Homework 4

36-465/665, Spring 2021

Due at 6 pm on Thursday, 4 March 2021

1. *U statistics* You'll remember that, in theoretical statistics, a "statistic" is a quantity which is a function of the data, and of the data alone. When we have observations $X_1, \ldots X_n$, a U**-statistic** of **order** r is a statistic with a particular form:

$$U = \frac{1}{\binom{n}{r}} \sum_{\{i_1, i_2, \ldots i_r\}} h(X_{i_1}, X_{i_2}, \ldots X_{i_r})$$

(The sum is taken over un-ordered subsets of r distinct integers in $1:n$.) The function h is called the **kernel** of the U-statistic. In words, a U statistic is what we get when we take all groups of r data points in our sample, calculate the kernel on each group, and average. U statistics are important in statistical theory because (i) they give unbiased estimates of $\mathbb{E}\left[h(X_1, \ldots X_r)\right]$ (hence "U"), (ii) the *minimum-variance* unbiased estimate of $\mathbb{E}\left[h(X_1, \ldots X_r)\right]$ is (almost always) a U statistics, and (iii) lots of interesting and important parameters can be written as expectations of functions. We won't prove (i)–(iii), but they should make the following better motivated.

   a. (3) Assume that $X_1, \ldots X_n$ are IID with mean μ and variance σ^2. Show that $\sigma^2 = \mathbb{E}\left[\frac{1}{2}(X_1 - X_2)^2\right]$.

   b. (3) Write out a formula for the U statistic we'd use to estimate σ^2. *Hint*: (a).

   c. (5) Now that you have some intuition about a particular U statistic, think about a more general problem. Take a U statistic of order r, where the kernel function h is bounded between $a > -\infty$ and $b < \infty$. Show that $U(X_1, \ldots X_n)$ has the finite-difference property, and find the maximum difference that can be made by changing X_i. *Hints*: the answer should be the same for all i (why?); the answer should involve the bounds a and b on the kernel function, the sample size n, and the order r. It should be decreasing in n.

   d. (5) Using (c), give a bound on $\mathbb{P}\left(|U - \mathbb{E}\left[h(X_1, \ldots X_r)\right]| \geq \epsilon\right)$. The answer should involve a, b, n and r (and possibly mathematical constants such as e, π, or 137).

   e. (5) When can you apply your result from (d) to the U statistic you built in (b)? Explain.

2. *The Glivenko-Cantelli theorem.* I am asking you once again to suppose that $X_1, \ldots X_n$ are all IID. Since they're IID, the same the same cumulative distribution function (CDF) $F(a) \equiv \mathbb{P}(X \leq a)$. The **empirical CDF** is $\hat{F}_n(a) \equiv n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq a\}$, i.e., the fraction of data points $\leq a$. The **Glivenko-Cantelli theorem** says that the empirical CDF converges *uniformly* to the true CDF. Specifically, for any $\epsilon > 0$,

$$\mathbb{P}\left(\max_a \left|\hat{F}_n(a) - F(a)\right| \geq \epsilon\right) \to 0$$

This result is sometimes called the "fundamental theorem of statistics", because it tells us that we can, in fact, learn probability distributions from observational data. We will now prove it, assuming for simplicity that the true CDF is continuous. The technique is related to the "covering" tactic introduced in lecture 7 (slides 12ff.), but not quite the same. (The case of general, discontinuous CDFs is an extra credit problem below.)

   a. (5) Show that, at any fixed a, $\mathbb{E}\left[\hat{F}_n(a)\right] = F(a)$.

b. (5) Use (a) and the bounded-difference inequality to show that, for any fixed a, $\mathbb{P}\left(\left|\hat{F}_n(a) - F(a)\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2\right)$.

c. (5) Explain how to locate points $-\infty = b_0 < b_1 < b_2 < \ldots < b_{q+1} = \infty$ so that $F(b_i) - F(b_{i-1}) \leq \epsilon$. Explain how q is related to ϵ. *Hint*: use the assumption on F, and the general properties of CDFs.

d. (4) Explain why

$$\mathbb{P}\left(\max_{i \in 0:(q+1)} |\hat{F}_n(b_i) - F(b_i)| \geq \epsilon\right) \leq 2q\exp\left(-2n\epsilon^2\right)$$

In particular, explain why there's a factor of q and not $q+2$.

e. (3) Show that any point a can be "bracketed" by a pair b_{i-1}, b_i, where $F(b_i) - F(a) \leq \epsilon$ and $F(a) - F(b_{i-1}) \leq \epsilon$.

f. (5) Using the previous part, show that for any a,

$$\hat{F}_n(a) - F(a) \leq \hat{F}_n(b_i) - F(b_i) + \epsilon \quad (1)$$
$$\hat{F}_n(a) - F(a) \geq \hat{F}_n(b_{i-1}) - F(b_{i-1}) - \epsilon \quad (2)$$

Hint: It's important that both F and $\hat{F}_n$ are non-decreasing functions.

g. (5) Using the previous part, show that

$$\max_a |\hat{F}_n(a) - F(a)| \leq \max_{i \in 0:(q+1)} |\hat{F}_n(b_i) - F(b_i)| + \epsilon$$

h. (5) Conclude by showing that

$$\mathbb{P}\left(\max_a |\hat{F}_n(a) - F(a)| \geq \epsilon\right) \leq 2q(\epsilon/2)\exp\left(-n\epsilon^2/2\right)$$

3. *High-dimensional geometric weirdness, part I: hyper-balls are trippy.* The joint distribution of n IID variables can be thought of as the distribution of a single point in an n-dimensional space, and so it's influenced by the intrinsic geometry of n-dimensional space. When n is large, this geometry is very strange, even alien, to those of us raised on merely three-dimensional Euclidean space. Many of the deviation inequalities we've painfully acquired are actually natural consequences of this mind-bending geometry. To explore this, let's start by considering the n-dimensional unit ball, i.e., the set of all vectors $(x_1, x_2, \ldots x_n)$ such that $\|x\| = \sqrt{\sum_{i=1}^n x_i^2} \leq 1$.

   a. (3) Fix any $\epsilon > 0$. What fraction of the volume of the sphere is within ϵ of its surface? Plot this with $\epsilon = 0.01$ as n grows from 2 to 1000. *Hints*: (i) the volume of an n-dimensional sphere of radius r is $\frac{\sqrt{\pi^n}}{\Gamma(\frac{n}{2}+1)} r^n$; (ii) what fraction of the volume is *not* ϵ-close to the surface?

   b. (3) Suppose a point is drawn uniformly at random from within the n-dimensional ball. Show that the probability that its distance from the center of the ball is at least r goes to 1 as n increases, for any $0 < r < 1$.

   c. (5) Explain the statement "averaging over a high-dimensional ball is basically the same as averaging over a high-dimensional sphere".

4. *High-dimensional geometric weirdness, part II: hyper-cubes are also trippy.* The (Cartesian) coordinates of a point in a ball have weird, annoying correlations, and we might worry that's what's driving the results in Q3. So consider the unit-side n-dimensional cube centered at the origin, i.e., the set of all n-dimensional vectors where for each $i \in 1:n$, $-1/2 \leq x_i \leq 1/2$.

   a. (2) Show that for any fixed ϵ, the volume within ϵ of the surface $\to 1$ as $n \to \infty$.

   b. (3) Explain the statement: "most of a high-dimensional cube is near its surface".

c. (5) Consider generating n IID variables $U_i \sim \text{Unif}(0,1)$. What is the probability that at least one U_i is $\leq \epsilon$ or $\geq 1-\epsilon$, as a function of ϵ and n? What does this have to do with early parts of this question?

5. *High-dimensional geometric weirdness, part III: high-dimensional Gaussians are very definite about distance.* Maybe the uniform distribution on the hyper-cube is weird because it doesn't favor a central location. Let's consider a high-dimensional Gaussian instead, so suppose $Z_i \sim \mathcal{N}(0,1)$, for $i \in 1:n$, and all the Z_i are IID.

   a. (1) Show that $\mathbb{E}\left[\|Z\|^2\right] = n$.

   b. (2) Explain why $\|Z\|^2 \sim \chi^2_n$.

   c. (2) Show that $\mathbb{P}\left(\left|\frac{\|Z\|^2}{n} - 1\right| \geq \epsilon\right) \leq 2\exp\left(-n\epsilon^2/8\right)$. *Hint: HW3Q4.*

   d. (5) Explain "A univariate standard Gaussian is a smear around 0; a high-dimensional Gaussian is a hard spherical shell".

6. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra credit (5 total): Refer to Q2 for general setting and notation, In that problem, we assumed that $F(a)$ was continuous in a. This meant that $\mathbb{P}(X = a) = 0$ for each a. If $\mathbb{P}(X = c) = p > 0$, then we need F to be *dis*-continuous at c. We'll say that F is "continuous on the right, limited on the left"[1], so for $\delta \downarrow 0$, $F(c+\delta) \to F(c)$, while $\lim F(c-\delta) = F(c) - p$. Write $c_1, c_2, \ldots$ for the locations of the jumps, of sizes $p_1, p_2, \ldots$. Again, fix on any $\epsilon > 0$ (and < 1).

a. (1) Explain why F can make at most $1/\epsilon$ jumps of size ϵ or more.

b. (2) Say that F makes exactly $m(\epsilon)$ jumps of size ϵ or more. Explain how to locate points $-\infty = b_0 < b_1 < b_2 < \ldots < b_{q+1} = \infty$ so that $F(b_i) - F(b_i) \leq \epsilon$. How is q related to ϵ? (The answer may involve m.)

c. (2) Find a bound on $\mathbb{P}\left(\max_a |\hat{F}_n(a) - F(a)| \geq \epsilon\right)$. Be explicit (using ECb) about all factors involving ϵ. *Hint*: What, if anything, has to change in the line of argument in Q2?

[1] Often referred to by the French acronym "cadlag", from ≪ continues à droite, limites à gauche ≫. This is because French mathematicians were some of the first to consider such questions, and no English phrase lends itself to such a pronouncable acronym for this rather complicated property. (The closest anyone came is "rcll" for "right-continuous, left-limited", which is fatally deficient in vowels.)