4
Inequalities

4.1 Probability Inequalities

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

4.1 Theorem (Markov's inequality)**.** *Let X be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}. \tag{4.1}$$

PROOF. Since $X > 0$,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \\ &\geq \int_t^\infty x f(x)dx \geq t \int_t^\infty f(x)dx = t\,\mathbb{P}(X > t) \end{aligned}$$ ∎

4.2 Theorem (Chebyshev's inequality). *Let* $\mu = \mathbb{E}(X)$ *and* $\sigma^2 = \mathbb{V}(X)$. *Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{4.2}$$

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

PROOF. We use Markov's inequality to conclude that

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

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

4.3 Example. Suppose we test a prediction method, a neural net for example, on a set of n 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. How likely is $\overline{X}_n$ to not be within ϵ of p? We have that $\mathbb{V}(\overline{X}_n) = \mathbb{V}(X_1)/n = p(1-p)/n$ and

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

since $p(1-p) \leq \frac{1}{4}$ for all p. For $\epsilon = .2$ and $n = 100$ the bound is .0625. ∎

Hoeffding's inequality is similar in spirit to Markov's inequality but it is a sharper inequality. We present the result here in two parts.

4.4 Theorem (Hoeffding's Inequality). *Let* $Y_1, \ldots, Y_n$ *be independent observations such that* $\mathbb{E}(Y_i) = 0$ *and* $a_i \leq Y_i \leq b_i$. *Let* $\epsilon > 0$. *Then, for any* $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^n e^{t^2(b_i - a_i)^2/8}. \tag{4.3}$$

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

$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{4.4}$$

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

4.6 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Let $n = 100$ and $\epsilon = .2$. We saw that Chebyshev's inequality yielded

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq .0625.$$

According to Hoeffding's inequality,

$$\mathbb{P}(|\overline{X}_n - p| > .2) \leq 2e^{-2(100)(.2)^2} = .00067$$

which is much smaller than .0625. ■

Hoeffding's inequality gives us a simple way to create a **confidence interval** for a binomial parameter p. We will discuss confidence intervals in detail later (see Chapter 6) but here is the basic idea. Fix $\alpha > 0$ and let

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

By Hoeffding's inequality,

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

Let $C = (\overline{X}_n - \epsilon_n, \overline{X}_n + \epsilon_n)$. Then, $\mathbb{P}(p \notin C) = \mathbb{P}(|\overline{X}_n - p| > \epsilon_n) \leq \alpha$. Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$, 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.

The following inequality is useful for bounding probability statements about Normal random variables.

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

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}.$$

4.2 Inequalities For Expectations

This section contains two inequalities on expected values.

4.8 Theorem (Cauchy-Schwartz inequality). *If X and Y have finite variances then*

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \tag{4.5}$$

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 and $g''(x) \geq 0$ for all x, then g is convex. It can be shown that if g is convex, then g lies above any line that touches g at some point, called a tangent line. 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 $g(x) = \log x$.

4.9 Theorem (Jensen's inequality). *If g is convex, then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X). \tag{4.6}$$

If g is concave, then

$$\mathbb{E}g(X) \leq g(\mathbb{E}X). \tag{4.7}$$

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

$$\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}X). \quad \blacksquare$$

From Jensen's inequality we see that $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$ and if X is positive, then $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$. Since log is concave, $\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$.

4.3 Bibliographic Remarks

Devroye et al. (1996) is a good reference on probability inequalities and their use in statistics and pattern recognition. The following proof of Hoeffding's inequality is from that text.

4.4 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 4.4. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= \mathbb{P}\left(t\sum_{i=1}^n Y_i \geq t\epsilon\right) = \mathbb{P}\left(e^{t\sum_{i=1}^n Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^n Y_i}\right) = e^{-t\epsilon}\prod_i \mathbb{E}(e^{tY_i}). \qquad (4.8)\end{aligned}$$

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 $\mathbb{E}(Y_i) = 0$ to get

$$\mathbb{E}e^{tY_i} \leq -\frac{a_i}{b_i - a_i}e^{tb_i} + \frac{b_i}{b_i - a_i}e^{ta_i} = e^{g(u)} \qquad (4.9)$$

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,

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

The result follows from (4.8). ∎

Proof of Theorem 4.5. Let $Y_i = (1/n)(X_i - p)$. Then $\mathbb{E}(Y_i) = 0$ and $a \leq Y_i \leq b$ where $a = -p/n$ and $b = (1-p)/n$. Also, $(b-a)^2 = 1/n^2$. Applying Theorem 4.4 we get

$$\mathbb{P}(\overline{X}_n - p > \epsilon) = \mathbb{P}\left(\sum_i Y_i > \epsilon\right) \leq e^{-t\epsilon}e^{t^2/(8n)}.$$

The above holds for any $t > 0$. In particular, take $t = 4n\epsilon$ and we get $\mathbb{P}(\overline{X}_n - p > \epsilon) \leq e^{-2n\epsilon^2}$. By a similar argument we can show that $\mathbb{P}(\overline{X}_n - p < -\epsilon) \leq e^{-2n\epsilon^2}$. Putting these together we get $\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$. ∎

4.5 Exercises

1. Let $X \sim \text{Exponential}(\beta)$. Find $\mathbb{P}(|X - \mu_X| \geq k\sigma_X)$ for $k > 1$. Compare this to the bound you get from Chebyshev's inequality.

2. Let $X \sim \text{Poisson}(\lambda)$. Use Chebyshev's inequality to show that $\mathbb{P}(X \geq 2\lambda) \leq 1/\lambda$.

3. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Bound $\mathbb{P}(|\overline{X}_n - p| > \epsilon)$ using Chebyshev's inequality and using Hoeffding's inequality. Show that, when n is large, the bound from Hoeffding's inequality is smaller than the bound from Chebyshev's inequality.

4. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

   (a) Let $\alpha > 0$ be fixed and define

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

   Let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Define $C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$. Use Hoeffding's inequality to show that

   $$\mathbb{P}(C_n \text{ contains } p) \geq 1 - \alpha.$$

   In practice, we truncate the interval so it does not go below 0 or above 1.

   (b) (Computer Experiment.) Let's examine the properties of this confidence interval. Let $\alpha = 0.05$ and $p = 0.4$. Conduct a simulation study to see how often the interval contains p (called the coverage). Do this for various values of n between 1 and 10000. Plot the coverage versus n.

   (c) Plot the length of the interval versus n. Suppose we want the length of the interval to be no more than .05. How large should n be?

5. Prove Mill's inequality, Theorem 4.7. Hint. Note that $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$. Now write out what $\mathbb{P}(Z > t)$ means and note that $x/t > 1$ whenever $x > t$.

6. Let $Z \sim N(0,1)$. Find $\mathbb{P}(|Z| > t)$ and plot this as a function of t. From Markov's inequality, we have the bound $\mathbb{P}(|Z| > t) \leq \frac{\mathbb{E}|Z|^k}{t^k}$ for any $k > 0$. Plot these bounds for $k = 1, 2, 3, 4, 5$ and compare them to the true value of $\mathbb{P}(|Z| > t)$. Also, plot the bound from Mill's inequality.

7. Let $X_1, \ldots, X_n \sim N(0,1)$. Bound $\mathbb{P}(|\overline{X}_n| > t)$ using Mill's inequality, where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Compare to the Chebyshev bound.