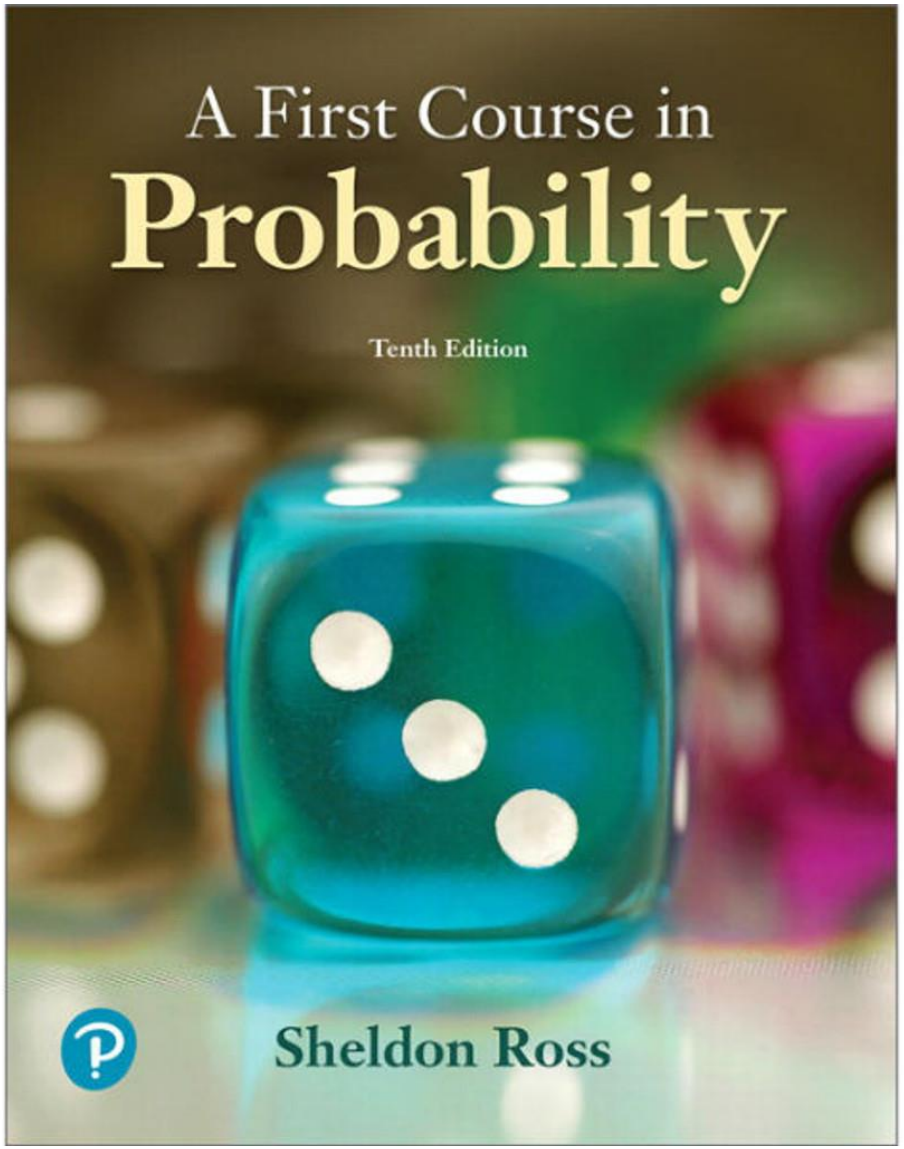
Lecture 14.2

Covariance

PURDUE UNIVERSITY | Department of Mathematics



Today's reading: 7.4

Next class: 7.7

HW11 is your LAST homework assignment! Available now, due 4/25

Course evaluations are now open.
Please do one!

Today's draft problem A

To be presented by today's A draftee.

Consider n independent flips of a coin having probability p of landing on heads. Say that a changeover occurs whenever an outcome differs from the one preceding it. (For instance, if $n = 5$ and the outcome is $HHTHT$, then there are 3 changeovers.) Find the expected number of changeovers.

Hint: Express the number of changeovers as the sum of $n - 1$ Bernoulli random variables.

Today's draft problem B

To be presented by today's B draftee.

Let $ABCD$ be the unit square where $A = (0,0), B = (1,0), C = (1,1), D = (0,1)$. Let $\alpha, \beta, \gamma, \delta$ be uniformly distributed on the intervals AB, BC, CD, DA. Let S be the area of the quadrilateral $\alpha\beta\gamma\delta$. Find $E[S]$.

Hint: $S = \frac{\det(\gamma-\alpha, \delta-\beta)}{2}$

Covariance - definition

Definition

If X and Y are random variables, then their *covariance* $\mathrm{Cov}(X, Y)$ is defined to be

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

Equivalently (I find this more useful most of the time):

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

Why are these equivalent?

Hint: a three word phrase that begins with an L...

Covariance – why?

Key fact

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Intuition: the covariance is a quantitative measure of how much two random variables fail to be independent.

BUT BEWARE: *dependent* random variables can *sometimes* have 0 covariance.

For example: let X be uniformly distributed on the finite set $\{-1,0,1\}$ and let Y be the indicator for the event $X = 0$.

Covariance – useful formal properties

Proposition 4.2

i. $Cov(X,Y) = Cov(Y,X)$
ii. $Cov(X,X) = Var(X)$
iii. $Cov(aX,Y) = a\,Cov(X,Y)$
iv. $Cov\left(\sum_{i=0}^{n} X_i, \sum_{j=0}^{m} Y_j\right) = \sum_{i=0}^{n}\sum_{j=0}^{m} Cov(X_i, Y_j)$

To put (i), (iii) and (iv) succinctly: covariance is a symmetric, bilinear operation on pairs of random variables (kind of like an inner product!).

Let's prove (i), (iii) and (iv). (ii) is an easy exercise you should be able to do.

Variance of sums

Using Proposition 4.2, we can prove the following useful identity for variance of sums

If $X_1, X_2, \ldots, X_n$ are random variables then

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j}\sum Cov(X_i, X_j)$$

To prove it, we use (ii) and (iv) from the proposition.

Example 4a

Let $X_1, X_2, \ldots, X_n$ be iid with mean μ and variance σ^2. If $\bar{X}$ is the sample mean, then the random variables $X_i - \bar{X}, \quad i = 1,2, \ldots, n$ are called the *deviations* and the random variable

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

is called the *sample variance*.

Find $Var(\bar{X})$ and $E[S^2]$.

(Hint: the latter's answer will "explain" why we divide by $n-1$ and not n.)

Example 4c – population sampling

Suppose there are N people (where N is large) and each has an unknown "preference" for a presidential candidate that can be represented by a real number v_i (for example, maybe each v_i equals either 0 or 1, where $v_i = 1$ means the person will vote for the candidate, and $v_i = 0$ means they will not). A polling firm has enough resources to poll $n < N$ to learn what each of their preferences is. Assume that when they randomly select the n people to poll, any of the $\binom{N}{n}$ possibilities are equally likely. If T is the sum total of all of the polled preferences, determine $E[T]$ and $Var(T)$.

(Note: in the case each $v_i = 0 \text{ or } 1$, then T/n could be used to estimate the true fraction of all the N people who will vote for the candidate. This is something political campaigns and journalists pay good money to actually implement!)

Friday's draft problem A

To be presented by Friday's A draftee.

9. A fair coin is tossed twice independently. Let X and Y be the indicator random variables of the events that "H" appear in the 1st and 2nd toss respectively.

(a) (5pts) Compute the covariance of $X+Y$ and $X-Y$ (simplify).

(b) (5pts) Are $X+Y$ and $X-Y$ independent? Justify your answer clearly.

Friday's draft problem B

To be presented by Friday's B draftee.

Suppose X and Y are jointly continuously distributed with PDF

$$f(x,y) = \begin{cases} \dfrac{2e^{-2x}}{x}, & 0 \le x < \infty, 0 \le y \le x \\ 0, & else \end{cases}$$

Compute $Cov(X,Y)$.