MMaD: Lecture 4 handout

Combing probability distributions

In general, if X and Y are independent random variables, then:

$$\text{Mean}(X \pm Y) = \text{Mean}(X) \pm \text{Mean}(Y)$$

and

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

Note that the variances are never subtracted!

Combining normal distributions

Applying this to normally-distributed variables in particular, adding and subtracting them will give a new variable that also obeys a normal distribution.

If we have two normally-distributed variables:

$$X \sim N(\mu_x, \sigma_x^2) \quad \text{and} \quad Y \sim N(\mu_y, \sigma_y^2)$$

And if we combine them to create new variables:

$$S = X + Y \quad \text{and} \quad D = X - Y$$

Then:

$$S \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2) \quad \text{and} \quad D \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$$

Example 2

A company assembles replica swords that consist of two components: the blade and the hilt (i.e. the handle).

- The steel blades B (excluding the tang that overlaps with the hilt) are manufactured with lengths (in cm) given by a normal distribution $B \sim N(85, 4)$.
- The wooden hilts H have length obeying $H \sim N(20, 1)$.

What is the probability that a particular sword exceeds 106 cm in total length?

Outliers and Chauvenet's criterion

When collecting data, we may find that some values lie very far outside of the "expected" range, which we call **outliers**. This may be due to an error in the measurement, or a random fluctuation.

Chauvenet's criterion provide a systematic method of classifying outliers. It **only** applies to data samples that are known to have been taken from a **normally-distributed population**.

The idea is to take a value from our sample and determine how many data points in a sample of this size we would expect to be as far from the mean as this value we are testing. If it is less than one, we would not expect such an unlikely value to occur, and so can consider it an outlier.

Procedure for applying Chauvenet's criterion in EXCEL

Given a set of N data points $\{x_i\}$ drawn from a normal distribution:

1. Determine the mean of the sample:

   $$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

   and the sample standard deviation σ.

2. For each data point x_i, determine how many standard deviations D away it is from the mean, using:

   $$D = \frac{|x_i - \mu|}{\sigma}$$

3. Calculate the probability $P_{x<D}$ in the standard normal distribution using the Excel function `NORMSDIST(D)`

4. Calculate:

   $$P_{x>D} = 1 - P_{x<D}$$

   This is the probability of a point being at least D standard deviations *above* the mean.

5. Calculate the number of data points in a sample of size N that we would expect to be *beyond* this, using:

   $$N_E = N \times P_{x>D}$$

6. If this value is less than 0.5, reject data point x_i as an outlier.