1 Standardized Random Variables

If X is a random variable with expectation $\mathbb{E}[X] = \mu$ and standard deviation $\sigma > 0$, then the standardized version of this random variable is given by

$$\widetilde{X} = \frac{X - \mu}{\sigma}.$$

The standardized RV is a shifted and rescaled version of X such that

$$\mathbb{E}\left[\widetilde{X}\right] = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbb{E}[X]-\mu}{\sigma} = 0.$$
$$\mathrm{Var}(\widetilde{X}) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{\mathrm{Var}(X)}{\sigma^2} = 1.$$

Note that the distribution of a standardized random variable has the same probabilities as the original random variable, but the values in the range are rescaled.

That is, whereas the distribution of X is $\{(a, \mathbb{P}[X=a]) : a \in \mathscr{A}\}$, the distribution of $\widetilde{X}$ is $\{(\frac{a-\mu}{\sigma}, \mathbb{P}[X=a]) : a \in \mathscr{A}\}$

If we think of a collection of data as samples from a distribution, standardizing that data is conceptually equivalent to standardizing the equivalent underlying random variable model. Standardization gives us a way to reason about our data in a scale-free manner.

For example, suppose we want to compare grade distributions on exams between two different schools with completely different grading scales. Standardizing allows us to make easier apples-to-apples comparisons.

Standardization is also often used when training machine learning models. By keeping properties of our data on the same scale, we get better behavior out of the numerical methods used to train ML models.

As we will see below in these notes, standardization also allows us to convert an important quantity called the "covariance" of two random variables into its scale-free equivalent called the "correlation".

2 Covariance and Correlation

In the previous notes, we showed that for random variables X, Y, we have:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]).$$

The expression $\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ is a measure of association between X, Y, and is called the *covariance*.

Definition 21.1 (Covariance)**.** *The* covariance *of random variables X and Y, denoted* $\text{cov}(X,Y)$*, is defined as*

$$\text{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y],$$

where $\mu_X = \mathbb{E}[X]$ *and* $\mu_Y = \mathbb{E}[Y]$.

Exercise: Check that the above two expressions for $\text{cov}(X,Y)$ are indeed equal.

Remarks. We note some important facts about variance and covariance.

1. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{cov}(X,Y)$.
2. If X, Y are independent, then $\text{cov}(X,Y) = 0$. However, the converse is **not** true.
3. $\text{cov}(X,X) = \text{Var}(X)$.
4. $\text{cov}(aX+b, cX+d) = ac\,\text{cov}(X,X) = ac\,\text{Var}(X)$.
5. Covariance is *bilinear*; i.e., for any collection of random variables $\{X_1,\ldots,X_n\}, \{Y_1,\ldots,Y_m\}$ and fixed constants $\{a_1\ldots,a_n\}, \{b_1,\ldots,b_m\}$,

$$\text{cov}\left(\textstyle\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \,\text{cov}(X_i, Y_j).$$

While the *sign* of $\text{cov}(X,Y)$ (positive or negative) is informative of how X and Y are associated, its magnitude is difficult to interpret. A statistic that is easier to interpret is *correlation*:

Definition 21.2 (Correlation)**.** *Suppose X and Y are random variables with $\sigma(X) > 0$ and $\sigma(Y) > 0$. Then, the correlation of X and Y is defined as*

$$\text{Corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Correlation is more useful than covariance because the former always ranges between -1 and $+1$, as the following theorem shows:

Theorem 21.1. *For any pair of random variables X and Y with $\sigma(X) > 0$ and $\sigma(Y) > 0$,*

$$-1 \le \text{Corr}(X,Y) \le +1.$$

Proof. Let $\mathbb{E}[X] = \mu_X$ and $\mathbb{E}[Y] = \mu_Y$, and define $\widetilde{X} = (X-\mu_X)/\sigma(X)$ and $\widetilde{Y} = (Y-\mu_Y)/\sigma(Y)$. Then, $\mathbb{E}[\widetilde{X}^2] = \mathbb{E}[\widetilde{Y}^2] = 1$, so

$$\begin{aligned} 0 \le \mathbb{E}[(\widetilde{X}-\widetilde{Y})^2] &= \mathbb{E}[\widetilde{X}^2] + \mathbb{E}[\widetilde{Y}^2] - 2\,\mathbb{E}[\widetilde{X}\widetilde{Y}] = 2 - 2\,\mathbb{E}[\widetilde{X}\widetilde{Y}] \\ 0 \le \mathbb{E}[(\widetilde{X}+\widetilde{Y})^2] &= \mathbb{E}[\widetilde{X}^2] + \mathbb{E}[\widetilde{Y}^2] + 2\,\mathbb{E}[\widetilde{X}\widetilde{Y}] = 2 + 2\,\mathbb{E}[\widetilde{X}\widetilde{Y}], \end{aligned}$$

which implies $-1 \le \mathbb{E}[\widetilde{X}\widetilde{Y}] \le +1$. Finally, note that

$$\text{Corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma(X)\sigma(Y)} = \mathbb{E}\left[\widetilde{X}\widetilde{Y}\right].$$

Therefore, we can deduce that $-1 \le \text{Corr}(X,Y) \le +1$. □

Note that the above proof shows that $\text{Corr}(X,Y) = +1$ if and only if $\mathbb{E}[(\widetilde{X} - \widetilde{Y})^2] = 0$, which implies $\widetilde{X} = \widetilde{Y}$ with probability 1. Similarly, $\text{Corr}(X,Y) = -1$ if and only if $\mathbb{E}[(\widetilde{X} + \widetilde{Y})^2] = 0$, which implies $\widetilde{X} = -\widetilde{Y}$ with probability 1. In terms of the original random variables X, Y, this means the following: if $\text{Corr}(X,Y) = \pm 1$, then there exist constants a and b such that, with probability 1,

$$Y = aX + b,$$

where $a > 0$ if $\text{Corr}(X,Y) = +1$ and $a < 0$ if $\text{Corr}(X,Y) = -1$. (The values a, b depend on $\mu_X, \mu_Y, \sigma(X), \sigma(Y)$; in particular, $a = \pm\frac{\sigma(Y)}{\sigma(X)}$ according to whether $\text{Corr}(X,Y) = \pm 1$.) Thus X and Y differ just by a scaling factor and a shift.

3 Joint Distributions: conditional expectation

Recall the following definition that we presented previously.

Definition 21.3. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X = a, Y = b]) : a \in \mathscr{A}, b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X = a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X = a, Y = b].$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

From a joint distribution, we can compute the *conditional probability* of $X = x$ given that $Y = y$ from the definition of conditional probability as follows:

$$\mathbb{P}[X = x \mid Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]}.$$

The *conditional expectation* of X given $Y = y$ is defined naturally as follows:

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathscr{A}} x \cdot \mathbb{P}[X = x \mid Y = y].$$

That is, $\mathbb{E}[X \mid Y = y]$ is simply the expectation of X given that $Y = y$.

3.1 Total Expectation

Suppose we roll a fair four sided die, then flip a number of fair coins equal to the result of that die roll. Let X be the number of heads, and suppose we want to $\mathbb{E}[X]$.

The *Law of Total Expectation* (LoTE) says that given any random variable X and events $A_1, A_2, \ldots, A_n$ that partition the sample space Ω, we have that

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i]\, \mathbb{P}[A_i].$$

We can think of this sum as splitting the sample space into partitions (events) and summing the expectation of X in each partition, weighted by the probability of that event occurring.

For our motivating example, if we define A_i as the event where the die roll comes up i, then $\mathbb{E}[X \mid A_i]$ is $\frac{i}{2}$, and $\mathbb{P}[A_i] = \frac{1}{4}$. Thus the LoTE gives us:

$$\mathbb{E}[X] = \frac{1}{4} \times \left(\frac{1}{2} + \frac{2}{2} + \frac{3}{2} + \frac{4}{2} \right) = \frac{5}{4}.$$

3.2 Iterated Expectation and Wald's identity

We can instead conceptualize our motivating example in terms of joint random variables X, the number of heads, and Y, the result of the die roll. In that case, we can rewrite the LoTE as:

$$\mathbb{E}[X] = \sum_{y \in range(Y)} \mathbb{E}[X \mid Y = y] \, \mathbb{P}[Y = y]. \tag{1}$$

Here,

We can also think of the die and sum example above as an iterated expectation, where the events $Y = y_1, Y = y_2, \ldots$ partition the sample space, where $\{y_1, y_2, \ldots\}$ are all the possible values of Y. In this case, $\mathbb{E}[X \mid Y = y]$ is a function of Y: it takes inputs $y \in Y$ and outputs $f(y) = \mathbb{E}[X \mid Y = y]$. So $f(Y) = \mathbb{E}[X \mid Y]$ is itself a random variable.

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_{y \in \mathscr{B}} \mathbb{E}[X \mid Y = y] \, \mathbb{P}[Y = y]. \tag{2}$$

For example, To compute this expectation, we can apply the law of total expectation as described bove.

In this case, the sum given in the law of total expectation would have four terms, one for each possible roll of the dice.

We observe that another way to write the Law of Total Expectation is as an iterated expectation, first over

If Y is a random variable, then the events $Y = y_1, Y = y_2, \ldots$ partition the sample space, where $\{y_1, y_2, \ldots\}$ are all the possible values of Y. In this case, $\mathbb{E}[X \mid Y = y]$ is a function of Y: it takes inputs $y \in Y$ and outputs $f(y) = \mathbb{E}[X \mid Y = y]$. So $f(Y) = \mathbb{E}[X \mid Y]$ is itself a random variable.

Finally, we discuss *the law of iterated expectations*, also called *the law of total expectation* which is

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_{y \in \mathscr{B}} \mathbb{E}[X \mid Y = y] \, \mathbb{P}[Y = y]. \tag{3}$$

This simple concept can be quite useful. For example, consider choosing an integer N at random and forming a random variable $Y = X_1 + \cdots + X_N$ where the X_i are independent and identically distributed. Note the *number* of terms is a random variable here! We wish to compute $\mathbb{E}[Y]$ and assuming that the random

variables X_i is independent of the value of N.

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid N]] \\
&= \sum_n \mathbb{E}[Y \mid N = n]\,\mathbb{P}[N = n] \\
&= \sum_n n\,\mathbb{E}[X_1]\,\mathbb{P}[N = n] \\
&= \mathbb{E}[X_1] \sum_n n\,\mathbb{P}[N = n] \\
&= \mathbb{E}[X_1]\,\mathbb{E}[N]
\end{aligned}$$

The third line follows from $Y = X_1 + \cdots + X_n$ and the fact that X_1 are identically distributed and are independent of the value of N. Thus, we have $\mathbb{E}[Y] = \mathbb{E}[X_1]\,\mathbb{E}[N]$ which is the basic form of *Wald's identity.*

This can be useful for modelling the total time to serve customers in a time interval, where we have "Poisson" arrivals, and each customer's service time is from the same distribution. That is, we have a Poisson random variable, $N \sim \text{Poisson}(\lambda)$, that determines the number of customers and each X_i is a random variable that corresponds to the time needed to serve customer i.

To conclude, we note that the law of iterated expectations is sometimes called the tower rule as one can extend the concept to more than two random variables, e.g., $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X \mid Y, Z]]]$, where the outer expectations are over the values of Y and Z analogously to (3).