Homework 4.2
Stat 605, Fall 2018

Instructor: Brian Van Koten

Due Thursday, October 4

In class, I mentioned that by making an analogy with inner products of vectors in $\mathbb{R}^3$ one might interpret $\rho(X, Y)$ as the cosine of the angle between $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$. Therefore, one might say that correlation measures the degree to which the deviations of X and Y from their expectations point in the same direction. The following problem is intended to clarify a subtly different statement about correlation: Correlation measures the strength of the linear relationship between two random variables.

Problem 1 (1+2+2 points)**.** *Let* $X, Y : (\Omega, \mathscr{F}, \mathbb{P}) \to (\mathbb{R}, \mathscr{B})$ *be random variables.*

1. *Write down an* L^2*-orthonormal basis for* $\operatorname{span}\{Y, \mathbf{1}\}$*. Here,* $\mathbf{1}$ *denotes the constant random variable with value* 1*. Note that* $\operatorname{span}\{Y, \mathbf{1}\}$ *is the set of all linear functions* $\alpha Y + \beta$ *of* Y*.*

2. *Write down a formula for the orthogonal projection of* X *onto* $\operatorname{span}\{Y, \mathbf{1}\}$*. Try to make the correlation* $\rho(X, Y)$ *appear in this formula.*

3. *Define*

$$d(X, \operatorname{span}\{Y, \mathbf{1}\}) = \min_{Z \in \operatorname{span}\{Y, \mathbf{1}\}} \|X - Z\|_2.$$

Derive a formula expressing $d(X, \operatorname{span}\{Y, \mathbf{1}\})$ *in terms of* $\rho(X, Y)$ *and* $\operatorname{var}(X)$*.* Note: *When statisticians say things about* R^2 *values and the proportion of the variation in* X *explained by a linear regression on* Y*, they are using a version of this formula.*

Recall that for $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (U, \mathscr{S})$ a random variable, we define the σ-algebra generated by X to be

$$\sigma(X) = \{X^{-1}(B); B \in \mathscr{S}\}$$

(I believe that this definition first appeared in our discussion of independent random variables.) The σ-algebra $\sigma(X)$ is sometimes called the information contained in X. To see why, work out the following problem.

Problem 2 (2+4 points)**.** *For $\mathscr{G}$ a σ-algebra on Ω, we call a map $H : \Omega \to \mathbb{R}$ $\mathscr{G}$-measurable if $H^{-1}(B) \in \mathscr{G}$ for all $B \in \mathscr{B}$. We say that a function $H : \Omega \to \mathbb{R}$ is $\mathscr{G}$-simple if it is a $\mathscr{G}$-measurable simple function.*

1. *Show that the set of all $\mathscr{G}$-measurable maps is the smallest class containing the simple functions and closed under pointwise limits.* Note: *Most of the work here has been done in class.*

2. *Show that Y is $\sigma(X)$-measurable if and only if $Y = f(X)$ for some measurable $f : (\mathbb{R}, \mathscr{B}) \to (\mathbb{R}, \mathscr{B})$. In particular, if two random variables contain the same information, then they are functions of each other.*

We will see that many properties of a random variable depend only on the information which it contains. For example, $X \perp\!\!\!\perp Y$ implies that $X' \perp\!\!\!\perp Y'$ for any $\sigma(X)$-measurable X' and $\sigma(Y)$-measurable Y'. Later, we will see that the conditional expectation $\mathbb{E}[X|Y]$ depends only on X and the information contained in Y.

The results in this problem may possibly be useful below.

Problem 3 (1+1+1 points)**.** *A random variable X is said to be a Bernoulli trial with probability of success p if X takes only the values $\{0, 1\}$ and $\mathbb{P}[X = 1] = p$. Let $\{X_i\}_{i=1}^{\infty}$ be independent Bernoulli trials, and define*

$$T = \inf\{m : X_m = 1\}.$$

The distribution of T is called the geometric distribution*. Since T takes values in the countable set $\mathbb{N}$, this distribution is characterized by the* probability mass function *$f_T : \mathbb{N} \to \mathbb{R}$ by*

$$f_T(t) = \mathbb{P}[T = t].$$

1. *Find a formula expressing $f_T(t)$ in terms of p and t.*

2. *Compute $\mathbb{E}[T]$.*

3. *Compute* $\mathrm{var}(T)$*.*

This next problem is extra credit. If you like, you can use the result below even if you do not turn in a proof. It is helpful to introduce some new terminology here: We say that the functions $f, g : \mathbb{R} \to \mathbb{R}$ are asymptotic in the limit as x tends to c if

$$\lim_{x \to c} \frac{f(x)}{g(x)} = 1.$$

Here, c may be a real number or it may be infinity. When f is asymptotic to g in some limit, we write $f \sim g$. It is important to note that this definition is really only interesting when either f and g both tend to zero in the limit or both tend to infinity in the limit. Otherwise, at least if both limits exist, f is asymptotic to g if and only if the limits are equal.

Problem 4 (1 points)**.** *Show that $n \sum_{m=1}^{n} m^{-1} \sim n \log n$ as $n \to \infty$.*

Problem 5 (1 point). *Let* $\{Y_n\}$ *be a sequence of random variables and* v_n *a sequence of real numbers so that* $\operatorname{var}(Y_n)/v_n^2 \to 0$ *as* $n \to \infty$*. Show that*

$$\frac{Y_n - \mathbb{E}Y_n}{v_n} \xrightarrow{\text{P}} 0.$$

Problem 6 (5 points). *Let* $\{X_i\}_{i=1}^{\infty}$ *be independent random variables, each having the uniform distribution on the finite set* $\{1, 2, \ldots, n\}$*. Let*

$$\tau_k^n = \inf\{m : \#\{X_1, \ldots, X_m\} = k\}$$

be the least index m *for which the sequence* $X_1, \ldots, X_m$ *takes* k *distinct values. (Here, we use* $\#S$ *to denote the cardinality of the set* S*.) To make this more concrete, let us suppose that you like watching birds and that you will stop only once you have seen every bird on a list of* n *birds. If we assume that the birds are observed with equal probability and that the observations are independent, then* τ_n^n *is a good model for the total number of observations that you will make before stopping. Prove that*

$$\frac{\tau_n^n}{n \log n} \xrightarrow{\text{P}} 1.$$

Hint: Use the results of the previous three problems.