Extra problems

18.600 problems/remarks that did not make the cut in 2016

For those who can't get enough of this stuff... here are a few problems (or musings, or remarks or rough ideas for problems) that did not make the cut for being on a 2016 problem set, but that could possibly be revised and somehow made into problems in later years. We'll explore doomsday, smugness, relative entropy, life expectancy, mutual funds, charity, multi-armed bandits, and conspiracy theories. This will not be on the test. :)

PROBLEMS:

A. **Doomsday:** Many people think it is likely that intelligent alien civilizations exist *somewhere* (though perhaps so far separated from us in space in time that we will never encounter them). When a species becomes roughly as advanced and intelligent as our own, how long does it typically survive before extinction? A few thousand years? A few millions years? A few billion years? Closely related question: how many members of such a species typically get to exist before it goes extinct?

Let's consider a related problem. Suppose that one factory has produced a million baseball cards in 10,000 batches of 100. Each batch is numbered from 1 to 100. Another factory has produced a million baseball card in 1,000 batches of 1,000, each batch numbered from 1 to 1,000. A third factory produced a million baseball card in 100 batches of 10,000, with each batch numbered from one to 10,000. You chance upon a baseball card from one of these three factories, and *a priori* you think it is equally likely to come from each of the three factories. Then you notice that the number on it is 76.

(a) Given the number you have seen, what is the conditional probability that the card comes from the first factory? The second? The third?

Now consider the following as a variant of the card problem. Suppose that one universe contains 10^{20} intelligent beings, grouped into civilizations of size 10^{12} each. Another universe contains 10^{20} intelligent beings, grouped into civiliations of size 10^{15} each. A final universe contains 10^{20} intelligent beings, grouped into civilizations of size 10^{18} each. You pick a random one of these 3×10^{20} beings and learn that before this being was born, at most 10^{11} other beings were born in its civilization.

(b) What is the conditional probability that the being comes from the first universe?

Remark: The *doomsday argument* (google it) suggests that perhaps is it more likely that human civilization will disappear within thousands of years than that it is that it will last for millions of years for the following reason: *if* advanced civilizations lasted for millions of years (with perhaps 10 billion people born per century), then it would seem *coincidental* for us to find ourselves within the

first few thousand years. People disagree on what to make of this argument (what the Bayesian prior on civilization length should be, what to do with all the other information we have about our world, how to imagine alternative universes, etc.) However many people say we should do more to prepare for seemingly unlikely apocalyptic scenarios (giant asteroids, incurable plagues, etc.) to improve humanity's chance of surviving a few thousand (or perhaps million or billion) more years.

B. **Relative entropy and world view:** Suppose that there are n possible outcomes of an athletic tournament. I assign probabilities $p_1, p_2, \ldots, p_n$ to these outcomes and you assign probabilities $q_1, q_2, \ldots, q_n$ to the same outcomes. If the ith outcome occurs and $p_i > q_i$ then I will feel interpret this as evidence that my probability estimates are better than yours, and that perhaps I am smarter than you. In short, I will feel smug. Suppose that my precise smugness level in this situation is $\log(p_i/q_i)$. Then before the event occurs, my *expected* smugness level is $\sum p_i \log(p_i/q_i)$.

(a) Show that my *expected* smugness level is always non-negative, and that it is zero if and only $p_i = q_i$ for all i. (Hint: use some calculus to find the vector $(q_1, q_2, \ldots, q_n)$ that minimizes my expected smugness level.)

(b) Look up the term *relative entropy* and explain what it has to do expected smugness.

Remark: I expect an *infinite* amount of smugness if I assign positive probability to things that you assign zero probability. We sometimes say our probability distributions are *singular* when this is the case. As a practical matter, might be a bad thing if my professed probability distribution is close to singular with respect to yours on some of the great unknowns (e.g., likelihood that certain tax cuts help the economy or that certain public investments provide net benefits or that some religious or philosophical ideas are true). The discrepancies might make it hard for us to find *any* common political ground, even if we are both utilitarians seeking the greater good. On the other hand, discrepancies are betting/trading opportunities. If our probability differences are real (and not political smokescreen) perhaps we can make a policy bet in the form of a policy that funds your priorities if your predictions pan out and my priorities if my predictions pan out.

C. **Life expectancy:** There is a certain naive and straightforward way to compute life expectancy. Let p_n denote the percentage of people of age n who die that year, and compute the expected amount of time a single person would live if, conditioned on surviving to year n, the person died with probability p_n. In other words, if each $X_i \in \{H, T\}$ is an independent coin toss with parameter p_i, then you can let $K = \min\{j : X_j = H\}$ and the life expectancy is defined to be $E[K]$.

1. If you switch from years to months, would the life expectancy (computed using an average year from the list) typically be more or less than the life expectancy computed using the whole year (or can it be either)?

2. When you combine two populations into one, is it possible that life expectancy of the combined group is longer than that of either group separately?

3. A certain expensive state does a terrible job caring for its chronically ill—so terrible that most of them are unable to stay in the work force and are forced to leave the state to live in cheaper places. A new governor is elected and implements a plan to improve health care for the chronically ill to allow more of them to stay. Do you think this will increase or decrease the state's life expectancy?

4. Suppose a city opens a new world class cancer hospital, which really does treat cancer better than any other hospital in the world. Do you think this would increase or decrease the city's life expectancy?

5. A country has a major war — which leads to widespread hunger and disease and kills about 30 percent of the population — after which violence subsides. Assuming that the people who died were on average less physically resilient than those who lived, what effect will the war have on life expectancy statistics (which continue to be calculated annually) over the next few decades?

6. Suppose that a governor identifies people who (based on health and demographics) are more likely to die during the coming few years than others their age. The governer sets up incentives that encourage these people to leave the state. How much do you think a governor could improve a state's life expectancy using this scheme?

Remark: There are actually various ways to compute life expectancy, some of which involve *smoothing* out the p_n values in some way to reduce noise in the measurement. You can google this if you want to read more about the methodology.

D. **Friendship bias:** Consider a world where population growth is stable at zero. Each person being born should be a parent of two children on average. Explain why, if you pick a random person, you expect that that person's parents have more genetic offspring than that person will have. In other words, explained why the expected number of siblings of a randomly chosen person (where half siblings counted as half) is greater than one. Explain why, if you are a randomly chosen person, we expect that your average friend has more friends than you do. See

https://www.washingtonpost.com/graphics/business/wonkblog/majority-illusion/

E. **Survivor bias:** The story of Pedro the hedge fund manager, as presented in lecture, illustrates one reason why you should not *automatically* be impressed when a fund manager says "My fund beat the market every year for the last ten years." In principle, it *could* be that this person just invests in a market index fund and then at the end of the year wagers the money on a crazy gamble that grows the investment by 2 percent with roughly 98 percent probability and otherwise loses *everything*. (Alternatively, the fund manager could be buying risky bonds or other complicated instruments that *effectively* implement this kind of gamble—bringing likely gains but carrying large downside risk that the fund manager may or may not understand.)

On other hand, suppose that your fund managers are not doing anything this evil but are instead kind of picking stocks randomly. Then some of these funds will do better than others just by chance, and the ones that do worse will be quietly shut down. Due to so called *survivor bias* you will find that most of the funds that are marketed to you have a pretty good track over the last few years.

On Simple Stock Planet, there are 1024 stocks in the market. During the course of a year, all stock values are multiplied by C where C is a random constant (with expectation a bit more than one). After this is done, the value of each individual stock is individually multiplied by either .8. or 1.2 (each with equal likelihood, independently for each stock). A fund manager picks 64 of these stocks at random and invests one million dollars in each of them. Using the central limit theorem, estimate the probability that the fund will beat the market return rate by at least five percent (before management fees are taken out).

F. **Gompertz mortality:** Here is another life expectancy question. Denote by p_n the conditional probability that an American female dies during her nth year of life, given she has survived until age $n-1$. During a given year this can be estimated explicitly (by checking what fraction of people of a certain age die that year). See `https://www.ssa.gov/oact/STATS/table4c6.html` for the data. As n ranges from 20 to 80 it seems that p_n doubles every 8 to 10 years. The death rate for men is higher (and especially elevated for men in their 20's — maybe due to violence or risk taking?) but otherwise a similar pattern holds, with death rates doubling every 8 to 10 years in later life.

(a) Suppose that X is an exponential random variable with parameter η so that $P\{X > T\} = e^{-aT}$. Write $Y = (1/b)\log(X+1)$ and argue that

$$P\{Y \geq S\} = P\{X \leq e^{bS} - 1\} = e^{-\eta(e^{bS}-1)}.$$

Look up the *Gompertz-Makeham law of mortality* and the *Gompertz distribution* and verify that Y is a Gompertz random variable.

(b) Assuming time is measured in years, what values of η and b should we choose if we want death rate (per infinitesimal unit of time) to double every eight years, and to start out at a rate of 1/2000 per year. You can use `http://sundoc.bibliothek.uni-halle.de/habil-online/07/07H056/t3.pdf` if it helps.

(c) One striking thing about the actuarial tables is that (from teens until early 70's) the percentage of men of a given age who die is *much much* higher than the percentage of women who die — at least 50 percent higher, sometimes more than 100 percent higher. (The ratio is smaller in the 80's and 90's.) A very naive person might guess, based on this, that the life expectancy for women should be 50 percent higher than for men; but in fact, the difference is only about five years. Explain why, if we assume that death rates follow the Gompertz distribution, this is actually not so surprising.

REMARKS:

Remark: Over 1 million people in US in prison have been convicted of crimes, usually as the result of some kind of a plea bargain. If you could had a magic time machine and could go back and see what actually happened in each case, in what fraction of the cases would you determine (based on your own standard) that the person was actually guilty? In theory, the answer exists: it is rational number between zero and one. Call this number α. Your beliefs about α could affect your attitude about many things. You can inform your beliefs by looking at data (how many convictions are later overturned, how many convicts claim innocence in anonymous surveys, etc.) but you'll still have plenty of uncertainty.

Suppose I take seriously the possibility that at least 30 percent of the incarcerated convicts are innocent and I also take seriously the possibility that less than 1 percent are innocent. I assign each of these a subjective probability of greater than 10 percent. Can I reason and think effectively about criminal justice when, for every policy decision I make, I have to estimate its consequences under these two *completely different* realities? Or do I in practice have to convince myself of one of them and block the other out of my mind, just so I can think coherently enough to get things done? Generally speaking, if my brain is not powerful enough to do computations under *every* possible reality, should I just adopt a few plausible seeming assumptions (a "world view") and run with those?

Remark: People often say utility functions grow sublinearly, but is that true? Here is a naive story about charitable giving. Suppose your utility function is given by your own health/comfort plus a constant c times the sum of the health/comfort of all other humans on the planet. For example, if $c = .01$, then you are mostly selfish, but you would be willing to give up a comfort for yourself if it would enable more than 100 strangers to enjoy the same comfort. You'd give up your life if you could save more than 100 other lives. Utilitarians theorize that it is a good to thing that $c > 0$ (so that we help others when we can make a big difference) but maybe also a good thing that $c < 1$ (since a little selfishness might be efficient in practice). As you acquire more money, there may come some point at which you believe that the marginal value of another dollar to you (in added health/comfort) is *less than* c times the amount a dollar donated to a smart global charity (like those profiled at givewell.org) would increase health/comfort for others. After that point, in principle you should donate *all* of your additional money to charity. If this is indeed your plan, then your utility function might be roughly linear after that point, since the amount of good you do in a huge global effort is roughly linear in the amount you give.

Remark: Some economists say that in reality charitable giving should be modeled as a consumptive good (that happens to have positive externality — google *"warm glow giving"*) that has to compete with other consumption goods among even the very wealthy. This point of view might predict actual behavior better than what I sketched above.

Remark: Sometimes medical trials have to be stopped early if the results are *too* significant. You

need experimentation to become confident that one treatment option is better than the other, but once you start to become confident it becomes ethically problematic to keep testing the (apparently) less effective option. Imagine one has a rare disease, which comes up 100 times per year worldwide, and that there are only two treatment options (e.g., surgery and antibiotics) and two outcomes (success or failure). Let p_s and p_a denote success probabilities corresponding to the surgery and antibiotics options. At the beginning these numbers are unknown, but estimates of their values improve as further trials are done. How do we decide when to experiment further and when to just stick with the option that looks best so far? Look up *multi-armed bandit* to read more about the statistical theory behind questions like this.

Remark: The more people who know about a conspiracy, and the more time goes by, the more likely it is for the conspiracy to be (deliberately or accidentally) exposed. Taking a model that makes leaks a Poisson point process, the following url makes the case that several famous conspiracy theories are very unlikely to be correct.
`http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0147905`

Remark: Google the phrase *"than liberals"* and then google the phrase *"than conservatives"* (both in quotes) and scroll through a few pages of hits. If you consider yourself a liberal or conservative partisan, you'll find many surveys and studies that flatter your side and raise doubts the other side. But how many do you believe? Surely, there will be real discernable differences between two large and distinctive populations, but there is also a huge potential for bias in the storytelling, and many studies and surveys are designed with a particular story in mind.

Remark: Look up *prospect theory* (people take low probability events too seriously) and *hyperbolic discounting* (people undervalue long term rewards).

Remark: If you are reporting on science, you can avoid commitment by using extra qualifiers. Instead of an overwhelmingly forceful statement like "X is a bit more likely than Y based on what I know now," try "X *might be* a bit more likely than Y based on what I know now." Similarly: "This drug may lower your risk of heart attack" or "People who do A may be more likely to do B" or "Investing in X may increase your expected long term return" or "It is at least theoretically possible that $P[B|A] > P[B]$."

Remark: Be careful about independence assumptions. Imagine that ten students are accepted to a graduate engineering program at Princeton. They go to the Princeton open house together, have a wonderful time, and decide they want to attend graduate school together. Then they all decide to go to MIT — an outcome that would be unlikely (statistically) if they were deciding independently, and could be a surprise to both universities. There are many settings (e.g., online reviews) where people's influence on one another can lead to larger fluctuations in average scores.

Remark: A "rational" person (in the economic sense) has a utility function and a subjective probability measure, and makes decisions in order to optimize expected utility. A group of people,

each of whom is rational in this sense, may be decidely *irrational* as a group. Arrow's impossibility theorem (look it up) states that (under any reasonable voting scheme) a democratic group of people may prefer A to B and B to C and C to A. Political parties, companies, and entire countries can all be "irrational" to a greater extent than their individual members.

Remark: If I said "This restaurant is unsafe. I think I see a norovirus on the table right here!" you would be skeptical. In some sense it might be true — there might be photons bouncing off a norovirus and hitting my eye. But it is not true that I have a mechanism for reliably distinguishing these photons from photons that have not bounced off a norovirus. What I see cannot be reliable distinguished from noise. Related problems in statistics: when we see correlations and patterns in data, can we reliably distinguish them from noise? When somebody says "In my experience, celery makes people healthier" should we be as skeptical as we are about the norovirus, if our common sense estimate of the magnitude of celery's impact is such that it cannot possibly be discerned with haphazard personal observation?

Remark: Here is another article about the limitations of our nutrition research.
`http://www.nytimes.com/2015/10/27/upshot/surprising-honey-study-shows-woes-of-nutrition-research.html?rref=upshot` But enough complaining. Let's see what can be done better with the resources we actually have. We have a lot of experience in assessing the impact of advertising on purchasing. Can we better assess the impact of advertising on health? Instead of seeing whether people get healthier when assigned to consume blueberries in a trial, can we see if people get healthier when shown blueberries on billboards or sent blueberry text messages? Can we send a health-related message to a billion people and see if they wind up healthier than the billion who don't get the message? Any ideas, young hackers?

Remark: Here is an article about significance fudging.
`http://www.economist.com/blogs/freeexchange/2016/01/fudging-hell`

Remark: If you're trying to prove that something is good for you, and that thing is actually bad for you, you're more likely to get a statistically significant result in the direction you want if you use small sample sizes. (If the thing is actually good for you, then larger sample sizes are more likely to turn up the result you want, but they are also more expensive.)
`http://articles.mercola.com/sites/articles/archive/2013/02/13/publication-bias.aspx`