Features of Probability Distributions

Central Tendency: The Expected Value

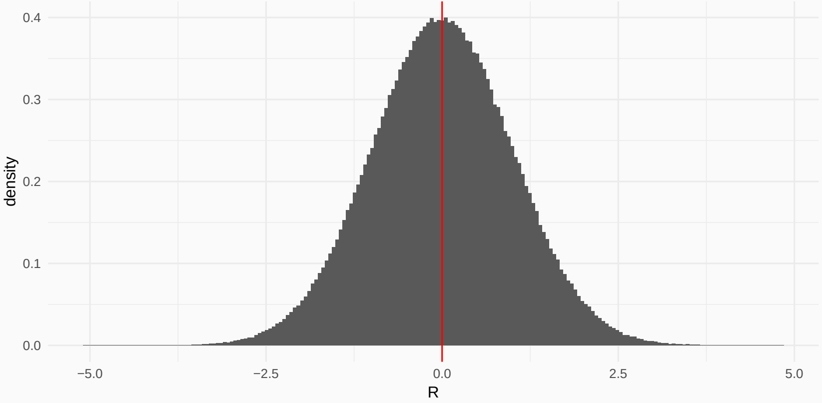
The most commonly used feature of a probability distribution is the *expected value*. It splits the probability under the *pdf* in half. It is calculated as a weighted average of the random variable:

$$\mathbb{E}[X] = \sum x_i f(x_i)$$

The expected value is the value from a probability distribution, for which the probabilities of another value being greater or smaller than it are exactly equal.

For example, we simulate a random variable R from a standard normal distribution. First, we simulate the variable and draw a histogram with the arithmetic mean:

```
set.seed(1337) #seed for reproducibility
data <- data.frame(R = rnorm(n = 1000000))
```



It looks like the area to the left of the mean is the same size at the area on the right side. Let's check that:

```
A <- numeric() # Variable for storing the area sizes

A[1] <- (sum (abs (data$R[data$R < mean(data$R)])) /sum(ab
A[2] <- (sum (abs (data$R[data$R > mean(data$R)])) /sum(ab

print(A)
```

```
## [1] 0.5007953 0.4992047
```

As we can see (and calculate), the area sizes are very similar, meaning that half of the probability mass is on each side of the mean. Notice that since we're in a simulation context, the areas are not *exactly* the same.

Useful Rules for Expected Values

For any constants a and b and a random variable X, the following rules apply:

- The expected value of a constant is the constant

$$\mathbb{E}[c] = c$$

- The expected value of a linear function of a random variable is the linear function of the expected value of the variable

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

- The expected value of a sum of random variables is the sum of the expected values of these random variables:

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} \mathbb{E}[a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

- For a variable Y that is independent of X it holds that:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[X]$$

- The population mean is often denoted by μ or μ_X instead of $\mathbb{E}[X]$

Variablity: Variance and Standard Deviation

The central tendendcy tells us, around which value the outcomes of the random variable cluster, but it is also important to get a measure on how far they *spread around the mean*. The variance and standard deviation are measures for this.

You can see this notion of *distance from the mean* in the definition of the variance:

$$Var[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

The deviation from the expected value $(X - \mathbb{E}[X])$ is squared in order to prevent perfect cancellation of deviations, since:

$$\mathbb{E}\left[(X - \mathbb{E}[X])\right] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

Useful Rules for Variances

For any constants a and b and a random variable X, the following rules apply:

- Alternative representation

$$Var[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

- Variance of a constant is zero:

$$Var[c] = 0$$

- Variance of a linear combination:

$$Var[aX + b] = a^2 Var[X]$$

- The population variance is often denoted by σ^2 or σ_X^2 instead of $Var[X]$.

The Standard Deviation

If we change the unit of measurement of X by $a = 1000$, e.g. from kilometers to meters, the variance increases linearly by $a^2 = 1000 * 1000$ (see rules above). Since this makes it difficult to compare the variability of different variables, we use the standard deviation, which is the positive part of the square root of the variance:

$$sd[X] = +\sqrt{Var[X]}$$

For any constants a and b and a random variable X, it has the following useful properties: - Standard deviation of a linear combination:

$$sd[aX + b] = |a| sd[X]$$

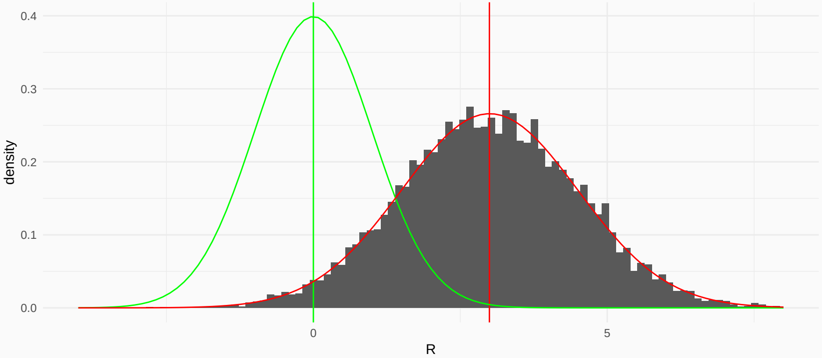
- The standard deviation is often denoted by σ or σ_X instead of $sd[X]$.

Standardisation of Variables

Using the properties of the expected value and variance, we can transform any variable to a standardised variable with mean 0 and standard deviation 1:

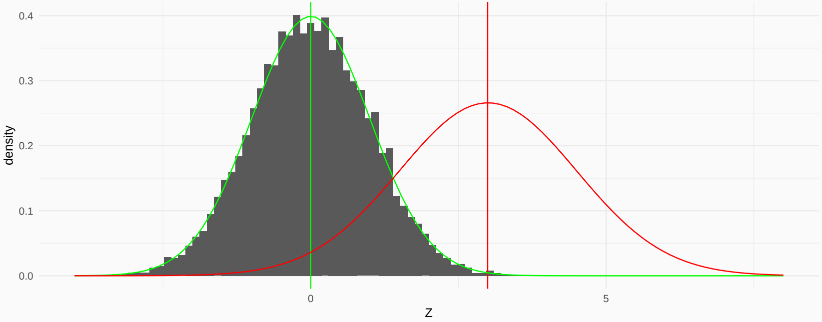
$$Z := \frac{X - \mu}{\sigma}$$

We can demonstrate this with a short simulation. In the following we simulate a random variable X from a normal distribution, that has mena 3 and standard deviation 1.5



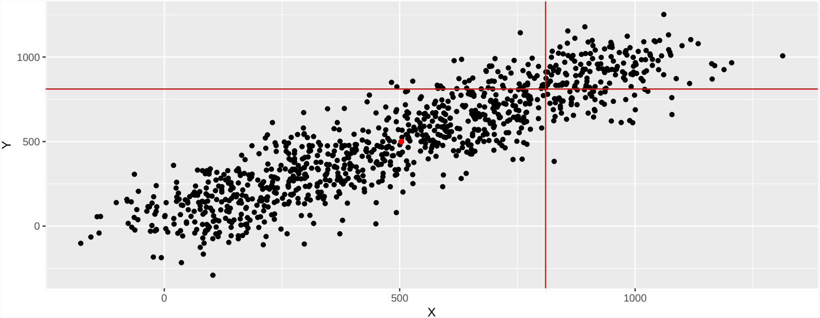
The green lines indicate a standard normal distribution with $\mu = 0$ and $\sigma = 1$. Our simulated varaible clearly deviates from it.

If we now transform X according to the formula above, the result yields:



Association: Covariance and Correlation

The covariance captures the joint variation of two random variables. It allows us to get a sense of the relationship of them, e.g. X is always high when Y is. This can be visualised graphically:



We can easily see that X and Y have a strong relationship. We can stat that: when the deviation of X (vertical line) from its mean (red dot) is high, the same can be said about the deviation of Y (horizontal line).

Expressing this relation of deviations from the mean mathematically yields us the formula of the covariance:

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

Useful Rules for Covariances

For any constants a and b and random variables X, and Y the following rules apply:

- Alternative representations:

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[(X - \mathbb{E}[X])Y] \\ &= \mathbb{E}[X(Y - \mathbb{E}[Y])] \end{aligned}$$

- Covariance given independence of X and Y:

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= 0 && \text{since } \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[X] \end{aligned}$$

- Variance of linear combinations:

$$Cov[a_1X + b_1, a_2Y + b_2] = a_1a_2Cov[X, Y]$$

- The boundary of the covariance of two random variables (its most extreme value) is given by the product of their standard deviations. This is also called the *Cauchy-Schwartz Inequality*:

$$|Cov[X, Y]| \leq sd[X]sd[Y]$$

- The population covariance is usually denoted by σ_{XY} instead of $Cov[X, Y]$