A (snippet from a) Crash Course in (discrete) Probability

Students

- Jonah Botvinick-Greenhouse (Amherst College)
- Connor Fitch (Bowdoin)
- Jonah Green (Lehman College)
- Mark Kong (Harvard)
- Taylor Meredith (NYU).
- Philip Speegle (U Alabama)
- Rachel Tan (UConn)
- Oliver Vandenberg (Kenyon College)

Staff

- Iddo Ben-Ari
- Hugo Panzo

Preface

These notes are one part of a collaborative writing project we ran in the first few weeks of the Markov Chains REU at UConn in Summer 2019.

In this project, notes for a crash course in probability were written and edited by the students and the staff. The structure of the crash-course and an initial template were prepared in advance by the staff. The students worked on filling the content with introductory and motivating text and explanations, giving examples and applications, writing proofs, revising their own and others' work. They also spiced it up with some cartoons.

The goal was to experience mathematical writing and collaboration on a writing project, reviewing own and other's work, and learning from feedback. It also helped filling gaps in preparation for our main research projects.

We present here the section on some common inequalities in probability.

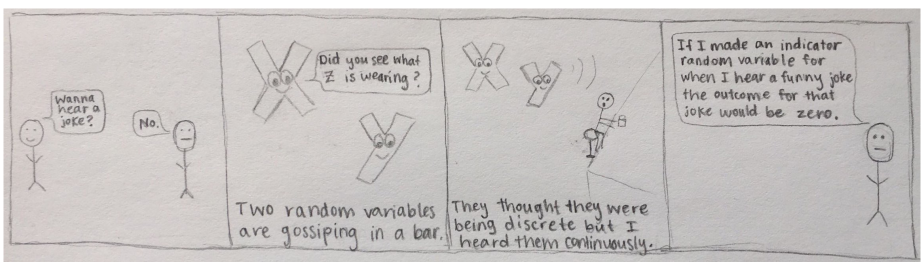

Figure 1: RV joke

1 Inequalities

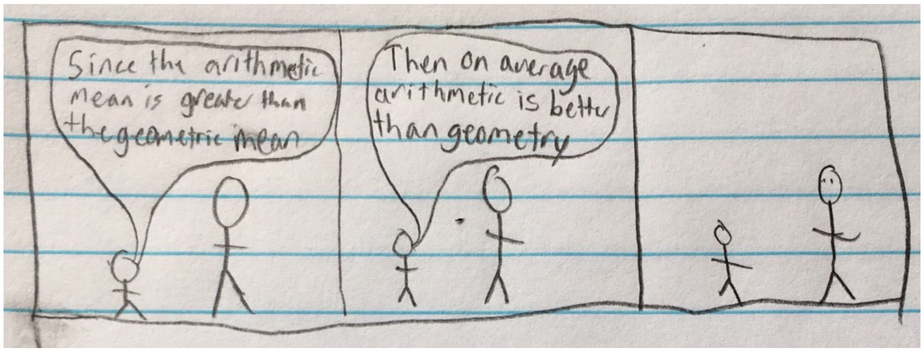

Figure 2: Inequality Joke

Inequalities are extremely important because they provide tools to control various quantities, usually when no exact expression is readily available. Two examples:

- Markov's inequality, Theorem 1.1, allows us to bound probabilities through the expectation.
- the Cauchy-Schwarz inequality, Theorem 1.2, allows us to bound the expectation of a product of random variables (obviously, unnecessary if they are independent, but non-trivial in general) in terms of the respective second moments.

We will detail four fundamental inequalities, go through their proofs, and demonstrate uses.

1.1 Markov's Inequality

According to the World Health Organization, an average (expectation) of $3,287$ deaths each day are caused by car crashes. Markov's inequality allows us to use this average and only the average to conclude that the probability of a catastrophic day with at least 10 times deaths than the average is no more than 10%, and a day with 100 times the average has probability less than 1%. These estimates are pretty crude, but with more information (like higher moments rather than expectation), Markov's inequality can be substantially refined. For some applications that are more "qualitative" in nature, we don't even need a sharp bound. An example is the Weak Law of Large Numbers.

Assume Y is a nonnegative random variable and let $y > 0$. Then

$$\begin{aligned}\mathbb{E}[|Y|] &\geq \mathbb{E}[|Y|\mathbf{1}_{\{|Y|\geq y\}}] \\ &\geq \mathbb{E}[y\mathbf{1}_{\{|Y|\geq y\}}] \\ &= yP(Y \geq y).\end{aligned}$$

This proves the following:

Theorem 1.1 (Markov's Inequality)**.** *Let Y be a nonnegative RV and let $y > 0$. Then*

$$P(Y \geq y) \leq \frac{\mathbb{E}[Y]}{y}.$$

Example 1.1. *According to the CDC's Vital and Health Statistics, Series 3, Number 39, the average height for a woman aged 20 and over living in the U.S. is 161.8 cm. Using Markov's inequality, we see that the probability that a woman is over 180 cm tall is no greater than 89.9 percent.*

Problem 1.1. *Let's assume all women 20 and up are at least 120cm tall. Use this assumption and the data from Example 1.1 to show that the proportion of women taller than 180cm is not larger than 70%. Still far off, but at least a number...*

Example 1.2. *According to the Social Security Administration, the life expectancy of men is* 84 *years. As a result of Markov's inequality we can see that the probability of a man living past the age of 100 is no greater than 84 percent.*

One important utility of Markov's Inequality is that you can precompose the random variable with a monotonically-increasing function before applying Markov's inequality.

For example, precomposing with the function $x \mapsto x^2$ gives us Chebyshev's inequality. Fix a RV X with finite second moment. Let $Y = (X - E[X])^2$. For $x > 0$, the event $\{Y \geq x^2\}$ coincides with the event $\{|X - E[X]| \geq x\}$. Thus,

$$P(|X - E[X]| \geq x) = P(Y \geq x^2) \leq \frac{E[Y]}{x^2} = \frac{\sigma_X^2}{x^2},$$

Where σ_X^2 is the variance of X. The inequality thus obtained is so important it even has a name:

Corollary 1.1 (Chebyshev's inequality)**.** *Let X be a random variable, with finite second moment. Then*

$$P(|X - \mathbb{E}[X]| \geq x) \leq \frac{\sigma_X^2}{x^2}.$$

One can repeat this trick. Here is another useful and celebrated example. Precomposing with the function $x \mapsto e^{sx}$ (where s is any constant, usually chosen later to maximize the strength of the inequality) gives bounds known as Chernoff bounds.

- Suppose $s > 0$. Let $Y = e^{sX}$. Again, Y is nonnegative. The event $\{X \geq x\}$ coincides with the event $\{Y \geq e^{sx}\}$. By Markov's inequality,

$$P(X \geq x) = P(Y \geq e^{sx}) \leq \frac{E[e^{sX}]}{e^{sx}}.$$

- Suppose now $s < 0$. Let $Y = e^{sX}$. Then the event $\{X \leq x\}$ coincides with the event $\{Y \geq e^{sx}\}$. Again, Markov's inequality gives

$$P(X \geq x) \leq \frac{E[e^{sX}]}{e^{sx}}.$$

Notice that $s \to E[e^{sX}]$ is the moment-generating function for the random variable X, and since the bounds obtained above are valid for any s suchthat $E[e^{sX}]$ is finite, we obtain the following, known as Chernoff bounds:

Corollary 1.2 (Chernoff Bound)**.** *Let Y be a random variable, then*

$$P(X \geq x) \leq \inf_{s>0} \frac{E[e^{sX}]}{e^{sx}}$$
$$P(X \leq x) \leq \sup_{s<0} \frac{E[e^{sX}]}{e^{sx}}.$$

Let's see how one can Chernoff's bounds to get some time estimates for probabilities otherwise elusive. Consider a Poisson RV X with parameter λ. The moment generating function for X is equal to

$$E[e^{sX}] = e^{\lambda(e^s-1)},$$

and its calculation is nothing but summation of the Taylor series for e^x. Finding tail probabilities like $P(X > x)$ for large x is harder and has no closed formula. By Chernoff's bounds, for $s > 0$,

$$P(X \geq x) \leq \exp(\lambda(e^s - 1) - sx).$$

Differentiate the exponent on the right-hand side with respect to s to find a minimum. The minimum is attained at $\lambda e^s = x$, and therefore the exponent is equal to $x - \lambda - x\ln(x/\lambda))$. Choosing $x = (1+r)\lambda$, this simplifies to $r\lambda - (1+r)\lambda\ln(1+r)$. Since $\lambda = E[X]$, we obtain the following bound:

$$P(X - E[X] \geq rE[X]) \leq (1+r)^{-\lambda}(\frac{e}{1+r})^{r\lambda}.$$

For a concrete example, ...

We used this applet to calculate the exact probability.

The Chernoff bound can be used to bound probabilities that would be difficult to compute through other methods. Take, for instance, a Poisson distribution with rate $\lambda = 15$. It is possible to calculate the probability that the random variable takes a value ≥ 20, but that would require that you calculate and take one minus the sum of the probabilities of observing values 0 through 19. While this is possible, it is tedious. You could instead employ the Chernoff bound since the MGF for a Poisson is known.

1.2 Cauchy-Schwarz Inequality

The Cauchy-Scwarz Inequality allows us to "decouple" expectations of products of random variables and provide bounds in terms of the second moments of the individual RVs. The inequality is based on the positivity of the square function (as well as positivity and linearity of expectation).

Theorem 1.2 (Cauchy-Schwarz Inequality)**.** *Let X and Y be random variables. Then,*

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Furthemore, equality holds if and only if one of the RVs is a constant multiple of the other with probability 1*.*

Proof. Clearly, it is enough to show assuming X and Y are nonnegative. We will work under this assumption. We may also assume $E[X^2]E[Y^2] > 0$ because otherwise at least one of the RVs is equal to zero with probability 1, and then the inequality is trivial (both sides are zero). We begin by analyzing the special case. Assume X', Y' are nonnegative RVs satisfying $E[(X')^2] = E[(Y')^2] = 1$. We have

$$0 \leq E[(X' - Y')^2] = 2 - E[X'Y'].$$

Therefore

$$E[X'Y'] \leq 1. \tag{1}$$

To complete the proof for the general case, assume X, Y are nonnegative RVs with $E[X^2]E[Y^2] > 0$. Let $X' = X/\sqrt{E[X^2]}$ and $Y' = Y/\sqrt{E[Y^2]}$. Then from (1) we have

$$\frac{E[XY]}{\sqrt{E[X^2]}\sqrt{E[Y^2]}} = E[X'Y'] \leq 1,$$

and the result follows. □

An important use of the Cauchy-Schwarz inequality is in the concepts of covariance and correlation of two RVs. The covariance of two RVs X and Y, $\text{Cov}(X, Y) = E[(X - E[X])\,(Y - E[Y])]$. This is a generalization of the notion of variance (in which case $Y = X$). Note that from linearity of the expectation,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

What does it mean in "real life"? Note that if X and Y are independent, then the covariance is zero. More generally, we say that X, Y are uncorrelated if $\text{Cov}(X, Y) = 0$. Independence implies being uncorrelated but the converse is not true in general.

By the Cauchy-Schwarz inequality,

$$|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y, \tag{2}$$

and an inequality holds if and only if for some constant c, $X = cY$ or $Y = cX$.

Example 1.3. *Suppose the gas price per gallon in the month of February is a RV X with expectation* \$2.99*, and that the amount of gallons per month consumed by a household during that month is a RV Y with expectation* 400*. The expected amount spent on gas by a family is then $E[XY]$, and, in general, it need not coincide with the product of the respective expectations, which is* $\$2.99 * 400 = 1196$*. The covariance gives us the difference between the former and the latter. Suppose now we also know that $\sigma_X = \$1$ and $\sigma_Y = 200$. Then by* (**??**) *we have*

$$|E[XY] - 1196| \leq 200.$$

So without additional information on the intricate details of how X and Y are dependent of each other, we have a bound on the average amount spent on gas by household.

It is often convenient to express correlation as a "unitless" or "normalized" number. We define the correlation of X and Y

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, y)}{\sigma_X \sigma_Y},$$

whenever $\sigma_X \sigma_Y > 0$ (the alternative is X or Y is a constant RV). From (2), we have

$$\mathrm{Cor}(X, Y) \in [-1, 1],$$

wirh corrlation equal to ± 1 if and only if $X = cY$ or $Y = cX$ for some constant c, and in this case the sign of c coincides with the sign of the correlation.

Cauchy-Schwarz can be seen as a special case of the more general Hölder's inequality:

Theorem 1.3 (Hölder's Inequality)**.** *Let X, Y be positive random variables. Then for any $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$|\mathbb{E}[XY]| \leq \mathbb{E}[X^p]^{1/p} \mathbb{E}[Y^q]^{1/q}.$$

(Cauchy-Schwarz is the special case where we apply Hölder's inequality with $p = q = 2$ to $|X|, |Y|$.)

In fact, Cauchy-Schwarz can be used to prove Hölder's inequality. The proof we present below is from *A proof of H ölder's inequality using the Cauchy-Schwarz inequality*, by Li and Shaw, Journal of Inequalities in Pure and Applied Mathematics. Vol. 7-(2), 2006.

In the proof, we will use multiple times the fact that a function (which is allowed to take on values of ∞) is convex if and only if it is midpoint-convex and continuous. Since this isn't a real analysis text, we will leave the proof of this as an exercise.

Proof. We will assume that X, Y are positive; it suffices to prove the inequality in this case, for if X, Y are any random variables and p, q are numbers such that the inequality makes sense (e.g. p and q are integers), we can get the desired inequality from Hölder's inequality for the random variables $|X|$ and $|Y|$.

Define the function
$$F(t) := \mathbb{E}[X^{pt}Y^{q(1-t)}].$$
Then $F(1) = \mathbb{E}[X^p], F(0) = \mathbb{E}[Y^q]$, and $F\left(\frac{1}{p}\right) = \mathbb{E}[XY]$. We then want to show
$$F\left(\frac{1}{p}\right) \leq F(0)^{1/p} + F(1)^{1/q} = F(0)^{\frac{1}{p}} + F(1)^{\left(1-\frac{1}{p}\right)}.$$
Observing the similarity to the definition of a convex function, we take logs of both sides to get the equivalent inequality
$$\log F\left(\frac{1}{p}\right) \leq \frac{1}{p}\log F(0) + \left(1 - \frac{1}{p}\right)\log F(1).$$
It is therefore sufficient to show that $\log F$ is convex.

To prove that F is midpoint-convex, we use the Cauchy-Schwarz inequality.

We want to show that, for any a, b,
$$\log F\left(\frac{a+b}{2}\right) \leq \frac{1}{2}\log F(a) + \frac{1}{2}\log F(b).$$
Exponentiating both sides, we want to show
$$F\left(\frac{a+b}{2}\right) \leq \sqrt{F(a)}\sqrt{F(b)}$$
or equivalently (by expanding the definition of F)
$$\mathbb{E}[X^{p\left(\frac{a+b}{2}\right)}Y^{q\left(1-\frac{a+b}{2}\right)}] \leq \sqrt{\mathbb{E}[X^{pa}Y^{q(1-a)}]}\sqrt{\mathbb{E}[X^{pb}Y^{q(1-b)}]}$$
which is precisely the statement of Cauchy-Schwarz applied to $X^{\frac{pa}{2}}Y^{\frac{q(1-a)}{2}}$ and $X^{\frac{pb}{2}}Y^{\frac{q(1-b)}{2}}$.

It remains to show that $\log F$ is continuous. It is equivalent to show that F is continuous (and positive, but F is clearly positive from the definition). Again using the equivalence of convexity with midpoint-convexity and continuity, it suffices to show that F is convex.

To show that F is convex, rewrite F as
$$F(t) = \mathbb{E}[Y^q(X^pY^{-q})^t].$$
Now observe that, for any fixed ω in the sample space Ω over which X, Y are defined, $Y(\omega)^q(X(\omega)^pY(\omega)^{-q})^t$ is a convex function in t. That is, for any $\alpha \in [0,1]$,
$$Y(\omega)^q(X(\omega)^pY(\omega)^{-q})^t \leq \alpha Y(\omega)^q + (1-\alpha)X(\omega)^p.$$
Taking the expected value of both sides gives the convexity of F. (Here we are implicitly using the fact that if A, B are random variables with $A \leq B$ with probability 1, then $\mathbb{E}[A] \leq \mathbb{E}[B]$; we leave the proof of this as an exercise.) $\square$

1.3 Jensen's Inequality

Jensen's inequality is a powerful tool allowing to bound expectation of convex functions of RVs. To a large extent, it can be viewed as a generalization of the idea of using the positivity of the square function to obtain Cauchy-Schwarz to more general functions. By Cauchy-Schwarz,

$$E[X]^2 \leq E[X^2].$$

Jensen's inequality generalizes this to any convex function ϕ:

$$\phi(E[X]) \leq E[\phi(X)].$$

Let X be a random variable with finite expectation. Consider the graph of a convex function ϕ.

Then consider the left-tangent line l of the graph of ϕ at $\mathbb{E}[X]$ - it lies below the graph of ϕ by the definition of convexity. This simple fact almost immediately implies our inequality.

Theorem 1.4 (Jensen's Inequality)**.** *Let X be a random variable and ϕ be a convex function. Then,*

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Observe that if φ is concave, then $\phi = -\varphi$ is convex. Therefore we have

$$E[\varphi(X)] \leq \varphi(E[X]).$$

Proof. Since ϕ is convex, if for all $x_0, x \in \mathbb{R}$,

$$\phi(x) \geq \phi(x_0) + \phi'_-(x_0) \cdot (x - x_0),$$

where ϕ'_- is the derivative of ϕ from the left. Let $x = X$ and $x_0 = \mathbb{E}[X]$, then:

$$\phi(X) \geq \phi(\mathbb{E}[X]) + \phi'_-(\mathbb{E}[X]) \cdot (X - \mathbb{E}[X])$$

Then take the expectation of both sides of the inequality.

$$\begin{aligned}\mathbb{E}[\phi(X)] &\geq \mathbb{E}[\phi(\mathbb{E}[X])] + \mathbb{E}[\phi'_-(\mathbb{E}[X]) \cdot (X - \mathbb{E}[X])] \\ &= \mathbb{E}[\phi(\mathbb{E}[X])] + \phi'_-(\mathbb{E}[X])\mathbb{E}[X - \mathbb{E}[X]] \\ &= \phi(\mathbb{E}[X]).\end{aligned}$$

□

Now let's look at a corollary of Jensen's Inequality.

Corollary 1.3. *Let $h(p) = (\mathbb{E}[|X|^p])^{\frac{1}{p}}$ where $p \geq 1$. If $p_1 \leq p_2$, then $h(p_1) \leq h(p_2)$.*

Proof. Let $0 < p_1 \le p_2$. Since $p_2/p_1 > 1$, the function $\phi(x) = x^{p_2/p_1}$ is convex. We have

$$\begin{aligned}(h(p_2))^{p_2} &= (\mathbb{E}[|X|^{p_2}]^{\frac{1}{p_2}})^{p_2} \\ &= \mathbb{E}[|X|^{p_2}] \\ &= \mathbb{E}[(|X|^{p_1})^{\frac{p_2}{p_1}}] \\ &\ge (\mathbb{E}[|X|^{p_1}])^{\frac{p_2}{p_1}} \\ &= h(p_1)^{p_2}.\end{aligned}$$

The result follows. $\square$

Note that in the proof above, we used Jensen's inequality to obtain $\mathbb{E}(|X|^{p_1})^{\frac{p_2}{p_1}} \ge (\mathbb{E}(|X|^{p_1}))^{\frac{p_2}{p_1}}$ by picking $\phi(Z) = |Z|^{\frac{p_2}{p_1}}$, which is a convex function, as $p_2 \ge p_1$. Then, $\mathbb{E}(|X|^{p_1})^{\frac{p_2}{p_1}} = \mathbb{E}(\phi(|X|^{p_1}) \ge \phi(\mathbb{E}(|X|^{p_1})) = (\mathbb{E}(|X|^{p_1}))^{\frac{p_2}{p_1}}$.

Example 1.4. *Consider you have been able to measure the kinetic energy due to their motion of stars the same mass as the sun in the galaxy. Then since kinetic energy is related to motion by $KE = \frac{1}{2}mv^2$ and in this case mass is constant. Then KE is related to velocity by a concave function with $v \propto \sqrt{KE}$. So by Jensen's inequality the average velocity will be less than the square root of the average kinetic energy.*

Example 1.5. *Show that the mean of squares exceeds the square of the mean for a set of natural numbers of the form $\{1, 2, ..., N\}$*

Here, $\phi(x) = x^2$, and each number in the set has weight (probability) of $\frac{1}{N}$.
This gives $\phi(\mathbb{E}[X]) = (\frac{1}{N}\sum_{i=1}^{N} x_i)^2$ and $\mathbb{E}[\phi(X)] = \frac{1}{N}\sum_{i=1}^{N} x_i^2$
Thus, by Jensen's inequality:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^2 \ge (\frac{1}{N}\sum_{i=1}^{N} x_i)^2.$$

Note that the function ϕ can be replaced by any convex function.

Jensen's inequality is very important because it can be used in many cases to simplify mathematical problems. Specifically, this inequality has been used in many proofs such as the arithmetic and geometric means inequality:

Example 1.6 (Arithmetic-Geometric Mean inequality)**.** *Suppose $x_1.x_2, \ldots, x_n$ are positive numbers and $w_1, \ldots, w_n$ are positive numbers, satisfying $\sum_{i=1}^{n} w_i = 1$. Then*

$$\sum_{i=1}^{n} w_i x_i \ge \prod_{i=1}^{n} x_i^{w_i}.$$

Le X be a RV with $P(X = \ln x_i) = w_i$. Since $x \to e^x$ is convex,

$$e^{E[X]} \le E[e^X]$$

The left-hand side is $\prod_{i=1^n} x_i^{w_i}$. The right-hand side is $\sum_{i=1}^{n} w_i x_i$.

One can leverage Jensen's inequality is by changing the probability measure. This will allow us to extend Jensen's inequality beyond the single RV setup. The key is a simple trick, change of measure: we will "absorb' one RV into a newly-formed probability measure.

To explain the notion of changing the measure, let Z be any positive random variable on a probability space $(\Omega, \mathcal{F}, P)$ with finite expectation. We can then define a new probability measure Q on $(\Omega, \mathcal{F})$ through

$$Q(A) := \frac{\mathbb{E}_P[Z \cdot 1_A]}{\mathbb{E}_P[Z]}$$

for any $A \in \mathcal{F}$. Note that here, to avoid confusion, we are using $\mathbb{E}_P$ to denote the expected value with respect to the probability measure P. We will also write $\mathbb{E}_Q$ for expectation with respect to Q. Checking that this satisfies the axioms of a probability measure is routine as is checking that if Y is any random variable, then $\mathbb{E}_Q[Y] = \frac{\mathbb{E}_P[Z \cdot Y]}{\mathbb{E}_P[Z]}$. Both are obtained through monotone convergence. The latter also uses the linearity of expectation. Note that if X is an indicator random variable for the event A, then Q is the measure induced by conditioning on A.

We can now apply Jensen's inequality with this new probability measure: for any convex function ϕ and any random variables W, $\phi(\mathbb{E}_Q[W]) \leq \mathbb{E}_Q[\phi(W)]$. Unwinding definitions,

$$\phi\left(\frac{\mathbb{E}_P[Z \cdot W]}{\mathbb{E}_P[Z]}\right) \leq \frac{\mathbb{E}_P[Z \cdot \phi(W)]}{\mathbb{E}_P[Z]}$$

or equivalently

$$\mathbb{E}_P[Z \cdot W] \leq \phi^{-1}\left(\frac{\mathbb{E}_P[Z \cdot \phi(W)]}{\mathbb{E}_P[Z]}\right)\mathbb{E}_P[Z], \tag{3}$$

whenever ϕ has an inverse.

Let's look at a special case. We'll use Jensen's inequality to derive Holder's inequality, Theorem 1.1. Let X and Y be two positive RVs. The choice of function in Jensen's inequality is easy. Fix $p > 1$ and let $\phi(x) = x^p$.

Next, looking carefully at (4), in order to derive Holder's inequality, Theorem 1.3, we want to come up with two RVs Z, W such that

1. $ZW = XY$; and
2. Z is only a function of X; and
3. $Z\phi(W)$ is only a function of Y.

Let's try to "solve". From the first two constraints, $Z = XY/W$ is only a function of X, therefore there exists some function h such that $W = Y/h(X)$. Rewritten, $Z = Xh(X)$. Plug this into the third constraint to have $Xh(X)\phi(Y/h(X))$ is only a function of Y, that is, $Xh(X)/h(X)^p$ must be a constant c. In light of this we select $h(x) = x^{1/(p-1)}$ and all constraints are satisfied, with the constant $c = 1$. Let's write down (4) in terms of X an Y. The lefthand side is, by construction, equal to $E[XY]$. Now $Z\phi(W) = \phi(Y) = Y^p$, and $Z = X^{\frac{p}{p-1}}$. Putting it all together, we obtained:

$$\mathbb{E}_P[X \cdot Y] \leq \left(\frac{\mathbb{E}_P[Y^p]}{\mathbb{E}_P[X^{\frac{p}{p-1}}]} \right)^{1/p} \mathbb{E}_P[X^{\frac{p}{p-1}}]. \tag{4}$$

Setting $q = p/(1-p)$, we observe $1/q + 1/p = 1$, and so the righthand side becomes

$$(\mathbb{E}_P[Y^p])^{1/p}(\mathbb{E}_P[X^q])^{1/q},$$

and we're all set. Nice?

1.4 Harris Inequality

Harris' inequality is the intuitive statement that that two increasing functions of the same RV are positively correlated.

Theorem 1.5 (Harris Inequality)**.** *Suppose f and g are non-decreasing functions. Let X be a random variable. Then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)],$$

whenever both sides are defined.

Why does the inequality hold? To get some insight, let's look at a discrete version. Suppose $a_1, a_2, \ldots, a_n$ is a nondecreasing sequence $a_1 \leq a_2 \leq \cdots \leq a_n$ and so are $b_1, b_2, \ldots, b_n$. Let's look at the average product.

$$\frac{1}{n} \sum_{i=1}^{n} a_i b_i.$$

What would happen to this product if we start moving things around? Let's play! Fix $i_1 < i_2$. Let's "move mass" from the "larger" b_{i_2} to the "smaller" b_{i_1} by setting $b'_{i_2} = b_{i_2} - \epsilon$ and $b'_{i_1} = b_{i_1} + \epsilon$ with $\epsilon \in (0, b_{i_2})$. Then the net change in the average product is $\frac{\epsilon}{n}(a_{i_1} - a_{i_2}) \leq 0$. Aha! Repeating this in a clever way until all b'_i are the same we keep lowering the average sum while preserving the average value of the b'_i at every step. Clearly the resulting constant sequence $b'_1, \ldots, b'_n$ has each element equal to the average of our original sequence, $\frac{1}{n}\sum_{i=1}^{n} b_i$. In other words

$$\frac{1}{n} \sum_{i=1}^{n} a_i b_i \geq \left(\frac{1}{n} \sum_{i=1}^{n} a_i \right) \left(\frac{1}{n} \sum_{i=1}^{n} b_i \right).$$

Now for a probabilistic proof.

Proof. Suppose X, Y are identical random variables. Since f, g are non-decreasing functions,

$$\mathbb{E}[(f(X) - f(Y))(g(X) - g(Y))] \geq 0$$

because if $f(X) > f(Y)$, $g(X) > g(Y)$ as well, and vice versa. If we expand the inequality above, we obtain

$$\mathbb{E}[f(X)g(X)] - \mathbb{E}[f(Y)g(X)] - \mathbb{E}[f(X)g(Y)] + \mathbb{E}[f(Y)g(Y)] \geq 0$$

Since X, Y are independent, we can write $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Therefore

$$\mathbb{E}[f(X)g(X)] + \mathbb{E}[f(Y)g(Y)] - \mathbb{E}[f(Y)]\mathbb{E}[g(X)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)] \geq 0$$

Since X and Y have the same distribution function, we replace Y with X and obtain

$$2\mathbb{E}[f(X)g(X)] - 2\mathbb{E}[f(X)]\mathbb{E}[g(X)] \geq 0$$

Therefore,

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$$

□

Example 1.7. *The tax rate for an income is the proportion of that income that is collected as tax. Suppose that the average tax rate per household is 5%, and that the average income per household is $100,000. Let's also assume that the tax rate is higher if the income is higher.*

Then, the Harris Inequality implies that the average tax collected per household is at least as large as the product of the expected tax rate times the expected income. Therefore, we should have at least $5,000 collected per household on average.

Just to show the power of the inequality, we will verify the trivial fact that the variance of a random variable is always positive or zero.

Example 1.8. *Use Harris's inequality to verify that for a random variable X, the variance* $Var(X) \geq 0$.

Proof. We know that for all X, $f(X) = X$ is monotonically increasing. Therefore, we can utilize the Harris inequality with f(X) = g(X) = X. Plugging this in, we see

$$\mathbb{E}[f(X)g(X)] \geq E[f(X)]E[g(X)]$$
$$\mathbb{E}[X^2] \geq E[X]^2$$

Subtracting $E[X^2]$ from both sides and utilizing the relationship between variance and expectation shows that the inequality holds.

□