CHEM 116 February 27, 2020

Unit 7, Lecture 3

Numerical Methods and Statistics

Companion Reading

Langley Chapter 4

1 Lecture Goals

1. Know how to compute covariance of two random variables
2. Know the difference between sample covariance and covariance
3. Know the difference between covariance and Correlation
4. Be able to interpret a covariance or correlation matrix

2 Descriptive Statistics of 2D Data

Now we will consider data with 2 dimensions. The underlying probability distributions that we assume describes this data now have 2 random variables. We'll return to probability theory to examine this situation and then use statistics to analyze 2D data.

3 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:

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \tag{1}$$

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \tag{2}$$

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. $\mathrm{Cov}(X, Y) = 0$, if X and Y are independent
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(x)$
3. $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$
4. $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$

5. $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$

3.1 Example

Let's say the normal body temperature for a person follows a $\mu = 98.6°\text{F}$ and $\sigma = 0.5°\text{F}$ normal distribution. A fever temperature is exactly $1.1 \times T$,where T is their body temperature. What's the covariance between fever and body temperature? We are saying that body temperature of a random person follows a random variable and fever is an analytic, deterministic function of body temperature. Now we're asking about their covariance.

$$\text{Cov}(F, T) = \text{Cov}(1.1 \times T, T) = 1.1 \times \text{Cov}(T, T) = 1. \text{Var}(T)$$
$$\text{Cov}(F, T) = 1.1 \times \sigma^2 = 0.275°\text{F}^2$$

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? Notice that now we have two *independent* random variables I and T, and some summation of the two of them: E.

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

$$\text{Cov}(E, T) = (0.5°\text{F})^2 = 0.25°\text{F}^2.$$

4 Sample Covariance

Just like sample variance and sample mean, there is a sample covariance. It is connected by the Law of Larger Numbers to covariance. To compute, 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.

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:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i}^{N} (x - \bar{x})(y - \bar{y}) \tag{3}$$

Following this notation, sometimes people write sample variance as σ_{xx} instead of σ_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.

4.1 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.

5 Sample Correlation

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

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4}$$

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.