Stat 2911 Lecture Notes

Class 9, 2017

Uri Keich

Covariance, Variance of a sum of RVs, Independence and 0 covariance, Variance of the binomial, hypergeometric, and negative binomial distributions (Rice 4.3)

$L^1 = \{ X : \sum_k |x_k| p_X(x_k) < \infty \}$ $\quad [X \in L^1 \iff E|X| < \infty]$

L^1 is the vector space of RVs with finite mean

$L^2 = \{ X : \sum_k x_k^2 p_X(x_k) < \infty \}$ $\quad [X \in L^2 \iff EX^2 < \infty \iff X^2 \in L^1]$

Generally, $L^2 \subsetneq L^1$, where

L^2 is the vector space of RVs with finite variance.

Generally, $X, Y \in L^1 \not\Rightarrow X \cdot Y \in L^1$

Claim. Suppose $X, Y \in L^1$ are ind then $XY \in L^1$

and $E(XY) = E(X)E(Y)$

Claim. If $X, Y \in L^2$ then $XY \in L^1$

Claim. If $X \in L^2$ $\alpha, \beta \in \mathbb{R}$ then

$$V(\alpha X + \beta) = \alpha^2 \sigma^2$$

Want $V(X+Y)$

Def. For jointly distributed RVs $X, Y \in L^1$ we define the covariance of X and Y as

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right],$$

provided $(X-\mu_x)(Y-\mu_y) \in L^1$ and where

$$\mu_x = E(X), \quad \mu_y = E(Y).$$

Note the symmetry: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

When does the covariance exist?

Claim. If $X, Y, X\cdot Y \in L^1$ then $\mathrm{Cov}(X,Y)$ exists and

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

Cor. If $X, Y \in L^2$ then the conclusion of the claim holds.

Proof. $(X-\mu_x)(Y-\mu_y) = XY - \mu_x Y - \mu_y X + \mu_x\mu_y$

Since $XY, X, Y, 1$ (constant) $\in L^1$ the RHS $\in L^1$

$\Rightarrow (X-\mu_x)(Y-\mu_y) \in L^1$ and its expected value is

$$\mathrm{Cov}(X,Y) = E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x\mu_y$$

$$= E(XY) - \mu_x\mu_y$$

□

Cor. If $X, Y \in L^1$ are ind. then

$$\mathrm{Cov}(X, Y) =$$

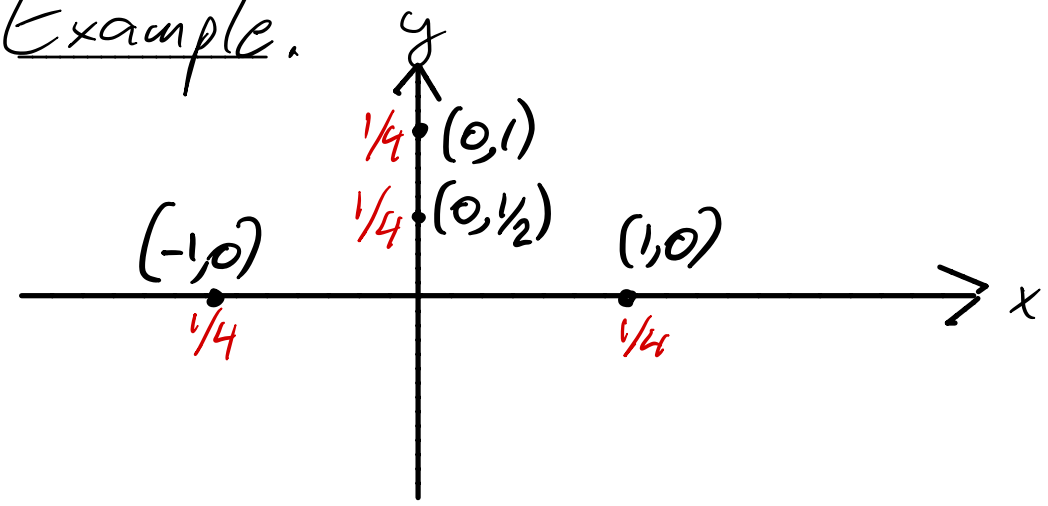
Or, independent RVs are <u>uncorrelated</u>

(having 0 covariance)

What about the converse?

Covariance is a measurement of the <u>linear</u> dependence between RVs, which is substantially more restricted than probabilistic dependence.

Example.



Can't predict Y using a linear function of X $\not\Rightarrow$ can't predict Y from X

The joint pmf p_{XY} assigns a mass of ¼ for the four depicted points (and o elsewhere)

Check: $\mathrm{Cov}(X, Y) = 0$ but X and Y are <u>not</u> ind.

$E(X+Y) = E(X)+E(Y)$, is $V(X+Y) = V(X)+V(Y)$?

Clearly we need $X, Y \in L^2$ for the RHS.

$\Rightarrow X+Y \in L^2$ so $V(X+Y)$ exists and is finite.

If $X = -Y$ then $V(X+Y) = V(X-X) =$

$V(X+Y)$ can be smaller than $V(X)+V(Y)$

If $X = Y$ then $V(X+Y) = V(2X) =$

$V(X+Y)$ can be larger than $V(X)+V(Y)$

Claim. If $X, Y \in L^2$ then

$$V(X+Y) = V(X)+V(Y)+2\mathrm{Cov}(X,Y)$$

Proof. Note that $\mathrm{Cov}(X,Y)$ is well defined.

$$E(X+Y)^2 = E(X^2+2XY+Y^2) = E(X^2)+2E(XY)+E(Y^2)$$

$$E^2(X+Y) = [E(X)+E(Y)]^2 = E^2(X)+2E(X)E(Y)+E^2(Y)$$

subtract

$$V(X+Y) = V(X)+V(Y)+2\mathrm{Cov}(X,Y)$$

$$= V(X)+V(Y)+\mathrm{Cov}(X,Y)+\mathrm{Cov}(Y,X)$$

By induction, if $X_1, \dots, X_n \in L^2$ then

$$V\left(\sum_1^n X_i\right) = \sum_1^n V(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)$$

For the proof you will need the linearity:

$$\operatorname{Cov}(X+Y, Z) = \operatorname{Cov}(X, Z) + \operatorname{Cov}(Y, Z)$$

Cor. If X_i are <u>ind.</u> L^2-RVs then

$$V\left(\sum_1^n X_i\right) = \sum_1^n V(X_i)$$

<u>Examples</u>.

1) X_i are iid Bernoulli(p) RVs

Let $S_n' = \sum_1^n X_i$ then

$$V(S_n') = \sum_1^n V(X_i)$$

$$= \sum_1^n p(1-p)$$

$$= np(1-p)$$

Note that $S_n' \sim$

2) Let X = # of red balls in a sample of m balls from an urn with r red balls and $n-r$ black balls. Then $X \sim hyper(n, r, m)$. What are $E(X)$, $V(X)$? We can compute those directly from the definitions, using the pmf. Alternatively:

For $i = 1, 2, \ldots, m$, let $X_i = 1\{i^{th}$ drawn ball is red$\}$.

Clearly, $X_i \sim Bernoulli(p_i)$, where

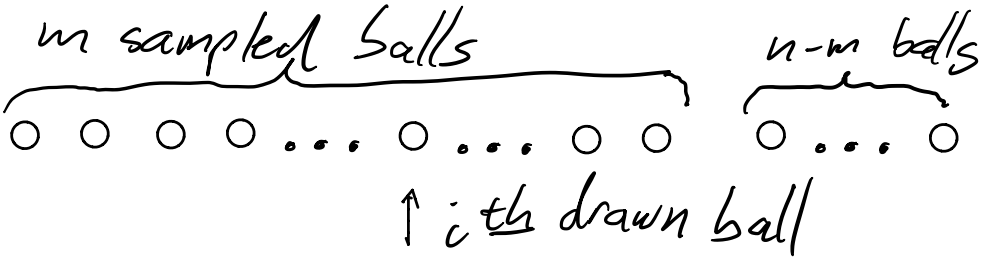
$p_1 =$

$p_2 = P(X_2 = 1 | X_1 = 1) P(X_1 = 1) + P(X_2 = 1 | X_1 = 0) P(X_1 = 0)$

$= \frac{r-1}{n-1} \cdot \frac{r}{n} + \frac{r}{n-1} \cdot \frac{n-r}{n} = \ldots =$

$p_3 = \ldots =$ why?

The sampling is consistent with randomly sorting the n balls in a row and picking the first m balls.



$\Rightarrow p_i = r/n$ symmetry

Clearly, $X =$

$\Rightarrow E(X) = \sum_1^m E(X_i) =$

Are X_i ind.?

The X_i are not ind.: $X_i = 1$ implies that the proportion of remaining red balls has decreased

$$V(\bar{X}) = \sum_{1}^{m} V(X_i) + \sum_{i \neq j}' Cov(X_i, X_j)$$

$$V(X_i) = r/n \left(1 - r/n\right)$$

$$Cov(X_i, X_j) = E(X_i X_j) - \underbrace{E(X_i)E(X_j)}_{(r/n)^2}$$

$$E(X_i X_j) = 1 \cdot P(X_i = 1, X_j = 1)$$

$$= \frac{r(r-1)}{n(n-1)} \qquad \forall\, i \neq j$$

$$\forall\, i \neq j \Rightarrow Cov(X_i, X_j) = \frac{r(r-1)}{n(n-1)} - \frac{r^2}{n^2} = \frac{r}{n}\left(\frac{r-1}{n-1} - \frac{r}{n}\right)$$

$$= -\frac{r}{n}\left(1 - \frac{r}{n}\right)\frac{1}{n-1}$$ (does the sign make sense?)

$$\Rightarrow V(\bar{X}) = m\frac{r}{n}\left(1 - \frac{r}{n}\right) + m(m-1)\left(-\frac{r}{n}\right)\left(1 - \frac{r}{n}\right)\frac{1}{n-1}$$

$$= m\frac{r}{n}\left(1 - \frac{r}{n}\right)\left(1 - \frac{m-1}{n-1}\right)$$

It is instructive to compare this expression to sampling with replacement.

Let Y = # of red balls in a sample of m balls taken with replacement where $m \le n$.

Clearly, $Y \lor$

Therefore,

$$E(Y) = \qquad E(\overline{X})$$

$$V(Y) = \qquad m\frac{r}{n}\left(1-\frac{r}{n}\right)\left(1-\frac{m-1}{n-1}\right) = V(\overline{X})$$

and $>$ if $m > 1$

Does it make sense?

$m=1$: $V(\overline{X}) = V(Y) = \frac{r}{n}\left(1-\frac{r}{n}\right)$

as both $\overline{X}$ and Y are Bernoulli(r/n) RVs

$m=n$: $V(\overline{X}) =$

Fix m and let $n \to \infty$ s.t. $r_n/n \xrightarrow[n]{} p$, then

$$V(\overline{X}_n) = m\frac{r_n}{n}\left(1-\frac{r_n}{n}\right)\left(1-\frac{m-1}{n-1}\right) \xrightarrow[n]{} mp(1-p)$$

$$V(Y_n) = m\frac{r_n}{n}\left(1-\frac{r_n}{n}\right) \xrightarrow[n]{} mp(1-p)$$

Makes sense?

3) $N \sim$ negative binomial (r, p) [NB(r,p)]

Claim. If X_i are iid geometric(p) RVs then

$$\sum_1^r X_i \sim NB(r,p).$$

Proof. Algebraic, using convolutions (problem set).

Alternatively, consider X_i as counting the number of additional trials to the i^{th} success.

By definition,

$$N = \sum_1^r X_i \sim NB(r,p).$$

At the same time it is intuitively clear that the X_i are iid geometric(p) RVs. □

It follows that

$$E(N) = \sum_1^r E(X_i)$$

$$= r/p.$$

$$V(N) = \sum_1^r V(X_i)$$

$$= \frac{r(1-p)}{p^2}.$$

Alternatively, compute using the NB(r,p) part.