Probability Inequalities
10/36-705 Intermediate Statistics
Lecture Notes 2

Xuezhe Ma

1 Basic Probability Inequalities

Ineuqlities are useful for bounding quantities that might otherwise be hard to compute. They will also be used in the theory of convergence. In this lecture, we will briefly describe some important and widely used probability inequalities.
We begin with the Markov's Inequality which is widely used to link **cdf** with expectation.

Theorem 1 (**Markov's Inequality**). *Let X be a non-negative random variable and suppose that $\mathrm{E}[X]$ exits. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathrm{E}[X]}{t} \tag{1}$$

Proof. Since $X > 0$, we have

$$\begin{aligned} \mathrm{E}[X] &= \int_0^\infty xp(x)dx = \int_0^t xp(x)dx + \int_t^\infty xp(x)dx \\ &\geq \int_t^\infty xp(x)dx \geq t\int_t^\infty p(x)dx \\ &= t\mathbb{P}(X > t) \end{aligned}$$

□

Markov's Inequality requires non-negative random variables. The following theorem provides a general case for Markov's Inequality by exploiting the commonly used exponential trick, a.k.a *Chernoff's method*:

Theorem 2 (**Chernoff's Method**). *Let X be a random variable. Then,*

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathrm{E}[e^{tX}] \tag{2}$$

Proof. For any $t > 0$,

$$\mathbb{P}(X > \epsilon) = \mathbb{P}(e^{tX} > e^{t\epsilon})$$

Then, from Markov's Inequality, for any $t > 0$

$$\mathbb{P}(X > \epsilon) \leq e^{-t\epsilon}\mathrm{E}[e^{tX}]$$

Thus,

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathrm{E}[e^{tX}]$$

□

Based on Markov's Inequality, we derive the Chebyshev's Inequality whcih links **cdf** with variance,

Theorem 3 (Chebyshev's Inequality). *Let* $\mathrm{E}[X] = \mu$ *and* $\mathrm{Var}[X] = \sigma^2$. *Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{3}$$

Proof. From Markov's Inequality,

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}((X - \mu)^2 \geq t^2) \leq \frac{\mathrm{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

□

Base on the Chernoff's Method, we have the Gausian Tail Inequality and further introduce sub-Gaussian random variables.

Theorem 4 (Gaussian Tail Inequality). *Let* $X \sim N(\mu, \sigma^2)$. *Then*

$$\mathbb{P}(|X - \mu| > \epsilon) \leq 2e^{-\epsilon^2/(2\sigma^2)} \tag{4}$$

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, *then*

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2e^{-n\epsilon^2/(2\sigma^2)} \tag{5}$$

Proof. Since $X \sim N(\mu, \sigma^2)$, we have that the mgf of X is:

$$M_X(t) = \mathrm{E}[e^{tX}] = e^{t\mu + t^2\sigma^2/2}.$$

By applying the Chernoff method, we have

$$\mathbb{P}(X - \mu > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathrm{E}[e^{t(X-\mu)}] = \inf_{t \geq 0} e^{-t\epsilon + t^2\sigma^2/2},$$

which is minimized when $t = \epsilon/\sigma^2$ which in turn yields the tail bound,

$$\mathbb{P}(X - \mu > \epsilon) \leq e^{-\epsilon^2/(2\sigma^2)}$$

By symmetry,

$$\mathbb{P}(|X - \mu| > \epsilon) \leq 2e^{-\epsilon^2/(2\sigma^2)}.$$

As $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, we have $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2e^{-\epsilon^2/(2\sigma^2/n)} = 2e^{-n\epsilon^2/(2\sigma^2)} \tag{6}$$

□

For better comprehension, we also give the Mill's Inequality here, which gives a sharper bound than the Gaussian Tail Inequality,

Theorem 5 (**Mill's Inequality**). *Let* $X \sim N(0, 1)$. *Then*

$$\mathbb{P}(|X| > \epsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2/2}}{\epsilon} \tag{7}$$

If $X_1, \ldots, X_n \sim N(0, 1)$, *then*

$$\mathbb{P}(|\overline{X}_n| > \epsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-n\epsilon^2/2}}{\sqrt{n}\epsilon} \tag{8}$$

Proof. The density of X is $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, so we have $p'(x) = -xp(x)$. Hence,

$$\begin{aligned} \mathbb{P}(X > \epsilon) &= \int_\epsilon^\infty p(x)dx = \int_\epsilon^\infty \frac{xp(x)}{x} dx \\ &\leq -\frac{1}{\epsilon} \int_\epsilon^\infty p'(x)dx = -\frac{1}{\epsilon} p(x)|_\epsilon^\infty = \frac{1}{\epsilon} p(\epsilon) \\ &= \frac{1}{\sqrt{2\pi}\epsilon} e^{-\epsilon^2/2} \end{aligned}$$

By symmetry,

$$\mathbb{P}(|X| > \epsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2/2}}{\epsilon}$$

As $X_1, \ldots, X_n \sim N(0, 1)$, we have $\overline{X}_n \sim N(0, \frac{1}{n})$. Thus, $\sqrt{n}\overline{X}_n \sim N(0, 1)$.

$$\mathbb{P}(|\overline{X}_n| > \epsilon) = \mathbb{P}(|\sqrt{n}\overline{X}_n| > \sqrt{n}\epsilon) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-n\epsilon^2/2}}{\sqrt{n}\epsilon}$$

□

Sub-Gaussian Random Variables From the proof of the Gaussian Tail inequality, we see that a central property of Gaussian random variables which yields the bound of tail is

$$\mathrm{E}[e^{t(X-\mu)}] = e^{t^2\sigma^2/2},$$

for all $t \in \mathcal{R}$. Based on this property, we can define a class of random variables whose tails decay faster than a Gaussian. Formally,

Definition 1 (**Sub-Gaussion Random Variables**). A random variable X with mean μ is *sub-Gaussian* if there exists a positive number σ such that for all $t \in \mathcal{R}$,

$$\mathrm{E}[e^{t(X-\mu)}] \leq e^{t^2\sigma^2/2}. \tag{9}$$

Gaussian random variables with variance σ^2 satisfy the above condition with equality, so a σ-sub-Gaussian random variable basically just has an mgf that is dominated by a Gaussian with variance σ.
It is straightforward to go through the above tail bound to conclude that for a sub-Gaussian random variable, we have the same two-sided exponential tail bound,

$$\mathbb{P}(|X-\mu| > \epsilon) \leq 2e^{-\epsilon^2/(2\sigma^2)}.$$

2 Hoeffding's Inequality

Hoeffding's inequality is similar in spirit to Markov's inequality but it is a sharper inequality, extending Markov's Inequality to multivariate case. Before describing Hoeffding's Inequality, we begin with the following lemma:

Lemma 6. *Suppose that* $\mathrm{E}[X] = 0$ *and* $a \leq X \leq b$*. Then, for any* $t > 0$*,*

$$\mathrm{E}[e^{tX}] \leq e^{t^2(b-a)^2/8} \tag{10}$$

Proof. Since $a \leq X \leq b$, we can write $X = a(1-Z) + bZ$, where $Z = (Z - a)/(b-a)$ and $Z \in [0,1]$. Define function $g(X) = e^{tX}$. Then, we have that g is a convex function. We have

$$e^{tX} = g(X) = g(a(1-Z)+bZ) \leq (1-Z)g(a)+Zg(b) = \frac{b-X}{b-a}e^{ta}+\frac{X-a}{b-a}e^{tb}$$

Taking expectations of both sides and using the fact that $\mathrm{E}[X] = 0$, we get

$$\mathrm{E}[e^{tX}] \leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{h(\omega)}$$

where $\omega = t(b-a)$, $h(\omega) = \gamma\omega + \log(1+\gamma-\gamma e^{\omega})$ and $\gamma = a/(b-a)$. One can verity that $h(0) = h'(0) = 0$, and $h''(x) \leq 1/4, \forall x > 0$. By Taylor's theorem, there exists a $\xi \in (0, \omega)$ such that

$$h(\omega) = h(0) + \omega h'(0) + \frac{\omega^2}{2} h''(\xi) \leq \frac{\omega}{8} = \frac{t^2(b-a)^2}{8}$$

Hence,

$$\mathrm{E}[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

□

Now we describe Hoeffding's Inequality:

Theorem 7 (**Hoeffding's Inequality**). *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathrm{E}[X_i] = \mu$ and $a_i \leq X_i \leq b_i$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-nt\epsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8} \tag{11}$$

Corollary 8. *Let $X_1, \ldots, X_n$ be iid random variables such that $\mathrm{E}[X_i] = \mu$ and $a \leq X_i \leq b$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{12}$$

Proof. Without loss of generality, we assume $\mu = 0$. We have

$$\mathbb{P}(|\overline{X}_n| \geq \epsilon) = \mathbb{P}(\overline{X}_n \geq \epsilon) + \mathbb{P}(-\overline{X}_n \geq \epsilon)$$

Using Chernoff's method, for any $t > 0$,

$$\mathbb{P}(\overline{X}_n \geq \epsilon) = \mathbb{P}(\sum_{i=1}^{n} X_i \geq n\epsilon) \leq e^{-tn\epsilon} \mathrm{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] = e^{-tn\epsilon} \prod_{i=1}^{n} \mathrm{E}[e^{tX_i}]$$

From Lemma 6, $\mathrm{E}[e^{tX_i}] \leq e^{t^2(b_i-a_i)^2/8}$. So

$$\mathbb{P}(\overline{X}_n \geq \epsilon) \leq e^{-tn\epsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8}$$

By symmetry,

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-nt\epsilon} \prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8}$$

When $X_1, \ldots, X_n$ are iid random variables, for nay $t > 0$

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-nt\epsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8} = 2e^{-nt\epsilon} e^{t^2 n(b-a)^2/8}$$

This is minimized by setting $t = 4\epsilon/(b-a)^2$, giving

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

□

3 Bernstein's Inequality

Hieffding's bound depends only on the bounds of the random variable but not explicily on the variance of the random variables. The bound $b - a$ provides a (possibly loose) upper bound on the variance. One might at least hope that if the random variables are bounded, and additionally have *small variance*, we might be able to improve Hoeffding's bound.
Such inequality is typically known as Bernstein's inequality.

Theorem 9 (Bernstein's Inequality). *Let $X_1, \ldots, X_n$ be iid random variables such that $\mathrm{E}[X_i] = \mu$, $\mathrm{Var}[X_i] = \sigma^2$ and $a \leq X_i \leq b$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2 + (b-a)\epsilon/3}} \tag{13}$$

It is not hard to verify that with small ϵ and σ^2, Bernstein's inequality provides a sharper bound than Hoeffding's inequality.

4 McDiarmid's Inequality

So far we have focused on sums of random variables. McDiarmid's inequality, a.k.a. the Bounded Difference inequality, extends Hoeffding's inequality to more general functions $g(x_1, \ldots, x_n)$.

Theorem 10 (McDiarmid's Inequality). *Let $X_1, \ldots, X_n$ be independent random variables. Suppose that for $i = 1, \ldots, n$,*

$$\sup_{x_1, \ldots, x_n, x_i'} \left| g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i$$

Then,

$$\mathbb{P}\left(g(X_1, \ldots, X_n) - \mathrm{E}[g(X_1, \ldots, X_n)] \geq \epsilon\right) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n} c_i^2} \tag{14}$$

Proof. Denote $Z_i = \mathrm{E}[g(X_1, \ldots, X_n)|X_1, \ldots, X_i] - \mathrm{E}[g(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}]$. Then,

$$g(X_1, \ldots, X_n) - \mathrm{E}[g(X_1, \ldots, X_n)] = \sum_{i=1}^{n} Z_i$$

and $\mathrm{E}[Z_i|X_1, \ldots, X_{i-1}] = 0$. Using Chernoff's method, for any $t > 0$,

$$\mathbb{P}(g(X_1, \ldots, X_n) - \mathrm{E}[g(X_1, \ldots, X_n)] \geq \epsilon) = \mathbb{P}\left(\sum_{i=1}^{n} Z_i\right) \leq e^{-t\epsilon}\mathrm{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right]$$

Moreover, we have

$$\sup\{Z_i|X_1, \ldots, X_{i-1}\} - \inf\{Z_i|X_1, \ldots, X_{i-1}\} \leq c_i$$

From Lemma 6, for any $t > 0$

$$\mathrm{E}\left[e^{tZ_i}|X_1, \ldots, X_{i-1}\right] \leq e^{t^2 c_i^2/8}$$

Then,

$$\begin{array}{rcl} \mathrm{E}\left[e^{t\sum_{i=1}^{n} Z_i}\right] & = & \mathrm{E}\left[e^{t\sum_{i=1}^{n-1} Z_i}\mathrm{E}\left[e^{tZ_n|X_1,\ldots,X_{n-1}}\right]\right] \\ & \leq & e^{t^2 c_n^2/8}\mathrm{E}\left[e^{t\sum_{i=1}^{n-1} Z_i}\right] \\ & \vdots & \\ e^{t^2\sum_{i=1}^{n} c_i^2} & & \end{array}$$

By taking $t = \frac{4\epsilon}{\sum_{i=1}^{n} c_i^2}$,

$$\mathbb{P}\left(g(X_1, \ldots, X_n) - \mathrm{E}[g(X_1, \ldots, X_n)] \geq \epsilon\right) \leq e^{-2\epsilon^2/\sum_{i=1}^{n} c_i^2}$$

□

5 Bounds on Expected Values and Variances

In this section, we visit the inequalities that provide bounds for the expected values and variances.

Theorem 11 (Cauchy-Schwartz Inequality). *If* $\mathrm{Var}[X] < \infty$ *and* $\mathrm{Var}[Y] < \infty$, *then*

$$\mathrm{E}[|XY|] \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[Y^2]} \tag{15}$$

Theorem 12 (**Jensen's Inequality**). *If g is convex, then*

$$\mathrm{E}[g(x)] \geq g(\mathrm{E}[X]) \tag{16}$$

If g is concave, then

$$\mathrm{E}[g(x)] \leq g(\mathrm{E}[X]) \tag{17}$$

Now we consider bounding the maximum of a set of random variables.

Theorem 13. *Let $X_1, \ldots, X_n$ be random variables. Suppose there exists $\sigma > 0$ such that for any $t > 0$,*

$$\mathrm{E}[e^{tX_i}] \leq e^{t^2\sigma^2/2}$$

Then,

$$\mathrm{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \sigma\sqrt{2\log n} \tag{18}$$

Proof. By Jensen's Inequality, for any $t > 0$,

$$\begin{aligned} e^{\mathrm{E}\left[t\max_{1\leq i\leq n} X_i\right]} &\leq \mathrm{E}\left[e^{t\max_{1\leq i\leq n} X_i}\right] \\ &= \mathrm{E}\left[\max_{1\leq i\leq n} e^{tX_i}\right] \\ &\leq \sum_{i=1}^{n} \mathrm{E}\left[e^{tX_i}\right] \\ &\leq ne^{t^2\sigma^2/2} \end{aligned}$$

Thus,

$$\mathrm{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}$$

The result follows by taking $t = \sqrt{2\log n}/\sigma$. □

In order to bound variances, we describe the Bhatia–Davis inequality.

Theorem 14 (**Bhatia–Davis Inequality**). *Suppose that $a \leq X \leq b$ and $\mathrm{E}[X] = \mu$. Then,*

$$\mathrm{Var}[X] \leq (b-\mu)(\mu-a) \leq \frac{1}{4}(b-a)^2 \tag{19}$$

Proof. for every $x \in [a, b]$, we have $0 \leq (x-a)(b-x)$, which gives us

$$x^2 \leq (b+a)x - ab.$$

Then we have

$$\begin{aligned} \mathrm{Var}[X] &= \textstyle\int_a^b x^2 p(x)dx - \mu^2 \\ &\leq \textstyle\int_a^b ((b+a)x - ab)p(x)dx - \mu^2 \\ &= (a+b)\mu - ab - \mu^2 = (b-\mu)(\mu-a) \\ &\leq \tfrac{1}{4}(b-a)^2 \end{aligned}$$

□