33-765 — Statistical Physics

Department of Physics, Carnegie Mellon University, Spring Term 2020, Deserno

Problem sheet #1

1. Simpson's paradox *(6 points)*

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 $\mu_{\rm A}$ and $\phi_{\rm A}$, 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 $f_{\rm A}$ and $f_{\rm B}$ for women are *larger* than those for men ($m_{\rm A}$ and $m_{\rm B}$), 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!

2. Sick Bayes *(5 points)*

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!

3. Characteristic functions and the amazing Central Limit Theorem *(9 points)*

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

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

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 exists, prove that

$$\mu_n \;=\; {\rm i}^{-n}\left[\frac{\partial^n}{\partial k^n}\tilde{p}_X(k)\right]_{k=0} . \tag{2}$$

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

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 $p_{X+Y}(z) = \int {\rm d}x\; p_X(x)\,p_Y(z-x)$ and that $\tilde{p}_{X+Y}(k) = \tilde{p}_X(k)\;\tilde{p}_Y(k)$.

4. Let $X_1,\dots,X_n$ be n *independent* random variables with *identical distribution* $p_X(x)$, 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

$$Z_n \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i \;=\; \frac{X_1+X_2+\cdots+X_n \;-\; n\,\mu_1}{\sigma\,\sqrt{n}} . \tag{3}$$

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

$$\lim_{n\to\infty}\tilde{p}_{Z_n}(k) \;=\; {\rm e}^{-\frac{1}{2}k^2} \qquad \text{and hence} \qquad p_{Z_n}(x) \;\longrightarrow\; \frac{1}{\sqrt{2\pi}}{\rm e}^{-\frac{1}{2}x^2} \;\equiv\; \mathcal{G}_{(0,1)}(x) . \tag{4}$$

Hint: The proof follows swiftly from what you've worked out so far; you will also need a cute representation for the exponential function: $\lim_{n\to\infty}[1+x/n+o(x/n)]^n = {\rm e}^x$, *where* $o(z)$ *is any term that satisfies* $\lim_{z\to 0} o(z)/z = 0$.

This is (a version of) the amazing **Central Limit Theorem**: The distribution of the $\sqrt{n}$-normalized sum of the centered X_i becomes a Gaussian with zero mean and unit variance, *independent of the actual distribution of the* X_i (as long as their variance is finite). It also implies that for increasing n the p-density of the arithmetic mean, $\overline{X_i} = \frac{1}{n}(X_1+\cdots+X_n) = \mu + \frac{\sigma}{\sqrt{n}}Z_n$ converges against, $\mathcal{G}_{(\mu,\sigma/\sqrt{n})}(x)$, a Gaussian centered around μ with variance $\sigma/\sqrt{n}$. Hence, the *error of the mean* also becomes Gaussian and decreases like $1/\sqrt{n}$. The Central Limit Theorem explains why the Gaussian distribution is "normal": It naturally emerges once you do averaging. This is also why it appears all over the place.