Lecture 23

23.1 Pearson's theorem.

Today we will prove one result from probability that will be useful in several statistical tests.

Let us consider r boxes $B_1, \ldots, B_r$ as in figure 23.1

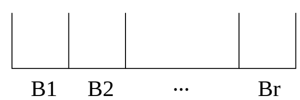

Figure 23.1:

Assume that we throw n balls $X_1, \ldots, X_n$ into these boxes randomly independently of each other with probabilities

$$\mathbb{P}(X_i \in B_1) = p_1, \ldots, \mathbb{P}(X_i \in B_r) = p_r,$$

where probabilities add up to one $p_1 + \ldots + p_r = 1$. Let ν_j be a number of balls in the jth box:

$$\nu_j = \#\{\text{balls } X_1, \ldots, X_n \text{ in the box } B_j\} = \sum_{l=1}^{n} I(X_l \in B_j).$$

On average, the number of balls in the jth box will be np_j, so random variable ν_j should be close to np_j. One can also use Central Limit Theorem to describe how close ν_j is to np_j. The next result tells us how we can describe in some sense the closeness of ν_j to np_j simultaneously for all $j \le r$. The main difficulty in this Thorem comes from the fact that random variables ν_j for $j \le r$ are not independent, for example, because the total number of balls is equal to n,

$$\nu_1 + \ldots + \nu_r = n,$$

i.e. if we know these numbers in $n-1$ boxes we will automatically know their number in the last box.

Theorem. *We have that the random variable*

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \to \chi^2_{r-1}$$

converges in distribution to χ^2_{r-1} distribution with $(r-1)$ degrees of freedom.

Proof. Let us fix a box B_j. The random variables

$$I(X_1 \in B_j), \ldots, I(X_n \in B_j)$$

that indicate whether each observation X_i is in the box B_j or not are i.i.d. with Bernoully distribution $B(p_j)$ with probability of success

$$\mathbb{E}I(X_1 \in B_j) = \mathbb{P}(X_1 \in B_j) = p_j$$

and variance

$$\mathrm{Var}(I(X_1 \in B_j)) = p_j(1-p_j).$$

Therefore, by Central Limit Theorem we know that the random variable

$$\begin{aligned} \frac{\nu_j - np_j}{\sqrt{np_j(1-p_j)}} &= \frac{\sum_{l=1}^n I(X_l \in B_j) - np_j}{\sqrt{np_j(1-p_j)}} \\ &= \frac{\sum_{l=1}^n I(X_l \in B_j) - n\mathbb{E}}{\sqrt{n\mathrm{Var}}} \to N(0,1) \end{aligned}$$

converges to standard normal distribution. Therefore, the random variable

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to \sqrt{1-p_j}N(0,1) = N(0, 1-p_j)$$

converges to normal distribution with variance $1-p_j$. Let us be a little informal and simply say that

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to Z_j$$

where random variable $Z_j \sim N(0, 1-p_j)$.

We know that each Z_j has distribution $N(0, 1-p_j)$ but, unfortunately, this does not tell us what the distribution of the sum $\sum Z_j^2$ will be, because as we mentioned above r.v.s ν_j are not independent and their correlation structure will play an important role. To compute the covariance between Z_i and Z_j let us first compute the covariance between

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \text{ and } \frac{\nu_j - np_j}{\sqrt{np_j}}$$

which is equal to

$$\mathbb{E}\frac{\nu_i - np_j}{\sqrt{np_i}}\frac{\nu_j - np_j}{\sqrt{np_j}} = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - \mathbb{E}\nu_i np_j - \mathbb{E}\nu_j np_i + n^2p_ip_j)$$
$$= \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - np_inp_j - np_jnp_i + n^2p_ip_j) = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - n^2p_ip_j).$$

To compute $\mathbb{E}\nu_i\nu_j$ we will use the fact that one ball cannot be inside two different boxes simultaneously which means that

$$I(X_l \in B_i)I(X_l \in B_j) = 0. \tag{23.1}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\nu_i\nu_j &= \mathbb{E}\Big(\sum_{l=1}^n I(X_l \in B_i)\Big)\Big(\sum_{l'=1}^n I(X_{l'} \in B_j)\Big) = \mathbb{E}\sum_{l,l'} I(X_l \in B_i)I(X_{l'} \in B_j)\\
&= \underbrace{\mathbb{E}\sum_{l=l'} I(X_l \in B_i)I(X_{l'} \in B_j)}_{\text{this equals to 0 by (23.1)}} + \mathbb{E}\sum_{l\neq l'} I(X_l \in B_i)I(X_{l'} \in B_j)\\
&= n(n-1)\mathbb{E}I(X_l \in B_j)\mathbb{E}I(X_{l'} \in B_j) = n(n-1)p_ip_j.
\end{aligned}$$

Therefore, the covariance above is equal to

$$\frac{1}{n\sqrt{p_ip_j}}\Big(n(n-1)p_ip_j - n^2p_ip_j\Big) = -\sqrt{p_ip_j}.$$

To summarize, we showed that the random variable

$$\sum_{j=1}^r \frac{(\nu_j - np_j)^2}{np_j} \to \sum_{j=1}^r Z_j^2.$$

where random variables $Z_1, \ldots, Z_n$ satisfy

$$\mathbb{E}Z_i^2 = 1 - p_i \text{ and covariance } \mathbb{E}Z_iZ_j = -\sqrt{p_ip_j}.$$

To prove the Theorem it remains to show that this covariance structure of the sequence of Z_i's will imply that their sum of squares has distribution χ^2_{r-1}. To show this we will find a different representation for $\sum Z_i^2$.

Let $g_1, \cdots, g_r$ be i.i.d. standard normal sequence. Consider two vectors

$$\vec{g} = (g_1, \ldots, g_r) \text{ and } \vec{p} = (\sqrt{p_1}, \ldots, \sqrt{p_r})$$

and consider a vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$, where $\vec{g}\cdot\vec{p}=g_1\sqrt{p_1}+\ldots+g_r\sqrt{p_r}$ is a scalar product of $\vec{g}$ and $\vec{p}$. We will first prove that

$$\vec{g}-(\vec{g}\cdot\vec{p})\vec{p} \text{ has the same joint distribution as } (Z_1,\ldots,Z_r). \tag{23.2}$$

To show this let us consider two coordinates of the vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$:

$$i^{th}: g_i-\sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_i} \quad \text{and } j^{th}: g_j-\sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_j}$$

and compute their covariance:

$$\begin{aligned}&\mathbb{E}\Big(g_i-\sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_i}\Big)\Big(g_j-\sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_j}\Big)\\ &=-\sqrt{p_i}\sqrt{p_j}-\sqrt{p_j}\sqrt{p_i}+\sum_{l=1}^n p_l\sqrt{p_i}\sqrt{p_j}=-2\sqrt{p_ip_j}+\sqrt{p_ip_j}=-\sqrt{p_ip_j}.\end{aligned}$$

Similarly, it is easy to compute that

$$\mathbb{E}\Big(g_i-\sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_i}\Big)^2=1-p_i.$$

This proves (23.2), which provides us with another way to formulate the convergence, namely, we have

$$\sum_{j=1}^r\Big(\frac{\nu_j-np_j}{\sqrt{np_j}}\Big)^2\to\sum_{i=1}^r(i^{th}\text{ coordinate})^2$$

where we consider the coorinates of the vector $\vec{g}-(\vec{g}\cdot\vec{p})\vec{p}$. But this vector has a simple geometric interpretation. Since vector $\vec{p}$ is a unit vector:

$$|\vec{p}|^2=\sum_{l=1}^r(\sqrt{p_i})^2=\sum_{l=1}^r p_i=1,$$

vector $\vec{V}_1=(\vec{p}\cdot\vec{g})\vec{p}$ is the projection of vector $\vec{g}$ on the line along $\vec{p}$ and, therefore, vector $\vec{V}_2=\vec{g}-(\vec{p}\cdot\vec{g})\vec{p}$ will be the projection of $\vec{g}$ onto the plane orthogonal to $\vec{p}$, as shown in figures 23.2 and 23.3.

Let us consider a new orthonormal coordinate system with the last basis vector (last axis) equal to $\vec{p}$. In this new coordinate system vector $\vec{g}$ will have coordinates

$$\vec{g}'=(g_1',\ldots,g_r')=\vec{g}V$$

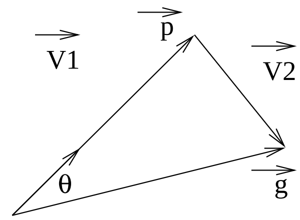

Figure 23.2: Projections of $\vec{g}$.

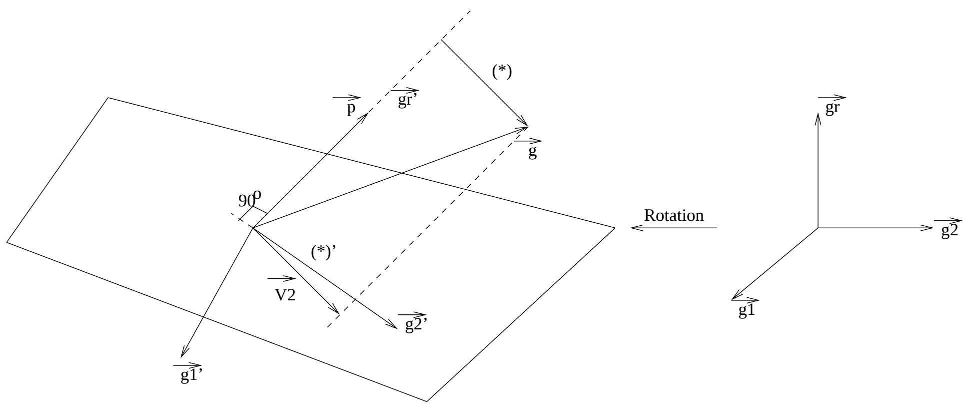

Figure 23.3: Rotation of the coordinate system.

obtained from $\vec{g}$ by orthogonal transformation V that maps canonical basis into this new basis. But we proved a few lectures ago that in that case $g_1', \ldots, g_r'$ will also be i.i.d. standard normal. From figure 23.3 it is obvious that vector $\vec{V_2} = \vec{g} - (\vec{p} \cdot \vec{g})\vec{p}$ in the new coordinate system has coordinates

$$(g_1', \ldots, g_{r-1}', 0)$$

and, therefore,

$$\sum_{i=1}^{r} (i^{th} \text{ coordinate})^2 = (g_1')^2 + \ldots + (g_{r-1}')^2.$$

But this last sum, by definition, has χ^2_{r-1} distribution since $g_1', \cdots, g_{r-1}'$ are i.i.d. standard normal. This finishes the proof of Theorem.

$\square$