Combining distributions & Chauvenet's criterion

Dr Gavin M Abernethy

Contents

Today we will cover. . .

- Combining probability distributions.
- Rules for combining normal distributions.
- What are outliers in data?
- Classifying outliers using Chauvenet's criterion.

Combining Data Sets

In some cases, we wish to ask questions concerning combinations of data sets.

For example, say one variable is the diameter of a bored hole, and another is the width of the bolt that needs to fit within it.

If we know the mean and standard deviation of the holes and of the bolts, can we determine how many pairs of one bolt and one hole are likely to fit?

To answer such questions, we need to be able to combine probability distributions.

Combining Data Sets

Consider two sets of metal bars, A and B. Set A are known to have mean $\bar{a}$ and variance σ_a^2, while set B has mean $\bar{b}$ and variance σ_b^2.

One bar from set A and one from set B are chosen at random. What is the distribution of the length of the two bars combined?

It turns out that the new mean is the sum of the two means:

$$\bar{l} = \bar{a} + \bar{b}$$

And the new variance is the sum of the two variances:

$$\sigma_l^2 = \sigma_a^2 + \sigma_b^2$$

Adding and Subtracting Probability Distributions

Combining random variables

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!

Adding and Subtracting Normal Distributions

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

Combining normally-distributed variables

If:

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

And if:

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

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 1

Two companies produce ball bearings whose masses obey normal distributions. Company A's products have mass in grams given by $A \sim N(6.1, 0.5^2)$, while Company B's products have mass $B \sim N(5.8, 0.5^2)$. What is the probability distribution of the *difference* in the masses of ball bearings between the companies?

Let variable D be the difference, in terms of how much larger a ball bearing made by Company A is than one made by company B:

$$D = A - B$$

Example 1

Two companies produce ball bearings whose masses obey normal distributions. Company A's products have mass in grams given by $A \sim N(6.1, 0.5^2)$, while Company B's products have mass $B \sim N(5.8, 0.5^2)$. What is the probability distribution of the *difference* in the masses of ball bearings between the companies?

Let variable D be the difference, in terms of how much larger a ball bearing made by Company A is than one made by company B:

$$D = A - B$$

Hence, subtract the means, but add the variances:

$$D \sim N(6.1 - 5.8, 0.5^2 + 0.5^2)$$

which can be simplified to either:

$$D \sim N(0.3, 0.5) \qquad \text{or} \qquad D \sim N(0.3, 0.707^2)$$

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?

Example 2 - Solution (I/III)

Let variable S be the length of a sword given by the combination of a blade B and hilt H. So:

$$S = B + H$$

where

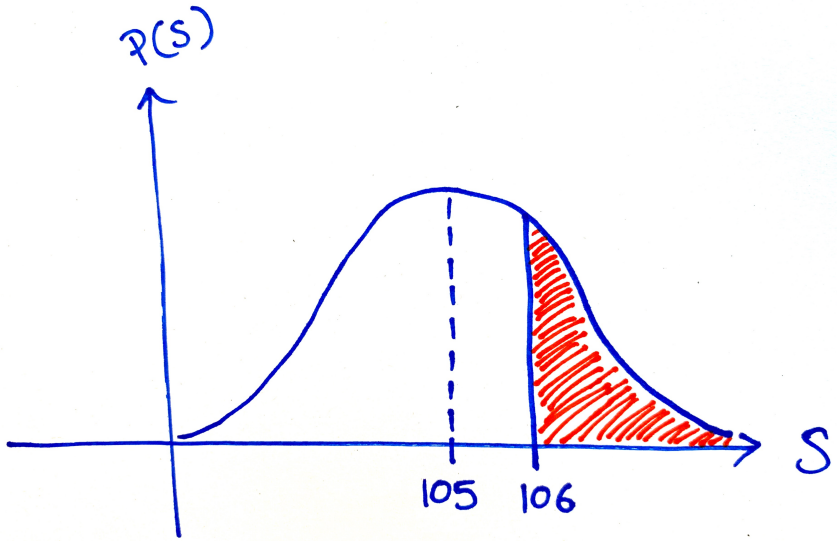
$$B \sim N(85, 4) \quad \text{and} \quad H \sim N(20, 1)$$

To *add* two normally-distributed variables, add the means and variances to obtain a new normal distribution:

$$S \sim N(85 + 20, 4 + 1) \implies S \sim N(105, 5)$$

Example 2 - Solution (II/III)

Then we want to find $P(S > 106)$:



Convert $s = 106$ to its z-value in the standard normal distribution:

$$z = \frac{|s - \mu|}{\sigma} = \frac{106 - 105}{\sqrt{5}} = \frac{1}{\sqrt{5}} = 0.4472$$

So 106 is 0.45 standard deviations above the mean to 2 d.p.

Example 2 - Solution (III/III)

From the standard normal distribution table, we find that this corresponds to an area of 0.1736. Hence:

$$P(105 < S < 106) = 0.1736$$

and so

$$\begin{aligned} P(S > 106) &= P(S > 105) - P(105 < S < 106) \\ &= 0.5 - 0.1736 \\ &= 0.3264 \end{aligned}$$

So approximately 33% of swords will exceed 106 cm in total length.

Chauvenet's criterion

When collecting data, we may find that some values lie very far outside of the "expected" range, called **outliers**. This could be due to an error in the measurement, or simply a random fluctuation. There are statistical procedures that can help us decide whether a data point should be rejected from the data set or not. This can be controversial as some scientists believe that one should never discard any measurements. Chauvenet's theorem provide one systematic method of doing so.

Chauvenet's theorem is a method to classify data points as outliers. It **only** applies to samples that are known to have been taken from a **normally-distributed population**.

Chauvenet's criterion

The process involves checking whether a data point lies beyond a certain distance (more than so many standard deviations) away from the mean. It is a way of asking "is this sample *representative* of the population from which is was drawn"?

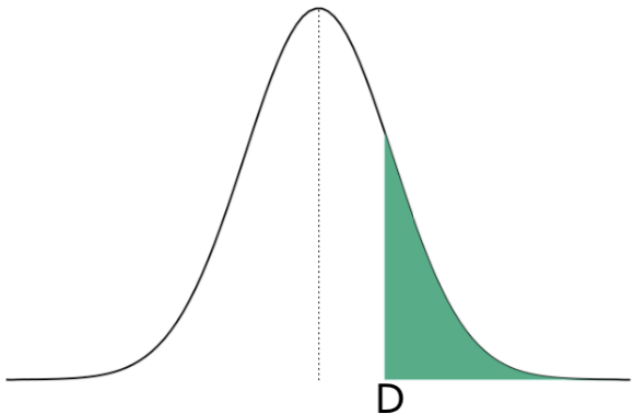
Given a particular data point x, the number of standard deviations σ it is from the mean μ is given by:

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

Chauvenet's criterion

To calculate the number of expected points that far from the mean, we first need to know the probability ($P_{x>D}$) of finding a point beyond D in the *standard* normal distribution.

We can do this using the Excel function NORMSDIST, which returns one minus this area (i.e. the size of the white area).



Chauvenet's criterion

Then the number N_E of points, from our sample of size N, that are expected to lie farther away than D above the mean is:

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

Finally:

Chauvenet's Criterion:

If N_E is less than 0.5, the data point is rejected.

If $N_E < 0.5$, that would mean that (when we consider the probability of also being at least D standard deviations *below* the mean) that overall we expect *less than one point in a sample this size* to be that far from the mean, and so we may discount it as not being representative of the population.

Chauvenet's theorem for a set of N data points $\{x_i\}$:

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 from the mean it is, 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)

Chauvenet's Theorem for a set of N data points $\{x_i\}$:

4. Calculate:

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

So this gives us the probability of a point being at least D standard deviations *above* the mean.

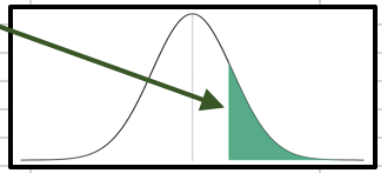
5. Calculate the number of data points in a sample of size N that we would expect to be beyond this, using:

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

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

Example: MMaDLectureExampleChauvenetsTheorem.xlsx

	A	B	C	D	E	F	G	H
1	Data	Deviation from Mean	Number of SD's from Mean	Area below	Area of outliers	No. expected values outside	Accept/Reject	Data, outliers removed
2	6.01	=A2-$B$18	=ABS(A2-$B$18)/$B$19	=NORMSDIST(C2)	=1-D2	=E2*12	=IF(F2<0.5,"Reject","Accept")	=IF(G2="Accept",A2,"")
3	0.03	=A3-$B$18	=ABS(A3-$B$18)/$B$19	=NORMSDIST(C3)	=1-D3	=E3*12	=IF(F3<0.5,"Reject","Accept")	=IF(G3="Accept",A3,"")
4	6.77	=A4-$B$18	=ABS(A4-$B$18)/$B$19	=NORMSDIST(C4)	=1-D4	=E4*12	=IF(F4<0.5,"Reject","Accept")	=IF(G4="Accept",A4,"")
5	7.93	=A5-$B$18	=ABS(A5-$B$18)/$B$19	=NORMSDIST(C5)	=1-D5	=E5*12	=IF(F5<0.5,"Reject","Accept")	=IF(G5="Accept",A5,"")
6	7.48	=A6-$B$18	=ABS(A6-$B$18)/$B$19	=NORMSDIST(C6)	=1-D6	=E6*12	=IF(F6<0.5,"Reject","Accept")	=IF(G6="Accept",A6,"")
7	6.34	=A7-$B$18	=ABS(A7-$B$18)/$B$19	=NORMSDIST(C7)	=1-D7	=E7*12	=IF(F7<0.5,"Reject","Accept")	=IF(G7="Accept",A7,"")
8	6.72	=A8-$B$18	=ABS(A8-$B$18)/$B$19	=NORMSDIST(C8)	=1-D8	=E8*12	=IF(F8<0.5,"Reject","Accept")	=IF(G8="Accept",A8,"")
9	7.45	=A9-$B$18	=ABS(A9-$B$18)/$B$19	=NORMSDIST(C9)	=1-D9	=E9*12	=IF(F9<0.5,"Reject","Accept")	=IF(G9="Accept",A9,"")
10	7.03	=A10-$B$18	=ABS(A10-$B$18)/$B$19	=NORMSDIST(C10)	=1-D10	=E10*12	=IF(F10<0.5,"Reject","Accept")	=IF(G10="Accept",A10,"")
11	13.01	=A11-$B$18	=ABS(A11-$B$18)/$B$19	=NORMSDIST(C11)	=1-D11	=E11*12	=IF(F11<0.5,"Reject","Accept")	=IF(G11="Accept",A11,"")
12	6.63	=A12-$B$18	=ABS(A12-$B$18)/$B$19	=NORMSDIST(C12)	=1-D12	=E12*12	=IF(F12<0.5,"Reject","Accept")	=IF(G12="Accept",A12,"")
13	7.16	=A13-$B$18	=ABS(A13-$B$18)/$B$19	=NORMSDIST(C13)	=1-D13	=E13*12	=IF(F13<0.5,"Reject","Accept")	=IF(G13="Accept",A13,"")
14								
15								
16	Original	Data		Data with outliers	removed			
17	Number of points	12		Mean	=AVERAGE(H2:H13)			
18	Mean	=AVERAGE(A2:A13)		Std Dev	=STDEV.P(H2:H13)			
19	Std Dev	=STDEV.P(A2:A13)						
20								
21								
22								



NORMSDIST(x) - where x is the number of standard deviations from the mean: This command gives the area under the standard normal curve in the range from $-\infty$ to x.