MATHEMATICS

# Find the covariances of a multinomial distribution

Asked 4 years, 1 month ago Active 1 year, 11 months ago Viewed 8k times

8

5

If $(X_1, \cdots, X_n)$ is a vector with multinomial distribution, proof that $\text{Cov}(X_i, X_j) = -rp_i p_j$, $i \neq j$ where $r$ is the number of trials of the experiment, $p_i$ is the probability of success for the variable $X_i$.

$$fdp = f(x_1, \dots x_n) = \frac{r!}{x_1! x_2! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

if $x_1 + x_2 + \cdots + x_n = r$

I'm trying to use the property: $\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$ and find that $E[X_i] = rp_i$, but I don´t know the efficient way to calculate $E[X_i X_j]$.

probability combinatorics statistics

edited Feb 24 '16 at 1:46

asked Feb 24 '16 at 1:18

User 2014
**636** 5 16

What is $E(X_i X_j)$ for $r = 1$? – A.S. Feb 24 '16 at 1:37

1 As what A.S. hinted, one common trick is to express $X_i = \sum_{k=1}^{r} Y_{i,k}, X_j = \sum_{l=1}^{r} Y_{j,l}$ and use linearity of covariance. By independence across different multinomial trials, you only left the calculate the case with $Cov[Y_{i,k}, Y_{j,k}]$. But those $Y$ are indicators only (i.e. the $r = 1$ case mentioned by A.S.) which is easy to calculate. – BGM Feb 24 '16 at 3:43

## 1 Answer

Active Oldest Votes

13

We can use indicator random variables to help simplify the covariance expression. We can interpret the problem as $r$ independent rolls of an $n$ sided die. Let $X_i$ be the number of rolls that result in side $i$ facing up, and let $I_k^{(i)}$ be an indicator equal to 1 when roll $k$ is equal to $i$ and 0 otherwise. Then, we can express $X_i$ and $X_j$ as follows:

$$X_i = \sum_{k=1}^{r} I_k^{(i)} \quad \text{and} \quad X_j = \sum_{k=1}^{r} I_k^{(j)}$$

Let's re-write the covariance using indicators:

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$$

Let's compute the first term:

$$E[X_i X_j] = E\Big[(\sum_{k=1}^{r} I_k^{(i)})(\sum_{l=1}^{r} I_l^{(j)})\Big] = \sum_{k=l} E[I_k^{(i)} I_l^{(j)}] + \sum_{k \neq l} E[I_k^{(i)} I_l^{(j)}] =$$
$$= 0 + \sum_{k \neq l} E[I_k^{(i)}]E[I_l^{(j)}] = \sum_{k \neq l} p_i p_j = (r^2 - r) p_i p_j$$

where we expanded the product of sums, used linearity of expectation and the fact that when $k = l$ we can't simultaneously roll $i$ and $j$ on the same trial $k = l$ (making the product of indicators zero), finally we applied independence of rolls that enabled us to write it as a product of probabilities. Let's compute the remaining term:

$$E[X_i] = E[\sum_{k=1}^{r} I_k^{(i)}] = \sum_{k=1}^{r} E[I_k^{(i)}] = r p_i$$

Therefore, the covariance equals:

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = (r^2 - r) p_i p_j - r^2 p_i p_j = -r p_i p_j$$

Notice that $\text{Cov}(X_i, X_j) = -r p_i p_j < 0$ is negative, this makes sense intuitively since for a fixed number of rolls $r$, if we roll many outcomes $i$, this reduces the number of possible outcomes $j$, and therefore $X_i$ and $X_j$ are negatively correlated!

answered Apr 3 '18 at 21:41

Vadim Smolyakov

394 2 8