**CHEM 116** February 9, 2018

# Unit 6, 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. We won't see how statistics of data and probability theory connect for a little while.

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.

We assume for these things that they are coming from a distribution. [Draw a picture]

## 1.1 Average

An average is a single value description 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 $\sigma_x$, the square root of the variance. It's more often used in engineering because it's in the same units as the sample mean.

### 1.1.3 Mode

The mode is the most frequently observed value in the sample set. When data is not discrete, the mode is ill-defined. It's generally only useful for data from sets or integers.

For example, the mode from 5,27,23,4,5,28,19,5,30,2 is 5.

### 1.2 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. Using a sample mean implies your data is normally distributed. A median value is an average that makes no assumptions. It is generally a safer statistic to use, 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

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

$$\text{median} = 12$$

## 2 Covariance

Covariance describes the relationship between two random variables changing. A positive covariance means the both move in the same direction. A negative covariance means they move in opposite directions. The magnitude of the covariance contains information about both the two random variables' variances and the relationship between the two. It's defined as:

$$\operatorname{Cov}(X, Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right] \tag{3}$$

$$\operatorname{Cov}(X, Y) = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] \tag{4}$$

**This is for working with random variables, not data!**. We can use some math and such to come up with the following properties of covariance:

1. $\operatorname{Cov}(X, Y) = 0$, if $X$ and $Y$ are independent
2. $\operatorname{Cov}(X, Y) \leq \sqrt{\operatorname{Var}(x)\operatorname{Var}(y)}$
3. $\operatorname{Cov}(X, X) = \operatorname{Var}(x)$
4. $\operatorname{Cov}(aX, bY) = ab\operatorname{Cov}(X, Y)$
5. $\operatorname{Cov}(X + a, Y + b) = \operatorname{Cov}(X, Y)$
6. $\operatorname{Cov}(W + X, Y + Z) = \operatorname{Cov}(W, Y) + \operatorname{Cov}(W, Z) + \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z)$

### 2.1 Example

Let's say body temperature has a $\mu = 98.6°\text{F}$ and $\sigma = 0.5°\text{F}$. A fever temperature is $1.1 \times T$ ,where $T$ is body temperature. What's the covariance between fever and body temperature?

$$\operatorname{Cov}(F, T) = \operatorname{Cov}(1.1 \times T, T) = 1.1 \times \operatorname{Cov}(T, T) = 1.\operatorname{Var}(T)$$

$$\operatorname{Cov}(F, T) = 1.1 \times \sigma^2 = 0.55°\text{F}$$

After exercise, your temperature is $E = T + I$, where $I$ is an exponentially distributed random variable with $\lambda = 0.25$. What's the covariance between body temperature and post-exercise body temperature?

$$\text{Cov}(E, T) = \text{Cov}(T + I, T) = \text{Cov}(T, T) + \text{Cov}(I, T) = \text{Var}(T)$$

$$\text{Cov}(E, T) = 0.5°\text{F}.$$

# 3 Sample Covariance

Just like sample variance and sample mean, there is a sample covariance. You must have $N$ sets of pairs of data to compute sample covariance. This is the first time we've been working in pairs by the way. The data muse be matched, meaning you are measuring two random variables in the same sample space. It might be that your sample space is a product space; but there must be some pairing in the data.

Examples of invalid 'pairs':

1. How much it snowed today and the total snowfall of the week
2. You have two groups. Group A gets a drug and group B gets a placebo. You match each the people up in the group. Are these paired data?

Examples of valid 'pairs':

1. You have people try exercise for 5 weeks and then stop for 5 weeks. You take their weights after each 5 week period.
2. You measure a planet's diameter and brightness.

The formula is for sample covariance with your $N$ paired data is:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i}^{N} (x - \bar{x})(y - \bar{y}) \tag{5}$$

Following this notation, sometimes people write sample variance as $\sigma_{xx}$ instead of $\sigma_x^2$. The reason that $N - 1$ is not $N - 2$ is that $N$ is the number of *pairs* of data points. That means that we only remove one degree of freedom when we calculate the mean of $x$ and $y$.

## 3.1 Covariance Matrix

You can write out all covariances/variances in a matrix like so:

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$$

This is called a covariance matrix. The diagonals are variances and the off-diagonals are covariances. The covariance can be larger, depending on the number of random variables, but it's always square.

# 4 Sample Correlation

One of the properties of covariance, and thus sample covariance, is that $\mathrm{Cov}(X, Y) \leq \sqrt{\mathrm{Var}(x)\,\mathrm{Var}(y)}$. That means there is a maximum value. And of course the minimum magnitude is 0. Thus we can rescale covariance to get correlation:

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{6}$$

This is the equation for sample correlation. It runs from $-1$ to 1 and removes the variance of the two random variables from the equation.