain $r = 0.2182$ and by using kappa we obtain $\hat{k} = 0.5412245$. Thus, $\hat{\tau} = 2 * \hat{k} - 1 = 2 * 0.5412245 - 1 = 0.082449$

(c) The test statistic is

$$\frac{r * \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.2182 * \sqrt{50-2}}{\sqrt{1-(0.2182)^2}} = \frac{0.2182 * 6.9282}{0.9524} = 1.5873$$

and the significance probability is

$$\mathbf{p} = 2 * pt(1.5873, 48) = 1.872064$$

We fail to reject the hypothesis $\rho = 0$

(d) $kappa.p.sim(Data, 20000) = 0.3985$. Because $\mathbf{p} > \alpha = 0.05$, we fail to reject the null hypothesis of no monotonic association.