Chapter 4

Variances and covariances

4.1 Overview

The expected value of a random variable gives a crude measure for 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 about its "center of location" we need some measure of spread (or concentration) around that value. For many distributions the simplest measure to calculate is the variance (or, more precisely, the square root of the variance).

Definition. The ***variance*** of a random variable X with expected value $\mathbb{E}X = \mu$ is defined as $\text{var}(X) = \mathbb{E}\left((X-\mu)^2\right)$. The square root of the variance of a random variable is called its ***standard deviation***, sometimes denoted by $\text{sd}(X)$.

The variance of a random variable X is unchanged by an added constant: $\text{var}(X+C) = \text{var}(X)$ for every constant C, because $(X+C) - \mathbb{E}(X+C) = X - \mathbb{E}X$, the C's cancelling. It is a desirable property that the spread should not be affected by a change in location. However, it is also desirable that multiplication by a constant should change the spread: $\text{var}(CX) = C^2\text{var}(X)$ and $\text{sd}(CX) = |C|\text{sd}(X)$, because $(CX - \mathbb{E}(CX))^2 = C^2(X - \mathbb{E}X)^2$. In summary: for constants a and b,

$$\text{var}(a+bX) = b^2\text{var}(X) \qquad \text{and} \qquad \text{sd}(a+bX) = |b|\text{sd}(X).$$

Remark. Try not to confuse properties of expected values with properties of variances: for constants a and b we have $\text{var}(a+bX) = b^2\text{var}(X)$ but $\mathbb{E}(a+bX) = a+b\mathbb{E}X$. Measures of location (expected value) and spread (standard deviation) should react differently to linear transformations of the variable. As another 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 $\text{var}(X \mid \text{information}) = 0$.

It is a common blunder 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.

There is an enormous probability literature that deals with approximations to distributions, and bounds for probabilities, expressible in terms of expected values and variances. One of the oldest and simplest examples, the Tchebychev inequality, is still useful, even though it is rather crude by modern standards.

Example <4.1> The Tchebychev inequality: $\mathbb{P}\{|X-\mu| \geq \epsilon\} \leq \text{var}(X)/\epsilon^2$, where $\mu = \mathbb{E}X$ and $\epsilon > 0$.

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

The Tchebychev inequality gives the right insight when dealing with sums of random variables, for which variances are easy to calculate. Suppose $\mathbb{E}Y = \mu_Y$ and $\mathbb{E}Z = \mu_Z$. Then

$$\begin{aligned}\text{var}(Y+Z) &= \mathbb{E}\left[Y-\mu_Y+Z-\mu_Z\right]^2 \\ &= \mathbb{E}\left[(Y-\mu_Y)^2 + 2(Y-\mu_Y)(Z-\mu_Z) + (Z-\mu_Z)^2\right] \\ &= \text{var}(Y) + 2\text{cov}(Y,Z) + \text{var}(Z)\end{aligned}$$

where $\text{cov}(Y,Z)$ denotes the ***covariance*** between Y and Z:

$$\text{cov}(Y,Z) := \mathbb{E}\left[(Y-\mu_Y)(Z-\mu_Z)\right].$$

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

More generally, for constants a, b, c, d, and random variables U, V, Y, Z,

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

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 becomes

$$\begin{aligned}&\mathbb{E}\left[(aU' + bV')(cY' + dZ')\right]\\ &\quad = \mathbb{E}\left[ac\,U'Y' + bc\,V'Y' + ad\,U'Z' + bd\,V'Z'\right]\\ &\quad = ac\,\mathbb{E}(U'Y') + bc\,\mathbb{E}(V'Y') + ad\,\mathbb{E}(U'Z') + bd\,\mathbb{E}(V'Z').\end{aligned}$$

The expected values in the last line correspond to the four covariances.

Sometimes it is easier to subtract off the expected values at the end of the calculation, by means of the formulae $\operatorname{cov}(Y,Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z)$ and, as a particular case, $\operatorname{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$. Both formulae follow via an expansion of the product:

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

Rescaled covariances define correlations, a concept that is much abused by those who do not understand probability.

Definition. The ***correlation*** between Y and Z is defined as

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

The random variables Y and Z are said to be ***uncorrelated*** if $\operatorname{corr}(Y,Z) = 0$.

Remark. Strictly speaking, the variance of a random variable is not well defined unless it has a finite expectation. Similarly, we should not talk about $\operatorname{corr}(Y,Z)$ unless both random variables have well defined variances for which $0 < \operatorname{var}(Y) < \infty$ and $0 < \operatorname{var}(Z) < \infty$.

Example <4.2> When well defined, correlations always lie between $+1$ and -1.

Variances for sums of uncorrelated random variables grow more slowly than might be anticipated. If Y and Z are uncorrelated, the covariance term drops out from the expression for the variance of their sum, leaving $\text{var}(Y+Z) = \text{var}(Y)+\text{var}(Z)$. Similarly, if $X_1, \ldots, X_n$ are random variables for which $\text{cov}(X_i, X_j) = 0$ for each $i \neq j$ then

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

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. These facts lead to a useful concentration property.

Example <4.3> Concentration of averages around expected value

Zero correlation is often deduced from independence. A pair of random variables X and Y is said to be ***independent*** if *every event determined by X is independent of every event determined by Y*. For example, independence implies that events such as $\{X \leq 5\}$ 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, $\sin(X)$ would be independent of e^Y, and so on. For the purposes of Stat241, you should not fret about the definition of independence: Just remember to explain why you regard some pieces of information as irrelevant when you calculate conditional probabilities and conditional expectations.

For example, suppose a random variable X can take values $x_1, x_2, \ldots$ and that X is independent of another random variable Y. Consider the expected value of a product $g(X)h(Y)$, for any functions g and h. Calculate by conditioning on the possible values taken by X:

$$\mathbb{E}g(X)h(Y) = \sum\nolimits_i \mathbb{P}\{X = x_i\}\mathbb{E}(g(X)h(Y) \mid X = x_i).$$

Given that $X = x_i$, we know that $g(X) = g(x_i)$ but we get no help with understanding the behavior of $h(Y)$. Thus, independence implies

$$\mathbb{E}(g(X)h(Y) \mid X = x_i) = g(x_i)\mathbb{E}(h(Y) \mid X = x_i) = g(x_i)\mathbb{E}h(Y).$$

Deduce that

$$\mathbb{E}g(X)h(Y) = \sum\nolimits_i \mathbb{P}\{X = x_i\}g(x_i)\mathbb{E}h(Y) = \mathbb{E}g(X)\mathbb{E}h(Y).$$

Put another way, if X and Y are independent random variables

$$\text{cov}\left(g(X), h(Y)\right) = \mathbb{E}\left(g(X)h(Y)\right) - \left(\mathbb{E}g(X)\right)\left(\mathbb{E}h(Y)\right) = 0.$$

That is, each function of X is uncorrelated with each function of Y. In particular, if X and Y are independent then they are uncorrelated. The converse is not usually true: uncorrelated random variables need not be independent.

Example <4.4> An example of uncorrelated random variables that are dependent

The concentration phenomenon can also hold for averages of dependent random variables.

Example <4.5> Comparison of spread in sample averages for sampling with and without replacement: the Decennial Census.

As with expectations, variances and covariances can also be calculated conditionally on various pieces of information. The conditioning formula in the final Example has the interpretation of a decomposition of "variability" into distinct sources, a precursor to the statistical technique know as the "analysis of variance".

Example <4.6> An example to show how variances can sometimes be decomposed into components attributable to difference sources. (Can be skipped.)

4.2 Things to remember

- $\mathbb{E}g(X)h(Y) = \mathbb{E}g(X)\mathbb{E}h(Y)$ if X and Y are independent random variables
- the definitions of variance and covariance, and their expanded forms $\text{cov}(Y, Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z)$ and $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$
- $\text{var}(a + bX) = b^2\text{var}(X)$ and $\text{sd}(a + bX) = |b|\text{sd}(X)$ for constants a and b.

- For constants a, b, c, d, and random variables U, V, Y, Z,

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

- Sampling without replacement gives smaller variances than sampling with replacement.

4.3 The examples

<4.1> **Example.** The Tchebychev inequality asserts: for a random variable X with expected value μ,

$$\mathbb{P}\{|X-\mu| > \epsilon\} \le \operatorname{var}(X)/\epsilon^2 \qquad \text{for each } \epsilon > 0.$$

The inequality becomes obvious if we write F for the event $\{|X-\mu| > \epsilon\}$. First note that $\mathbb{I}_F \le |X-\mu|^2/\epsilon^2$: when $\mathbb{I}_F = 0$ the inequality holds for trivial reasons; and when $\mathbb{I}_F$ takes the value one, the random variable $|X-\mu|^2$ must be larger than ϵ^2. It follows that

$$\mathbb{P}\{|X-\mu| > \epsilon\} = \mathbb{P}F = \mathbb{E}\mathbb{I}_F \le \mathbb{E}|X-\mu|^2/\epsilon^2.$$

□

<4.2> **Example.** *When well defined, correlations always lies between* $+1$ *and* -1.
Suppose

$$\begin{aligned}\mathbb{E}Y = \mu_Y \qquad &\text{and} \qquad \operatorname{var}(Y) = \sigma_Y^2\\ \mathbb{E}Z = \mu_Y \qquad &\text{and} \qquad \operatorname{var}(Z) = \sigma_Z^2\end{aligned}$$

Define standardized variables

$$Y' = \frac{Y-\mu_Y}{\sigma_Y} \qquad \text{and} \qquad Z' = \frac{Z-\mu_Z}{\sigma_Z}\,.$$

Note that $\mathbb{E}Y' = \mathbb{E}Z' = 0$ and $\operatorname{var}(Y') = \operatorname{var}(Z') = 1$. Also

$$\operatorname{corr}(Y,Z) = \operatorname{cov}(Y'Z') = \mathbb{E}(Y'Z').$$

Use the fact that variances are always nonnegative to deduce that

$$0 \le \operatorname{var}(Y'+Z') = \operatorname{var}(Y') + 2\operatorname{cov}(Y',Z') + \operatorname{var}(Z') = 2 + 2\operatorname{cov}(Y',Z'),$$

which rearranges to $\operatorname{cov}(Y',Z') \ge -1$. Similarly

$$0 \le \operatorname{var}(Y'-Z') = \operatorname{var}(Y') - 2\operatorname{cov}(Y',Z') + \operatorname{var}(Z') = 2 - 2\operatorname{cov}(Y',Z'),$$

which rearranges to $\operatorname{cov}(Y',Z') \le +1$. □

<4.3> **Example.** Suppose $X_1, \ldots, X_n$ are uncorrelated random variables, each with expected value μ and variance σ^2. By repeated application of the formula for the variance of a sum of variables with zero covariances,

$$\text{var}\,(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2.$$

Typically the X_i would come from repeated independent measurements of some unknown quantity. The random variable $\overline{X} = (X_1 + \cdots + X_n)/n$ is then called the ***sample mean***.

The variance of the sample mean decreases like $1/n$,

$$\text{var}(\overline{X}) = (1/n)^2 \text{var}\,(X_1 + \cdots + X_n) = \sigma^2/n.$$

From the Tchebychev inequality,

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

In particular, for each positive constant C,

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

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. □

<4.4> **Example.** Consider 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:

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

Nevertheless, the sum and difference are not independent. For example,

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

but

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

□

<4.5> **Example.** Until quite recently, in the Decennial Census of Housing and Population the Census Bureau would obtain some more detailed about the population via information from a more extensive list of questions sent to only a random sample of housing units. For an area like New Haven, about 1 in 6 units would receive the so-called "long form".

For example, one question on the long form asked for the number of rooms in the housing unit. We could imagine the population of all units numbered $1, 2, \ldots, N$, with the ith unit containing y_i rooms. Complete enumeration would reveal the value of the ***population average***,

$$\bar{y} = \frac{1}{N}\left(y_1 + y_2 + \cdots + y_N\right).$$

A sample can provide a good estimate of $\bar{y}$ with less work.

Suppose a sample of n housing units is selected from the population without replacement. (For the Decennial Census, $n \approx N/6$.) The answer from each unit is a random variable that could take each of the values $y_1, y_2, \ldots, y_N$, each with probability $1/N$.

> **Remark.** It might be better to think of a random variable that takes each of the values $1, 2, \ldots, N$ with probability $1/N$, then take the corresponding number of rooms as the value of the random variable that is recorded. Otherwise we can fall into verbal ambiguities when many of the units have the same number of rooms.

That is, the sample consists of random variables $Y_1, Y_2, \ldots, Y_n$, for each of which

$$\mathbb{P}\{Y_i = y_j\} = \frac{1}{N} \qquad \text{for } j = 1, 2, \ldots, N.$$

Notice that

$$\mathbb{E}Y_i = \frac{1}{N}\sum\nolimits_{j=1}^{N} y_j = \bar{y},$$

and consequently, the sample average $\bar{Y} = (Y_1 + \cdots + Y_n)/n$ also has expected value $\bar{y}$. Notice also that each Y_i has the same variance,

$$\operatorname{var}(Y_i) = \frac{1}{N}\sum\nolimits_{j=1}^{N}\left(y_j - \bar{y}\right)^2,$$

a quantity that I will denote by σ^2.

If the sample is taken without replacement—which, of course, the Census Bureau had to do, if only to avoid media ridicule—the random variables are

dependent. For example, in the extreme case where $n = N$, we would necessarily have

$$Y_1 + Y_2 + \cdots + Y_N = y_1 + y_2 + \cdots + y_N,$$

so that Y_N would be a function of the other Y_i's, a most extreme form of dependence. Even if $n < N$, there is still some dependence, as you will soon see.

Sampling with replacement would be mathematically simpler, for then the random variables Y_i would be independent, and, as in Example <4.3>, we would have $\operatorname{var}\left(\bar{Y}\right) = \sigma^2/n$. With replacement, it is possible that the same unit might be sampled more than once, especially if the sample size is an appreciable fraction of the population size. There is also some inefficiciency in sampling with replacement, as shown by a calculation of variance for sampling without replacement:

$$\begin{aligned}
\operatorname{var}\left(\bar{Y}\right) &= \mathbb{E}\left(\bar{Y} - \bar{y}\right)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(Y_i - \bar{y}\right)\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum\nolimits_{i=1}^{n}\left(Y_i - \bar{y}\right)^2 + 2\sum\nolimits_{1\le i<j\le n}(Y_i - \bar{y})(Y_j - \bar{y})\right) \\
&= \frac{1}{n^2}\left(\sum\nolimits_{i=1}^{n}\mathbb{E}\left(Y_i - \bar{y}\right)^2 + 2\sum\nolimits_{1\le i<j\le n}\mathbb{E}\left((Y_i - \bar{y})(Y_j - \bar{y})\right)\right) \\
&= \frac{1}{n^2}\left(\sum\nolimits_{i=1}^{n}\operatorname{var}(Y_i) + \sum\nolimits_{1\le i\ne j\le n}\operatorname{cov}(Y_i, Y_j)\right)
\end{aligned}$$

What formula did I just rederive?

There are n variance terms and $n(n-1)$ covariance terms. We know that each Y_i has variance σ^2, regardless of the dependence between the variables. The effect of the dependence shows up in the covariance terms. By symmetry, $\operatorname{cov}(Y_i, Y_j)$ is the same for each pair $i < j$, a value that I will denote by c. Thus, for sampling without replacement,

$$(*) \qquad \operatorname{var}\left(\bar{Y}\right) = \frac{1}{n^2}\left(n\sigma^2 + n(n-1)c\right) = \frac{\sigma^2}{n} + \frac{(n-1)c}{n}.$$

We can calculate c directly, from the fact that the pair (Y_1, Y_2) takes each of $N(N-1)$ pairs of values (y_i, y_j) with equal probability. Thus

$$c = \operatorname{cov}(Y_1, Y_2) = \frac{1}{N(N-1)}\sum\nolimits_{i\ne j}(y_i - \bar{y})(y_j - \bar{y}).$$

If we added the "diagonal" terms $(y_i - \bar{y})^2$ to the sum we would have the expansion for the product

$$\sum\nolimits_{i=1}^{N}\left(y_i - \bar{y}\right)\sum\nolimits_{j=1}^{N}\left(y_j - \bar{y}\right),$$

which equals zero because $N\bar{y} = \sum_{i=1}^{N} y_i$. The expression for the covariance simplifies to

$$c = \text{cov}(Y_1, Y_2) = \frac{1}{N(N-1)} \left(0^2 - \sum\nolimits_{i=1}^{N} (y_i - \bar{y})^2 \right) = -\frac{\sigma^2}{N-1}.$$

Substitution in formula $(*)$ then gives

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1} \right) = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

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. Indeed, when $n = N$ we know that $\bar{Y}$ equals the population mean, $\bar{y}$, a constant. A random variable that always takes the same constant value has zero variance. Thus the right-hand side of $(*)$ must reduce to zero when we put $n = N$, which gives a quick method for establishing the equality $c = -\sigma^2/(N-1)$, without all the messing around with sums of products and products of sums. □

<4.6> **Example.** Consider a two stage method for generating a random variable Z. Suppose we have k different random variables $Y_1, \ldots, Y_k$, with $\mathbb{E}Y_i = \mu_i$ and $\text{var}(Y_i) = \sigma_i^2$. Suppose also that we have a random method for selecting which variable to choose: a random variable X that is independent of all the Y_i's, with $\mathbb{P}\{X = i\} = p_i$ for $i = 1, 2, \ldots, k$, where $p_1 + p_2 + \cdots + p_k = 1$. If X takes the value i, define Z to equal Y_i.

The variability in Z is due to two effects: the variability of each Y_i; and the variability of X. Conditional on $X = i$, we have Z equal to Y_i, and

$$\begin{aligned} \mathbb{E}\left(Z \mid X = i\right) &= \mathbb{E}(Y_i) = \mu_i \\ \text{var}\left(Z \mid X = i\right) &= \mathbb{E}\left((Z - \mu_i)^2 \mid X = i\right) = \text{var}(Y_i) = \sigma_i^2. \end{aligned}$$

From the first formula we get

$$\mathbb{E}Z = \sum\nolimits_i \mathbb{P}\{X = i\} \mathbb{E}\left(Z \mid X = i\right) = \sum\nolimits_i p_i \mu_i,$$

a weighted average of the μ_i's that I will denote by $\bar{\mu}$. A similar conditioning exercise gives

$$\text{var}(Z) = \mathbb{E}\left(Z - \bar{\mu}\right)^2 = \sum\nolimits_i p_i \mathbb{E}\left((Z - \bar{\mu})^2 \mid X = i \right).$$

If we could replace the $\bar{\mu}$ in the ith summand by μ_i, the sum would become a weighted average of conditional variances. To achieve such an effect, rewrite $(Z-\bar{\mu})^2$ as

$$(Z-\mu_i+\mu_i-\bar{\mu})^2 = (Z-\mu_i)^2 + 2(\mu_i-\bar{\mu})(Z_i-\mu_i) + (\mu_i-\bar{\mu})^2.$$

Taking conditional expectations, we then get

$$\begin{aligned}&\mathbb{E}\left((Z-\bar{\mu})^2 \mid X=i\right)\\ &\quad= \mathbb{E}\left((Z-\bar{\mu}_i)^2 \mid X=i\right) + 2(\mu_i-\bar{\mu})\mathbb{E}\left(Z-\mu_i \mid X=i\right) + (\mu_i-\bar{\mu})^2.\end{aligned}$$

On the right-hand side, the first term equals σ_i^2, and the middle term disappears because $\mathbb{E}(Z \mid X=i) = \mu_i$. With those simplifications, the expression for the variance becomes

$$\operatorname{var}(Z) = \sum\nolimits_i p_i\sigma_i^2 + \sum\nolimits_i p_i(\mu_i-\bar{\mu})^2.$$

If we think of each Y_i as coming from a separate "population", the first sum represents the component of variability within the populations, and the second sum represents the variability between the populations.

The formula is sometimes written symbolically as

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

where $E(Z \mid X)$ denotes the random variable that takes the value μ_i when X takes the value i, and $\operatorname{var}(Z \mid X)$ denotes the random variable that takes the value σ_i^2 when X takes the value i. □