CSE 103: Probability and statistics **Fall 2013**

Topic 1 — Introduction

The goal of these notes is to give you a high-level preview of some of the main concepts that you will learn in this course. We will explain these concepts verbally, with very little math. Later on we will give each concept a precise mathematical definition and show you how to use these definitions to solve practical problems. It is my experience that having a rough intuition about concepts makes it easier to absorb the formal math.

1.1 Statistics

One of the fundamental concepts in probability and statistics is that of *sampling*.

Suppose we own a web site and that we make money by placing advertisements on the pages we host. The way web advertisement works, we get paid if a visitor to the page clicks on the advertisement, indicating that she is interested in the product or service that is advertised (or maybe simply that they are bored...). Such an event is called a "click-through" or CT.

For simplicity, let's consider a single static page, with a single location on the page for presenting the advertisement and just two alternative ads to place in this location, which we will call ad A and ad B. Suppose we don't have any additional information about the specific visitor.

How can we figure out which is the best advertisement to put on the page?

One thing we can do is this: each time the web page is presented, choose one of the possible ads in an alternating manner (B,A,B,A,B,...) and present the page to the visitor. This is referred to as producing an *impression* of the ad. We then see if the visitor clicks on the advertisement, producnig a CT. The fraction of visits to a page that a random visitor clicks on the ad is called the "click through rate" or CTR. If the paymnent for all ads are equal then the ad with the highest CTR is the best one to put on the page.

This is an example of using *Statistics*. A single page-visit, together with a flag indicating whether or not the ad was clicked is called an *observation* and the sequence of n observations is called a *sample*. Statistics is a set of methods for drawing conclusions from random samples. The process that is generating the observations is called a *Stochastic Process*. The conclusions we draw are stated as properties of this stochastic process. In our case, we conclude that one ad has a higher CTR than the other.

Suppose we have 10,000 impressions for each of the two ads, and suppose we got 200 CTs for ad A 200 and 150 CTs for ad B. It seems reasonable to conclude that ad A is better. However, how confident are we of this conclusions? Are we confident enough to never present ad B again? Or do we need to collect more outcomes for both A and B?

Suppose instead that we collected 1,000,000 impressions and got the 20,000 CTs for A and 10,000 CTs for B. The ratios are the same:

$$\frac{200}{10,000} = \frac{20,000}{1,000,000} = 0.02, \quad \frac{150}{10,000} = \frac{15,000}{1,000,000} = 0.015$$

However, intuitively, when the size of the sample is one million, we are more confident that ad A is indeed the better one. How can we justify this intuition?

To explain and quantify this intuition we use probability theory.

1.2 Probability Theory

Probability theory is a branch of mathematics which is the foundation for statistics. This is similar to the way in which discrete math and the theory of computation are the foundation for writing correct and efficient computer programs.

The theory of computation does not partain to a specific programming language or computer hardware. Similarly, probability theory does *not* deal with observations of real-world events, that is the job of Statistics. Instead, probability theory uses simplified mathematical models of the underlying stochastic process.

Specifically, in our case, we can model the process generating the click-through observation using two biased coins. An *unbiased coin* is a symmetric coin for which the probability of landing "heads" is equal to the probability of landing "tails". A *biased* coin is one where these probabilities are different. In our case the coin for ad A has probability p_A of landing "heads" which is equal to the probability that a random visitor will click on the ad. Similarly p_B corresponds to the probability that the coin B lands "heads" which is equal to the probability that a random visitor will click on the ad B. More succinctly, P_A is the CTR of ad A and P_B is the CTR is ad B.

The question that we want to answer is which probability is larger, is $P_A > P_B$ or is the reverse true: $P_B \leq P_B$?

Suppose we are are in the first case in which we presented each advertisement $10,000$ times and got 200 click-throughs for ad A and 150 for ad B. Clearly, it seems that $P_A > P_B$. But how *confident* can we be that this is indeed the case?

To answer this question we consider two alternative *statistical hypotheses*: the first is that $P_A > P_B$, the second is that $P_A \leq P_B$.

We then find the two settings of P_A and P_B that conform with each hypothesis and give the highest probability for the data. For the hypothesis $P_A > P_B$ we get the highest probability when $P_A = 0.02, P_B = 0.015$. For the alternative hypothesis we use $P_A = 0.0175, P_B = 0.0175$, i.e. we place the two probabilities midway between the two observed rates.[1].

We now compute the probability of the observed data corresponding to each of these two settings, which correspond to the two alternative hypotheses. Clearly the probability of the data under the first settings will be higher than the probability of the second, so if we take the ratio of the two probabilities we expect to get a number smaller than 1. The question is, how much smaller than 1?

Computing this ratio is not a trivial matter. You will learn how to do it later in the course. For now, I'll tell you that when the number of impressions for each ad is 10,000 the ratio is 0.6766, while when the number of impressions is 1,000,000 the ratio is smaller than $10^{-16} = 1/10,000,000,000,000,000$. In other words, in the first case we cannot conclude which ad is better with any degree of confidence, while in the second case we can be pretty sure that the first ad is better.

Note how dramatic is the improvement in confidence, increasing the sample size from 10,000 to 1,000,000 transforms a relative gap of 5% from an insignificant gap to one from which can conclude something almost surely. Keep this very small number 10^{-16} in mind, we will get back to it later in the course.

1.3 Low probability vs. certainty

In computer science we are used to giving absolute guarantees: we expect that hardware and software will give the correct answer each and every time. The whole concept of "debugging" a program is based on the assumption that every mistake can be traced back to a particular part of the computer code and thereby eliminated.

[1] Strictly speaking the mid-point is not the highest probability setting, but the difference is small and we ignore it here.

However, in the real world, many computer errors cannot be reconstructed, explained, or corrected. The more realistic goal is becoming to have computer systems that minimize the number of errors without necessarily correcting all of the bugs.

Certainty is golden, however, in the real world, having a guarantee that some error will occur no more than once every 10^{16} attempts is a very strong guarantee. Increasing n from $1,000,000$ to $100,000,000$ will reduce the probability to 1 in 10^{64}.

1.4 Average and Mean

The number of observations, or the size of the sample n is of central importance in probability and in statistics. In general, as n increases we reach the "law of large numbers" which tells us that averages converge to the means.

People sometimes use "mean" and "average" interchangeably. However in probability and statistics the difference is very important. Going back to the click-through example above. consider the sequence of presentations of the ad A, each resulting either in a click (1) or in a non-click (0). For example, suppose the sequence is:

$$0,0,0,1,0,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,...$$

The *average* of the first 5 outcomes is $1/5$. If we compute the average of the first n outcome for $n = 1,2,3,...$ we get a sequence of *running averages*:

$$0,0,0,1/3,1/4,1/5,1/6,1/7,2/8,3/9,3/10,3/11,...$$

As the sequence of outcomes is random, the sequence of averages is also random. By this we mean that if we collected a difference sequence of observations from the same source, we are very likely to get a different sequence. In the language of probability, we say that the outcomes and the averages are *random variables*.

The terms: "empirical", "observational", "sample", "random variable" indicate that the quantity we are referring to is likely to change if we repeat the same experiment.

The mean, on the other hand, is *not* a random variable. The mean is a *constant property of the source*, in this case, the mean is equal to P_A, P_B which are the CTRs for ad A and ad B respectively. In fact P_A and P_B are the quantities that determine which ad is more profitable.[2]

The law of large numbers connects the sequence of running averages and the mean. What the law of large number says (roughly) is that the sequence of averages converges to the mean. In other words, if n is large enough, the difference between the average and the mean is going to be small with high probability.

1.5 Monte-carlo simulations

As I said above, computing the probability flipping a biased coin n times will result in k heads and $n-k$ tails is not trivial and requires some knowledge of probability theory. However, symbolic derivation (also called "closed form solution") is not the only way to arrive at the answer. There is another way which is called "monte-carlo simulations".

A monte-carlo simulation is a computer program that simulates the process of generating the outcomes. It uses "pseudo-random number generators" about which we will learn later on. For now it suffices to describe the pseudo-random number generator as a function `random(p)`. Every time `random(p)` is called it returns a bit whose value is 1 with probability p and 0 with probability $1-p$.

[2] Other terms that are used to refer to the average are the *empirical mean* or the *sample mean*. These can be easily confused with the regular mean. However, they are random variables, not constants.

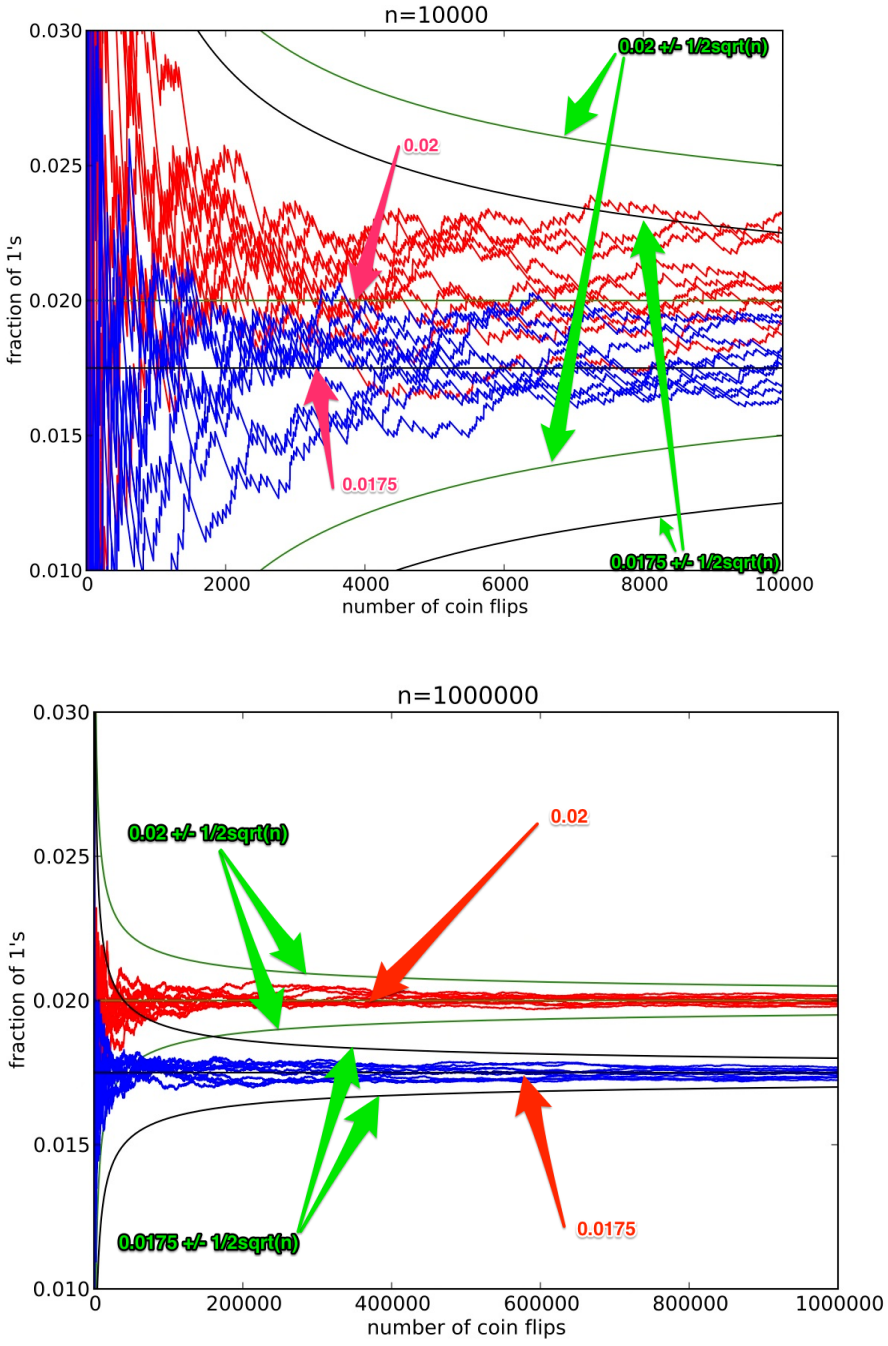

Figure 1.1. Trajectories of running averages of random sequences generated by a biased coin with sides that say "0" and "1". There are two sets of 10 trajectories. The red trajectories correspond to sequences of random coin flips where the probability of 1 is 0.02. The blue trajectories correspond to sequences of random coin flips where the probability of a 1 is 0.0175

Using a random number generator we can create samples of the type described above, compute running averages of these sample, and plot them, see Figure 1.1.[3] These plots demostrate why there is such a qualitative difference between a sample of size $10,000$ and a sample of size $1,000,000$. The red trajectories correspond to sequences generated by a coin with bias $P = 0.02$ while the blue trajectories were generated by a coin with bias $P = 0.0175$ (the most likely value for P_A which allows for $P_A \leq P_B$). Looking at the bottom figure, corresponding to $n = 1,000,000$ we see that the sequences all converge to their mean, as is predicted by the law of large numbers. However, if we look at the top figure, where $n = 10,000$, we see that the red trajectories and the blue trajectories are not well separated. The red and the blue lines cross each other many times.

What does this mean for drawing conclusions about whether $P_A > P_B$? The one experiment that we did corresponds to two sequences of length $10,000$ one for each ad. In this figure are focusing on the sequence for ad A. We know that this sequence ended up at $k = 200$ when $n = 10,000$. We want to compare the probability that this sequence was generated by a coin with bias $P_A = 0.02$ to the probability that it was generated by a coin with bias $P_A = 0.0175$. What we see from the sample is that the probabilities are comparable. It is hard to say what is the ratio of the probabilities, but if instead of generating 10 trajectories we generated 10,000 trajectories we could probably give an accurate estimate of the ratio.

Compare that situatin to the one when $n = 1,000,000$ and you see that now the separation between the two sets of trajectory is perfect. This means that the probability of a blue trajectories having $k = 20,000$ is miniscule.

These graphs give us a useful intuition for the behaviour of running averages of coin flips. We can clearly see the effect of the law of large numbers.

What's more, probability theory tells us that the rate at which the running averages converge to the mean is $O(1/\sqrt{n})$. We demonstrate this in the figures by drawing arround the horizontal line representing the means the envelope which represents the rate at which a random sequence is expected to converge to the mean. You can see that there is a nice fit between the random trajectories and the envelope: none of the trajectories escape out of the envelope.

You might think: if I can do a monte-carlo simulation why do I need probability theory? In fact, in many practical problems monte carlo simulations play an important role. However, recall that probability theory tells you that when $n = 1,000,000$ the probability of the hypothesis $P_A = P_B = 0.0175$ is smaller than 10^{-16}. If you wanted to prove this using monte-carlo, you had to generate at least 10^{16} sequences of lengh $1,000,000$, that is a pretty large computer job! Then consider $n = 100,000,000$ which gives rise to a probability of 10^{-64}, and how much resources will is take to do this monte-carlo simulation. On the other hand, using probability theory and a statistical table you can compute these probabilities quite precisely with no computer at all!

1.6 Summary

In this chapter I have introduced you to many new terms without giving you formal definitions. We will define each term in a precise mathematical way in the coming weeks. I hope that this introduction will help you make sense of the math. The math is necessary if you want to compute probabilities, especially small ones, but keeping the applications in mind will help you develop an intuition for what to expect from stochastic processes.

Here is a list of the probability and statistics terms that we touched upon are:

1. Probability Theory
2. Coin flip, biased and unbiased coins.

[3] The code is available via GitHub: `https://github.com/yoavfreund/CSE103-code/blob/master/MonteCarlo/monteCarlo.py`

3. Statistics
4. Outcome, Sample, sample size.
5. Observational, Empirical.
6. Confidence.
7. Average vs. Mean.
8. Random variable.
9. Monte-carlo simulation.
10. Pseudo-Random number generator.