Chapter 4

Variances and covariances

The expected value of a random variable gives a crude measure of the "center of location" of the distribution of that random variable. For instance, if the distribution is symmetric about a value μ then the expected value equals μ.

To refine the picture of a distribution distributed about its "center of location" we need some measure of spread (or concentration) around that value. The simplest measure to calculate for many distributions is the VARIANCE. There is an enormous body of probability literature that deals with approximations to distributions, and bounds for probabilities and expectations, expressible in terms of expected values and variances.

•variance

<4.1> **Definition.** *The* VARIANCE *of a random variable* X *with expected value* $\mathbb{E}X = \mu_X$ *is defined as* $\text{var}(X) = \mathbb{E}\left((X-\mu_X)^2\right)$. *The* COVARIANCE *between random variables* Y *and* Z*, with expected values* μ_Y *and* μ_Z*, is defined as* $\text{cov}(Y,Z) = \mathbb{E}\left((Y-\mu_Y)(Z-\mu_Z)\right)$. *The* CORRELATION *between* Y *and* Z *is defined as*

•covariance

•correlation

$$\text{corr}(Y,Z) = \frac{\text{cov}(Y,Z)}{\sqrt{\text{var}(Y)\text{var}(Z)}}$$

The square root of the variance of a random variable is called its STANDARD DEVIATION. □

•standard deviation

As with expectations, variances and covariances can also be calculated conditionally on various pieces of information.

Try not to confuse properties of expected values with properties of variances. For example, if a given piece of "information" implies that a random variable X must take the constant value C then $\mathbb{E}(X \mid \text{information}) = C$, but $\text{var}(X \mid \text{information}) = 0$. More generally, if the information implies that X must equal a constant then $\text{cov}(X,Y) = 0$ for every random variable Y. (You should check these assertions; they follow directly from the Definition.)

Notice that $\text{cov}(X,X) = \text{var}(X)$. Results about covariances contain results about variances as special cases.

A few facts about variances and covariances

Write μ_Y for $\mathbb{E}Y$, and so on, as above.

(i) $\text{cov}(Y,Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z)$ and, in particular, $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$:

$$\begin{aligned}\text{cov}(Y,Z) &= \mathbb{E}\left(YZ - \mu_Y Z - \mu_Z Y + \mu_Y\mu_Z\right)\\ &= \mathbb{E}(YZ) - \mu_Y\mathbb{E}Z - \mu_Z\mathbb{E}Y + \mu_Y\mu_Z\\ &= \mathbb{E}(YZ) - \mu_Y\mu_Z\end{aligned}$$

(ii) For constants a,b,c,d, and random variables U,V,Y,Z,

$$\begin{aligned}&\text{cov}(aU+bV, cY+dZ)\\ &\qquad = ac\,\text{cov}(U,Y) + bc\,\text{cov}(V,Y) + ad\,\text{cov}(U,Z) + bd\,\text{cov}(V,Z)\end{aligned}$$

It is easier to see the pattern if we work with the centered random variables $U' = U - \mu_U, \ldots, Z' = Z - \mu_Z$. For then the left-hand side of the asserted equality expands to

$$\mathbb{E}\left((aU' + bV')(cY' + dZ')\right) = \mathbb{E}(ac\,U'Y' + bc\,V'Y' + ad\,U'Z' + bd\,V'Z')$$
$$= ac\,\mathbb{E}(U'Y') + bc\,\mathbb{E}(V'Y') + ad\,\mathbb{E}(U'Z') + bd\,\mathbb{E}(V'Z')$$

The expected values in the last line correspond to the covariances on the right-hand side of the asserted equality.

As particular cases of fact (ii) we get two useful identities.

- Put $a = b = c = d = 1$ and $U = Y$ and $V = Z$ to get
  $$\text{var}(Y + Z) = \text{var}(Y) + 2\text{cov}(Y, Z) + \text{var}(Z)$$
  It is easy to confuse the formula for $\text{var}(Y + Z)$ with the formula for $\mathbb{E}(Y + Z)$. When in doubt, rederive.
- Put $U = Y = 1$, and $a = c$, and $b = d$, and $V = Z$:
  $$\text{var}(c + dZ) = d^2\text{var}(Z) \qquad \text{for constants } c \text{ and } d.$$
  Notice how the constant c disppeared, and the d turned into d^2. Many students confuse the formula for $\text{var}(c + dZ)$ with the formula for $\mathbb{E}(c + dZ)$. Again, when in doubt, rederive.

> You will find it easy to confuse variances with expectations. For example, it is a common blunder for students to confuse the formula for the variance of a difference with the formula $\mathbb{E}(Y - Z) = \mathbb{E}Y - \mathbb{E}Z$. If you ever find yourself wanting to assert that $\text{var}(Y - Z)$ is equal to $\text{var}(Y) - \text{var}(Z)$, think again. What would happen if $\text{var}(Z)$ were larger than $\text{var}(Y)$? Variances can't be negative.

Uncorrelated versus independent

•independent

Two random variables X and Y are said to be INDEPENDENT if "every event determined by X is independent of every event determined by Y". For example, independence of the random variables implies that the events $\{X \leq 5\}$ and $\{5Y^3 + 7Y^2 - 2Y^2 + 11 \geq 0\}$ are independent, and that the events $\{X \text{ even}\}$ and $\{7 \leq Y \leq 18\}$ are independent, and so on. Independence of the random variables also implies independence of functions of those random variables. For example, $\sin(X)$ must be independent of $\exp(1 + \cosh(Y^2 - 3Y))$, and so on.

<4.2> **Example.** Suppose a random variable X has a discrete distribution. The expected value $\mathbb{E}(XY)$ can then be rewritten as a weighted sum of conditional expectations:

$$\mathbb{E}(XY) = \sum_x \mathbb{P}\{X = x\}\mathbb{E}(XY \mid X = x) \qquad \text{by rule E4 for expectations}$$
$$= \sum_x x\mathbb{P}\{X = x\}\mathbb{E}(Y \mid X = x).$$

If Y is independent of X, the information "$X = x$" does not help with the calculation of the conditional expectation:

$$\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y) \qquad \text{if } Y \text{ is independent of } X.$$

The last calculation then simplifies further:

$$\mathbb{E}(XY) = (\mathbb{E}Y)\sum_x x\mathbb{P}\{X = x\} = (\mathbb{E}Y)(\mathbb{E}X) \qquad \text{if } Y \text{ independent of } X.$$

It follows that $\text{cov}(X, Y) = 0$ if Y is independent of X. □

•uncorrelated

A pair of random variables X and Y is said to be UNCORRELATED if $\text{cov}(X, Y) = 0$. The Example shows (at least for the special case where one random variable takes only a discrete set of values) that independent random variables are uncorrelated. The converse assertion—that uncorrelated should imply independent—is not true in general, as shown by the next Example.

<4.3> **Example.** For two independent rolls of a fair die, let X denote the value rolled the first time and Y denote the value rolled the second time. The random variables X and Y are independent, and they have the same distribution. Consequently $\text{cov}(X, Y) = 0$, and $\text{var}(X) = \text{var}(Y)$.

The two random variables $X + Y$ and $X - Y$ are uncorrelated:

$$\begin{aligned}\text{cov}(X+Y, X-Y) &= \text{cov}(X, X) + \text{cov}(X, -Y) + \text{cov}(Y, X) + \text{cov}(Y, -Y)\\ &= \text{var}(X) - \text{cov}(X, Y) + \text{cov}(Y, X) - \text{var}(Y)\\ &= 0.\end{aligned}$$

The two random variables $X + Y$ and $X - Y$ are not independent:

$$\mathbb{P}\{X + Y = 12\} = \mathbb{P}\{X = 6\}\mathbb{P}\{Y = 6\} = \frac{1}{36}$$

but

$$\mathbb{P}\{X + Y = 12 \mid X - Y = 5\} = \mathbb{P}\{X + Y = 12 \mid X = 6, Y = 1\} = 0$$

□

If Y and Z are uncorrelated, the covariance term drops out from the expression for the variance of their sum, leaving

$$\text{var}(Y + Z) = \text{var}(Y) + \text{var}(Z) \qquad \text{for } Y \text{ and } Z \text{ uncorrelated.}$$

Similarly, if $X_1, \dots, X_n$ are random variables for which $\text{cov}(X_i, X_j) = 0$ for each $i \neq j$ then

$$\text{var}(X_1 + \ldots + X_n) = \text{var}(X_1) + \ldots + \text{var}(X_n) \qquad \text{for "pairwise uncorrelated" rv's.}$$

You should check the last assertion by expanding out the quadratic in the variables $X_i - \mathbb{E}X_i$, observing how all the cross-product terms disappear because of the zero covariances.

Probability bounds

•Tchebychev's inequality

TCHEBYCHEV'S INEQUALITY asserts that if a random variable X has expected value μ then, for each $\epsilon > 0$,

$$\mathbb{P}\{|X - \mu| > \epsilon\} \leq \text{var}(X)/\epsilon^2$$

The inequality becomes obvious if we define a new random variable Z that takes the value 1 when $|X - \mu| > \epsilon$, and 0 otherwise: clearly $Z \leq |X - \mu|^2/\epsilon^2$, from which it follows that $\mathbb{P}\{|X - \mu| > \epsilon\} = \mathbb{E}Z \leq \mathbb{E}|X - \mu|^2/\epsilon^2$.

In the Chapter on the normal distribution you will find more refined probability approximations involving the variance.

The Tchebychev bound explains an important property of sample means. Suppose $X_1, \dots, X_n$ are uncorrelated random variables, each with expected value μ and variance σ^2. Let $\overline{X}$ equal the average. Its variance decreases like $1/n$:

$$\text{var}(\overline{X}) = (1/n)^2 \text{var}\left(\sum_{i \leq n} X_i\right) = 1/n^2 \sum_{i \leq n} \text{var}(X_i) = \sigma^2/n.$$

From the Tchebychev inequality,

$$\mathbb{P}\{|\overline{X} - \mu| > \epsilon\} \leq (\sigma^2/n)/\epsilon^2 \qquad \text{for each } \epsilon > 0$$

In particular, for each positive C,

$$\mathbb{P}\{|\overline{X} - \mu| > C\sigma/\sqrt{n}\} \leq 1/C^2$$

For example, there is at most a 1% chance that $\overline{X}$ lies more than $10\sigma/\sqrt{n}$ away from μ. (A normal approximation will give a much tighter bound.) Note well the dependence on n.

Variance as a measure of concentration in sampling theory

<4.4> **Example.** Suppose a finite population of objects (such as human beings) is numbered $1, \ldots, N$. Suppose also that to each object there is a quantity of interest (such as annual income): object α has the value $f(\alpha)$ associated with it. The population mean is

$$\bar{f} = \frac{f(1) + \ldots + f(N)}{N}$$

•random sample

Often one is interested in $\bar{f}$, but one has neither the time nor the money to carry out a complete census of the population to determine each $f(\alpha)$ value. In such a circumstance, it pays to estimate $\bar{f}$ using a RANDOM SAMPLE from the population.

The Bureau of the Census uses sampling heavily, in order to estimate properties of the U.S. population. The "long form" of the Decennial Census goes to about 1 in 6 households. It provides a wide range of sample data regarding both the human population and the housing stock of the country. It would be an economic and policy disaster if the Bureau were prohibited from using sampling methods. (If you have been following Congress lately you will know why I had to point out this fact.)

The mathematically simpler method to sample from a population requires each object to be returned to the population after it is sampled ("sampling with replacement"). It is possible that the same object might be sampled more than once, especially if the sample size is an appreciable fraction of the population size. It is more efficient, however, to sample without replacement, as will be shown by calculations of variance.

Consider first a sample $X_1, \ldots, X_n$ taken with replacement. The X_i are independent random variables, each taking values $1, 2, \ldots, N$ with probabilities $1/N$. The random variables $Y_i = f(X_i)$ are also independent, with

$$\mathbb{E}Y_i = \sum_\alpha f(\alpha)\mathbb{P}\{X_i = \alpha\} = \bar{f}$$

$$\text{var}(Y_i) = \sum_\alpha \left(f(\alpha) - \bar{f}\right)^2 \mathbb{P}\{X_i = \alpha\} = \frac{1}{N}\sum_\alpha \left(f(\alpha) - \bar{f}\right)^2$$

Write σ^2 for the variance.

The sample average,

$$\overline{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n),$$

has expected value

$$\mathbb{E}\overline{Y} = \frac{1}{n}(\mathbb{E}Y_1 + \ldots + \mathbb{E}Y_n) = \bar{f} \qquad \text{(no independence needed here)}$$

and variance

$$\begin{aligned}\text{var}(\overline{Y}) &= \frac{1}{n^2}\text{var}(Y_1 + \ldots + Y_n) \\ &= \frac{1}{n^2}\left(\text{var}(Y_1) + \ldots + \text{var}(Y_n)\right) \qquad \text{by independence} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

The sample average $\overline{Y}$ concentrates around $\bar{f}$ with a standard deviation $\sigma/\sqrt{n}$ that tends to zero as n gets larger.

Now consider a sample $X_1, \ldots, X_n$ taken without replacement. By symmetry, each X_i has the same distribution as before, and

$$\mathbb{E}Y_i = \bar{f} \qquad \text{and} \qquad \text{var}(Y_i) = \sigma^2.$$

This time the dependence between the X_i has an important effect on the variance of $\overline{Y}$.

By symmetry, for each pair $i \neq j$, the pair (X_i, X_j) takes each of the $N(N-1)$ values (α, β), for $1 \leq \alpha \neq \beta \leq N$, with probabilities $1/N(N-1)$. Consequently, for $i \neq j$,

$$\text{cov}(Y_i, Y_j) = \frac{1}{N(N-1)} \sum_{\alpha \neq \beta} \big((f(\alpha) - \bar{f})(f(\beta) - \bar{f})\big)$$

If we added the 'diagonal' terms $(f(\alpha) - \bar{f})^2$ to the sum we would have the expansion for

$$\sum_\alpha \big(f(\alpha) - \bar{f}\big) \sum_\beta \big(f(\beta) - \bar{f}\big),$$

which equals zero because $N\bar{f} = \sum_a f(\alpha)$. The expression for the covariance simplifies to

$$\text{cov}(Y_i, Y_j) = \frac{1}{N(N-1)} \left(0^2 - \sum_\alpha (f(\alpha) - \bar{f})^2\right) = -\frac{\sigma^2}{N-1}$$

If N is large, the covariance between any Y_i and Y_j, for $i \neq j$, is small; but there a lot of them:

$$\begin{aligned} \text{var}(\overline{Y}) &= \frac{1}{n^2}\text{var}(Y_1 + \ldots + Y_n) \\ &= \frac{1}{n^2}\mathbb{E}\big((Y_1 - \bar{f}) + \ldots + (Y_n - \bar{f})\big)^2 \end{aligned}$$

If you expand out the last quadratic you will get n terms of the form $(Y_i - \bar{f})^2$ and $n(n-1)$ terms of the form $(Y_i - \bar{f})(Y_j - \bar{f})$ with $i \neq j$. Take expectations:

$$\begin{aligned} \text{var}(\overline{Y}) &= \frac{1}{n^2}\left(n\text{var}(Y_1) + n(n-1)\text{cov}(Y_1, Y_2)\right) \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) \\ &= \frac{N-n}{N-1}\frac{\sigma^2}{n} \end{aligned}$$

•correction factor

Compare with the σ^2/n for $\text{var}(\overline{Y})$ under sampling with replacement. The CORRECTION FACTOR $(N-n)/(N-1)$ is close to 1 if the sample size n is small compared with the population size N, but it can decrease the variance of $\overline{Y}$ appreciably if n/N is not small. For example, if $n \approx N/6$ (as with the Census long form) the correction factor is approximately 5/6. If $n = N$, the correction factor is zero. That is, $\text{var}(\overline{Y}) = 0$ if the whole population is sampled. Why? What value must $\overline{Y}$ take if $n = N$? (As an exercise, see if you can use the fact that $\text{var}(\overline{Y}) = 0$ when $n = N$ to rederive the expression for $\text{cov}(Y_1, Y_2)$.) □

You could safely skip the rest of this Chapter, at least for the moment.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

Variances via conditioning

Suppose $\mathcal{F} = \{F_1, F_2, \ldots\}$ is a partition of the sample space into disjoint events. Let X be a random variable. Write y_i for $\mathbb{E}(X \mid F_i)$ and v_i for

$$\text{var}(X \mid F_i) = \mathbb{E}(X^2 \mid F_i) - (\mathbb{E}(X \mid F_i))^2.$$

Define two new random variables by the functions

$$\begin{aligned} Y(s) &= y_i \qquad \text{for } s \in F_i \\ V(s) &= v_i \qquad \text{for } s \in F_i \end{aligned}$$

Notice that $\mathbb{E}(Y \mid F_i) = y_i$, because Y must take the value y_i if the event F_i occurs. By rule E4 for expectations,

$$\begin{aligned}\mathbb{E}Y &= \sum_i \mathbb{E}(Y \mid F_i)\mathbb{P}F_i \\ &= \sum_i y_i \mathbb{P}F_i \\ &= \sum_i \mathbb{E}(X \mid F_i)\mathbb{P}F_i \\ &= \mathbb{E}X \qquad \text{by E4 again}\end{aligned}$$

Similarly,

$$\begin{aligned}\operatorname{var}(Y) &= \mathbb{E}Y^2 - (\mathbb{E}Y)^2 \\ &= \sum_i \mathbb{E}(Y^2 \mid F_i)\mathbb{P}F_i - (\mathbb{E}Y)^2 \\ &= \sum_i y_i^2 \mathbb{P}F_i - (\mathbb{E}X)^2\end{aligned}$$

The random variable V has expectation

$$\begin{aligned}\mathbb{E}V &= \sum_i \mathbb{E}(V \mid F_i)\mathbb{P}F_i \\ &= \sum_i v_i \mathbb{P}F_i \\ &= \sum_i \left(\mathbb{E}(X^2 \mid F_i) - y_i^2\right)\mathbb{P}F_i \\ &= \mathbb{E}X^2 - \sum_i y_i^2 \mathbb{P}F_i\end{aligned}$$

Add, cancelling out the common sum, to get

$$\operatorname{var}(Y) + \mathbb{E}V = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \operatorname{var}(X),$$

a formula that shows how a variance can be calculated using conditioning.

When written in more suggestive notation, the conditioning formula takes on a more pleasing appearance. Write $\mathbb{E}(X \mid \mathcal{F})$ for Y and $\operatorname{var}(X \mid \mathcal{F})$ for V. The symbolic "$\mid \mathcal{F}$" can be thought of as conditioning on the "information one obtains by learning which event in the partition $\mathcal{F}$ occurs". The formula becomes

$$\operatorname{var}\left(\mathbb{E}(X \mid \mathcal{F})\right) + \mathbb{E}\left(\operatorname{var}(X \mid \mathcal{F})\right) = \operatorname{var}(X)$$

and the equality $\mathbb{E}Y = \mathbb{E}X$ becomes

$$\mathbb{E}\left(\mathbb{E}(X \mid \mathcal{F})\right) = \mathbb{E}X$$

In the special case where the partition is defined by another random variable Z taking values $z_1, z_2, \ldots$, that is, $F_i = \{Z = z_i\}$, then the conditional expectation is usually written as $\mathbb{E}(X \mid Z)$ and the conditional variance as $\operatorname{var}(X \mid Z)$. The expectation formula becomes

$$\mathbb{E}\left(\mathbb{E}(X \mid Z)\right) = \mathbb{E}X$$

The variance formula becomes

$$\operatorname{var}\left(\mathbb{E}(X \mid Z)\right) + \mathbb{E}\left(\operatorname{var}(X \mid Z)\right) = \operatorname{var}(X)$$

When I come to make use of the formula I will say more about its interpretation.