

Some Basic Statistics

Chapter 3

It is essential that any brewing operation is founded on reliable data. We must be able to measure things if we are going to interpret what is going on, respond to problems, and make whatever corrections are necessary in our single-minded determination to assure consistent excellence in our beers.

We need two things if we are to make reliable measurements: carefully defined methods and the ability to understand the statistical significance of what we are measuring. In this chapter, we address statistics.

Mean

Let us start with the mean. You might call it the average. Let us say that we found ourselves in a small island nation that boasted 15 breweries, each of which produced four types of beer. Armed with our trusty gas chromatograph (what self-respecting brewing scientist does not travel with one of those in his or her luggage?), we measure the alcohol content by volume (ABV) in each of those beers and we find these values:

5.0 (9 beers), 3.9 (3 beers), 7.2 (2 beers), 2.5 (1 beer) , 5.1 (9 beers), 5.5 (13 beers), 6.2 (5 beers), 9.0 (1 beer), 1.3 (1 beer), 8.2 (1 beer), 5.2 (8 beers), 7.8 (1 beer), 4.3 (6 beers).

We get the mean by adding them all together and dividing the sum by the total number of measurements (60), giving us a value of 5.26.

If we call the mean m , then we can explain this calculation using a simple formula:

$$m = \frac{1}{n} \sum x_i$$

where Σ is “the sum of” (i.e., all the values added together), x_i refers to each individual measurement, and n is the total number of measurements.

Range

Now, if you look at these 60 numbers, you can see that the average ABV is at the middle of quite a huge range of strengths, with a low of 1.3% (no thank you) and a high of 9.0% (okay, but just the one, please). Just by knowing the mean, we really don’t know much about our population of samples. The easiest way is to plot them out (Fig. 3-1). You can then see that actually the greatest number of beers have an ABV of 5.5%, which is above the mean. However, the greater preponderance of beers with a lower ABV renders the mean a little less than this.

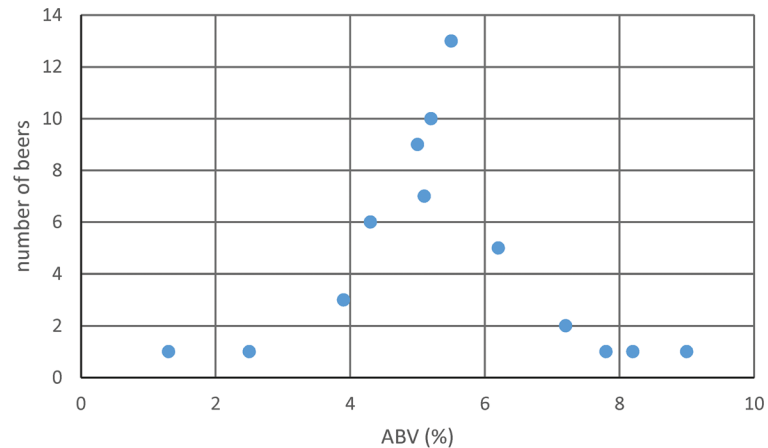


Fig. 3-1. Data distribution about a mean. ABV = alcohol content by volume. (Courtesy C. W. Bamforth—© ASBC)

So how can we express the range in a formula?

The mean absolute deviation (MAD) is given by

$$\text{MAD} = \frac{1}{n} \sum |x_i - m|$$

The $|$ notation indicates that the $+$ and $-$ values are all rendered positive, irrespective of whether the value is lower or higher than the mean. In other words, what we have is the total of all the differences between each value and the mean divided by the total number of measurements. Table 3-1 gives us the necessary values.

Let us put all those values in the right hand column into our MAD formula, and we get a value for MAD of 0.70.

Table 3-1.

Spread of Data in Calculating Mean Absolute Deviation^a

ABV ^b	Difference from the Mean	Total Beers at This ABV	$\sum x_i - m $ for This ABV ^c
1.3	-3.96	1	3.96
2.5	-2.76	1	2.76
3.9	-1.36	3	4.08
4.3	-0.96	6	5.76
5.0	-0.26	9	2.34
5.1	-0.16	9	1.44
5.2	-0.06	8	0.48
5.5	0.24	13	3.12
6.2	0.94	5	4.70
7.2	1.94	2	3.88
7.8	2.54	1	2.54
8.2	2.94	1	2.94
9.0	3.74	1	3.74

^a Courtesy C. W. Bamforth—© ASBC.

^b Alcohol content by volume.

^c x_i = each individual measurement, and m = mean.

The other way to overcome the problem of negative values is to use squared values, and now we have the mean square deviation (MSD):

$$\text{MSD} = \frac{1}{n} \sum [x_i - m]^2$$

Let's do that in Table 3-2. If we calculate MSD according to our formula, we get 1.3.

Table 3-2.

Spread of Data in Calculating Mean Square Deviation^a

ABV ^b	Difference from the Mean ($x_i - m$)	Rounded Square of This ($x_i - m$) ²	Total Beers at This ABV	$\Sigma (x_i - m)^2$ for This ABV ^c
1.3	-3.96	15.6816	1	15.6816
2.5	-2.76	7.6176	1	7.6176
3.9	-1.36	1.8496	3	5.5488
4.3	-0.96	0.9216	6	5.5296
5.0	-0.26	0.0676	9	0.6084
5.1	-0.16	0.0256	9	0.2304
5.2	-0.06	0.0036	8	0.0288
5.5	+0.24	0.0576	13	0.7488
6.2	+0.94	0.8836	5	4.4180
7.2	+1.94	3.7636	2	7.5272
7.8	+2.54	6.4516	1	6.4516
8.2	+2.94	8.6436	1	8.6436
9.0	+3.74	13.9876	1	13.9876

^a Courtesy C. W. Bamforth—© ASBC.

^b Alcohol content by volume.

^c x_i = each individual measurement, and m = mean.

Standard Deviation

And so we arrive at the term that statisticians tend to use and that you are more likely to have encountered, namely, “standard deviation.” We go to the extra stage of taking the square root of what we have just calculated. By doing this, we get a number that is of the same order

of magnitude as the values that we are actually measuring, in this example ABV.

The formula for standard deviation, usually referred to as σ , is

$$\sigma = \sqrt{\frac{1}{n-1} \sum [x_i - m]^2}$$

You will note that rather than n , we are now using $n - 1$. Don't ask! It is no biggie. So our value for σ in this example is 1.1.

If you look at our ABV range (a low of 1.3% and a high of 9.0%), you can see that a standard deviation of 1.1 is not the same value you would get from subtracting the lowest value from the highest one. In short, a small standard deviation signifies that the mean is the average of very similar values. A large standard deviation indicates substantial data spread. The fact that the greater proportion of values in our data set were in the region close to the mean explains why the standard deviation is a smaller number than it would have been if, say, we had a range of beers with equal numbers of samples at each ABV. If we had the latter situation, then the standard deviation would be very close to being the number that, if you subtracted from or added it to the mean value, would have given you the two values at the extremes of the data range.

Normal Distribution

For many data sets, particularly if there are very large numbers of samples, the plotted values adopt what is known as the “normal distribution” (Fig. 3-2). For example, if you measured the ABV of all the beers across the world, then the mean would (I guess—I haven't done this) be around 4.5%. There would be genuine alcohol-free beers at one end and products like Snake Venom (67.5% ABV) at the other. However, the vast majority (I venture) would be in the range of 0.5 to 8.5% ABV, with a progressively decreasing minority down from and above the extremes. The height or weight of the male and female populations is a great example of a normal distribution, though the mean for each is higher than it was a century ago. The shape of the data set is probably the same, but heights on average have increased and weights most certainly have. However, let us stick to beer.

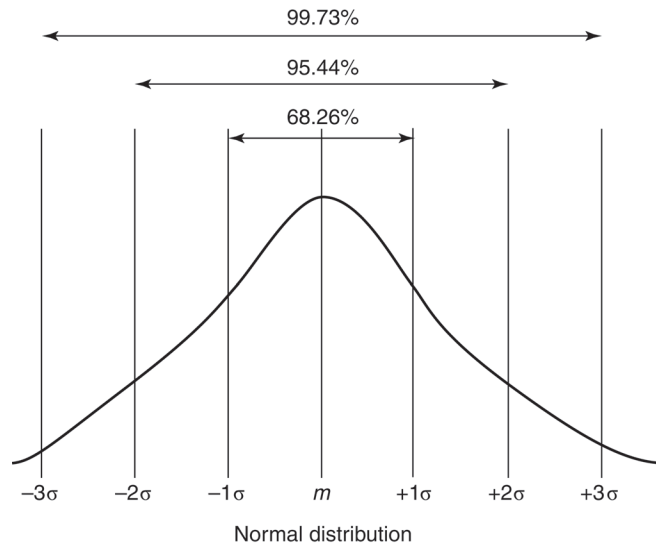


Fig. 3-2. A normal distribution. σ = standard deviation, and m = mean. (Courtesy C. W. Bamforth—© ASBC)

Folks call these plots “bell curves”—look at the shape. The important thing to observe is that 68% (roughly two-thirds) of the measured values are within one standard deviation (σ) above or below the mean (m), while 19 of 20 measurements (95%—actually just a tad higher) are within two standard deviations above or below the mean. Almost all, but not quite (99.7%), are within three standard deviations above and three standard deviations below the mean. I address the significance and application of this in Chapter 4.