# Probability and Stochastic Processes I - Lecture 18

Michael Evans
University of Toronto
https://utstat.utoronto.ca/mikevans/stac62/staC622023.html

2023

- there are several important inequalities we need to know: Markov's inequality, Cauchy-Schwartz inequality and Jensen's inequality

**Markov's Inequality**

**Proposition III.7.1** *(Markov's Inequality)* If $X$ is a nonnegative r.v. and $x > 0$, then

$$P(X \geq x) \leq \frac{E(X)}{x}$$

with equality iff $P(X = x) = 1 - P(X = 0)$.

Proof: We have

$$P(X \geq x) = E(I_{\{X \geq x\}}) \leq E\left(\frac{X}{x} I_{\{X \geq x\}}\right) = \frac{E\left(XI_{\{X \geq x\}}\right)}{x} \leq \frac{E(X)}{x}.$$

If $P(X = x) = 1 - P(X = 0)$, then $P_X$ is concentrated on $\{0, x\}$ and so $E(X) = xP(X = x) = xP(X \geq x)$. Conversely, if $E(X) = xP(X \geq x)$ at $x > 0$, then

$$0 = E(X) - E(xI_{\{X \geq x\}}) = E(XI_{\{X<x\}}) + E((X-x)I_{\{X \geq x\}})$$

and since $XI_{\{X<x\}}$ and $(X-x)I_{\{X \geq x\}}$ are both nonnegative r.v.'s this implies

$$E(XI_{\{X<x\}}) = E((X-x)I_{\{X \geq x\}}) = 0$$

which implies $1 = P(XI_{\{X<x\}} = 0) = P((X-x)I_{\{X \geq x\}} = 0)$ which implies $P(0 < X < x) = 0$ and $P(X > x) = 0$ which implies $P(X = x) = 1 - P(X = 0)$. ■

- **note** - Markov's inequality gives bounds on *tail probabilities* of $X$

**Exercise III.7.1** If $X$ is a r.v., then determine an upper bound for $P(\exp(tX) \geq k)$ when $k > 0$.

**Exercise III.7.2** If $X$ is a r.v. and $k > 0$, then prove $P(|X| \geq k) \leq E(|X|)/k$ and also $P(|X| \geq k) \leq E(X^2)/k^2$. If $X \sim$ exponential$(1)$ which inequality is sharper. What is the exact value of $P(X \geq 2)$ when $X \sim$ exponential$(1)$ and compare this with the bounds.

**Corollary III.7.2** *(Chebyshev's Inequality)* If $X$ has mean $\mu$ and variance $\sigma^2$, then for $k > 0$

$$P(|X - \mu| \geq k\sigma) \leq 1/k^2$$

with equality iff $P(X \in \{\mu - k\sigma, \mu + k\sigma\}) = 1 - P(X = \mu)$.

Proof: Since $|X - \mu|$ is nonnegative we can apply Markov and obtain

$$P(|X - \mu| \geq k\sigma) = P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E((X - \mu)^2)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

and the equality result follows as with Markov. ■

- note - $P(|X - \mu| \geq k\sigma) = P(X \geq \mu + k\sigma) + P(X \leq \mu - k\sigma)$ so Chebyshev is a bound on two tail probabilities of $X$

**Example III.7.1** *5 sigma*

- $P(|X - \mu| \geq 5\sigma) \leq 1/25 = 0.04$ while if $X \sim N(\mu, \sigma^2)$, then $P(|X - \mu| \geq 5\sigma) = 5.733031e - 07$ ■

**Corollary III.7.3** *(Chernoff Bounds)* If $E(\exp\{tX\})$ is finite for all $t \in (a, b)$ where $a < 0 < b$, then

$$\begin{array}{ll} P(X \geq x) \leq \inf_{t\in(0,b)} \left\{E\left(e^{tX}\right)e^{-tx}\right\} & \text{if} \quad x > 0 \\ P(X \leq x) \leq \inf_{t\in(a,0)} \left\{E\left(e^{tX}\right)e^{-tx}\right\} & \text{if} \quad x < 0 \end{array}$$

Proof: When $x > 0$, then for every $t \in (0, b)$, by Markov's inequality

$$\begin{aligned} P(X \geq x) &= P(tX \geq tx) = P(\exp\{tX\} \geq \exp\{tx\}) \\ &\leq E(\exp\{tX\})/\exp\{tx\}. \end{aligned}$$

When $x < 0$, then for every $t \in (a, 0)$, by Markov's inequality

$$\begin{aligned} P(X \leq x) &= P(tX \geq tx) = P(\exp\{tX\} \geq \exp\{tx\}) \\ &\leq E(\exp\{tX\})/\exp\{tx\}.\blacksquare \end{aligned}$$

**Example III.7.2** *Standard Normal*

- suppose $X \sim N(0,1)$ then

$$E\left(e^{tX}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2}\, dx = \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2}\, dx = e^{t^2/2}$$

so for $x > 0$

$$1 - \Phi(x) = P(X \geq x) \leq \inf_{t>0} e^{t^2/2 - tx} = e^{-x^2/2}$$

since $t^2/2 - tx$ is minimized at $t = x$

- note - $X_+$ has mean $E(X_+) = (2\pi)^{-1/2} \int_0^{\infty} xe^{-x^2/2}\, dx = (2\pi)^{-1/2}$ so using Markov's inequality when $x > 0$, then

$$1 - \Phi(x) = P(X \geq x) = P(X_+ \geq x) \leq (2\pi)^{-1/2}/x$$

but $e^{-x^2/2}/(1/x) = xe^{-x^2/2} \to 0$ as $x \to \infty$ so the Chernoff bound is better but even better bounds can be obtained ■

## Cauchy-Schwartz Inequality

**Proposition III.7.3** *(Cauchy-Schwartz Inequality)* If $E(X^2) < \infty, E(Y^2) < \infty$, then

$$|E(XY)| \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}$$

with equality iff $Y = cX$ (or $X = cY$) wp1 with $c = 0$ when $P(Y = 0) = 1$ $(P(X = 0) = 1)$ and $c = E(XY)/E(X^2)$ otherwise.

Proof: If $E(X^2) = 0$, then $P(X = 0) = 1$ which implies $P(XY = 0) = 1$ so $E(XY) = 0$ and $X = 0Y$ so the result follows. So assume hereafter that $E(X^2) > 0, E(Y^2) > 0$.

For any $c \in R^1$

$$\begin{aligned} 0 &\leq (Y - cX)^2 = Y^2 - 2cXY + c^2X^2 \text{ which implies} \\ 0 &\leq E(Y^2) - 2cE(XY) + c^2E(X^2) \end{aligned}$$

which is a convex parabola in $c$ with minimum at $c = E(XY)/E(X^2)$ so

$$0 \leq E(Y^2) - 2\frac{(E(XY))^2}{E(X^2)} + \frac{(E(XY))^2}{E(X^2)} = E(Y^2) - \frac{(E(XY))^2}{E(X^2)}$$

which gives the inequality. Equality occurs iff, when $c = E(XY)/E(X^2)$, $0 = E((Y - cX)^2)$ which occurs iff

$$1 = P((Y - cX)^2 = 0) = P(Y - cX = 0) = P(Y = cX). \blacksquare$$

**Corollary III.7.4** *(Correlation Inequality)* If $0 < \sigma_X^2 < \infty, 0 < \sigma_Y^2 < \infty$, then

$$-1 \leq \rho_{XY} = Corr(X, Y) \leq 1$$

with equality iff $Y \overset{\text{wp1}}{=} \mu_Y + \sigma_Y(X - \mu_X)/\sigma_X$ when $\rho_{XY} = 1$ and $Y \overset{\text{wp1}}{=} \mu_Y - \sigma_Y(X - \mu_X)/\sigma_X$ when $\rho_{XY} = -1$.

Proof: In CS inequality replace $X$ by $(X - \mu_X)/\sigma_X$ and $Y$ by $(Y - \mu_Y)/\sigma_Y$ so $E((X - \mu_X)^2/\sigma_X^2) = E((Y - \mu_Y)^2/\sigma_Y^2) = 1$ and so

$$|\rho_{XY}| = \left| E\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right| \leq 1$$

with equality iff

$$\left(\frac{Y - \mu_Y}{\sigma_Y}\right) \overset{\text{wp1}}{=} c\left(\frac{X - \mu_X}{\sigma_X}\right)$$

where

$$c = \frac{E\left(\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right)}{E\left(\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right)} = \rho_{XY}$$

which implies

$$Y \overset{\text{wp1}}{=} \mu_Y + \sigma_Y \rho_{XY}\left(\frac{X-\mu_X}{\sigma_X}\right)$$

and $\rho_{XY} = \pm 1$. ■

**note** - a measure of the *total variation* in $Y$ is given by

$$Var(Y) = E((Y-\mu_Y)^2)$$

- if we approximate $Y$ by $a + bX$ for some constants $a$ and $b$ then the amount of variation in $Y$ that is not explained (the *residual variation*) by $a + bX$ is

$$E((Y-a-bX)^2)$$

**Definition III.7.1** The *best affine predictor (linear regression)* of $Y$ from $X$ is given by $a + bX$ where $a, b$ are constants that minimize $E((Y-a-bX)^2)$. ■

**Exercise III.7.3** Assume $0 < \sigma_X^2 < \infty, 0 < \sigma_Y^2 < \infty$. Show that if $a, b$ minimize $E((Y - a - bX)^2)$, then $a_*, b_*$ with $a_* = a - \mu_Y + b\mu_X, b_* = b$ minimizes $E(((Y - \mu_Y) - a_* - b_*(X - \mu_X))^2)$ over all constants $a_*, b_*$.

**Exercise III.7.4** (i) Assume $\mu_X = \mu_Y = 0$ and $0 < \sigma_X^2 < \infty, 0 < \sigma_Y^2 < \infty$. For all constants $a, b$, and putting $c_{XY} = \sigma_Y \rho_{XY} / \sigma_X$, prove $E(Y - c_{XY} X) = 0, Cov(Y - c_{XY} X, a + bX) = 0$ and

$$E((Y - a - bX)^2 = Var(Y - c_{XY} X) + a^2 + (b - c_{XY})^2 Var(X).$$

Use this to prove that $c_{XY} X$ is the best affine predictor of $Y$ from $X$.
(ii) Combine (i) and Exercise III.7.3 to determine the best affine predictor of $Y$ from $X$ when the assumption of 0 means is not made.
(iii) Show that the proportion of the total variation in $Y$ explained by the best affine predictor from $X$ is given by $\rho_{XY}^2$.
(iv) When

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left( \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho_{XY} \\ \sigma_X \sigma_Y \rho_{XY} & \sigma_Y^2 \end{pmatrix} \right),$$

show that $E_{Y \mid X}(Y \mid x)$ equals the best affine predictor of $Y$ from $X$.