CHEM 116 January 16, 2020

Unit 1, Lecture 1

Numerical Methods and Statistics

Companion Reading

- Bulmer, Pages 12-25
- Langley, Pages 11-27

1 Course Structure

1.1 Goal of course

Real world problems aren't perfect, they require statistics. Real world problems require computers. This course will equip you for that.

1.2 Components of Course

Probability Theory: The analysis of randomness of sets and variables.

Statistics: The analysis of data using probability theory

Numerical Methods: The solution of numerical problems with algorithms

1.3 Course Textbooks

Principles of Statistics A book concerned with the theory of statistics. It has hilarious examples and seems to be set in the Charles Dickens universe

Practical Statistics Explained A practical handbook concerned with the application of statistics

Code Academy A simple introduction to python: www.codecademy.com/learn/python

Hacker Rank Progressively difficult Python exercises: hackerrank.com

Python in Easy Steps 2nd edition Accessible book covering Python basics. Bright colors, friendly cartoons that talk to you in the margins.

1.4 Grading

This course is primarily electronic so homework and projects are more heavily weighted.

1.5 Policies

See syllabus

1.6 Example problems

- Assessing significance. For example, is a new safety protocol significantly effective?
- Fitting data to kinetic rate laws
- Solving equations numerically ($\sin(x) = x$)
- Create a website containing equations, graphs and text for distributing work

1.7 Projects

The goal of the projects are for you to demonstrate that you know how to apply numerical methods and statistics on open-ended problems. Your project should be on something interesting to you and your target audience should be someone who knows little about statistics or numerical methods. You should be demonstrating how numerical methods and statistics can solve complex problems.

1.8 Python

1. Good introductory language, used by 8/10 top CS departments
2. Wide use out of chemical engineering. Python is in top 4 programming languages, along with C, C++, Java.
3. Less used than MatLab and Excel in chemical engineering, but growing
4. Many libraries in engineering and significantly more than MatLab outside of engineering
5. These libraries allow us to mix cool things together, like live websites with instrument data
6. Heavily used in machine learning, deep learning, and AI
7. Programs will be easy distribute, easy to install, easy to view by people with no programming experience

2 Probability Theory

2.1 Sample spaces

Sets, ordered sets, integers, real numbers. We denote a sample space with Q and the number of elements in it with $|Q|$.

Examples:

1. $\{A, B\}$ (set)
2. The roll of a die: $1 - 6$ (integers, which is an ordered set, which is a set)
3. Cards in a 52 card deck (ordered set, which is a set)
4. All possible molecules formed in a chemical reaction (set)
5. Electronic configurations of a molecule (set, possibly ordered with energy)
6. Flow rate into a tank (real number)
7. Temperature (real number)
8. Value of cryptographic key (integer)
9. Pet type owned (cat, dog, fox) (set)

Think about what is ordered, what is a set, what are real numbers, etc

2.2 Probability of sample spaces

Probability assigns a number P to each sample x in the sample space Q. The only requirement is that the sum of $P(x)$ over the sample space is 1. That condition is called normalization.

Notice that even if $\sum_X P(x) > 1$, as long as it's finite we can *normalize* the probability by dividing it by a constant such that its sum is 1. Example:

$$P(\text{die roll}) = \text{value of die} \tag{1}$$

is not normalized, because $\sum_X P(x) = 21$. Normalizing it gives:

$$P(\text{die roll}) = \frac{\text{value of die}}{21} \tag{2}$$

2.3 Probability Algebra - For Samples

OR

The probability of sample A or sample B being drawn is exactly:

$$P(A \text{ or } B) = P(A) + P(B) \tag{3}$$

AND

If we draw two samples sequentially (!) and independently (!):

$$P(A \text{ and } B) = P(A)P(B) \tag{4}$$

For now, independence between trials means the outcome of one doesn't affect the probability of the other. An example would be rolling two dice. Each die's roll is independent of the other. Another example would be measuring the height of two people. Generally, sequential trials are independent. One counter example is drawing cards from a deck of cards. If you draw an ace first, there are less aces in the deck on your next draw so the two card draws are not independent.

NOT

$$P(\sim A) = 1 - P(A) \tag{5}$$

These statements allow us to bridge probabilities together:

```
Draw an ace of spades AND roll a 2 OR roll a 4 AND NOT own a cat
```

Notice that AND is used to bridge together independent samples, OR is used to bridge together multiple possibilities, and NOT is used to "invert" probabilities.

What is wrong with this statement?

```
Draw an ace of spades OR roll a 2
```

You cannot join samples which are from different sample spaces

2.4 Events

Forget what you think the word event means. Events mean a specific thing in probability theory. Choose a set of elements in your sample space. An event occurs if your sample (e.g., the card you draw) is in that set of elements of your sample space. Because a "set" can include any number of elements, you can have events that overlap, events that occupy an entire sample space or be generally messy. Here are some examples:

- Roll an odd number. The sample space is $\{1, 2, 3, 4, 5, 6\}$, the set defining the event is $\{1, 3, 5\}$
- Draw a 9 from a deck of cards. The sample space is 52 cards, the set defining the event is $\{9\}$
- Roll a number less than 3. The sample space is $\{1, 2, 3, 4, 5, 6\}$, the set defining the event is $\{2, 1\}$

The probability of an event is the sum of the probabilities of its elements. For example, If we are rolling a fair die (all probabilities equal), the probability of rolling an odd number is:

$$E = \{1, 3, 5\}, \quad P(E) = \sum_{e \in E} P(e) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \tag{6}$$

AND and **NOT** apply to probabilities of events. **OR** only applies if the events do not overlap (mutually exclusive). For example if event A is roll an odd number $A = \{1, 3, 5\}$ and event B is roll a 3 ($B = \{3\}$), event A includes event B meaning the normal **OR** does not apply. To deal with this, generally you redefine your event or you compute the intersection $P(A \cap B)$ and use:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B) \tag{7}$$

$P(A \cap B)$ doesn't mean the probability of A AND B occurring like above because it is about events whereas the **AND** rule above is about samples. Think instead that $P(A \cap B)$ is the probability of A and B occurring simultaneously. Alternatively, $P(A \cap B)$ is the sum of probabilities of the elements in the sample space that overlap in A and B.

2.5 Independent samples and combination vs permutation

Permutation A particular sequence of elements of a sample space (e.g., the alphabet: ABDEFGH...). **Order matters**

Combination A collection of elements of a sample space. **Order does not matter**

When sampling multiple independent observations, or samples, the **AND** rule applies to permutations. A **combination** of events have no ordering - order does not matter. A particular sequence from those events is called a **permutation** - order matters

Sometimes we do not care about order, like when rolling two dice: $3, 2$ is the same as $2, 3$. There are two permutations, and we must consider both to get the probability of the whole 2 observation combination. This can be done with the **OR** rule:

$$P(3, 2) = P(3) \times P(2) + P(2) \times P(3) = 2 \times P(3) \times P(2) \tag{8}$$

So, to get the probability of a permutation we can just use the **AND** rule. To get the probability of a combination, we must use the **AND** rule for each permutation and use the **OR** rule to combine each permutation that is possible for the combination. For example, a combination of 1,2,3 is the sum of the probabilities of each possible permutation of rolling a 1,2,3:

1. $P(\text{perm} : 1, 2, 3) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$
2. $P(\text{perm} : 1, 3, 2) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$
3. $P(\text{perm} : 2, 3, 1) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$
4. $P(\text{perm} : 2, 1, 3) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$
5. $P(\text{perm} : 3, 1, 2) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$
6. $P(\text{perm} : 3, 2, 1) = P(1) \cdot P(2) \cdot P(3) = \frac{1}{6^3}$

So the probability of the combination of 1,2,3 is $\frac{6}{6^3}$.

2.6 Uniform Probability Sample Spaces

If all samples have equal probabilities probabilities, the probability of a sample is

$$P(x) = \frac{1}{|Q|} \tag{9}$$

where $|Q|$ is the sample space size.

The probability of an *event* occurring is the number of samples in the event, q, divided by the size of the sample space, $|Q|$:

$$P(\text{event}) = \frac{q}{|Q|} \tag{10}$$

For example, the probability of rolling an odd number is 3/6.

The probability of a *combination* occurring is then the number of permutations, n, times the probability of a single permutation:

$$P(\text{combination}) = nP(\text{permutation}) \tag{11}$$

See the above permutation example to see this. Notice this is about multiple observations, whereas the previous two equations are about single observations. For example, consider the probability of rolling an even number and an odd number with a die within two rolls:

$$P(\text{permutation}) = P(\text{even event}) \times P(\text{odd event}) \tag{12}$$

$$P(\text{permutation}) = \frac{3}{6} \times \frac{3}{6} = \frac{1}{4} \tag{13}$$

The number of permutations for rolling an odd and even number within two rolls is two: even then odd or odd then even.

$$P(\text{combination}) = nP(\text{permutation}) = 2 \times \frac{1}{4} = \frac{1}{2} \tag{14}$$

2.7 Concepts to Consider

Independence : For now it means the trials don't affect one another. One way to tell is if the trials can be permuted without changing probability, they are likely independent.

OR : Cannot combine statements in different samples spaces, whereas **AND** can cross sample spaces.

Normalization : As long as your made-up probability measure is finite everywhere, it can be normalized.

Combination vs Permutation : A permutation is a particular ordering of a combination.

Event vs. Observation : I've used observation here to indicate the generation of a sample (outcome). The generation of the sample may correspond to an event, but don't confuse the word 'event' with meaning we generated a sample. An event is a set of samples, where if any elements of the set occur then we say the event occured. An example of an event is the set of $1, 5$ for rolling a die. The observation, or generation of a sample, is the outcome of a die roll (e.g., rolling a 1).