Inequalities

Probability I (BST 230)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Introduction

Markov's inequality
- Markov's inequality
- Chebyshev's inequality
- Chernoff's bound

Jensen's inequality
- Jensen's inequality
- Weighted AM-GM inequality
- Hoeffding's inequality

L^p norm inequalities
- L^p spaces
- Hölder's inequality
- Cauchy–Schwarz inequality
- Minkowski's inequality

Outline

Introduction

Markov's inequality
- Markov's inequality
- Chebyshev's inequality
- Chernoff's bound

Jensen's inequality
- Jensen's inequality
- Weighted AM-GM inequality
- Hoeffding's inequality

L^p norm inequalities
- L^p spaces
- Hölder's inequality
- Cauchy–Schwarz inequality
- Minkowski's inequality

Introduction

- Earlier we saw Boole's inequality and Bonferroni's inequality.
- There are many useful inequalities in probability theory.
- Inequalities are useful because it is usually easier to bound some quantity of interest than to characterize it exactly.
- And often, a decent bound is all that is needed to show what you want to show.
- *Note: To simplify things, in this set of slides we will generally assume that all expectations are finite.*

Introduction

- For example, suppose you are manufacturing widgets.
- Each widget can be defective in one of three ways, denoted by events A_1, A_2, A_3.
- You have data on the probability of each type of defect, $P(A_k)$, but you don't have any data on the joint probability of these events.
- Fortunately, you can still bound the probability of any type of defect occurring by using Boole's inequality:

$$P(A_1 \cup A_2 \cup A_3) \leq P(A_1) + P(A_2) + P(A_3).$$

- If each $P(A_k)$ is small, then you can guarantee that the probability of any defect occuring is small.

Recall: Boole's and Bonferroni's inequalities

- *Boole's inequality (a.k.a. union bound)*: For any $A_1, A_2, \ldots$,

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} P(A_i).$$

- *Bonferroni's inequality*: For any $A_1, A_2, \ldots$,

$$P\Big(\bigcap_{i=1}^{\infty} A_i\Big) \geq 1 - \sum_{i=1}^{\infty} P(A_i^c).$$

Outline

Introduction

Markov's inequality
- Markov's inequality
- Chebyshev's inequality
- Chernoff's bound

Jensen's inequality
- Jensen's inequality
- Weighted AM-GM inequality
- Hoeffding's inequality

L^p norm inequalities
- L^p spaces
- Hölder's inequality
- Cauchy–Schwarz inequality
- Minkowski's inequality

Markov's inequality

- This is one of the simplest but most useful inequalities in probability theory.

- *Markov's inequality*: If X is a nonnegative random variable and $a > 0$, then
$$P(X \geq a) \leq \frac{\mathrm{E}X}{a}.$$

- Proof: Since $1 \geq \mathbb{1}(X \geq a)$,
$$\begin{aligned} \mathrm{E}X &\geq \mathrm{E}X\mathbb{1}(X \geq a) \\ &\geq \mathrm{E}a\mathbb{1}(X \geq a) \\ &= aP(X \geq a). \end{aligned}$$

  Dividing both sides by a yields the result.

Example: Investing returns

- You invest \$1000 dollars in a holding where the annual returns are $\text{Pareto}(\alpha, c)$ distributed with $\alpha = 2$ and $c = 1/4$.

- More precisely, after n years, your investment is worth

$$Y_n = 1000 X_1 X_2 \cdots X_n$$

dollars, where $X_1, \ldots, X_n \sim \text{Pareto}(\alpha, c)$ independently.

- Recall that the pdf of $\text{Pareto}(\alpha, c)$ is

$$p(x) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}(x > c).$$

- Is this a good investment?

  Group exercise (5 minutes): First guess using your intuition. Then try to show something formally.

Corollaries of Markov's inequality

1. For any r.v. X and any $a > 0$,
$$P(|X| \geq a) \leq \frac{\mathrm{E}|X|}{a}.$$
2. For any r.v. X, any $a \in \mathbb{R}$, and any monotone increasing function $g(x) \geq 0$ such that $g(a) > 0$,
$$P(X \geq a) \leq \frac{\mathrm{E}g(X)}{g(a)}.$$
3. *Chebyshev's inequality*: For any r.v. X and any $a > 0$,
$$P\big(|X - \mathrm{E}X| \geq a\big) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Chebyshev's allows us to bound the probability that a r.v. is a certain distance from its mean.

Group exercise (5 minutes): Try to show 2 and 3 using Markov's inequality.

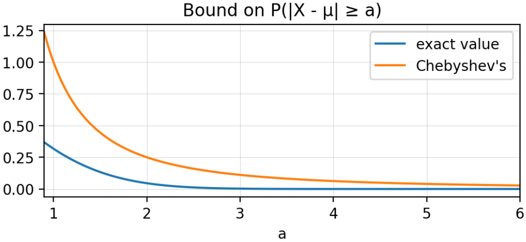
Example: Tail bounds for normal distributions

- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and we want to bound the probability that X is far from its mean.
- The exact expression involves the standard normal cdf $\Phi(x)$:

$$P\big(|X-\mu| \geq a\big) = P\left(\left|\frac{X-\mu}{\sigma}\right| \geq a/\sigma\right) = 2\Phi(-a/\sigma)$$

  for $a > 0$. However, $\Phi(x)$ does not have a simple closed form.
- Meanwhile, Chebyshev's inequality easily yields

$$P\big(|X-\mu| \geq a\big) \leq \frac{\mathrm{Var}(X)}{a^2} = \frac{\sigma^2}{a^2}.$$



Chernoff's bound

- This is surprisingly powerful corollary of Markov's inequality. It yields an exponentially decaying bound as a grows, compared to the $1/a$ in Markov's inequality.

- *Chernoff's bound*: For any r.v. X and any $a \in \mathbb{R}$,
$$P(X \geq a) \leq \inf_{t>0} e^{-ta} \operatorname{E} \exp(tX).$$

- Proof: For all $t > 0$,
$$\begin{aligned} P(X \geq a) &= P(tX \geq ta) \\ &= P(\exp(tX) \geq \exp(ta)) \\ &\leq \frac{\operatorname{E} \exp(tX)}{\exp(ta)} \\ &= e^{-ta} \operatorname{E} \exp(tX). \end{aligned}$$
  Since the left-hand side doesn't depend on t, the inequality holds when taking the infimum of the right-hand side over t.

Example: Tail bounds for normal distributions

- Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. By Chernoff's bound,

$$\begin{aligned} P(X - \mu \geq a) &\leq \inf_{t>0} e^{-ta}\, \mathrm{E} \exp\big(t(X-\mu)\big) \\ &= \inf_{t>0} e^{-ta}\, \exp\big(\tfrac{1}{2}\sigma^2 t^2\big) \\ &= \inf_{t>0} \exp\big(-ta + \tfrac{1}{2}\sigma^2 t^2\big) \end{aligned}$$

  using the formula for the mgf of $X - \mu \sim \mathcal{N}(0, \sigma^2)$.

- To minimize $f(t) = -ta + \frac{1}{2}\sigma^2 t^2$, we set

$$0 = f'(t) = -a + \sigma^2 t$$

  and solve to get $t = a/\sigma^2$. Plugging this in yields

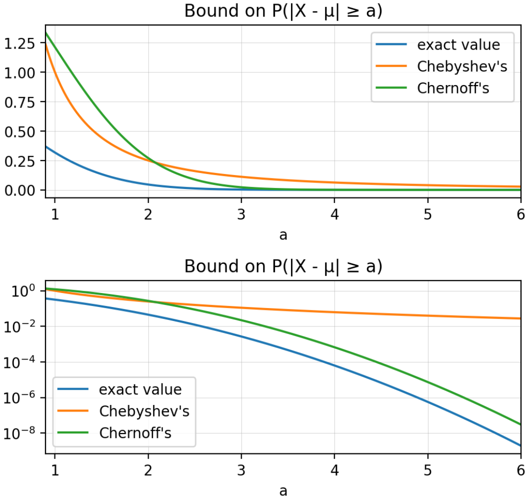
$$P(X - \mu \geq a) \leq \exp(-a^2/\sigma^2 + \tfrac{1}{2}a^2/\sigma^2) = \exp(-\tfrac{1}{2}a^2/\sigma^2).$$

- By symmetry, $P(-(X-\mu) \geq a) \leq \exp(-\frac{1}{2}a^2/\sigma^2)$. Thus,

$$P(|X - \mu| \geq a) \leq 2\exp(-\tfrac{1}{2}a^2/\sigma^2).$$

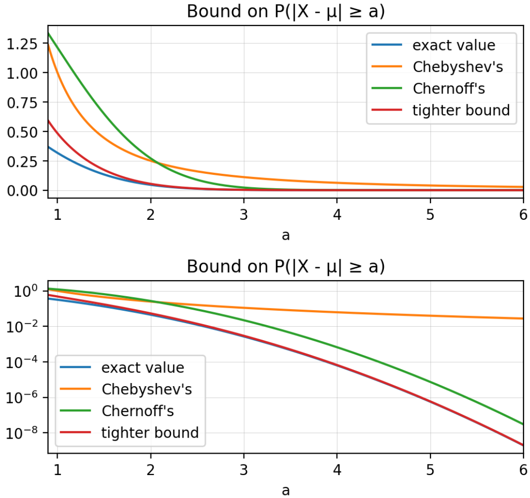
Example: Tail bounds for normal distributions

- Chebyshev's inequality: $P\big(|X-\mu| \geq a\big) \leq \sigma^2/a^2$.
- Chernoff's bound: $P(|X-\mu| \geq a) \leq 2\exp(-\frac{1}{2}a^2/\sigma^2)$.



Example: Tail bounds for normal distributions

- Chebyshev's inequality: $P\big(|X-\mu| \geq a\big) \leq \sigma^2/a^2$.
- Chernoff's bound: $P(|X-\mu| \geq a) \leq 2\exp(-\frac{1}{2}a^2/\sigma^2)$.
- A tighter bound: $P(|X-\mu| \geq a) \leq \sqrt{\frac{2\sigma^2}{\pi a^2}}\exp(-\frac{1}{2}a^2/\sigma^2)$.



Outline

Introduction

Markov's inequality
- Markov's inequality
- Chebyshev's inequality
- Chernoff's bound

Jensen's inequality
- Jensen's inequality
- Weighted AM-GM inequality
- Hoeffding's inequality

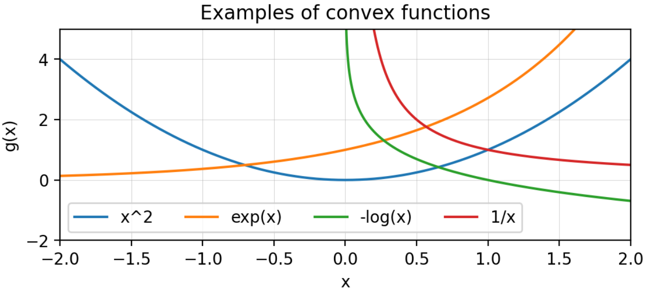
L^p norm inequalities
- L^p spaces
- Hölder's inequality
- Cauchy–Schwarz inequality
- Minkowski's inequality

Convex functions

- A function $g : \mathcal{X} \to \mathbb{R}$ is *convex* if
$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y)$$
for all $x, y \in \mathcal{X}$ and all $t \in (0, 1)$.

- A function $g : \mathcal{X} \to \mathbb{R}$ is *concave* if $-g$ is convex.

- Intuition: Convex functions curve upwards, concave functions curve downwards.



Properties of convex functions

- Suppose $g : \mathcal{X} \to \mathbb{R}$ is twice-differentiable at all $x \in \mathcal{X}$. Then g is convex if and only if

$$\frac{\partial^2}{\partial x^2} g(x) \geq 0$$

for all $x \in \mathcal{X}$.

- Suppose $g : \mathcal{X} \to \mathbb{R}$ is a convex function. For any $x_0 \in \mathcal{X}$, there exist $a, b \in \mathbb{R}$ such that

$$ax + b \leq g(x)$$

for all $x \in \mathcal{X}$ and

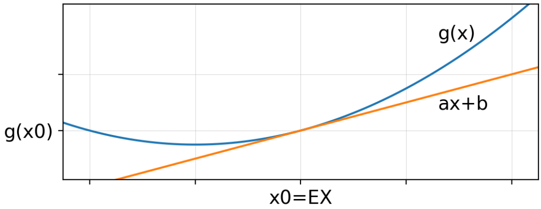
$$ax_0 + b = g(x_0).$$

Jensen's inequality

- This is a key inequality with many important consequences.

- *Jensen's inequality*: Let X be a r.v. with range $\mathcal{X}$.
  If $g : \mathcal{X} \to \mathbb{R}$ is a convex function then

$$g(\mathrm{E}X) \leq \mathrm{E}g(X).$$



- Proof: Define $x_0 = \mathrm{E}X$. Since g is convex, there exist $a, b \in \mathbb{R}$ such that $ax + b \leq g(x)$ for all $x \in \mathcal{X}$ and $ax_0 + b = g(x_0)$. Therefore,

$$g(\mathrm{E}X) = g(x_0) = ax_0 + b = \mathrm{E}(aX + b) \leq \mathrm{E}g(X).$$

Jensen's inequality: Examples

- Examples of Jensen's inequality:
  - $|\mathrm{E}X| \leq \mathrm{E}|X|$.
  - $(\mathrm{E}X)^k \leq \mathrm{E}X^k$ for all $k \in \{2, 4, 6, \ldots\}$.
  - If $X \geq 0$ then $(EX)^r \leq \mathrm{E}X^r$ for all $r \geq 1$.
  - $\exp(t\mathrm{E}X) \leq \mathrm{E}\exp(tX)$ for $t > 0$.
  - If $X > 0$ then $1/\mathrm{E}X \leq \mathrm{E}(1/X)$.
  - If $X > 0$ then $-\log(\mathrm{E}X) \leq -\mathrm{E}\log(X)$.

Weighted AM-GM inequality

- The inequality of arithmetic means and geometric means is a classic result that is easily proved using Jensen's inequality.

- *Weighted AM-GM inequality*: For any $x_1, \ldots, x_n \geq 0$ and $w_1, \ldots, w_n \geq 0$ such that $\sum_{i=1}^{n} w_i = 1$,

$$w_1 x_1 + \cdots + w_n x_n \geq x_1^{w_1} \cdots x_n^{w_n}.$$

Group exercise (5 minutes): Try to show this using Jensen's inequality.

Hoeffding's inequality

- This is an interesting application of Jensen's inequality.

- *Hoeffding's inequality*: Suppose $X_1, \ldots, X_n$ are independent r.v.s such that $r_i \leq X_i \leq s_i$, and denote $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then for all $a > 0$,

$$P(|\overline{X} - \mathrm{E}\overline{X}| > a) \leq 2\exp\left(-\frac{2a^2 n}{\frac{1}{n}\sum_{i=1}^n (s_i - r_i)^2}\right).$$

- Like Chernoff's bound, Hoeffding's provides an exponentially decaying bound as a grows. An advantage of Hoeffding's is that the mgf doesn't need to be known to get an explicit bound. On the other hand, the r.v.s need to be bounded.

- For instance, if $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(q)$, then

$$P(|\overline{X} - q| > a) \leq 2\exp(-2a^2 n).$$

Outline

Introduction

Markov's inequality
- Markov's inequality
- Chebyshev's inequality
- Chernoff's bound

Jensen's inequality
- Jensen's inequality
- Weighted AM-GM inequality
- Hoeffding's inequality

L^p norm inequalities
- L^p spaces
- Hölder's inequality
- Cauchy–Schwarz inequality
- Minkowski's inequality

L^p spaces

- L^p spaces are nice classes of functions that come up a lot.
- For $p \geq 1$, the L^p *norm* of a random variable X is $(\mathrm{E}|X|^p)^{1/p}$.
- Examples:
  - The L^1 norm is simply $\mathrm{E}|X|$.
  - If $\mathrm{E}X = 0$ then the L^2 norm is $(\mathrm{E}|X|^2)^{1/2} = \sqrt{\mathrm{Var}(X)}$.
- The set of r.v.s X such that $(\mathrm{E}|X|^p)^{1/p} < \infty$ is denoted L^p.
- That is, $X \in L^p$ means that $(\mathrm{E}|X|^p)^{1/p} < \infty$.
- Note that $(\mathrm{E}|X|^p)^{1/p} < \infty$ iff $\mathrm{E}|X|^p < \infty$. The purpose of the $1/p$ is that it makes it have the properties of a "norm".

Hölder's inequality

- *Hölder's inequality*: For any random variables X and Y, if $p, q > 1$ such that
$$\frac{1}{p} + \frac{1}{q} = 1$$
then
$$\mathrm{E}|XY| \leq (\mathrm{E}|X|^p)^{1/p}(\mathrm{E}|Y|^q)^{1/q}.$$

Proof: By the weighted AM-GM inequality with $n = 2$, $w_1 = 1/p$, and $w_2 = 1/q$,
$$\frac{1}{p}\frac{|X|^p}{\mathrm{E}|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}|Y|^q} \geq \frac{|XY|}{(\mathrm{E}|X|^p)^{1/p}(\mathrm{E}|Y|^q)^{1/q}}.$$
Taking the expectation of both sides yields
$$1 = \frac{1}{p} + \frac{1}{q} \geq \frac{\mathrm{E}|XY|}{(\mathrm{E}|X|^p)^{1/p}(\mathrm{E}|Y|^q)^{1/q}}.$$

Corollaries of Hölder's inequality

- The Cauchy–Schwarz inequality is an important special case of Hölder's inequality.

- *Cauchy–Schwarz inequality*: For any r.v.s X and Y,
$$\mathrm{E}|XY| \leq (\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}.$$
Proof: Apply Hölder's with $p = q = 2$.

- *Lyapunov's inequality*: If $1 \leq r < s < \infty$, then
$$(\mathrm{E}|X|^r)^{1/r} \leq (\mathrm{E}|X|^s)^{1/s}.$$
Thus, if $X \in L^s$ then $X \in L^r$ for all $r \in [1, s)$.

  Proof: Apply Hölder's to the random variables $|X|^r$ and $Y = 1$ with $p = s/r$ (and $q = 1/(1 - 1/p)$) to get
$$\mathrm{E}|X|^r \leq (\mathrm{E}|X|^{rp})^{1/p} = (\mathrm{E}|X|^s)^{r/s}.$$
Raising both sides to the power of $1/r$ yields the result.

Corollaries of Hölder's inequality

- *Covariance inequality*: If X and Y have means μ_X, μ_Y and variances σ_X^2, σ_Y^2, then
$$|\mathrm{Cov}(X, Y)| \le \sigma_X \sigma_Y.$$
Proof: By Jensen's and the Cauchy–Schwarz inequality,
$$\begin{aligned} |\mathrm{Cov}(X, Y)| &= |\mathrm{E}(X - \mu_X)(Y - \mu_Y)| \\ &\le \mathrm{E}|(X - \mu_X)(Y - \mu_Y)| \\ &\le (\mathrm{E}|X - \mu_X|^2)^{1/2} (\mathrm{E}|Y - \mu_Y|^2)^{1/2} = \sigma_X \sigma_Y. \end{aligned}$$
- This shows that $-1 \le \rho_{X,Y} \le 1$ where $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$.
- *Minkowski's inequality*: For any r.v.s X and Y and any $p \ge 1$,
$$(\mathrm{E}|X + Y|^p)^{1/p} \le (\mathrm{E}|X|^p)^{1/p} + (\mathrm{E}|Y|^p)^{1/p}.$$
Proof: See Casella & Berger, Theorem 4.7.5.
- Minkowski's establishes the triangle inequality for L^p norms.