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.

Theorem 1 (The Gaussian Tail Inequality) Let $X \sim N(0,1)$. Then

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

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

$$P(|X_n| > \epsilon) \leq \frac{2}{\sqrt{n\epsilon}}e^{-n\epsilon^2/2}.$$

Proof. The density of $X$ is $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$. Hence,

$$P(X > \epsilon) = \int_{\epsilon}^{\infty} \phi(s)ds \leq \frac{1}{\epsilon} \int_{\epsilon}^{\infty} s \phi(s)ds = -\frac{1}{\epsilon} \int_{\epsilon}^{\infty} \phi'(s)ds = \frac{\phi(\epsilon)}{\epsilon} \leq \frac{e^{-\epsilon^2/2}}{\epsilon}.$$

By symmetry, $P(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}$.

Now let $X_1, \ldots, X_n \sim N(0,1)$. Then $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i \sim N(0,1/n)$. Thus, $\overline{X}_n \overset{d}{=} n^{-1/2}Z$ where $Z \sim N(0,1)$ and

$$P(|\overline{X}_n| > \epsilon) = P(n^{-1/2}|Z| > \epsilon) = P(|Z| > \sqrt{n}\epsilon) \leq \frac{2}{\sqrt{n\epsilon}}e^{-n\epsilon^2/2}.$$
**Theorem 2 (Markov’s inequality)** Let $X$ be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,

$$\Pr(X > t) \leq \frac{\mathbb{E}(X)}{t}. \quad (1)$$

**Proof.** Since $X > 0$,

$$\mathbb{E}(X) = \int_0^\infty x p(x) dx = \int_0^t x p(x) dx + \int_t^\infty x p(x) dx \geq \int_t^\infty x p(x) dx \geq t \int_t^\infty p(x) dx = t \Pr(X > t).$$

$\square$

**Theorem 3 (Chebyshev’s inequality)** Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. Then,

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

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

**Proof.** We use Markov’s inequality to conclude that

$$\Pr(|X - \mu| \geq t) = \Pr(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}. \quad \text{(The second part follows by setting } t = k\sigma. \square)$$

If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ then and $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then, $\text{Var}(\overline{X}_n) = \text{Var}(X_1)/n = p(1-p)/n$ and

$$\Pr(|\overline{X}_n - p| > \epsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2} \quad \text{since } p(1-p) \leq \frac{1}{4} \text{ for all } p.$$

# 2 Hoeffding’s Inequality

Hoeffding’s inequality is similar in spirit to Markov’s inequality but it is a sharper inequality. We begin with the following important result.

**Lemma 4** Suppose that $\mathbb{E}(X) = 0$ and that $a \leq X \leq b$. Then

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}.$$
Recall that a function $g$ is **convex** if for each $x, y$ and each $\alpha \in [0, 1]$, 

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

**Proof.** Since $a \leq X \leq b$, we can write $X$ as a convex combination of $a$ and $b$, namely, $X = \alpha b + (1 - \alpha)a$ where $\alpha = (X - a)/(b - a)$. By the convexity of the function $y \to e^{ty}$ we have

$$e^{tx} \leq \alpha e^{tb} + (1 - \alpha)e^{ta} = \frac{X - a}{b - a} e^{tb} + \frac{b - X}{b - a} e^{ta}.$$

Take expectations of both sides and use the fact that $\mathbb{E}(X) = 0$ to get

$$\mathbb{E}e^{tx} \leq -\frac{a}{b - a} e^{tb} + \frac{b}{b - a} e^{ta} = e^{g(u)} \quad (3)$$

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

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2} g''(\xi) = \frac{u^2}{2} g''(\xi) \leq \frac{u^2}{8} = \frac{t^2(b - a)^2}{8}.$$

Hence, $\mathbb{E}e^{tx} \leq e^{g(u)} \leq e^{t^2(b-a)/8}$. \(\square\)

Next, we need to use Chernoff’s method.

**Lemma 5** Let $X$ be a random variable. Then

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

**Proof.** For any $t > 0$,

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

Since this is true for every $t \geq 0$, the result follows. \(\square\)

**Theorem 6 (Hoeffding’s Inequality)** Let $Y_1, \ldots, Y_n$ be iid observations such that $\mathbb{E}(Y_i) = \mu$ and $a \leq Y_i \leq b$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{Y}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}. \quad (4)$$

**Corollary 7** If $X_1, X_2, \ldots, X_n$ are independent with $\mathbb{P}(a \leq X_i \leq b) = 1$ and common mean $\mu$, then, with probability at least $1 - \delta$,

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n \log \left( \frac{2}{\delta} \right)}}. \quad (5)$$
Proof. Without loss of generality, we assume that $\mu = 0$. First we have

$$\mathbb{P}(|Y_n| \geq \epsilon) = \mathbb{P}(Y_n \geq \epsilon) + \mathbb{P}(Y_n \leq -\epsilon) = \mathbb{P}(Y_n \geq \epsilon) + \mathbb{P}(-Y_n \geq \epsilon).$$

Next we use Chernoff’s method. For any $t > 0$, we have, from Markov’s inequality, that

$$\mathbb{P}(Y_n \geq \epsilon) = \mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq n\epsilon\right) = \mathbb{P}\left(e^{t \sum_{i=1}^{n} Y_i} \geq e^{tn\epsilon}\right) \leq e^{-t\epsilon n} \mathbb{E}\left(e^{t \sum_{i=1}^{n} Y_i}\right) = e^{-t\epsilon n} \prod_{i} \mathbb{E}(e^{t Y_i}) = e^{-t\epsilon n} (\mathbb{E}(e^{t Y_i}))^n.

From Lemma 4, $\mathbb{E}(e^{t Y_i}) \leq e^{t^2(b-a)^2/8}$. So

$$\mathbb{P}(Y_n \geq \epsilon) \leq e^{-t\epsilon n} e^{t^2 n(b-a)^2/8}.$$

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

$$\mathbb{P}(Y_n \geq \epsilon) \leq e^{-2n\epsilon^2/(b-a)^2}.$$

Applying the same argument to $\mathbb{P}(-Y_n \geq \epsilon)$ yields the result. □

Example 8 Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. From, Hoeffding’s inequality,

$$\mathbb{P}(|X_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

3 The Bounded Difference Inequality

So far we have focused on sums of random variables. The following result extends Hoeffding’s inequality to more general functions $g(x_1, \ldots, x_n)$. Here we consider McDiarmid’s inequality, also known as the Bounded Difference inequality.
Theorem 9 (McDiarmid) Let $X_1, \ldots, X_n$ be independent random variables. Suppose that
\[
\sup_{x_1, \ldots, x_n, x'_i} \left| g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n) \right| \leq c_i \tag{6}
\]
for $i = 1, \ldots, n$. Then
\[
P \left( g(X_1, \ldots, X_n) - \mathbb{E}(g(X_1, \ldots, X_n)) \geq \epsilon \right) \leq \exp \left\{ -\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right\}. \tag{7}
\]

Proof. Let $V_i = \mathbb{E}(g|X_1, \ldots, X_i) - \mathbb{E}(g|X_1, \ldots, X_{i-1})$. Then $g(X_1, \ldots, X_n) - \mathbb{E}(g(X_1, \ldots, X_n)) = \sum_{i=1}^n V_i$ and $\mathbb{E}(V_i|X_1, \ldots, X_{i-1}) = 0$. Using a similar argument as in Hoeffding’s Lemma we have,
\[
\mathbb{E}(e^{tV_i}|X_1, \ldots, X_{i-1}) \leq e^{t^2 c_i^2/8}. \tag{8}
\]
Now, for any $t > 0$,
\[
P \left( g(X_1, \ldots, X_n) - \mathbb{E}(g(X_1, \ldots, X_n)) \geq \epsilon \right) = P \left( \sum_{i=1}^n V_i \geq \epsilon \right)
= P \left( e^{t \sum_{i=1}^n V_i} \geq e^{t \epsilon} \right) \leq e^{-t \epsilon} \mathbb{E} \left( e^{t \sum_{i=1}^n V_i} \right)
= e^{-t \epsilon} \mathbb{E} \left( e^{t \sum_{i=1}^{n-1} V_i} \mathbb{E} \left( e^{V_n} \mid X_1, \ldots, X_{n-1} \right) \right)
\leq e^{-t \epsilon} e^{t^2 c_1^2/8} \mathbb{E} \left( e^{t \sum_{i=1}^{n-1} V_i} \right)
\vdots
\leq e^{-t \epsilon} e^{t^2 \sum_{i=1}^{n-1} c_i^2}.
\]
The result follows by taking $t = 4\epsilon / \sum_{i=1}^n c_i^2$. □

Example 10 If we take $g(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^n x_i$ then we get back Hoeffding’s inequality.

Example 11 Suppose we throw $m$ balls into $n$ bins. What fraction of bins are empty? Let $Z$ be the number of empty bins and let $F = Z/n$ be the fraction of empty bins. We can write $Z = \sum_{i=1}^n Z_i$ where $Z_i = 1$ if bin $i$ is empty and $Z_i = 0$ otherwise. Then
\[
\mu = \mathbb{E}(Z) = \sum_{i=1}^n \mathbb{E}(Z_i) = n(1 - 1/n)^m = ne^{m \log(1-1/n)} \approx ne^{-m/n}
\]
and $\theta = \mathbb{E}(F) = \mu/n \approx e^{-m/n}$. How close is $Z$ to $\mu$? Note that the $Z_i$’s are not independent so we cannot just apply Hoeffding. Instead, we proceed as follows.
Define variables $X_1, \ldots, X_m$ where $X_i = 1$ if ball $i$ falls into bin $i$. Then $Z = g(X_1, \ldots, X_m)$. If we move one ball into a different bin, then $Z$ can change by at most 1. Hence, (6) holds with $c_i = 1$ and so

$$\mathbb{P}(|Z - \mu| > t) \leq 2e^{-2t^2/m}.$$  

Recall that the fraction of empty bins is $F = Z/m$ with mean $\theta = \mu/n$. We have

$$\mathbb{P}(|F - \theta| > t) = \mathbb{P}(|Z - \mu| > nt) \leq 2e^{-2n^2t^2/m}.$$  

4 Bounds on Expected Values

**Theorem 12 (Cauchy-Schwarz inequality)** If $X$ and $Y$ have finite variances then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \quad (9)$$

The Cauchy-Schwarz inequality can be written as

$$\text{Cov}^2(X,Y) \leq \sigma_X^2\sigma_Y^2.$$  

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 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 $g(x) = \log x$.

**Theorem 13 (Jensen’s inequality)** If $g$ is convex, then

$$\mathbb{E}g(X) \geq g(\mathbb{E}X). \quad (10)$$

If $g$ is concave, then

$$\mathbb{E}g(X) \leq g(\mathbb{E}X). \quad (11)$$

**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,

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

$\Box$
Example 14 From Jensen’s inequality we see that $\mathbb{E}(X^2) \geq (\mathbb{E} X)^2$.

Example 15 (Kullback Leibler Distance) Define the Kullback-Leibler distance between two densities $p$ and $q$ by

$$D(p,q) = \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx.$$ 

Note that $D(p,p) = 0$. We will use Jensen to show that $D(p,q) \geq 0$. Let $X \sim p$. Then

$$-D(p,q) = \mathbb{E} \log \left( \frac{q(X)}{p(X)} \right) \leq \log \mathbb{E} \left( \frac{q(X)}{p(X)} \right) = \log \int p(x) \frac{q(x)}{p(x)} dx = \log \int q(x) dx = \log(1) = 0.$$ 

So, $-D(p,q) \leq 0$ and hence $D(p,q) \geq 0$.

Example 16 It follows from Jensen’s inequality that 3 types of means can be ordered. Assume that $a_1, \ldots, a_n$ are positive numbers and define the arithmetic, geometric and harmonic means as

$$a_A = \frac{1}{n} (a_1 + \ldots + a_n),$$

$$a_G = (a_1 \times \ldots \times a_n)^{1/n},$$

$$a_H = \frac{1}{\frac{1}{a_1} + \ldots + \frac{1}{a_n}}.$$ 

Then $a_H \leq a_G \leq a_A$.

Suppose we have an exponential bound on $\mathbb{P}(X_n > \epsilon)$. In that case we can bound $\mathbb{E}(X_n)$ as follows.

Theorem 17 Suppose that $X_n \geq 0$ and that for every $\epsilon > 0$,

$$\mathbb{P}(X_n > \epsilon) \leq c_1 e^{-c_2 n \epsilon^2} \quad (12)$$

for some $c_2 > 0$ and $c_1 > 1/e$. Then,

$$\mathbb{E}(X_n) \leq \sqrt{\frac{C}{n}}. \quad (13)$$

where $C = (1 + \log(c_1))/c_2$.

Proof. Recall that for any nonnegative random variable $Y$, $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y \geq t) dt$. Hence, for any $a > 0$,

$$\mathbb{E}(X_n^2) = \int_0^\infty \mathbb{P}(X_n^2 \geq t) dt = \int_0^a \mathbb{P}(X_n^2 \geq t) dt + \int_a^\infty \mathbb{P}(X_n^2 \geq t) dt \leq a + \int_a^\infty \mathbb{P}(X_n^2 \geq t) dt.$$
Equation (12) implies that $P(X_n > \sqrt{t}) \leq c_1 e^{-c_2 nt}$. Hence,

$$
\mathbb{E}(X_n^2) \leq a + \int_a^\infty P(X_n^2 \geq t)dt = a + \int_a^\infty P(X_n \geq \sqrt{t})dt \leq a + c_1 \int_a^\infty e^{-c_2 nt}dt = a + \frac{c_1 e^{-c_2 nt}}{c_2 n}.
$$

Set $a = \log(c_1)/(nc_2)$ and conclude that

$$
\mathbb{E}(X_n^2) \leq \frac{\log(c_1)}{nc_2} + \frac{1}{nc_2} = 1 + \frac{\log(c_1)}{nc_2}.
$$

Finally, we have

$$
\mathbb{E}(X_n) \leq \sqrt{\mathbb{E}(X_n^2)} \leq \sqrt{\frac{1 + \log(c_1)}{nc_2}}.
$$

□

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

**Theorem 18** Let $X_1, \ldots, X_n$ be random variables. Suppose there exists $\sigma > 0$ such that $\mathbb{E}(e^{tX_i}) \leq e^{t^2\sigma^2/2}$ for all $t > 0$. Then

$$
\mathbb{E} \left( \max_{1 \leq i \leq n} X_i \right) \leq \sigma \sqrt{2 \log n}. \quad (14)
$$

**Proof.** By Jensen’s inequality,

$$
\exp \left\{ t \mathbb{E} \left( \max_{1 \leq i \leq n} X_i \right) \right\} \leq \mathbb{E} \left( \exp \left\{ t \max_{1 \leq i \leq n} X_i \right\} \right) = \mathbb{E} \left( \max_{1 \leq i \leq n} \exp \{ tX_i \} \right) \leq \sum_{i=1}^n \mathbb{E} (\exp \{ tX_i \}) \leq ne^{t^2\sigma^2/2}.
$$

Thus,

$$
\mathbb{E} \left( \max_{1 \leq i \leq n} X_i \right) \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}.
$$

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

## 5 $O_P$ and $o_P$

In statistics, probability and machine learning, we make use of $o_P$ and $O_P$ notation.

Recall first, that $a_n = o(1)$ means that $a_n \to 0$ as $n \to \infty$. $a_n = o(b_n)$ means that $a_n/b_n = o(1)$.

$a_n = O(1)$ means that $a_n$ is eventually bounded, that is, for all large $n$, $|a_n| \leq C$ for some $C > 0$. $a_n = O(b_n)$ means that $a_n/b_n = O(1)$.  

8
We write \( a_n \sim b_n \) if both \( a_n/b_n \) and \( b_n/a_n \) are eventually bounded. In computer science this is written as \( a_n = \Theta(b_n) \) but we prefer using \( a_n \sim b_n \) since, in statistics, \( \Theta \) often denotes a parameter space.

Now we move on to the probabilistic versions. Say that \( Y_n = o_P(1) \) if, for every \( \epsilon > 0 \),
\[
\mathbb{P}(|Y_n| > \epsilon) \to 0.
\]

Say that \( Y_n = o_P(a_n) \) if, \( Y_n/a_n = o_P(1) \).

Say that \( Y_n = O_P(1) \) if, for every \( \epsilon > 0 \), there is a \( C > 0 \) such that
\[
\mathbb{P}(|Y_n| > C) \leq \epsilon.
\]

Say that \( Y_n = O_P(a_n) \) if \( Y_n/a_n = O_P(1) \).

Let’s use Hoeffding’s inequality to show that sample proportions are \( O_P(1/\sqrt{n}) \) within the true mean. Let \( Y_1, \ldots, Y_n \) be coin flips i.e. \( Y_i \in \{0, 1\} \). Let \( p = \mathbb{P}(Y_i = 1) \). Let
\[
\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.
\]

We will show that: \( \hat{p}_n - p = o_P(1) \) and \( \hat{p}_n - p = O_P(1/\sqrt{n}) \).

We have that
\[
\mathbb{P}(|\hat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} \to 0
\]
and so \( \hat{p}_n - p = o_P(1) \). Also,
\[
\mathbb{P}(\sqrt{n}|\hat{p}_n - p| > C) = \mathbb{P}\left(\frac{|\hat{p}_n - p|}{\sqrt{n}} > \frac{C}{\sqrt{n}}\right) \leq 2e^{-2C^2} < \delta
\]
if we pick \( C \) large enough. Hence, \( \sqrt{n}(\hat{p}_n - p) = O_P(1) \) and so
\[
\hat{p}_n - p = O_P\left(\frac{1}{\sqrt{n}}\right).
\]

Make sure you can prove the following:
\[
\begin{align*}
O_P(1)o_P(1) &= o_P(1) \\
O_P(1)O_P(1) &= O_P(1) \\
o_P(1) + O_P(1) &= O_P(1) \\
O_P(a_n)o_P(b_n) &= o_P(a_nb_n) \\
O_P(a_n)O_P(b_n) &= O_P(a_nb_n)
\end{align*}
\]