CHEM 116 February 24, 2018

Unit 7, Lecture 1

Numerical Methods and Statistics

Companion Reading

Langley Chapter 4

1 Descriptive Statistics of Data

Recall the Lagnely is concerned with practical statistics and Bulmer with probability and theoretical statistics. Now we are going to begin to describe data. Statistics of data and probability theory connect via the Law of Large Numbers and Central Limit Theorem, which we'll learn about later.

The word **statistic** as a singular is a quantitative value that describes some property of a sample. For example, a sample mean is a statistic. Our assumption when computing statistics of data is that the data can be described with probability distribution, albeit with unknown parameters.

1.1 Average

An average is a single value description (statistic) of data that represents its center. The word average has no mathematical definition. There are multiple 'averages' in statistics and you must choose one and be specific when describing it. For probability distributions (PDF/PMF), expected value is almost always used for the average. For data, there are multiple types:

1.1.1 Sample Mean

The most common, the sample mean is what most think of for average:

$$\bar{x} = \frac{1}{N}\sum_x x \tag{1}$$

For example, the sample mean of $2, 5, 11$ is 6. We use the word 'sample' in sample mean to denote it came from data.

1.1.2 Sample Variance / Standard Deviation

Sample variance is similar to variance and describes the spread of data. Its equation is:

$$\sigma_x^2 = \frac{1}{N-1}\sum_x (x - \bar{x})^2 \tag{2}$$

Notice that the denominator has an $N-1$. Read your textbooks to find out more about that. Briefly, if it's not there the sample variance doesn't match up with the variance from a PDF.

The standard deviation is σ_x, the square root of the variance. It's more often used in engineering because it's in the same units as the sample mean.

1.2 Law of Large Numbers

The Law of Large Numbers says that if you gather enough data that comes from a particular probability distribution, then its expected value and sample mean will converge to the same value. For example, if you flip a fair coin 10 times you wouldn't expect to see exactly 5 heads and 5 tails. However, as the number of flips gets larger, you would expect the fraction of heads to approach 0.5. Mathemtically, we say that the expected value converges to the sample mean *in distribution.* In distribution means that if we had an infinite number of data points, the sample mean and expected value would be the same. The Law of Large Numbers is the connection between stastics and probability theory. It implies that sample variance converges to variance in distribution and connects many other such statistics to probability expressions.

1.3 Mode

Another type of average is the **mode**. It is the most frequently observed value in the sample set. When data is not discrete, the mode is ill-defined. The mode is useful on categorical data, like letters or type of pet owned. Sample mean cannot be calculated with such data, but mode is defined there. Let's see an example for integers:

The mode from 5,27,23,4,5,28,19,5,30,2 is 5 because it appears more than any other number.

1.4 Median

If your samples are sorted into a list, the median is the value at the center of that list. It represents the midpoint of your sample. Sample *means* are suspectible to skew and outliers. A median value is not sensitive to those. It is generally a better statistic to use in all cases, but challenging to compute. If the sample size is an even number, the median is the arithmetic mean of the two center elements.

An example from previously: 5,27,23,4,5,28,19,5,30,2

$$2, 4, 5, 5, \underbrace{5, 19}_{\text{median elements}}, 23, 27, 28, 30$$

$$\text{median} = 12$$

1.5 Quartiles

The median indicates the divsion between the lower 50% of values and the upper 50% of values. We can define two more numbers that indicate the lower 25% and upper 25%. These dividing numbers are called quartiles and are indicated as Q_1, Q_2 (also the median), and Q_3. The quartiles need not be actual values in the data, just like how sometimes the median is the arithmetic mean if we have an even number of values. For example, consider the numbers 4, 6, 13, 22. The median, which is also Q_2, is $(6+13)/2 = 9.5$. Q_1 is the median of the values below Q_2: the median of 4, 6. Thus $Q_1 = 5$. Similarily, Q_4 is the median of 13 and 22: 17.5. Now arrange the data and quartiles: $4, |\underbrace{5}_{Q_1}|, 6, |\underbrace{9.5}_{Q_2}|, 13, |\underbrace{17.5}_{Q_3}, |22$. Notice how the quartiles split the data into 4 groups, each containing 25% of the values.

Let's do one more example with an odd number of data points. Consider: 8, 1, -2, 3, 5, 11, 3. First rearrange:

$$-2, 1, 3, 3, 5, 8, 11$$

We can see that 3 is the median. Now to compute Q_1, just like in the above example *we do not include the median in the calculation.* So we get that Q_1 is the median of -2, 1, 3. $Q_1 = 1$. Similarily, $Q_3 = 8$.

1.6 Quantiles

Quartiling can be generalized to split up the data into any sized grouping. This is called *quantiling*. For example, you could have 100 splits in your dataset indicating the 100 percentiles. You would have a value showing where the bottom 1% is and another showing where the next 1% is, all the way to top 1% (99th percentile). Quantiling can split data into thirds, tens, 100 groups, etc.