STAT 390 A
Statistical Methods in Engineering and Science

Week 5 Lectures – Part 2 – Spring 2023

Joint Distributions, Covariance, and Correlation of Random Variables

Alexander Giessing

Department of Statistics

University of Washington

April 28, 2023

Outline

Example: Calculating the Volume of Hand-Blown Vases

- Consider a particularly simple cylindrical model of a hand-blown glass vase of height H and radius R (in cm).
- Since the vase is hand-blown H and R are not constant but random variables.
- *Since the volume $V = \pi H R^2$ is random, what is* $\mathrm{E}[V]$*?*

  (Why would one care absout this ...? *Answer:* Logistics, shipping, packaging, etc.)
- **Naive approach:** If we had the density f_V we could compute

$$\mathrm{E}[V] = \int_{-\infty}^{\infty} v f_V(v) dv.$$

- **Cleverer approach:** Obtain joint ddensity f of H and R and compute

$$\mathrm{E}[V] = \mathrm{E}[\pi H R^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \pi h r^2 f(h, r) dh dr.$$

Change-of-variable formula (Revisited)

More generally we have the following result:

Two-dimensional Change-of-variable formula

Let X and Y be random variables and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function.

If X and Y are discrete random variables with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, then

$$\mathrm{E}[g(X,Y)] = \sum_i \sum_j g(a_i, b_j) P(X = a_i,\ Y = b_j).$$

If X and Y are continuous random variables with joint probability density fucntion f, then

$$\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dx dy.$$

Linearity of Expectation (Revisited)

Linearity of Expectations of Two Random Variables

For all $r, s, t \in \mathbb{R}$ and random variables X and Y, one has

$$\mathrm{E}[rX + sY + t] = r\mathrm{E}[X] + s\mathrm{E}[Y] + t.$$

- Iterating above result yields, for random variables $X_1, \ldots X_n$ and $s_1, \ldots, s_n, t \in \mathbb{R}$:

$$\mathrm{E}[s_1 X_1 + \ldots s_n X_n + t] = s_1 \mathrm{E}[X_1] + \ldots s_n \mathrm{E}[X_n] + t.$$

Linearity of Expectation (Revisited)

(Derivation.)

Example: Short-Cut to the Expected Value of $Bin(n, p)$

Let $X \sim Bin(n, p)$. In last week's lab you've calculated

$$\mathrm{E}[X] = \sum_{k=0}^{n} kP(X_k = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \ldots = np.$$

This computation was not straightforward. Let's apply the linearity of expectation to calculate the expected value in an alternative way.

- Recall that $X = Y_1 + \ldots Y_n$, where $Y_i \sim_{iid} Ber(p)$ (Week 3 Lectures, Part 2, Slide 7!) Therefore,

$$\mathrm{E}[X] = \mathrm{E}[Y_1] + \ldots + \mathrm{E}[Y_n] \overset{(a)}{=} n\mathrm{E}[Y_1] = np,$$

  where (a) holds because $Y_1, \ldots Y_n$ are identically distributed (and hence $\mathrm{E}[Y_1] = \ldots \mathrm{E}[Y_n]$).

- What is the mean of $X = \sum_{i=1}^{n} Y_i$ if $Y_i \sim Ber(2^{-i})$ for $i = 1, \ldots n$? *(Verify that* $\mathrm{E}[X] = 1 - 2^{-n}$*!)*

Covariance of two Random Variables

We have just shown that for any two random variables X and Y,

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

Does there exist a similar result for the variance of two (or more) random variables?

- Let X and Y be arbitrary random variables. Then, direct calculations yield

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}\left[(X+Y-\mathrm{E}[X+Y])^2\right] \\
&= \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] + \mathrm{E}\left[(Y-\mathrm{E}[Y])^2\right] \\
&\quad + 2\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\underbrace{\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right]}_{\mathrm{Cov}(X,Y)}.
\end{aligned}$$

- The quantity $\mathrm{Cov}(X,Y)$ measures in some sense the way in which X and Y influence each other.

Covariance of two Random Variables (Cont.)

COVARIANCE

Let X ann Y be two random variables. The covariance between X and Y is defined as

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

- Straightforward algebra yields the following alternative expression:

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

- Loosely speaking, if
  - $\mathrm{Cov}(X, Y) > 0$, then a large (or small) value $X - \mathrm{E}[X]$ entails a large (or small) value of $Y - \mathrm{E}[Y]$. We say that X and Y are **positively correlated**.
  - $\mathrm{Cov}(X, u) < 0$, then a small (or large) value $X - \mathrm{E}[X]$ entails a large (or small) value of $Y - \mathrm{E}[Y]$. We say that X and Y are **negatively correlated**.
  - $\mathrm{Cov}(X, Y) = 0$, then X and Y we say that **uncorrelated**.

Two-Dim. Scatterplot and Correlation

(Sketches.)

Independence versus uncorrelatedness

INDEPENDENCE VERSUS UNCORRELATEDNESS

If two random variables X and Y are independent, then X and Y are uncorrelated.

- Let X and Y be two independent random variables. Then, X and Y have nothing to do with each other, we expect that they are uncorrelated.
- Indeed, for simplicity, consider the discrete case:

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_i \sum_j a_i b_j P(X = a_i,\, Y = b_j) \\
&= \sum_i \sum_j a_i b_j P(X = a_i) P(Y = b_j) \\
&= \left(\sum_i a_i P(X = a_i) \right) \left(\sum_j b_j P(Y = b_j) \right) \\
&= \mathrm{E}[X]\mathrm{E}[Y].
\end{aligned}$$

$$\Longrightarrow \ \mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = 0.$$

Example: Uncorrelated but dependent random variables

Let (X, Y) *form a unit perfect circle with center at the origin, i.e.* $X = \sqrt{1 - Y^2}$ *and* $Y \sim Unif(-1, 1)$.

- Then $\mathrm{Cov}(X, Y) = 0$ but X and Y are obviously not independent! *(Derivation.)*
- **Intuition:** Correlation measures only linear dependence between random variables.
- A perfect circle induces a highly nonlinear dependence between (X, Y) which correlation cannot capture!
- More examples of this sort on Wikipedia!

Example: Short-Cut to the Variance of $Bin(n, p)$

Let $X \sim Bin(n, p)$. In last week's lab you've calculated

$$\text{Var}(X) = np(1 - p).$$

This computation was not straightforward. Let's apply what we've just learnt sabout independence and uncorrelatedness to derive the variance in an alternative way.

- Recall that $X = Y_1 + \ldots Y_n$, where $Y_i \sim_{iid} Ber(p)$. Therefore,

$$\begin{aligned}\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} Y_i\right) &\overset{(a)}{=} \text{Var}(Y_1) + \ldots + \text{Var}(Y_n) \\ &\overset{(b)}{=} n\text{Var}(Y_1) = np(1 - p),\end{aligned}$$

  where (a) holds because $Y_1, \ldots, Y_n$ are independent (and hence uncorrelated) and (b) holds because $Y_1, \ldots Y_n$ are identically distributed (and hence $\text{Var}(Y_1) = \ldots \text{Var}(Y_n)$).

Three Important Properties of the Covariance

Let $a, b, c, d \in \mathbb{R}$ be arbitrary and X, Y, Z be random variables.

- $\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

(Derivation: Straightforward algebra, I suggest you try it out yourself.)

Correlation Coefficient

Major disadvantage of covariance: It depends on the units of the random variables!

- Suppose we were to model temperature (in Fahrenheit) and hours of day light by random variables T and H. For scientific purposes we may prefer to use Celsius instead of Fahrenheit. Recall that

$$C = \frac{5}{9} \times (T - 32).$$

- The covariance between temperature and hours of day light is therfore

$$\mathrm{Cov}(C, H) = \mathrm{Cov}\left(\frac{5}{9} \times (T - 32), H\right) = \ldots = \frac{5}{9} \times \mathrm{Cov}(T, H)$$

- Thus, by converting the units from Fahrenheit to Celsius the covariance between temperature and hours of day light has decreased (by the factor 5/9). That's disturbing.

Correlation Coefficient (Cont.)

Correlation Coefficient

Let X ann Y be two random variables. The correlation coefficient between X and Y is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- $\rho(X,Y)$ remains unaffected by a change of units and is therefore **dimensionless**.
- For $r, s, t, u \in \mathbb{R}$ and $r, t \neq 0$,

$$\rho(rX+s, tY+u) = \begin{cases} -\rho(X,Y) & \text{if } rt < 0, \\ \rho(X,Y) & \text{if } rt > 0. \end{cases}$$

- Random variables X and Y are "most correlated" if $X = Y$ or $X = -Y$ and we have

$$-1 \leq \rho(X,Y) \leq 1.$$

Correlation Coefficient (Derivation of $|\rho(X,Y)| \leq 1$)

Let $a, b \in \mathbb{R}$ be arbitrary and compute

$$0 \overset{(a)}{\leq} \operatorname{Var}(aX - bY) \overset{(b)}{=} a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) - 2ab\operatorname{Cov}(X,Y),$$

where (a) holds because variances are non-negative, i.e.

$$\operatorname{Var}(aX - bY) = \operatorname{E}\Big[\underbrace{\big(aX - bY - \operatorname{E}[aX - bY]\big)^2}_{\geq 0}\Big] \geq 0,$$

and (b) will be shown in Homework 6. Re-arrange above inequality to obtain

$$2ab\operatorname{Cov}(X,Y) \leq a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y).$$

Since $a, b \in \mathbb{R}$ arbitrary, we can now set $a = 1/\sqrt{\operatorname{Var}(X)}$ and $b = 1/\sqrt{\operatorname{Var}(Y)}$. Then,

$$2\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \leq \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)} + \frac{\operatorname{Var}(Y)}{\operatorname{Var}(Y)} = 2.$$

This proves that $\rho(X,Y) \leq 1$. To show that $\rho(X,Y) \geq -1$, repeat the argument with $a = -1/\sqrt{\operatorname{Var}(X)}$ and $b = 1/\sqrt{\operatorname{Var}(Y)}$.