University of Illinois
Spring 2017 — Econ 574 — Department of Economics
Roger Koenker

Lecture 2
"One too many inequalities"

In lecture 1 we introduced some of the basic conceptual building materials of the course. In this lecture we introduce some basic tools for handling these materials, these tools consist of a sequence of inequalities which play a fundamental role in subsequent developments. Most of these results may be found in any introductory test; a very comprehensive appendix on probabilistic inequalities may be found in Shorack and Wellner (1986, *Empirical Processes*).

1. Classical Inequalities

We begin with the Markov inequality which may be regarding as a basic "inequality generating theorem."

Theorem 1. *Let Z be a real r.v. and g a nonnegative, nondecreasing function on the support of Z, i.e., a set B such that $P(Z \in B) = 1$, then,*

$$P(Z \geq a) \leq \frac{Eg(Z)}{g(a)}$$

Proof. The proof is very simple. By hypothesis,

$$g(a)I(Z \geq a) \leq g(Z)I(Z \geq a) \leq g(Z).$$

This can be interpreted as holding for each $\omega \in \Omega$, with $Z = Z(\omega)$. Taking expectations yields the result. ■

Examples: We will use the common convention that $t^+ = \max\{0, t\}$ denotes the positive part of t

1. Markov. $Z = |X|$, $g(t) = t^+ \Rightarrow P(|X| \geq a) \leq \frac{E|X|}{a}$

2. Chebychev. $Z = |X|$, $g(t) = t^2 \Rightarrow P(|X| \geq a) \leq \frac{EX^2}{a^2}$

3. Bernstein. $Z = X$, $g(t) = e^{st} \Rightarrow P(X \geq a) \leq \frac{Ee^{sZ}}{e^{sa}}$

Generally speaking, these inequalities are rather crude, but one can construct examples for which they are actually sharp. For example, in the case of the Markov inequality, suppose,

$$X = \begin{cases} a & \quad\quad \mu/a \\ & w.p. \\ 0 & \quad\quad 1 - \mu/a \end{cases}$$

Then, $EX = \mu$, and obviously $P(X \geq a) = \frac{EX}{a}$.

The next important inequality is the Cauchy Schwartz (or correlation) inequality.

Theorem 2. *Let $X = (X_1, \ldots, X_p)$ be a p-vector of real r.v.'s and $U = EXX'$. The matrix U is symmetric, nonnegative definite with singularity ($|U| = 0$) iff there exists a p-vector $\alpha \neq 0$ such that*

$$(*) \qquad E(\alpha' X)^2 = 0.$$

Proof. Since E is applied componentwise (and multiplication commutes) symmetry is immediate. Nonnegative definiteness follows from

$$\alpha' U \alpha = E(\alpha' X)^2 \geq 0.$$

If equality holds, then clearly (*) holds and U is singular, since $U\alpha = 0$. On the other hand, if U is singular, there must exist $\alpha \neq 0$ such that $U\alpha = 0$, so equality holds. ■

Corollary 1. *(Cauchy-Schwartz) For r.v.'s X_1, X_2*

$$(EX_1X_2)^2 \leq EX_1^2 EX_2^2$$

and centering X_1, X_2 at their respective means yields,

$$(\mathrm{Cov}(X_1, X_2))^2 \leq V(X_1)V(X_2)$$

Proof. Specializing the previous result to $p = 2$, and recalling that $|U|$ may be expressed as the product of its eigenvalues which are nonnegative we have

$$0 \leq |U| = EX_1^2 EX_2^2 - (EX_1X_2)^2$$

and we obtain the second form by centering. ■

Remark: Note that the latter form implies the well-loved correlation inequality,

$$0 \leq \rho^2 \equiv \frac{\mathrm{Cov}(X_1, X_2)^2}{V(X_1)V(X_2)} \leq 1.$$

The simple result $V(X) \geq 0$ which implies

$$EX^2 \geq (EX)^2$$

can be thought of as a special case of the following general result. It plays a crucial role in establishing consistency of the maximum likelihood estimator, among many other applications.

Theorem 3. *(Jensen) If X and $g(X)$ are integrable r.v.'s and $g(\cdot)$ is convex, then*

$$g(EX) \leq Eg(X)$$

Proof. Convexity of g implies that for any ξ there exists a line L through the point $(\xi, g(\xi))$ such that the graph of g is above the line, i.e.,

$$g(x) \geq g(\xi) + \lambda(x - \xi).$$

Let $\xi = EX$, then for all x,

$$g(x) \geq g(EX) + \lambda(x - EX).$$

Note that λ depends on ξ, but not on x. Now let $x = X$ and we get

$$g(X) \geq g(EX) + \lambda(X - EX).$$

And taking expectations yields the result. This can be interpreted geometrically in $\mathbb{R}^p$ in terms of supporting hyperplanes. ■

Corollary 2. *(Liapounov) $(E|X|^r)^{1/r}$ is $\nearrow$ in r for $r \geq 0$.*

Proof. By Jensen's inequality, since $|X|^r$ is convex in $|X|$ for $r \geq 1$ we have $(E|X|)^r \leq E|X|^r$ so $E|X| \leq (E|X|^r)^{1/r}$. Now replace $|X|$ by $|X|^q$ for $0 < q < r$, so

$$(E|X|^q)^{1/q} \leq (E|X|^{rq})^{1/rq} \equiv (E|X|^s)^{1/s}$$

where $s = rq$, for $0 < q < s < \infty$ since $r \geq 1$ so $q \leq rq = s$. ■

Remark: This is usually called the moment inequality, obviously it implies that if X "has a r^{th} moment", it also "has a q^{th} moment" for $q < r$ in the sense that the r^{th} is finite if the q^{th} is.

2. Exponential Inequalities

An important class of further *exponential inequalities* play a critical role in the asymptotic theory of empirical processes. They involve approximating the tail probabilities of sums of independent r.v.s. See, Pollard (Appendix B) for further details.

Lemma 1. *(Feller, I 3rd ed, p. 175) As $x \to \infty$ $1 - \Phi(x) \sim x^{-1}\phi(x)$ or more precisely, for any $x > 0$*

$$(x^{-1} - x^{-3})\phi(x) \leq 1 - \Phi(x) < x^{-1}\phi(x)$$

Proof. Clearly for $x > 0$,

$$(1 - 3x^{-4})\phi(x) \leq \phi(x) \leq (1 + x^{-2})\phi(x)$$

the result follows by integrating from x to ∞. Hint: It is obviously easier to just differentiate the expression in the statement of the theorem and then use: $\phi'(x) = -x\phi(x)$. ■

Now we would like to extend this result to sums of independent r.v.'s. By the Bernstein form of the Markov inequality we have for $S = \sum X_i$,

$$(*) \qquad P(S \geq \gamma) \leq e^{-t\gamma} E e^{tS} = e^{-t\gamma} \prod_{i=1}^{n} E e^{tX_i}$$

For $X_i \sim iid \quad \mathcal{N}(0,1)$ the situation is simple, we have for all i,

$$E e^{tX_i} = e^{\frac{1}{2}t^2}$$

We can choose t to minimize the bound, i.e.,

$$\min_t \left\{ \frac{n}{2} t^2 - t\gamma \right\} \Rightarrow t^* = \gamma / n$$

so we have the bound,

$$P(S \geq \gamma) \leq e^{-\gamma^2/2n}$$

For non-normal X_i we need to work a bit harder. Here is one of the first results along these lines.

Theorem 4. *(Hoeffding) Let $X_1, \ldots, X_n$ be independent r.v.'s with $EX_i = 0$ and for each i $P(X_i \in [a_i, b_i]) = 1$. For each $\gamma > 0$ and $S = \sum X_i$*

$$P(S \geq \gamma) \leq \exp\{-2\gamma^2 / \sum (b_i - a_i)^2\}$$

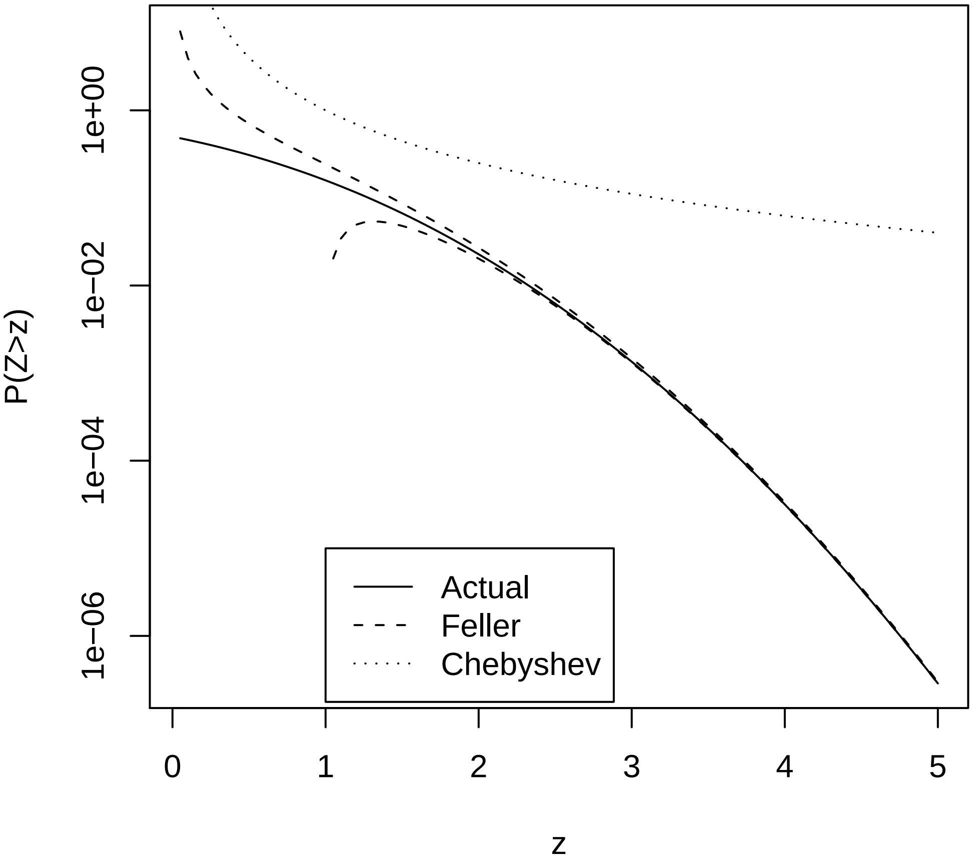

FIGURE 1. Approximation for the normal tail probability

Proof. By convexity

$$e^{tX} \leq e^{ta}\left(\frac{b-X}{b-a}\right) + e^{tb}\left(\frac{X-a}{b-a}\right)$$

Taking expectations, and using $EX = 0$

$$Ee^{tX} \leq e^{ta}\left(\frac{b}{b-a}\right) - e^{tb}\left(\frac{a}{b-a}\right)$$

Now set $\alpha = 1 - \beta = -a/(b-a)$ and $u = t(b-a)$ and rewrite this as

$$\log Ee^{tX} \leq \log(\beta e^{-\alpha u} + \alpha e^{\beta u})$$

or, factoring out $e^{-\alpha u}$, using $\alpha + \beta = 1$, and writing $\log Ee^{tX}$ as $\varphi(u)$ leaving the dependence of u on t implicit,

$$\varphi(u) \leq -\alpha u + \log(\beta + \alpha e^{u})$$

Expanding

$$\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2}u^2\varphi''(u^*)$$

where

$$\begin{aligned}\varphi'(u) &= -\alpha + \alpha/(\alpha + \beta e^{-u}) \\ \varphi''(u) &= \alpha\beta e^{-u}/(\alpha + \beta e^{-u})^2 \\ &= \left(\frac{\alpha}{\alpha + \beta e^{-u}}\right)\left(\frac{\beta e^{-u}}{\alpha + \beta e^{-u}}\right) \le \frac{1}{4}\end{aligned}$$

so

$$\varphi(u) = \frac{1}{2}u^2\varphi''(u^*) \le \frac{1}{8}u^2 = \frac{1}{8}t^2(b-a)^2$$

Applying this inequality to each X_i and then using the Bernstein inequality as in the normal case above, we have,

$$P(S \ge \gamma) \le e^{-t\gamma}\exp\{\frac{1}{8}t^2\sum(b_i - a_i)^2\}$$

Now minimizing with respect to t as before we have $t^* = 4\gamma/\sum(b_i - a_i)^2$ which yields the result. ■

Corollary 3. *Under the same conditions,*

$$P(|S| \ge \gamma) \le 2\exp\{-2\gamma^2/\sum(b_i - a_i)^2\}$$

Proof. Apply the same argument to $-X_i$ and combine. ■

Finally, we will state without proof the following generalization, see Pollard for details.

Theorem 5. *Let $X_1, \dots, X_n$ be independent r.v.'s with $EX_i = 0, EX_i^2 = \sigma_i^2$ and $P(|X_i| \le M) = 1$. Suppose $v = \sum\sigma_i^2$ then for each $\gamma > 0$,*

$$P(|S| > \gamma) \le 2\exp\{-\frac{1}{2}\gamma^2/(v + \frac{1}{3}M\gamma)\}$$

Remark: Note that for standard normal r.v.'s we don't need the $\frac{1}{3}M\gamma$ term.

3. Copulae and Fréchet Bounds

Given a pair, (X, Y) of random variables with joint distribution $G(x, y)$ and marginal distribution functions, $F(x)$ and $H(y)$, there is an associated copula function,

$$C(u, v) = G(F^{-1}(u), H^{-1}(v)).$$

Copulas are supposed to bind together two things: in linguistics, for example the subject and predicate of a sentence. In our setting F and H are bound together by G. The copula function $C : [0, 1]^2 \to [0, 1]$ provides a convenient way to describe dependence between X and Y that avoids the usual reliance on normal theory and global notions of correlation.

A first, basic question about copulas is: what kind of functions, C, can be valid copulae? Evidently, not just any function $F : [0, 1]^2 \to [0, 1]$ will do. Required properties of copulae are quite self evident: A copula function $C : [0, 1]^2 \to [0, 1]$ satisfies:

i. $C(u, 0) = C(0, v) = 0$, for all u and v in $[0, 1]$,
ii. $C(u, 1) = u$ and $C(1, v) = v$ for all u and v in $[0, 1]$,
iii. For all $u_1 < u_2$ and $v_1 < v_2$ in $[0, 1]$, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \ge 0$.

Note that $C(u,v)$ can be viewed as a distribution function for a pair of random variables, (U,V), that have uniform marginal distributions. The expression in (iii.) is simply the probability associated with the rectangle defined by the (u,v)'s and therefore must be non-negative. (Draw the picture!)

Theorem 6. *Let C be a copula, then for any $(u,v) \in [0,1]^2$*

$$\max\{u+v-1,0\} \leq C(u,v) \leq \min\{u,v\}.$$

Proof. Property (ii.) implies that $C(u,v) \leq C(u,1) \leq u$ and $C(u,v) \leq C(1,v) \leq v$ yielding the upper bound, and taking $u_1 = u$, $u_2 = 1$, $v_1 = v$, $v_2 = 1$ in (iii) implies that $C(u,v)-u-v+1 \geq 0$ and since $C(u,v) \geq 0$ we obtain the lower bound. ■

Given a copula function and the associated marginals, Sklar's theorem asserts that we can recover the joint distribution:

$$G(x,y) = C(F(x),H(y))$$

and thus we obtain the celebrated Fréchet bounds:

$$\max\{F(x)+H(y)-1,0\} \leq G(x,y) \leq \min\{F(x),H(y)\}.$$

These bounds play an important role in recent developments for partially identified models. Both bounds are sharp in the sense that they can be achieved. For the upper bound his occurs when X and Y are *comonotonic* that is when Y can be expressed as a deterministic, non-decreasing function of X. The lower bound is achieved when X and Y are countermonotonic, so Y is a deterministic, non-increasing function of X. These two very special cases correspond to the situations in which all of the mass of the copula function is concentrated on a curve connecting opposite corners of the unit square. These special cases correspond to rank correlation of +1 and -1 respectively. The other important special case is independent X and Y, which obviously corresponds to $C(u,v) = uv$.

An interesting, at least to me result about copulae is that one can construct conditional versions by the following differentiation trick.

Lemma 2. *(Bouyé and Salmon (2002)) $\partial C(u,v)/\partial u = F_{Y|X}(F_Y^{-1}(v)|F_X^{-1}(u))$.*

Proof. From

$$P(Y<y, X=x) = \partial F_{X,Y}(x,y)/\partial x$$

it follows that,

$$\begin{aligned} P(Y<y|X=x) &= (f_X(x))^{-1}\partial F_{X,Y}(x,y)/\partial x \\ &= (f_X(x))^{-1}\partial C(F_X(x),F_Y(y))/\partial x \\ &= (f_X(x))^{-1}\partial C(u,v)/\partial u\, \partial F_X(x)/\partial x, \end{aligned}$$

where the middle term is evaluated at $u = F_X(x)$ and $v = F_Y(y)$, and the last term cancels the first. ■

Examples: Roy's (1951) model of occupational choice is the Ur-text for the emerging literature on sample selection in micro-econometrics. In its simplest form individuals can be either fishers or hunters. Productivities in the two occupations have joint distribution function, $G(x,y)$ and marginal distributions, $F(x)$ and $H(y)$. Individuals, in accordance with their self interest choose their most productive occupation, so observed productivity is $Z = \max\{X,Y\}$, and thus has distribution function $\tilde{G}(z) = G(z,z)$. [Why?] Consequently, by Fréchet, the observed productivity probabilities obey,

$$\max\{0,F(z)+H(y)-1\} \leq G(x,y) \leq \min\{F(x),H(y)\}$$

As noted above, the upper bound corresponds to comonotonic productivities – good hunters make good fishers – while the lower bound holds when productivities are countermonotonic so the best hunters are the worst fishers and vice versa. In the absence of further evidence in which individuals are required to demonstrate their prowess in their "other" occupation, these bounds are all we are able to infer about the joint distribution. Very similar stories can be constructed for binary treatment models in which the occupations correspond to the control and treatment responses in settings in which participants are free to choose between the two options.

There is a rapidly expanding literature on copulae much of which has been motivated by finance where more general forms of dependence than the familiar normal, linear correlation models are needed. A standard reference is Nelsen (1999). In survival models with competing risks an important extension of the Fréchet bounds by Peterson (1976) has also proven to be useful. Jim Heckman has some nice notes on the Peterson results posted on the web.

4. References

Nelsen, R. B. (1999), *An Introduction to Copulas*, Springer.

Peterson, A.V. (1976) Bounds for a joint distribution function with fixed sub-distribution functions: Application to Competing Risks, *PNAS*, 73, 11-13,

Roy, A.D. (1951) Some Thoughts on the Distribution of Earnings *Oxford Economic Papers*, 3, 135-146.