Data Science Essentials

Probability and Random Variables

As data scientists, we're often concerned with understanding the qualities and relationships of a set of data points. For example, you may need to examine the data in a set of sales invoice records, in which each record contains data pointes relating to a sale; such as the store location, the amount spent, and the day on which the sale occurred.

Random Variables

Data points may be thought of outcomes of random variables, provided that they are expressed as numbers. So for example, the text values Monday, Tuesday, Wednesday, and so on are not outcomes of a random variable, but if we assign the value 1 to Monday, 2 to Tuesday, 3 to Wednesday, and so on, then we can consider these to be outcomes of a random variable.

Random variables can be discrete or continuous. Discrete variables have a countable number of distinct outcomes; for example, the number of cookies in a jar, or a numeric identifier associated with a day of the week. Continuous variables have real-valued outcomes, for example, the temperature on a given day or the volume of water flowing in a stream over a specified period.

"Probability" measures the likelihood of a random variable to take on a specified range of outcomes. For example, consider a random variable that is the store sales on Saturday at a particular store. The probability to have sales between $7000 and $9000 could be .43.

Probability is always expressed as a numeric value, and the total probability for all outcomes of any random variable is always 1 (i.e. 100%). For example, consider the following table, which shows a discrete random variable that represents the days of the week together with the probability that a particular sales transaction occurred on that day:

Day	Variable (X)	Probability (P)
Monday	1	1/8 (0.125 or 12.5%)
Tuesday	2	1/8 (0.125 or 12.5%)
Wednesday	3	1/8 (0.125 or 12.5%)
Thursday	4	1/8 (0.125 or 12.5%)
Friday	5	1/8 (0.125 or 12.5%)
Saturday	6	3/16 (0.1875 or 18.75%)
Sunday	7	3/16 (0.1875 or 18.75%)

This table describes the *probability mass function (PMF)* for the random variable X – it identifies the probability (P) for each outcome of random variable (X). So for example, the probability of the weekday of a sales transaction being Monday (or in other words, the probability of the value of the variable being 1) is 1/8 (which is the same as 0.125 or 12.5%). This is written as $P(X=1) = 1/8$. More generally, the formula is $P(X=x_n) = p_n$ (where x_n is the *n*th possible outcome and p_n is the corresponding probability.) Note that all of the probabilities add up to 1.

Summarizing Random Variables

When you know the PMF for a random variable, you can summarize the random variable by determining its *mean* and *variance*. The mean of the variable indicates its centrality – in other words, its average value. It's also known as the variables *expected value* and is represented by the μ symbol (*mu*). To calculate the mean, multiply each outcome by its probability and total the results, which expressed as a formula is $\mu = \sum P(x_n) \times x_n$. So in the example of the weekday variable above, the calculation is:

(0.125 x 1) + (0.125 x 2) + (0.125 x 3) + (0.125 x 4) + (0.125 x 5) + (0.1875 x 6) + (0.1875 x 7)

This yields the result 4.3125.

Variance indicates the spread of variable values from the mean. It is a squared value, represented by the symbol σ^2. To measure the variance, multiply the probability of each outcome by the (outcome minus the mean squared), and add together the results for each value. The formula is expressed as $\sigma^2= \sum P(x_n) \times [x_n - \mu]^2$. In the case of the weekdays example, it can be calculated as:

(0.125 x [1 – 4.3125]^2) + (0.125 x [2 – 4.3125]^2) + (0.125 x [3 – 4.3125]^2) + (0.125 x [4 – 4.3125]^2) + (0.125 x [5 – 4.3125]^2) + (0.1875 x [6 – 4.3125]^2) + (0.1875 x [7 – 4.3125]^2)

This yields the result 4.214844.

Since the variance is in squared units, it makes sense to calculate its square root – which is known as the *standard deviation* (or σ). In this case, the standard deviation from the mean for the week day variable is $\sqrt{4.214844}$, which yields 2.053008.

Probability Distributions

Now that you know how to work with random variables and probabilities, you can calculate more complicated probabilities of events.

For example, suppose you have five sales invoices. What is the probability that exactly two of those invoices will be for sales transactions that occurred on a Monday? To calculate this, you need to know two things: What is the probability that a random sales transaction occurred on a Monday, and how many possible ways are there to choose two Monday receipts out of five total?

First, let's simplify the problem and determine the probability of a single possible combination of five invoices that includes one Monday. Imagine that the first invoice you select is for a Monday sale and the next four aren't - you can represent this as YNNNN (where Y indicates a Monday invoice, and N indicates a non-Monday invoice). We know that the probability of a single sale being on a Monday (P(X)=1) is 0.125. We also know that the combined probability for any other value is 1 – 0.125 (because the probabilities for a variable always add up to 1). We can use this to calculate the probability of the first invoice being for a Monday (0.125^1) and the probability of the other four invoices being non-Mondays

([1 - 0.125]^4 and multiply the two together to get the overall probability for this particular combination (0.125 x 0.586182 = 0.073273.

Now let's look at a combination with two Mondays – for example, YYNNN. The probability for this particular combination of invoices is the probability for two Mondays and three non-Mondays, which can be calculated as 0.125^2 x [1 - 0.125]^3, which gives the result 0.010468.

So now we know the probability for one possible combination of invoices that contains two Mondays (YYNNN). However, we need to calculate the probability for all the other possible combinations (such as YNYNN, YNNYN, YNNNY, and so on). We can calculate this using factorials by dividing the total number of possible combinations of days by the number of combinations we want to include multiplied by the number of combinations we want to exclude. In other words, we can use the formula 5! / (2! x 3!), which gives the answer 10. For a small number like this, you can easily check by writing out all of the 10 possible combinations:

YYNNN, YNYNN, YNNYN, YNNNY, NYYNN, NYNYN, NYNNY, NNYYN, NNYNY, NNNYY

Now we can combine the calculations to find the probability of us having chosen two out of the five invoices having Monday dates, with the calculation to find the total number of possible combinations like this:

(0.125^2 x [1 - 0.125]^3) x (5 / (2! x 3!)),

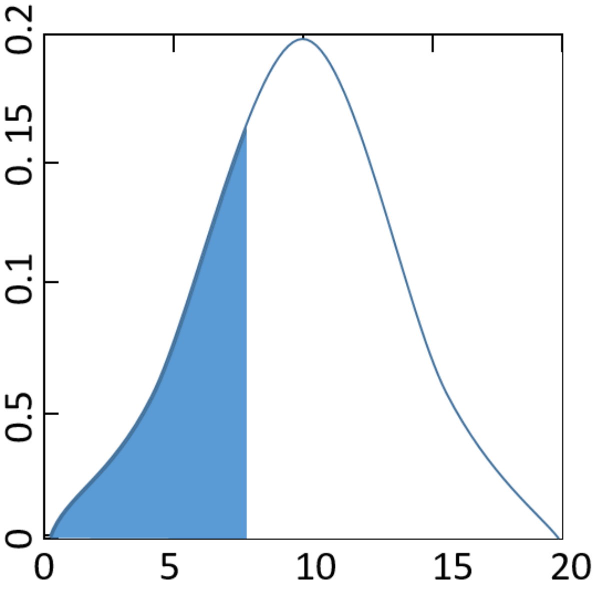
which gives the result 0.10468.

The formula for this calculation is more generally known as the *Binomial* distribution, and is often used to calculate probability distributions where the same independent test is repeated multiple times. Another similar distribution is the *Poisson* distribution, in which the probability of success decreases as the number of tests is increased.

Probability for Continuous Variables

Both the *Binomial* and *Poisson* distributions can be used to calculate probability for discrete random variables. For continuous variables, there are no discrete values in a PMF table, so the probability is expressed as a curve known as the *probability density function (PDF)*. Probabilities of the variable value being within a specified range are calculated based on the area under the curve, the total of which always adds up to 1.

For example, here's a PDF for a continuous variable that shows the probability that a variable value is less than 7.

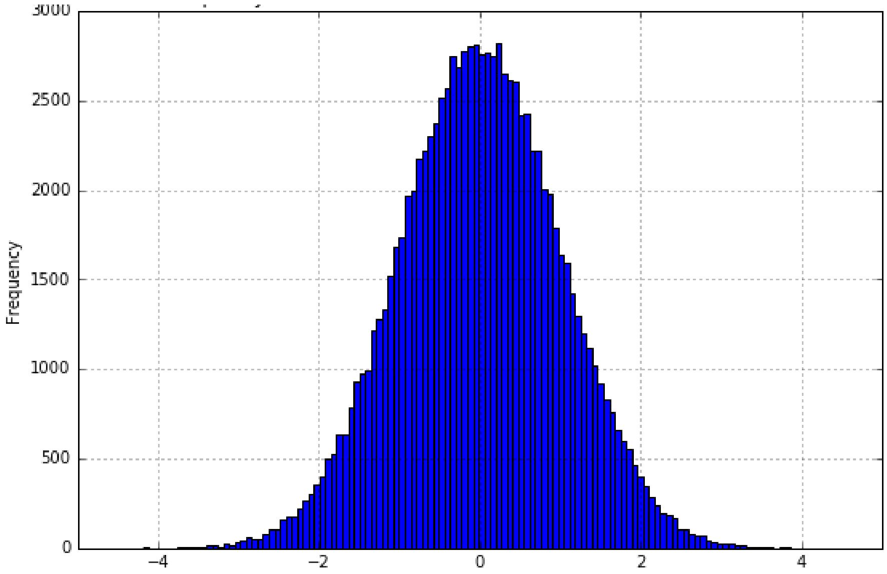


The curve in a PDF defines the a cumulative distribution function for the variable, in which the function of a value is less than the probability of the variable having that value; or expressed as a formula, $F(x) = P(X < x)$.

Central Limit Theorem

When an independent fair test is repeated multiple times, it turns out that there's a naturally occurring formula that determines the distribution of the results. This is called the *Central Limit Theorem*.

For example, suppose you performed the invoice selection test discussed earlier multiple times a day over a period of multiple days with the same set of invoices, and tabulated the mean (average) number of times the test succeeded (i.e. you select a batch of five invoices that includes two Monday invoices). You could record the results as a histogram that shows the frequency of successful tests each day. The histogram would look similar to the following:



The curve in this histogram represents a *normal distribution*, in which the most frequently occurring values are at the center of the distribution, and values become less frequent as they get closer to the edges of the range of possible values.