Probability and Statistics

Tsinghua Math Camp Prof. Paul Horn

Research Projects

Below are several projects you might think about as research projects in probability and statistics. These have many parts: some of them are reasonably simple, and others are quite a bit more involved. If one of them is interesting to you try to start to answer the easy questions and then move on to the more difficult parts. Many of the easy questions even have special cases you can think about to warm up. If it's *all* easy for you, then there are many interesting extensions you could think about. You can always talk to me about it.

1 The n-expectation problem

Question 1 *Let $m_1, m_2, m_3, \dots$ be a sequence of real numbers. What properties of (m_i) ensure that there a random variable X which satisfies:*

(a) X takes on (at most) n values.

(b) $\mathbb{E}[X^n] = m_n$. for $1 \leq n \leq \infty$.

As a start, I'd think about the $[n]$-expectation problem. Suppose X is a random variable taking values in $\{1, 2, \dots, n\}$. If $m_1, m_2, m_3, \dots$ are real numbers what properties of m_i ensure that $\mathbb{P}(X = i) = p_i$ can be chosen so that $\mathbb{E}[X^n] = m_n$.

For $n = 1$ this is really trivial (only the sequence $m_n = 1$ works) and this is very easy for $n = 2$. There are also some clear necessary conditions: clearly $1 \leq m_n \leq n$.

1. Can you find an answer for $n = 3$? For $n = 4$?

2. For general n?

Can you find non-trivial conditions that are necessary *or* sufficient, even if not both?

Here are some questions that one might want to ask:

1. How is m_n related to m_{n+1}? To m_{n+k}? Are there any restrictions? How much freedom is there?

2. What if the values that X takes on is not restricted to being just $\{1, \dots, n\}$ but is an arbitrary n element set. Again, for $n = 1$ the answer is trivial, but try it for $n = 2$ (and then for general n).

3. If this is too easy still, try and think about more general cases: Suppose you have a sequence $m_1, m_2, \dots$; when can you find a random variable X whose support is $[0, 1]$ (or $[0, \infty)$, or $(-\infty, \infty)$, or $\mathbb{N}$, or $\mathbb{Z}$, or ...) and so that $\mathbb{E}[X^n] = m_n$?

As a starting point, here's one non-trivial relation. Suppose that $m_i = \mathbb{E}[X^i]$. Then $m_2 \geq m_1^2$ – this is the same as saying that the variance is non-negative – so there are definitely conditions to be satisfied!

2 Keeping Close

One of the most important measures of a random variable is how close it is to it's expectation with high probability. A classical inequality of this type (that we'll see in class) is that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k) \leq \frac{\mathrm{var}(X)}{k^2}. \tag{1}$$

On the other hand, the central limit theorem (informally) says that many distributions are asymptotically normal. For a normal random variable, we see something much stronger: essentially that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k) \leq e^{-C \cdot k^2}. \tag{2}$$

This is *much* stronger than the previous example. We will see, in class, one way to get stronger bounds than (1), of the form (2), for sums of independent 0/1 valued random variables, but in many cases that's not good enough. This project asks you to look beyond and try to prove similar inequalities for different scenarios.

Some of these can be done by modifying our proof for independent random variables, some require some slightly different ideas, but the ideas are the same.

Here are some things you can try to do:

1. Suppose $X_1, \ldots, X_n$ are independent 0/1 valued random variables (if you like you may assume $\mathbb{P}(X_i = 0) = p$ is also fixed). Consider

   $$Z = X_1X_2 + X_2X_3 + \cdots + X_nX_1.$$

   Then

   $$\mathbb{E}[Z] = p^2n,$$

   The summands are not independent, however. (They are close!) Can you prove an improvement on (1). Can you prove something like (2)?

2. Suppose

   $$Z = f(X_1, \ldots, X_n).$$

   has the property that Z changes by at most one if you change any given X_i. (But it is not the sum of independent random variables. Can you prove any improvement on (1)? Something like (2)?

3. Suppose $X_1, \ldots, X_n$ are independent 0/1 valued random variables (if you like you may assume $\mathbb{P}(X_i = 0) = p$ is also fixed). Consider

   $$Z = \sum_{1 \leq i < j \leq n} X_iX_j$$

   Then

   $$\mathbb{E}[Z] = p^2\binom{n}{2},$$

   The summands are not independent, however. Can you prove an improvement on (1). Can you prove something like (2)?

4. Suppose $X_1, \ldots, X_n$ are independent 0/1 valued random variables (if you like you may assume $\mathbb{P}(X_i = 0) = p$ is also fixed). Consider

$$Z = \sum_{1 \le i < j \le n} X_i X_j X_k$$

Then

$$\mathbb{E}[Z] = p^3 \binom{n}{3},$$

The summands are not independent, however. Can you prove an improvement on (1). Can you prove something like (2)? (How about if you have more than 3 terms in each summand?)

5. If you can't find explicit solutions for $2 - 4$ (which I think are arranged in approximate order of difficulty, but your mileage may vary), can you find functions more like #1 that you can say something about?

3 Abnormal Statistics

.

Note: *There's a bit of a pun in the title. The various things that you are asked to look are related to not-quite-normal random variables. Your goal is to develop some tools and statistical tests to show this, and to (if you get far enough) compare them.*

This project has two parts, a theoretical part and a 'practical' part. Thinking somewhat about the theoretical part is probably useful for doing a good job on the 'practical' part, but there's probably enough in the 'practical' that you can get by without thinking *too* much about the practical.

Theory

The central limit theorem says that if $X_1, X_2, \ldots$ are iid random variables, then

$$\frac{\sum_{i=1}^n X_i - n\mathbb{E}[X_i]}{\sqrt{nvar(X_i)}} \to_d N(0,1).$$

For instance, if $X_1, X_2, \ldots,$ are iid, then the sample mean $(\frac{1}{N}\sum(X_i))$ is asymptotically normally distributed (with mean $\mathbb{E}[X_i]$ and variance $(X_i)/n$, in this case.

In this part of the project I'd like you to develop tests to determine how 'normal' the data is to start with. There are a lot of ways you can go with this: *Note:* The χ^2 test, discussed in class, is of this type – try different things. Be creative! It's best, however, if you can prove something about how good your test is, even if only when distinguishing from fairly routine distributions.

1. Pick your favorite distribution. Can you determine a test that tests the hypothesis H_0 : $(X_i)$is normal vs H_1 : (X_i) has (your favorite distribution)? This is especially interesting if you can give a sense of the power and significance of your test.

2. Can you find a test which works for *any* distribution, at least in theory?

3. Test it out – sample some data from a distribution and check. If you want to give your test a challege, perhaps pull it from the t-distribution. This is available in matlab (use the function `trnd`)

Practical

In this part I will describe for you a kind of random variables which have rather complicated distributions. Your goal is to determine whether or not this distributions are asymptotically normal (spoiler alert: they're not), and explain your reasoning. Use the tools of statistics to design a test to compare them to the normal distribution, and see if you can distinguish them from normal. If you get far enough, you can try and develop some further tests to compare the two distributions to each other. Unlike the other problems, there's a real computational aspect here: this problem will require you to do some programming (in python, matlab,... – you can choose what's best for you).

Note: If you're interested, I have another version of this Project which is a bit easier to generate if you have access to Matlab (and maybe Python) but is a bit harder to describe to this level of mathematics. (It involves looking at eigenvalues of matrices.)

Up-Paths

A permutation of $[n]$ is an ordering of the numbers from $1, \dots, n$. For instance 42531 is a permutation of [5]. A **up-path** in a permutation is a subsequence of the ordered numbers that are strictly increasing For instance **24**135 is a up-path in the permutation 24135.

We denote by $upl(\pi)$ the length of the longest up-path in a permutation π. For instance, $upl(24135) = 3$, as **24**135 is the longest up-path in 24135.

One direction you can go with this project is to try and understand the distribution of $upl(\pi)$ for a random permutation π. It's *very* hard to write down (explicitly) the limiting distribution as $n \to \infty$ – even finding the expectation is hard. Instead: try and consider the sample mean/variance/etc after you generate a significant amount of data. Can you describe a test that differentiates it from normal?

Note: In order to generate some good data it's much more important to generate the up paths of many permutations that are fairly long (as opposed to the up path length of all permutations that are fairly short.)

Here are some questions to start with:

1. What can you say about the expectation of $upl(\pi)$ for a random permutation of $[n]$. Finding a formula for this is very difficult, but can you examine your sample to try and determine this?

2. Is the nomralized distribution of $upl(\pi)$ approximately normal? (You may want to estimate the mean and variance experimentally as they are quite hard to find exactly.)

3. Can you say that the $upl(\pi)$ is close to it's expectation with high probability? What can you prove in this direction?

4. Build a notion of confidence intervals for things whose limiting distribution is that of hte *upl*. If you get this far, I can give you (or tell you how to generate) some data to look at as well, and you can try and determine whether it satisfies the same distribution as the *upl*.

Warning: While it is good to *think* a little about how you might compute the expectation, etc., really I want you to experiment to understand these parameters and understand how this converges to a distribution.

Computer Note:

Here's an easy way of generating a random permutation on a computer (assuming `rand()` generates a uniformly random number from $[0, 1]$). Generate n random numbers $[0, 1]$, $x_1, x_2, \ldots, x_n$. Then sort the numbers. If x_i is the jth biggest of the random numbers then the *ith* term of the permutation is j. So if I generate 3 random numbers $x_1 = 0.1235$, $x_2 = 0.54534$ and $x_3 = 0.3434$ then the permutation is 132 as x_1 is the smallest number, x_2 is the biggest number and x_2 is the second smallest.

If you have trouble finding a fast algorithm to compute the upl of a permutation, talk to me.