CHAPTER 4. Inequalities

4.1. Markov and Chebychev Inequalities

Inequalities are useful in probability for bounding quantities that might otherwise be hard to compute. They are also useful for developing the theory of convergence which is discussed in the next chapter. Our first inequality is Markov's inequality.

THEOREM. 4.1.1. (Markov's Inequality). Let X be a non-negative random variable (i.e. $P(X \geq 0) = 1$). Suppose that $E(X)$ exists. For any $t > 0$,

$$P(X > t) \leq \frac{E(X)}{t}.$$

PROOF. We have that $E(X) = \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \geq \int_t^\infty xf(x)dx \geq t\int_t^\infty f(x)dx = tP(X > t)$.

THEOREM. 4.1.2. (Chebyshev's inequality.) Let $\mu = E(X)$ and $\sigma^2 = Var(X)$. Then,

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

and

$$P(|Z| \geq k) \leq \frac{1}{k^2}$$

where $Z = (X - \mu)/\sigma$. In particular, $P(|Z| > 2) \leq 1/4$ and $P(|Z| > 3) \leq 1/9$.

PROOF. We use Markov's inequality to conclude that

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

The second part follows by setting $t = k\sigma$.

THEOREM. 4.1.3. *Let $X_1, \ldots, X_n$ be n independent random variables with common finite mean μ and common finite variance σ^2. Let*

$$Z_n = \frac{(\overline{X}_n - \mu)}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

where

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for $t > 0$,

$$P\left(|Z_n| > t\right) \leq \frac{1}{t^2}.$$

PROOF. This follows from the fact that $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ and Chebyshev's inequality.

EXAMPLE 4.1.4. Suppose we test a prediction method (a neural net for example) on a set $n = 10,000$ new test cases. Let $X_i = 1$ if the predictor is wrong and $X_i = 0$ if the predictor is right. Then $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ is the observed error rate. Each X_i may be regarded as a Bernoulli with unknown mean p. We would like to know the true, but unknown error rate p. Intuitively, we expect that $\overline{X}_n$ should be close to p. Now, $\mu = EX_1 = p$ and $\sigma = \sqrt{\mathrm{Var(X_1)}} = \sqrt{p(1-p)}$. Let us bound $P(|\overline{X} - p| > .01)$. Using Theorem 1.3:

$$\begin{aligned} P(|\overline{X} - p| > .01) &= P\left(\frac{\sqrt{n}|\overline{X}_n - \mu|}{\sigma} > \frac{.01\sqrt{n}}{\sigma}\right) \\ &= P\left(|Z_n| > \frac{1}{\sqrt{p(1-p)}}\right) \\ &\leq p(1-p) \\ &\leq \frac{1}{4} \end{aligned}$$

since $p(1-p) \leq \frac{1}{4}$ for all p.

4.2. Cauchy-Schwarz and Jensen

The next result is the Cauchy-Schwarz inequality. We have already seen this but we repeat it here for completeness.

THEOREM 4.2.1. (Cauchy-Schwarz inequality.) *If X and Y have finite variances then*

$$E\,|XY| \leq \left\{E(X^2)E(Y^2)\right\}^{1/2}.$$

Recall that a function g is *convex* if for each x, y and each $\alpha \in [0, 1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

If g is twice differentiable, then convexity reduces to checking that $g''(x) \geq 0$ for all x. It can be shown that if g is convex then it lies above any line that touches g at some point. A function g is *concave* if $-g$ is convex. Examples of convex functions are $g(x) = x^2$ and $g(x) = e^x$. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log x$.

THEOREM 4.2.2. (Jensen's Inequality.) *If g is convex then*

$$Eg(X) \geq g(EX).$$

If g is concave then

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

PROOF. Let $L(x) = a + bx$ be a line, tangent to $g(x)$ at the point $E(X)$. Since g is convex, it lies above the line $L(x)$. So,

$$\begin{aligned} Eg(X) &\geq EL(X) \\ &= E(a + bX) \\ &= a + bE(X) \\ &= L(E(X)) \\ &= g(EX). \end{aligned}$$

From Jensen's inequality we see that $EX^2 \geq (EX)^2$ and $E(1/X) \geq 1/E(X)$. Since log is concave, $E(\log X) \leq \log E(X)$. For example, suppose that $X \sim N(3, 1)$. Then $E(1/X) \geq 1/3$.

4.3. Hoeffding's Inequality

Recall Markov's inequality. If $X > 0$ and $t > 0$ then $P(X > t) < E(X)/t$. Hoeffding's Inequality is in the same spirit but it is a sharper inequality. We present the result here in two parts. The proofs are in the appendix of this Chapter.

THEOREM 4.3.1. *Let $Y_1, \ldots, Y_n$ be independent observations such that $E(Y_i) = 0$ and $a_i \leq Y_i \leq b_i$. Let $\epsilon > 0$. Then, for any $t > 0$,*

$$P\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}.$$

THEOREM 4.3.2. *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then, for any $\epsilon > 0$,*

$$P\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$.

EXAMPLE 4.3.3. (EXAMPLE 4.1.4 revisited.) Recall that $n = 10,000$ and $X_i \sim \text{Ber}(p)$. Using Chebyshev's inequality we found that

$$P(|\overline{X} - p| > .01) \leq \frac{1}{4}.$$

According to Hoeffding's inequality,

$$P(|\overline{X} - p| > .01) \leq 2e^{-2(.01)^2 n} = .27$$

which is roughly the same. In this case, there was not much difference. Often, Hoeffding's inequality gives tighter bounds. You will prove this in the homework.

As an aside, let us note that Hoeffding's inequality gives us a simple way to create a *confidence interval* for a binomial parameter p. We will discuss confidence intervals later but let is give the basic idea here. Fix $\alpha > 0$ and let

$$\epsilon_n = \left\{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)\right\}^{1/2}.$$

Hoeffding's inequality says that

$$P\left(|\overline{X}_n - p| > \epsilon_n\right) \leq 2e^{-2n\epsilon_n^2} = \alpha.$$

Let $C = [\overline{X}_n - \epsilon, \overline{X}_n + \epsilon]$. Then, $P(C \notin p) = P(|\overline{X}_n - p| > \epsilon) \leq \alpha$. Hence, $P(p \in C) \geq 1 - \alpha$ that is, the random interval C traps the true parameter value p with probability $1 - \alpha$; we call C a $1 - \alpha$ confidence interval. More on this later.

Appendix: Proof of Hoeffding's Inequality

We will make use of the exact form of Taylor's theorem: if g is a smooth function, then there is a number $\xi \in (0, u)$ such that $g(u) = g(0) + ug'(0) + \frac{u^2}{2} g''(\xi)$.

PROOF of Theorem 4.3.1. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned} P\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) &= P\left(t \sum_{i=1}^{n} Y_i \geq t\epsilon\right) \\ &= P\left(e^{t \sum_{i=1}^{n} Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon} E\left(e^{t \sum_{i=1}^{n} Y_i}\right) \\ &= e^{-t\epsilon} \prod_i E(e^{tY_i}). \end{aligned} \tag{1}$$

Since $a_i \leq Y_i \leq b_i$, we can write Y_i as a convex combination of a_i and b_i, namely, $Y_i = \alpha b_i + (1-\alpha) a_i$ where $\alpha = (Y_i - a_i)/(b_i - a_i)$. So, by the convexity of e^{ty} we have

$$e^{tY_i} \leq \frac{Y_i - a_i}{b_i - a_i} e^{tb_i} + \frac{b_i - Y_i}{b_i - a_i} e^{ta_i}.$$

Take expectations of both sides and use the fact that $E(Y_i) = 0$ to get

$$Ee^{tY_i} \leq -\frac{a_i}{b_i - a_i} e^{tb_i} + \frac{b_i}{b_i - a_i} e^{ta_i} = e^{g(u)} \tag{2}$$

where $u = t(b_i - a_i)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$ and $\gamma = -a_i/(b_i - a_i)$.

Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is a $\xi \in (0, u)$ such that

$$\begin{aligned} g(u) &= g(0) + ug'(0) + \frac{u^2}{2} g''(\xi) \\ &= \frac{u^2}{2} g''(\xi) \\ &\leq \frac{u^2}{8} = \frac{t^2 (b_i - a_i)^2}{8}. \end{aligned}$$

Hence,

$$Ee^{tY_i} \leq e^{g(u)} \leq e^{t^2 (b_i - a_i)^2/8}.$$

The result follows from (1).

PROOF of Theorem 4.3.2. Let $Y_i = (1/n)(X_i - p)$. Then $E(Y_i) = 0$ and $a \leq Y_i \leq b$ where $a = -p/n$ and $b = (1-p)/n$. Also, $(b-a)^2 = 1/n^2$. Applying the last Theorem we get

$$\begin{aligned} P(\overline{X}_n - p > \epsilon) &= P(\sum_i Y_i > \epsilon) \\ &\leq e^{-t\epsilon} e^{t^2/(8n)}. \end{aligned}$$

The above holds for any $t > 0$. In particular, take $t = 4n\epsilon$ and we get

$$P(\overline{X}_n - p > \epsilon) \leq e^{-2n\epsilon^2}.$$

By a similar argument we can show that

$$P(\overline{X}_n - p < -\epsilon) \leq e^{-2n\epsilon^2}.$$

Putting these together we get

$$P\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$