MATH 425 11/11/2011

Note Title 11/11/2011

The Pearson χ^2 test. We draw N balls with return. The balls are of k colors color i occurs with probability p_i. We got a_i balls of color i.

$$\chi^2 = \frac{(a_1 - p_1 N)^2}{p_1 N} + \cdots + \frac{(a_k - p_k N)^2}{p_k N}.$$

Claim: This has the distribution of a sum of squares of $(k-1)$ independent standard Gaussian variables.

To prove the claim, note that we have a sum of k random variables, but they are not independent. (For example, if $k=2$, a_2 is determined by a_1.)

How do we deal with sums of ℓ random variables which are not independent? How to quantify the lack of independence. Recall we observed that if X and Y are independent then

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

We define the covariance of X and Y as

$$\operatorname{cov}(X,Y) = E(X\cdot Y) - E(X)\cdot E(Y).$$

Note that

$$\operatorname{var}(X) = \operatorname{cov}(X,X).$$

This is kind of analogous to the dot product of vectors: $\vec{x}=(x_1,\dots,x_n)\in\mathbb{R}^n$, $\vec{y}=(y_1,\dots,y_n)\in\mathbb{R}^n$

$$\vec{x}\cdot\vec{y} = x_1y_1 + \cdots + x_ny_n.$$

$$\|\vec{x}\|^2 = \vec{x}\cdot\vec{x}$$

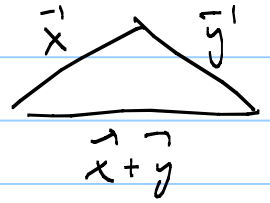
$$\|\vec{x}+\vec{y}\|^2 - \|\vec{x}\|^2 - \|\vec{y}\|^2 = (\vec{x}+\vec{y})\cdot(\vec{x}+\vec{y}) - \vec{x}\cdot\vec{x} - \vec{y}\cdot\vec{y}$$
$$= 2\vec{x}\cdot\vec{y}$$

Similarly in probability,

$$(*) \qquad \operatorname{cov}(X,Y) = \frac{1}{2}\left(\operatorname{var}(X+Y) - \operatorname{var}(X) - \operatorname{var}(Y)\right)$$

What is the angle α between two vectors $\vec{x}, \vec{y}$?

$$\cos\alpha = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|} \qquad \text{(theorem of cosines)}$$



In probability, this is called the correlation coefficient

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma(X)\,\sigma(Y)} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

In relevance to the χ^2, we want to look

at the variables

$$\frac{a_1 - p_1 N}{\sqrt{p_1 N}} \quad , \quad \dots \quad , \quad \frac{a_k - p_k N}{\sqrt{p_k N}} .$$

I want to compute all the variances and covariances of these variables.

a_i are binomial variables $X_{p_i, N}$

$$\operatorname{var}(a_i) = N p_i (1 - p_i)$$

$$\operatorname{var}\left(\frac{a_i - p_i N}{\sqrt{p_i N}} \right) = \frac{N p_i (1 - p_i)}{N p_i} = 1 - p_i$$

$i \neq j$ $\quad \mathrm{Cov}\left(\frac{a_i - p_i N}{\sqrt{p_i N}} , \frac{a_j - p_j N}{\sqrt{p_j N}} \right) = \frac{1}{\sqrt{p_i N}\sqrt{p_j N}} \mathrm{cov}(a_i, a_j)$

$= \frac{1}{N\sqrt{p_i p_j}} \mathrm{Cov}(a_i, a_j)$

probability
binomial $\widetilde{p_i + p_j}$, N trials

$\mathrm{cov}(a_i, a_j) = \frac{1}{2}\left(\mathrm{var}(a_i + a_j) - \mathrm{var}(a_i) - \mathrm{var}(a_j) \right)$

$= \frac{1}{2}\Big((p_i + p_j)(1 - p_i - p_j) N - p_i(1 - p_i) N$
$\qquad - p_j(1 - p_j) N \Big)$

$= -p_i p_j N$

$$\operatorname{cov}\left(\frac{a_i - p_i N}{\sqrt{p_i N}}, \frac{a_j - p_j N}{\sqrt{p_j N}}\right) = -\sqrt{p_i p_j}$$

$i \neq j$

$$\operatorname{cov}\left(\frac{a_i - p_i N}{\sqrt{p_i N}}, \frac{a_i - p_i N}{\sqrt{p_i N}}\right) = \operatorname{var}\left(\frac{a_i - p_i N}{\sqrt{p_i N}}\right) = 1 - p_i$$

If we have n vectors $\vec{v}_1, \dots, \vec{v}_n$, it is useful to form the matrix

$$\begin{pmatrix} \vec{v}_1 \cdot \vec{v}_1 & \cdots & \vec{v}_1 \cdot \vec{v}_n \\ \vdots & & \\ \vec{v}_n \cdot \vec{v}_1 & \cdots & \vec{v}_n \cdot \vec{v}_n \end{pmatrix}$$

(Gramm matrix)

In our case, let us make the matrix of covariances of $a_1, \dots, a_k$: $\begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{pmatrix}$

$$\begin{pmatrix} 1-p_1 & -\sqrt{p_1 p_2} & \cdots & -\sqrt{p_1 p_k} \\ -\sqrt{p_2 p_1} & 1-p_2 & \cdots & -\sqrt{p_2 p_k} \\ \vdots & & & \\ -\sqrt{p_k p_1} & -\sqrt{p_k p_2} & \cdots & 1-p_k \end{pmatrix} = I - \left(\sqrt{p_i p_j}\right)_{i,j} =$$

$$= I - \underbrace{\left(\sqrt{p_1}, \dots, \sqrt{p_k}\right)}_{1 \times k} \cdot \underbrace{\begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_k} \end{pmatrix}}_{k \times 1}$$

matrix product

$k \times k$

$$\| (\sqrt{p_1}, \dots, \sqrt{p_k}) \| = \sqrt{\sqrt{p_1}^2 + \dots + \sqrt{p_k}^2} = \sqrt{p_1 + \dots + p_k} = \sqrt{1} = 1$$

$v = (\sqrt{p_1}, \dots, \sqrt{p_k})$. I can change coordinates in a way preserving the dot product such that v becomes $(1, 0, \dots, 0)$.

So the in those coordinates becomes

$$I - (1,0,\dots,0)^T \cdot (1,0,\dots 0) = I - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & & 0 \\ & 1 & \\ & & \ddots \\ 0 & & & 1 \end{pmatrix}$$ $k-1$ here are my independent uncorrelated standard Gaussians

HW: Find a X^2 table on the internet.

(1) Suppose a standard die is cast 20 times, we get

3	1's
7	2's
2	3's
2	4's
2	5's
4	6's.

Can you conclude the die is biased with significance level 95% ?

(2) If we keep casting the die how many 6's in a row (from the first trial)

do we have to throw to conclude the die is biased with significance level 95%?

[Hint: degrees of freedoms can be 6 vs. not a 6]