33-765 Homework 1

Nathaniel D. Hoffman

January 21, 2020

1. Simpson's Paradox

A college offers two majors (A and B), to which both male and female students apply. The fraction of male and female students interested in major A is μ_A and ϕ_A, respectively, and to keep things simple, we assume that nobody applies to two majors. Show that the following can happen: in both majors the acceptance probabilities f_A and f_B for women are larger than those for men (m_A and m_B), and yet the overall acceptance rate for women is lower than that for men. Give a complete and precise characterization of the circumstances under which this situation occurs!

We are given $f_A > m_A$ and $f_B > m_B$ and we want to find criteria which allow for

$$f_A\phi_A + f_B\phi_B < m_A\mu_A + m_B\mu_B$$

In this scenario, Simpson's paradox typically results from the majority of female students applying to a much more competitive major. Even if the acceptance rate is higher, more men applying to the less competitive major means more of them will be accepted, making the sums over both majors seem to have the opposite correlation.
We can rewrite our inequality in terms of μ_A:

$$\mu_A > \frac{f_A\phi_A + f_B\phi_B - m_B\mu_B}{m_A}$$

Note that $\phi_A + \phi_B = 1$ and $\mu_A + \mu_B = 1$.

$$\begin{aligned}\mu_A &> \frac{f_A\phi_A + f_B\phi_B - m_B\mu_B}{m_A}\\ &> \frac{f_B\phi_B - m_B\mu_B}{m_A}\\ \mu_A &> \frac{f_B\phi_B - m_B}{m_A}\end{aligned}$$

where the last line comes from the fact that $\mu_B < 1$
If we rewrite in terms of ϕ_A:

$$\begin{aligned}\phi_A &< \frac{m_A\mu_A + m_B\mu_B - f_B\phi_B}{f_A}\\ &< \frac{m_A\mu_A + m_B\mu_B}{f_A}\\ \phi_A &< \frac{\mu_A m_A + m_B}{f_A}\end{aligned}$$

These inequalities are symmetric under interchange of $A \leftrightarrow B$. These just follow the condition that the total acceptance rates for females are lower than total acceptance rates for males. If we then factor in the requirements that $f_A > m_A$ and $f_B > m_B$, I honestly don't find anything useful, and I'm not really sure where to go from here. My reasoning in finding these inequalities in this form was to get conditions on ϕ_A and μ_A (and by definition ϕ_B and μ_B) and show that there was some hidden requirement for $f_A\phi_A > f_B\phi_B$ or something like that.

2. Sick Bayes

Consider a disease that exists with some small probability p in the general population. Assume that people can be checked for the disease with a test that correctly picks it up with a (large) probability α (which is often called the "sensitivity" of the test). Of course, any test also has a (hopefully small) false positive rate β. (Incidentally, $1-\beta$ is often called the "specificity" of the test). If a random person gets tested positive, what is the probability of them having the disease? How does one have to design such a test so that test-takers are not unnecessarily scared? Give an illustrative numerical example!

We want to find Pr(has disease | tests +). To do this, we can use Bayes' theorem:

$$\begin{aligned}\Pr(\text{has disease} \mid \text{tests} +) &= \frac{\Pr(\text{tests} + \mid \text{has disease})\Pr(\text{has disease})}{\Pr(\text{tests} +)} \\ &= \frac{(\alpha)(p)}{(\alpha p + \beta(1-p))}\end{aligned}$$

since the probability to test positive is the probability to be correctly diagnosed (αp) plus the probability to be misdiagnosed ($\beta(1-p)$).
To make a test where the test-takers are not unnecessarily scared, one must either increase the specificity to reduce the number of false positives, increase the sensitivity to reduce the number of missed diagnoses, or test multiple times to reduce the chance of being misdiagnosed. If the first two degrees of freedom are set, testing multiple times might be the only option. If you double-test people who tested positive, you run the risk of missing the false-negative cases (although if the sensitivity is high, this will likely not be a problem). Suppose $p = 0.01$, $\alpha = 0.99$, and $\beta = 0.01$. In other words, suppose 1% of the population has the disease, there is a test which is successful at diagnosing the disease 99% of the time, and it gives a false positive to 1% of people. Even with such a low false-positive rate, the above formula shows that the probability of actually having the disease given a positive test is only 50%. Intuitively, if you were to test a random population, most of them don't have the disease, so the specificity will play a bigger role in the results of the test. If you test twice, assuming the tests are independent of each other, the probability of having the disease given two positive tests becomes

$$\Pr(\text{has disease} \mid \text{tests} +_1 \text{tests} +_2) = \frac{\alpha^2 p}{\alpha^2 p + \beta^2(1-p)}$$

For the above values, people who test positive twice have a 99% probability of having the disease, so there is a much lower likelihood to scare patients with a false-positive.

3. Characteristic Functions and the Amazing Central Limit Theorem

The Fourier transform $\tilde{p}(k)$ of a probability density (henceforth: "p-density") $p(x)$ is also called the "characteristic function":

$$\tilde{p}(k) = \langle e^{\imath kx} \rangle = \int \mathrm{d}x\, p(x) e^{\imath kx} \qquad \left[\text{and hence: } p(x) = \frac{1}{2\pi} \int \mathrm{d}k\, \tilde{p}(k) e^{-\imath kx}\right].$$

1. Let X be a random variable whose p-density $p_X(x)$ has moments $\mu_n = \langle X^n \rangle$. If these moments μ_n exist, prove that

$$\mu_n = \imath^{-n} \left[\frac{\partial^n}{\partial k^n} \tilde{p}_X(k)\right]_{k=0}.$$

Let's begin by Taylor expanding the exponential in the definition of the characteristic function:

$$\tilde{p}_X(k) = \int \mathrm{d}x\, p_X(x) \sum_n \imath^n \frac{k^n x^n}{n!}$$

Taking the n-th k-derivative, we find

$$\begin{aligned}\frac{\partial^n}{\partial k^n} \tilde{p}_X(k) = \int \mathrm{d}x\, p_X(x) \Bigg(& \cdots + \imath^{n-1} \frac{x^{n-1}}{(n-1)!} \cancelto{0}{\frac{\partial^n}{\partial k^n} k^{n-1}} \\ & + \imath^n x^n \\ & + \imath^{n+1} x^{n+1} k^1 \\ & + \imath^{n+2} \frac{x^{n+2} k^2 (n-2)!}{n!} + \cdots \Bigg)\end{aligned}$$

Evaluating at $k = 0$ eliminates the higher order terms:

$$\left[\frac{\partial^n}{\partial k^n} \tilde{p}_X(k)\right]_{k=0} = \int \mathrm{d}x\, p_X(x) \imath^n x^n$$

finally, divide by $\imath^n$:

$$\imath^{-n} \left[\frac{\partial^n}{\partial k^n} \tilde{p}_X(k)\right]_{k=0} = \int \mathrm{d}x\, p_X(x) x^n = \mu_n$$

2. If $\tilde{p}_{aX}(k)$ is the characteristic function of the random variable aX (with some $a \in \mathbb{R}$), show that $\tilde{p}_{aX}(k) = \tilde{p}_X(ak)$.

$$\tilde{p}_{aX}(k) = \int \mathrm{d}x\, p_{aX}(ax) e^{\imath kax} = \int \mathrm{d}x\, p_X(x) e^{\imath(ka)x} = \tilde{p}_X(ak)$$

since

$$p_{aX}(ax) = \int_{\mathbb{R}} \mathrm{d}x\, \delta(ax - F(x)) p_X(x) = p_X\left(\frac{F(x)}{a}\right) = p_X\left(\frac{xa}{a}\right) = p_X(x)$$

by the transformation theorem.

3. Let X and Y be two independent random variables with p-densities $p_X(x)$ and $p_Y(y)$. Let $p_{X+Y}(z)$ be the p-density of $Z = X + Y$. Prove that $p_{X+Y}(z) = \int \mathrm{d}x\, p_X(x) p_y(z - x)$ and that $\tilde{p}_{X+Y}(k) = \tilde{p}_X(k)\tilde{p}_Y(k)$.

$$\tilde{p}_{X+Y}(k) = \left\langle e^{\imath k(x+y)} \right\rangle = \left\langle e^{\imath kx} e^{\imath ky} \right\rangle = \left\langle e^{\imath kx} \right\rangle \left\langle e^{\imath ky} \right\rangle = \tilde{p}_X(k)\tilde{p}_Y(k)$$

We can use the transformation theorem to get the other half of the problem. In the integral, we must now use the joint probability density of X and Y and the δ-function must “trigger” when the sum equals z. In this way, we are summing up all the ways to get to $x + y = z$ weighted by the probability of starting with those variables x and y:

$$\begin{aligned}
p_{X+Y}(z) &= \iint \mathrm{d}x\,\mathrm{d}y\,\delta(z - f(z))p_{X,Y}(x, y) \\
&= \iint \mathrm{d}x\,\mathrm{d}y\,\delta(z - x + y)p_{X,Y}(x, y) \\
&= \iint \mathrm{d}x\,\mathrm{d}y\,\delta(y - (z - x))p_{X,Y}(x, y) \\
&= \int \mathrm{d}x\,p_{X,Y}(x, z - x) \\
&= \int \mathrm{d}x\,p_X(x)p_Y(z - x)
\end{aligned}$$

We can do the last step because the variables are independent, so we can factor the joint probability: $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

4. Let $X_1, \cdots, X_n$ be n independent random variables with identical distributions $p_X(x)$, which has mean μ_1 and finite variance $\sigma^2 = \mu_2 - \mu_1^2$. Consider the centered and normalized random variables $Y_i = (X_i - \mu_1)/\sigma$ (which obviously have zero mean and unit variance) and the new (and seemingly curiously normalized) sum random variable

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i = \frac{X_1 + X_2 + \cdots + X_n - n\mu_1}{\sigma\sqrt{n}}.$$

If $\tilde{p}_{Z_n}(k)$ is the characteristic function of (the p-density of) Z_n, show that in the limit of large n you get

$$\lim_{n\to\infty} \tilde{p}_{Z_n}(k) = e^{-\frac{1}{2}k^2} \quad \text{and hence} \quad p_{Z_n}(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \equiv \mathcal{G}_{(0,1)}(x).$$

Part 3 showed that the characteristic function of a sum of random variables is a product of their characteristic functions:

$$\tilde{p}_{Z_n} = \prod_{i=1}^{n} \tilde{p}_{\frac{Y_i}{\sqrt{n}}}(k)$$

From part 2, the scalar multiplication of a variable multiplies the argument of its characteristic function:

$$\prod_{i=1}^{n} \tilde{p}_{\frac{Y_i}{\sqrt{n}}}(k) = \prod_{i=1}^{n} \tilde{p}_{Y_i}\left(\frac{k}{\sqrt{n}}\right) = \left[\tilde{p}_{Y_i}\left(\frac{k}{\sqrt{n}}\right)\right]^n$$

From part 1, the characteristic function can be written as a sum of its moments:

$$\tilde{p}_{Y_i}(k) = \sum_{m=0}^{\infty} \frac{(\imath k)^m}{m!}\mu_m$$

so

$$\tilde{p}_{Z_n}(k) = \left[\sum_{m=0}^{\infty} \frac{(\imath k)^m}{n^{m/2} m!}\mu_m\right]^n$$

Because the Y distributions are all (by definition) normalized and centered, we know two of these moments:

$$\tilde{p}_{Z_n}(k) = \left[\underbrace{\mu_0}_{1} + \frac{\imath k}{\sqrt{n}} \underbrace{\mu_1}_{0} + \frac{(\imath k)^2}{2n} \underbrace{\mu_2}_{1} + \underbrace{\frac{(\imath k)^3}{6n^{3/2}} \mu_3 + \cdots}_{\mathcal{O}\left(k^3/n^{3/2}\right)} \right]^n$$

Taking the limit as $n \to \infty$, we find the desired result, since the terms that we don't know are of order $\mathcal{O}\big(k^3/n^{3/2}\big)$:

$$\lim_{n\to\infty} \tilde{p}_{Z_n}(k) = \lim_{n\to\infty} \left[1 + \frac{-\frac{1}{2}k^2}{n} + \mathcal{O}\left(\frac{k^3}{n^{3/2}}\right) \right]^n = e^{-\frac{1}{2}k^2}$$

which Fourier-transforms to the desired Gaussian.