Chapter 5

Inequalities

5.1 The Markov and Chebyshev inequalities

As you have probably seen on today's front page: every person in the upper tenth percentile earns at least 12 times more than the average salary. In other words, if you pick up a random person, the probability that his salary is more than 12 times the average salary is more than 0.1. Put into a formal setting, denote by Ω the set of people endowed with a uniform probability and let $X(\omega)$ be the person's salary, then

$$P(X > 12\mathbb{E}[X]) \geq 0.1.$$

The following theorem will show that this is not possible.

Theorem 5.1 (Markov inequality) Let X be a random variable assuming non-negative values, i.e., $X(\omega) \geq 0$ *for every* $\omega \in \Omega$. *Then for every* $a \geq 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Comment: Andrey Andreyevich Markov (1856–1922) was a Russian mathematician notably known for his work on Stochastic processes.

Comment: Note first that this is a vacuous statement for $a < \mathbb{E}[X]$. For $a > \mathbb{E}[X]$ this inequality limits the probability that X assumes values larger than its expected value. This is the first time in this course that we derive an inequality. Inequalities, in general, are an important tool in analysis, where estimates (rather than exact identities) are often needed. The strength of such an inequality is that it holds for *every* random variable. As a result, this estimate may be far from being tight.

Proof: Let A be the event that X is at least a, that is

$$A = \{\omega \in \Omega \mid X(\omega) \ge a\}.$$

Define a new random variable $Y = a\,I_A$, where I_A is the indicator of A. That is,

$$Y(\omega) = \begin{cases} a & X(\omega) \ge a \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $X \ge Y$, so by the monotonicity of expectation

$$\mathbb{E}[X] \ge \mathbb{E}[Y] = a\,\mathbb{E}[I_A] = a\,P(A) = a\,P(X \ge a),$$

where we used the fact that I_A is a Bernoulli variable, hence, its expectation equals the probability that it is equal to 1. ■

A corollary of Markov's inequality is another inequality due to Chebyshev:

Theorem 5.2 (Chebyshev's inequality) Let X be a random variable with expected value μ and variance σ^2. Then, for every $a > 0$

$$P(|X - \mu| \ge a) \le \frac{\sigma^2}{a^2}.$$

Comment: Pafnuty Lvovich Chebyshev (1821–1894) was a Russian mathematician known for many contributions; he was Markov's advisor.

Comment: If we write $a = k\sigma$, this theorem states that the probability that a random variable assumes a value whose absolute distance from its expected value is more than k times its standard deviation is at most $1/k^2$,

$$P(|X - \mu| \ge k\sigma) \le \frac{1}{k^2}.$$

Proof: Define $Y = |X - \mu|^2$. Y is a non-negative random variable, satisfying, by definition, $\mathbb{E}[Y] = \sigma^2$. Applying Markov's inequality for Y,

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

■

Example: On average, John Doe drinks 25 liters of wine every week. What can be said about the probability that he drinks more than 50 liters of wine on a given week?

Here we apply the Markov inequality. If X is the amount of wine John Doe drinks on a given week, then

$$P(X > 50) \leq \frac{\mathbb{E}[X]}{50} = \frac{1}{2}.$$

Note that this may be a very loose estimate; it may well be that Joe Doe drinks exactly 25 liters of wine every week, in which case this probability is zero. ▲▲▲

Example: Let $X \sim \mathscr{B}(n, 1/2)$, The,

$$\mathbb{E}[X] = \frac{n}{2} \qquad \text{and} \qquad \text{Var}[X] = \frac{n}{4}.$$

If you toss a fair coin 10^6 times, the distribution of the number of Heads is $X \sim \mathscr{B}(10^6, 1/2)$; the probability of getting between 495,000 and 505,000 Heads can be bounded by

$$P(|X - \mathbb{E}[X]| < 5000) = 1 - P(|X - \mathbb{E}[X]| \geq 5000) \geq 1 - \frac{10^6/4}{5000^2} = 0.99$$

▲▲▲

Since a binomial variable $\mathscr{B}(n, p)$ can be represented as a sum of n Bernoulli variables, the following generalization is called for. Let $X_1, X_2, \dots, X_n$ by independent and identically distributed random variables, with

$$\mathbb{E}[X_1] = \mu \qquad \text{and} \qquad \text{Var}[X_1] = \sigma^2.$$

Define

$$X = \sum_{k=1}^{n} X_k.$$

Then, by the linearity of the expectation and the independence of the random variables,

$$\mathbb{E}[X] = n\mu \qquad \text{and} \qquad \mathrm{Var}[X] = n\sigma^2.$$

Chebyshev's inequality yields

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

This inequality starts getting useful when $a > \sqrt{n}\sigma$. For further reference, we note that in the particular case where $\mu = 0$, we get *a fortiori* that for every $a > 0$,

$$P(X \geq a) \leq \frac{n\sigma^2}{a^2}.$$

5.2 Hoeffding's inequality

We have just seen that if we consider a sum of independent random variables, we can use Chebyshev's inequality to bound the probability that this sum deviates by much from its expectation. We now ask: is this result tight?

It turns out there are much better bounds that we can get in this setting.

Theorem 5.3 (Hoeffding's inequality) Let $X_1, X_2, \dots, X_n$ be independent random variables (not necessarily identically-distributed) satisfying $\mathbb{E}[X_i] = 0$ and $|X_i| \leq 1$ for all $1 \leq i \leq n$ (that is, for every i and every $\omega \in \Omega$, $|X_i(\omega)| \geq 1$). Then, for every $a > 0$

$$P\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left(-\frac{a^2}{2n}\right).$$

Recall the definition of the moment generating function of X—$M_X(t) = \mathbb{E}[e^{tX}]$. The proof will use moment generating functions and Markov's inequality. We need the following lemma.

Lemma 5.1 Let X be a random variable satisfying $\mathbb{E}[X] = 0$ and $|X_i| \leq 1$. Then for all $t \in \mathbb{R}$

$$M_X(t) \leq \exp\left(\frac{t^2}{2}\right).$$

Proof: Fix $t \in \mathbb{R}$, and consider the function $f(x) = e^{-tx}$. This function is *convex* (the second derivative of f is everywhere positive), meaning that any line segment between two points on its graph lies above the graph; i.e., for every $a, b \in \mathbb{R}$ and $0 \leq \lambda \leq 1$,

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b).$$

In particular, taking $a = -1$ and $b = 1$ we get that for every t and every $0 \leq \lambda \leq 1$,

$$e^{t(\lambda-(1-\lambda))} \leq \lambda e^t + (1-\lambda)e^{-t}.$$

Setting $x = 2\lambda - 1$, i.e., $\lambda = (1+x)/2$ and $1 - \lambda = (1-x)/2$, we get that for any $-1 \leq x \leq 1$,

$$e^{tx} \leq \frac{1+x}{2}e^t + \frac{1-x}{2}e^{-t} = \frac{e^t + e^{-t}}{2} + x\frac{e^t - e^{-t}}{2}.$$

Since X is always between -1 and 1, it follows that

$$e^{tX} \leq \frac{e^t + e^{-t}}{2} + X\frac{e^t - e^{-t}}{2},$$

which is an inequality between random variables. By monotonicity of expectation, using the linearity of the expectation and the fact that $\mathbb{E}[X] = 0$,

$$M_X(t) = \mathbb{E}[e^{tX}] \leq \frac{e^t + e^{-t}}{2}.$$

It remains to show that for all $t \in \mathbb{R}$

$$\frac{e^t + e^{-t}}{2} \leq \exp\left(\frac{t^2}{2}\right).$$

This is done by comparison of the Taylor series around 0,

$$\frac{e^t + e^{-t}}{2} = \frac{1}{2}\sum_{k=0}^{\infty}\frac{t^k + (-t)^k}{k!} = \sum_{\ell=0}^{\infty}\frac{t^{2\ell}}{(2\ell)!} \leq \sum_{\ell=0}^{\infty}\frac{t^{2\ell}}{2^\ell \ell!} = \sum_{\ell=0}^{\infty}\frac{(t^2/2)^\ell}{\ell!} = \exp\left(\frac{t^2}{2}\right),$$

where we used the fact that $(2\ell)! \geq 2^\ell \ell!$. ■

Proof:(of Hoeffding's inequality) Let $X = \sum_{i=1}^n X_i$. First note that

$$M_X(t) = \mathbb{E}[\exp(t\sum_{i=1}^n X_i)] = \mathbb{E}[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n M_{X_i}(t)$$

where we used Proposition 4.3 and the independence of the random variables. Using the above lemma we get that

$$M_X(t) \leq \exp\left(\frac{nt^2}{2}\right).$$

For $t > 0$ the event $X \geq a$ is the same as the event $e^{tX} \geq e^{ta}$. Using Markov's inequality we get that for any $t > 0$

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = \frac{M_X(t)}{e^{ta}} \leq \exp\left(\frac{nt^2}{2} - ta\right).$$

The right-hand side depends on the parameter t, whereas the left-hand side does not; to get a tighter bound, we may choose t such to minimize the right-hand side. The minimum is achieved when $t = a/n$, hence

$$P(X \geq a) \leq \exp\left(\frac{n(a/n)^2}{2} - (a/n)a\right) = \exp\left(-\frac{a^2}{2n}\right).$$

■

Comment: Wassily Hoeffding (1914–1991) was a Finnish statistician. The inequality named after him was proved in 1963.

Comment: We have stated Hoeffding's inequality under the assumptions that the random variables are bounded by 1 in absolute value and the expectation is 0. In general, if we have bounded random variables we can apply an affine transformation to get a random variables which satisfy the condition of the theorem. For example, if $\mathbb{E}[X_i] = \mu$ and $|X_i| \leq b$, then

$$Y_i = \frac{X_i - \mu}{b + |\mu|}$$

satisfies,

$$\mathbb{E}[Y_i] = 0 \qquad \text{and} \qquad |Y_i| \leq \frac{|X_i| + |\mu|}{b + |\mu|} \leq 1.$$

Example: Going back to the last example from the previous section, if X_i are independent *Bernoulli*$\left(\frac{1}{2}\right)$ random variables, then $X = \sum_{i=1}^n X_i \sim \mathscr{B}\left(n, \frac{1}{2}\right)$. We define new random variables $Y_i = 2X_i - 1$ so that $|Y_i| \le 1$ and $\mathbb{E}[Y_i] = 0$. Furthermore, set $Y = \sum_{i=1}^{n} Y_i$, so that $Y = 2X - n$.

The event $X \ge \frac{n}{2} + 5\sqrt{n}$ equals the event $Y \ge 10\sqrt{n}$. Hoeffding's inequality gives us the bound

$$P\left(Y \ge 10\sqrt{n}\right) \le \exp\left(-\frac{100n}{2n}\right) = e^{-50}.$$

A similar bound holds for $X \le \frac{n}{2} - 5\sqrt{n}$. Put together,

$$P\left(\left|X - \frac{n}{2}\right| \ge 5\sqrt{n}\right) \le 2e^{-50}.$$

So, if you tossed a fair coin 10^6 times, the probability that you get between $495,000$ and $505,000$ heads is at least

$$1 - 2e^{-50} \approx 0.99999999999999999999961425...$$

Quite an improvement over our previous bound of 99 percent! ▲▲▲

5.3 Jensen's inequality

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is called *convex* if its graph between every two points lies under the secant, i.e., if for every $x, y \in \mathbb{R}$ and $0 < \lambda < 1$,

$$f(\lambda x + (1 - \lambda)y) \le \lambda f(x) + (1 - \lambda) f(y).$$

Proposition 5.1 (Jensen's inequality) If g is a convex real-valued function and X is a real-valued random variable, then

$$g(\mathbb{E}[X]) \le \mathbb{E}[g(X)],$$

provided that the expectations exist.

Proof: Let's start with a proof under the additional assumption that g is twice differentiable. Since g is convex, it follows that $g''(x) \ge 0$ for all x.

Taylor expanding g about $\mu = \mathbb{E}[X]$,

$$g(x) = g(\mu) + g'(\mu)\,(x-\mu) + \frac{1}{2}g''(\xi)(x-\mu)^2$$

for some ξ. The last term is non-negative, so we get that for all x

$$g(x) \geq g(\mu) + g'(\mu)\,(x-\mu),$$

and as an inequality between random variables:

$$g(X) \geq g(\mu) + g'(\mu)\,(X-\mu).$$

By the monotonicity of expectation,

$$\mathbb{E}[g(X)] \geq \mathbb{E}[g(\mu) + g'(\mu)\,(X-\mu)] = g(\mu) + g'(\mu)\,(\mathbb{E}[X]-\mu) = g(\mathbb{E}[X])$$

which is precisely what we need to show.

What about the more general case? Any convex function is continuous, and has one-sided derivatives with

$$g'_-(x) = \lim_{y\uparrow x}\frac{g(x)-g(y)}{x-y} \leq \lim_{y\downarrow x}\frac{g(x)-g(y)}{x-y} = g'_+(x).$$

For every $m \in [g'_-(\mu), g'_+(\mu)]$

$$g(x) \geq g(\mu) + m(x-\mu),$$

so the same proof holds with m replacing $g'(\mu)$. ■

Comment: Jensen's inequality is valid also for convex functions of several variables.

Example: Since exp is a convex function,

$$\exp(t\,\mathbb{E}[X]) \leq \mathbb{E}[e^{tX}] = M_X(t),$$

or,

$$\mathbb{E}[X] \leq \frac{1}{t}\log M_X(t),$$

for all $t > 0$. ▲▲▲

Example: Consider a discrete random variable X assuming the positive values $x_1, \dots, x_n$ with equal probability $1/n$. Jensen's inequality for $g(x) = -\log(x)$ gives,

$$-\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \le -\frac{1}{n}\sum_{i=1}^{n}\log(x_i) = -\log\left(\prod_{i=1}^{n} x_i^{1/n}\right).$$

Reversing signs, and exponentiating, we get

$$\frac{1}{n}\sum_{i=1}^{n} x_i \ge \left(\prod_{i=1}^{n} x_i\right)^{1/n},$$

which is the classical *arithmetic mean-geometric mean inequality*. In fact, this inequality can be generalized for arbitrary distributions, $p_X(x_i) = p_i$, yielding

$$\sum_{i=1}^{n} p_i x_i \ge \prod_{i=1}^{n} x_i^{p_i}.$$

▲ ▲ ▲

5.4 Kolmogorov's inequality

(Not taught in 2016-7)

The Kolmogorov inequality may first seem to be of similar flavor as Chebyshev's inequality, but it is considerably stronger. I have decided to include it here because its proof involves some interesting subtleties. First, a lemma:

Lemma 5.2 If X, Y, Z are random variables such that Y is independent of X and Z, then

$$\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\,\mathbb{E}[Y].$$

Proof: The fact that Y is independent of both X, Z implies that (in the case of discrete variables),

$$p_{X,Y,Z}(x,y,z) = p_{X,Z}(x,z)\,p_Y(y).$$

Now, for every z,

$$
\begin{aligned}
\mathbb{E}[XY|Z = z] &= \sum_{x,y} xy\, p_{X,Y|Z}(x, y|z) \\
&= \sum_{x,y} xy \frac{p_{X,Y,Z}(x, y, z)}{p_Z(z)} \\
&= \sum_{x,y} xy \frac{p_{X,Z}(x, z) p_Y(y)}{p_Z(z)} \\
&= \sum_{y} y\, p_Y(y) \sum_x x \frac{p_{X,Z}(x, z)}{p_Z(z)} \\
&= \mathbb{E}[Y]\, \mathbb{E}[X|Z = z].
\end{aligned}
$$

■

Theorem 5.4 (Kolmogorov's inequality) Let $X_1, \dots, X_n$ be independent random variables such that $\mathbb{E}[X_k] = 0$ and $\mathrm{Var}[X_k] = \sigma_k^2 < \infty$. Then, for all $a > 0$,

$$
P\left(\max_{1\le k\le n} |X_1 + \cdots + X_k| \ge a\right) \le \frac{1}{a^2} \sum_{i=1}^n \sigma_i^2.
$$

Comment: For $n = 1$ this is nothing but the Chebyshev inequality. For $n > 1$ it would still be Chebyshev's inequality if the maximum over $1 \le k \le n$ was replaced by $k = n$, since by independence

$$
\mathrm{Var}[X_1 + \cdots + X_n] = \sum_{i=1}^n \sigma_i^2.
$$

Proof: We introduce the notation $S_k = X_1 + \cdots + X_k$. This theorem is concerned with the probability that $|S_k| > a$ for *some* k. We define the random variable $N(\omega)$ to be the smallest integer k for which $|S_k| > a$; if there is no such number we set $N(\omega) = n$. We observe the equivalence of events,

$$
\left\{\omega :\ \max_{1\le k\le n} |S_k| > a\right\} = \left\{\omega :\ S_{N(\omega)}^2 > a^2\right\},
$$

and from the Markov inequality

$$P\left(\max_{1\le k\le n}|S_k|>a\right)\le\frac{1}{a^2}\mathbb{E}[S_N^2].$$

We need to estimate the right hand side. If we could replace

$$\mathbb{E}[S_N^2]\qquad\text{by}\qquad\mathbb{E}[S_n^2]=\operatorname{Var}[S_n]=\sum_{i=1}^n\sigma_i^2,$$

then we would be done.

The trick is to show that $\mathbb{E}[S_N^2]\le\mathbb{E}[S_n^2]$ by using conditional expectations. If

$$\mathbb{E}[S_N^2|N=k]\le\mathbb{E}[S_n^2|N=k]$$

for all $1\le k\le n$ then the inequality holds, since we have then an inequality between *random variables* $\mathbb{E}[S_N^2|N]\le\mathbb{E}[S_n^2|N]$, and applying expectations on both sides gives the desired result.

For $k=n$, the identity

$$\mathbb{E}[S_N^2|N=n]=\mathbb{E}[S_n^2|N=n],$$

hold trivially. Otherwise, we write

$$\begin{aligned}\mathbb{E}[S_n^2|N=k]&=\mathbb{E}[S_k^2|N=k]+\mathbb{E}[(X_{k+1}+\cdots+X_n)^2|N=k]\\&\quad+2\mathbb{E}[S_k(X_{k+1}+\cdots+X_n)|N=k]\end{aligned}$$

The first term on the right hand side equals $\mathbb{E}[S_N^2|N=k]$, whereas the second terms is non-negative. Remains the third term for which we remark that $X_{k+1}+\cdots+X_n$ is independent of both S_k and N, and by the previous lemma,

$$\mathbb{E}[S_k(X_{k+1}+\cdots+X_n)|N=k]=\mathbb{E}[S_k|N=k]\,\mathbb{E}[X_{k+1}+\cdots+X_n]=0.$$

Putting it all together,

$$\mathbb{E}[S_n^2|N=k]\ge\mathbb{E}[S_N^2|N=k].$$

Since this holds for all k's we have thus shown that

$$\mathbb{E}[S_N^2]\le\mathbb{E}[S_n^2]=\sum_{i=1}^n\sigma_i^2,$$

which completes the proof. ■