Notes on the Fundamental Limit Theorems

December 5, 2018

Abstract

These are some notes on the main limit theorems of probability theory.

0.1 The empirical mean

Let $X_{j\,j\in\mathbb{N}}$ be an infinite sequence of independent identically distributed real valued random variables on some probability space. The *empirical mean at N observations*, sometimes called the *sample mean*, is the random variable

$$\overline{X}_N(\omega) = \frac{1}{N}\sum_{j=1}^{N} X_j(\omega)\ . \tag{0.1}$$

The *Strong Law of Large Numbers* says that provided $\mathrm{E}|X_1| < \infty$,

$$\lim_{N\to\infty} \overline{X}_N(\omega) = \mathrm{E}X_1 \tag{0.2}$$

for almost every ω.

Two fundamental theorems give further information given stronger information on X_1. If it is also true that

$$\sigma^2 := \mathrm{E}(X_1 - \mathrm{E}X_1)^2 < \infty \tag{0.3}$$

the *Central Limit Theorem* says that for all $x \in \mathbb{R}$,

$$\lim_{N\to\infty} \Pr\left\{\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\frac{X_j - \mathrm{E}X_j}{\sigma} > x\right\} = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\mathrm{d}y\ .$$

That is, if we define $Z_N = \dfrac{1}{\sqrt{N}}\displaystyle\sum_{j=1}^{N}\frac{X_j - \mathrm{E}X_j}{\sigma}$

$$\overline{X}_N = \mathrm{E}X_1 + \frac{\sigma}{\sqrt{N}} Z_N$$

where Z_N is approximately *normally distributed*.

That is, the deviations of $\overline{X}_N$ from its mean $\mathrm{E}X_1$ are *typically* of order $N^{-1/2}$ for large N. The Central Limit Theorem gives a complete description of the *typical deviations*, and they depend on the distribution of X_1 only through its mean $\mathrm{E}X_1$ and is variances σ^2. In this sense the description is *universal*.

Cramér's Theorem (1938) describes *large deviations* from the mean, and in this case the description depends on the distribution of X_1 through more than only the mean and variance, but it provides more precise information. In these notes we prove and apply both of these theorems.

0.2 The Strong Law of Large Numbers

0.1 THEOREM (The Strong Law of Large Numbers)**.** *Let $\{X_j\}_j$ be sequence of independent identically distributed random variables on some probability space (Ω, P). Suppose that $\mathrm{E}(|X_1|) < \infty$, and put $\mu := \mathrm{E}(X_1)$. Then there is an even $A \subset \Omega$ with $P(A) = 1$ such that*

$$lim_{N\to\infty}\overline{X}_N(\omega) = \mu$$

for every $\omega \in A$.

0.2 Remark. An event E such that $P(E) = 1$ is called a *sure event*: The event E is sure to happen. The Strong Law of Large Numbers specifies conditions under which the convergence of the sample mean to the mean is sure to happen.

We will not prove Theorem 0.1 in full strength, to avoid technicalities that are better treated in a course that makes use of measure theory. However, we shall prove a version that is only slightly weaker, in that we make the stringer assumption that $\mathrm{E}(X_1^4) < \infty$, but we will also get some information on how fast the limit sets in. =

0.3 LEMMA. *Let $\{X_j\}_j$ be sequence of independent identically distributed random variables on some probability space (Ω, P). Suppose that $\mathrm{E}(X_1^4) = K < \infty$, and that $\mathrm{E}(X_1) = 0$. Then*

$$\mathrm{E}(\overline{X}_N^4) \le \frac{10K}{N^2} \ .$$

Proof. By the definition,

$$\mathrm{E}(\overline{X}_N^4) = \frac{1}{N^4}\sum_{i,j,k,\ell=1}^{N} \mathrm{E}(X_iX_jX_kX_\ell) \ .$$

By the independence, if i is different from each of j, k and ℓ, X_i and $X_jX_kX_\ell$ are independent and then

$$\mathrm{E}(X_iX_jX_kX_\ell) = \mathrm{E}(X_i)\mathrm{E}(X_jX_kX_\ell) = 0 \ ,$$

since $\mathrm{E}(X_i) = 0$. Therefore the only terms that contribute to the sum are those in which the indices are paired off. There are $\binom{4}{2} = 6$ ways to do this with two distinct pairs, and $\binom{4}{1} = 4$ ways to do this with two identical pairs. Moreover, for $i \neq j$

$$\mathrm{E}(X_i^2X_j^2) = \mathrm{E}(X_i^2)\mathrm{E}(X_j^2) \le (\mathrm{E}(X_i^4))^{1/2}(\mathrm{E}(X_j^4))^{1/2} = K \ ,$$

and so

$$\mathrm{E}(\overline{X}_N^4) \leq \frac{6KN(N-1)}{N^4} + \frac{4KN}{N^4} \leq \frac{10K}{N^2} \ .$$

□

Now pick $\epsilon > 0$. Then $|\overline{X}_N| \geq \epsilon$ if and only if $\overline{X}_N^4 \geq \epsilon^4$, and by Markov's inequality,

$$P(\overline{X}_N^4 \geq \epsilon^4) \leq \frac{\mathrm{E}(\overline{X}_N^4)}{\epsilon^4} \ .$$

Therefore, using the lemma proved just above, if we define $A_{N,\epsilon} := \{|\overline{X}_N| \geq \epsilon\}$, we have

$$P(A_{N,\epsilon}) \leq \frac{10K}{N^2\epsilon^4} \ .$$

and therefore that

$$\sum_{N=1}^{\infty} P(A_{N,\epsilon}) < \infty \ .$$

Now define $1_{A_{N.\epsilon}}$ to be the random variable

$$1_{A_{N.\epsilon}}(\omega) = \begin{cases} 1 & \omega \in A_{N,\epsilon} \\ 0 & \omega \notin A_{N,\epsilon} \end{cases} \ .$$

Define

$$B_\epsilon = \{\omega \ : \ \omega \in A_{N,\epsilon} \quad \text{for infinitely many values of } N\} \ .$$

For all $\omega \in B_\epsilon$,

$$\sum_{N=1}^{\infty} 1_{A_{N.\epsilon}}(\omega) = \infty \ .$$

But

$$\mathrm{E}\left(\sum_{N=1}^{\infty} 1_{A_{N,\epsilon}}\right) = \sum_{N=1}^{\infty} \mathrm{E}(1_{A_{N,\epsilon}}) = \sum_{N=1}^{\infty} P(A_{N,\epsilon}) < \infty \ .$$

This is impossible unless $P(B_\epsilon) = 0$. Therefore, define $A_\epsilon = B_\epsilon^c$, and we have $P(A_\epsilon) = 1$, and for all $\omega \in A_\epsilon$, $|\overline{X}_N(\omega)| > \epsilon$ for only finitely many values of N.

Finally, we define $A := \cap_{n=1}^{\infty} A_{1/n}$. Then by the axioms of probability, (or the Dominated Convergence Theorem in a measure theory based approach),

$$P(A) = \lim_{n\to\infty} P(A_{1/n}) = 1 \ .$$

By construction, for each $\omega \in A$ and each $n \in \mathbb{N}$, $|\overline{X}_N \geq 1/n$ for only finitely many values of N, and this means that

$$\lim_{N\to\infty} \overline{X}_N(\omega) = 0 \ .$$

Finally, if instead we have $\mathrm{E}(X_1) = \mu \neq 0$, define $\widetilde{X}_j = X_j - \mu$, and apply the analysis made above to $\{\widetilde{X}_j\}_{j\in\mathbb{N}}$.

0.3 The Central Limit Theorem

Let $\{X_j\}_{j\in\mathbb{N}}$ be an independent identically distributed sequence of random variables with zero mean and unit variance so that, for all j, $\mathrm{E}(X_j) = 0$ and $\mathrm{E}(X_j^2) = 1$. Define

$$\overline{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j \ .$$

Let $\{Y_j\}_{j\in\mathbb{N}}$ be an independent distributed sequence of standard norm random variables. Define

$$\overline{Y}_N = \frac{1}{N}\sum_{j=1}^{N} Y_j \ .$$

Recall the that sum of independent normal random variables is again normal. There for $\sum_{j=1}^{N} Y_j$ is normal. Since means add for sums of random variables, and variances add for sums of independent random variables. $\sum_{j=1}^{N} Y_j$ has mean zero and variance N. It follows that

$$\sqrt{N}\overline{Y}_N$$

is a standard normal variable.

In other words, $\dfrac{1}{\sqrt{N}}\displaystyle\sum_{j=1}^{N} Y_j$ is a standard normal random variables. The Central Limit Theorem, to be proved in this section, is often discussed in terms of this formula, normalized with the square root in the denominator. However, make the relation between the three main limit theorems discussed in these notes, we prefer to express them all in terms of the sample mean, so we divide our sums by N to obtain the sample mean, which in this case would have variance $1/N$, and then we multiply by $\sqrt{N}$ to bring the variance back up to 1.

Since $\sqrt{N}\overline{Y}_N$ is standard normal, for any bounded piecewise continuous function g,

$$\mathrm{E}(g(\sqrt{N}\overline{Y}_N)) = \int_{\mathbb{R}} g(y)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\mathrm{d}y \ .$$

0.4 THEOREM (Central Limit Theorem)**.** *Let $\{X_j\}_{j\in\mathbb{N}}$ be an independent identically distributed sequence of random variables with zero mean and unit variance so that, for all j, $\mathrm{E}(X_j) = 0$ and $\mathrm{E}(X_j^2) = 1$. Define*

$$\overline{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j \ .$$

Then for any bounded piecewise continuous function g,

$$\lim_{N\to\infty} \mathrm{E}(g(\sqrt{N}\overline{X}_N)) = \int_{\mathbb{R}} g(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x \ . \tag{0.4}$$

In particular, for any $a \in \mathbb{R}$, define

$$g(x) = \begin{cases} 1 & x \le a \\ 0 & x > a \ , \end{cases} \tag{0.5}$$

Then

$$\mathrm{E}(g(\sqrt{N}\overline{X}_N)) = P(\sqrt{N}\overline{X}_N \le a) \qquad \text{and} \qquad \int_{\mathbb{R}} g(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x = \Phi(a) \ .$$

Therefore, one special case of Central Limit Theorem is that for each $a \in \mathbb{R}$,

$$\lim_{N\to\infty} P(\sqrt{N}\overline{X}_N \le a) = \Phi(a) \ .$$

We shall first prove something a bit less: We will assume more about X_1 and g. About X_1 we assume that $\mathrm{E}(|X_1|^3) < \infty$. About g, we assume that it is three times continuously differentiable and g, g', g'' and g''' are all uniformly bounded function on R. The first assumption is not a serious restriction, and it could be dropped by making the proof slightly more technical, as is discussed below. The second assumption rules out functions like the one in (0.5), but we can approximate the function in (0.5) by nice smooth functions arbitrarily closely, so that as explained below, one we know the theorem for nice functions, it is easy to prove it in general.

0.5 LEMMA. *Let g be a function with three continuous derivatives such that for some finite constant C,*

$$\max\{g(x), g'(x), g''(x), g'''(x)\} \le C \tag{0.6}$$

for all x. Let $\{X_j\}_{j\in\mathbb{N}}$ be any sequence of independent, identically distributed random variables such that $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$. Suppose further that for some $K < \infty$, $\mathrm{E}|\widetilde{X_1}|^3 = K$. Then

$$\lim_{n\to\infty} \mathrm{E}g\left(\sqrt{N}\overline{X}_N)\right) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} g(x)e^{-x^2/2}\mathrm{d}x \ .$$

Proof. With $\{X_j\}_{j\in\mathbb{N}}$ and $\{Y_j\}_{j\in\mathbb{N}}$ as above, define $\widetilde{S}_n = \sum_{j=1}^n \widetilde{T}_j$, and put $S_n = \sum_{j=1}^n T_j$ as before. For each $k = 0, \ldots, n$, define

$$W_{N,k} := \frac{1}{\sqrt{N}}\left(\sum_{j=0}^{k} X_j + \sum_{j=k+1}^{N} Y_j\right)$$

Then $W_{N,0} = \sqrt{N}\,\overline{Y}_N$ and $W_{N,N} = \sqrt{N}\,\overline{X}_N$. Therefore, we have the telescoping sum

$$g\left(\sqrt{N}\,\overline{X}_N\right) - g\left(\sqrt{N}\,\overline{Y}_N\right) = g\left(W_{N,N}\right) - g\left(W_{N,0}\right) = \sum_{k=0}^{N-1}\left[g\left(W_{N,k+1}\right) - g\left(W_{N,k}\right)\right] \ .$$

By linearity of the expectation, and the fact that $\sqrt{N}\,\overline{Y}_N$ is standard normal,

$$\begin{aligned} g\left(\sqrt{N}\,\overline{X}_N\right) &= g\left(\sqrt{N}\,\overline{Y}_N\right) + \sum_{k=0}^{N-1}\mathrm{E}\left[g\left(W_{N,k+1}\right) - g\left(W_{N,k}\right)\right] \\ &= \int_{\mathbb{R}} g(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x + \sum_{k=0}^{N-1}\mathrm{E}\left[g\left(W_{N,k+1}\right) - g\left(W_{N,k}\right)\right] \ . \ . \end{aligned}$$

Therefore, it remains to be shown that

$$\lim_{N\to\infty}\left(\sum_{k=0}^{N-1}\mathrm{E}\left[g\left(W_{N,k+1}\right)-g\left(W_{N,k}\right)\right]\right)=0\ . \tag{0.7}$$

We accomplish this with a Taylor expansion. The sums defining $W_{N,k+1}$ and $W_{N,k}$ differ only in the $(k+1)$st term. Define

$$U_{N,k}=\frac{1}{\sqrt{N}}\left(\sum_{j=0}^{k}X_j+\sum_{j=k+2}^{N}Y_j\right)$$

Then

$$W_{N,k+1}=U_{N,k}+\frac{1}{\sqrt{N}}X_{k+1}\quad\text{and}\quad W_{n,k}=U_{n,k}+\frac{1}{\sqrt{N}}Y_{k+1}\ .$$

Therefore, by Taylor's Theorem,

$$\begin{aligned}g\left(W_{N,k+1}\right)=g\left(U_{n,k}\right)\ &+\ g'\left(U_{n,k}\right)\frac{X_{k+1}}{\sqrt{N}}\\ &+\ \frac{1}{2}g''\left(U_{n,k}\right)\left(\frac{X_{k+1}}{\sqrt{N}}\right)^2\\ &\pm\ C\left|\frac{X_{k+1}}{\sqrt{N}}\right|^3\ .\end{aligned}$$

Taking the expectation, since $\mathrm{E}(X_{k+1})=0$, $\mathrm{E}((X_{k+1})^2)=1$ and $\mathrm{E}(|X_{k+1}|^3)=K$, and using the independence of T_{k+1} and $U_{n,k}$, we have

$$\mathrm{E}g\left(W_{N,k+1}\right)=\mathrm{E}g\left(U_{N,k}\right)+\frac{1}{2}\mathrm{E}g''\left(U_{N,k}\right)\frac{1}{N}\pm\frac{C}{N^{3/2}}\ .$$

Since we also have $\mathrm{E}(Y_{k+1})=0$, $\mathrm{E}(Y_{k+1})^2=1$ and $\mathrm{E}|Y_{k+1}|^3=\frac{4}{\sqrt{2\pi}}=:K_0$, we have by the exact same reasoning that

$$\mathrm{E}g\left(W_{N,k}\right)=\mathrm{E}g\left(U_{n,k}\right)+\frac{1}{2}\mathrm{E}g''\left(U_{N,k}\right)\frac{1}{N}\pm\frac{CK_0}{N^{3/2}}\ .$$

In our expressions for $\mathrm{E}g\left(W_{N,k+1}\right)$ and $\mathrm{E}g\left(W_{N,k}\right)$, it is only the remainder terms that differ, and this shows shows that

$$\left|\mathrm{E}g\left(W_{N,k+1}\right)-\mathrm{E}g\left(W_{n,k}\right)\right|\le\frac{C(K+K_0)}{N^{3/2}}\ ,$$

and hence proves (0.7). □

To pass from the lemma that we have proved to the theorem we have stated, we just need to explain how the additional assumptions in the lemma can be relaxed. This uses simple results form analysis. The requirement that $\mathrm{E}|Y_j|^3<\infty$ was used only because we used the simplest form

of the remainder in Taylor's Theorem. Working harder with the integral form, one can show that nothing more is needed than $\mathrm{E}|Y_j|^2 < \infty$, which is already required to have a finite variance.

To relax the assumptions on g, let us fix and $a \in \mathbb{R}$, consider the function g defined in (0.5) as an example. Now fix any $\epsilon > 0$. By "rounding the corners" one can find functions g_0 and g_1 such that

$$0 \le g_0(x) \le g(x) \le g_1(x) \le 1$$

for all x, and $g_1(x) - g_0(x) = g(x)$ for all x with $|x - a| > \epsilon$, and finally such that g_0 and g_1 satisfy (0.6) for some finite C. Note that C will diverge to infinity as ϵ tends to zero, but we can use the same C value for *all* $a \in \mathbb{R}$ – C depends on how you smooth the jump, but not where it is.

We then have, using $\le g_0(x) \le g(x) \le g_1(x)$,

$$\mathrm{E}(g_0(\sqrt{N}\,\overline{X}_N) \le \mathrm{E}(g(\sqrt{N}\,\overline{X}_N) \le \mathrm{E}(g_1(\sqrt{N}\,\overline{X}_N)\ .$$

Taking N sufficiently large, the lemma assures us that

$$\mathrm{E}(g_1(\sqrt{N}\,\overline{X}_N) \le \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g_1(x) e^{-x^2/2} \mathrm{d}x + \epsilon$$

and

$$\mathrm{E}(g_0(\sqrt{N}\,\overline{X}_N) \ge \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g_0(x) e^{-x^2/2} \mathrm{d}x - \epsilon$$

But

$$\int_{\mathbb{R}} |g_1(x) - g_0(x)| e^{-x^2/2} \mathrm{d}x \le \frac{1}{\sqrt{2\pi}} 2\epsilon \le \epsilon\ .$$

Therefore, using $\le g_0(x) \le g(x) \le g_1(x)$ once more,

$$\mathrm{E}(g_1(\sqrt{N}\,\overline{X}_N) \le \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) e^{-x^2/2} \mathrm{d}x + 2\epsilon$$

and

$$\mathrm{E}(g_0(\sqrt{N}\,\overline{X}_N) \ge \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x) e^{-x^2/2} \mathrm{d}x - 2\epsilon\ .$$

Since epsilon is arbitrary, this proves (0.4) when g is given by (0.5), and moreover, it shows that the convergence is uniform in a. That is,

$$\lim_{N\to\infty} P\left(\overline{X}_N \le \frac{a}{\sqrt{N}}\right) = \Phi(a)\ , \tag{0.8}$$

with the convergence uniform in a.

0.6 EXAMPLE. *Let $\{X_j\}_{j\in N}$ be independent an identically distributed with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Then X_1 has zero mean and unit variance. Take $a = 3$. The Central Limit Theorem says that for large N,*

$$\lim_{N\to\infty} P\left(\overline{X}_N > \frac{3}{\sqrt{N}}\right) \approx 1 - \Phi(3) \approx 0.0013\ .$$

However, for $N = 10^4$, $\frac{3}{\sqrt{N}} = 0.03$. *Let* Z *be a random variable with* $P(Z = 1) = p$ *and* $P(Z = -1) = 1 - p$. *Then* $\mathrm{E}(Z) = 2p - 1$, *so we would have* $\mathrm{E}(Z) = 0.03$ *for* $p = 0.515$ *so there is a non-negligible chance that a fair coin might appear to have a significant bias even after* 10^4 *trials – assuming the the normal approximation is already valid. but for* $N = 10^6$, $\frac{3}{\sqrt{N}} = 0.003$, *and then it is quite unlikely that the sample mean differs from zero by more than* ± 0.003.

The Central Limit Theorem may be easily applied to any independent identically distributed sequence $\{Z_j\}_{j\in\mathbb{N}}$ when Z_1 has a finite second moment. Then the mean μ and variance σ^2 are also finite, and we define

$$X_j = \frac{Z_j - \mu}{\sigma} \ .$$

$$\overline{X}_N = \frac{\overline{Z}_N - \mu}{\sigma} \ ,$$

and $X_{j\,j\in\mathbb{N}}$ is an independent identically distributed sequence with zero mean and unit variance, and we may apply the Central Limit Theorem to it to conclude that

$$\lim_{N\to\infty} P\left(\frac{\overline{Z}_N - \mu}{\sigma} \le \frac{a}{\sqrt{N}}\right) = \Phi(a) \ ,$$

or equivalently,

$$\lim_{N\to\infty} P\left(\overline{Z}_N \le \mu + \frac{\sigma a}{\sqrt{N}}\right) = \Phi(a) \ ,$$

0.4 Cramér's Theorem

Cramér's Theorem (1938) describes *large deviations* from the mean, and in this case the description depends on the distribution of X_1 through more than merely its mean and variance.

For a real valued random variable X, define the function p_X on $\mathbb{R}$ by

$$p_X(\lambda) = \log\left(\mathrm{E}e^{\lambda X}\right) \tag{0.9}$$

and define $s_X(x)$ by

$$s_X(x) = \sup_{\lambda>0}\{\lambda x - p_X(\lambda)\} \ . \tag{0.10}$$

0.7 THEOREM (Cramér's Theorem)**.** *Let* $\{X_j\}_{j\in\mathbb{N}}$ *be an infinite sequence of independent identically distributed real values randoms variables on some probability space. For all* $x \in \mathbb{R}$, *the sequence* $\left\{\frac{1}{N}\log\Pr\left\{\overline{X}_N > x\right\}\right\}_{N\in\mathbb{N}}$ *converges as* $N \to \infty$, *and*

$$\lim_{N\to\infty} \frac{1}{N}\log\Pr\left\{\overline{X}_N > x\right\} = -s_{X_1}(x) \ , \tag{0.11}$$

and for all N,

$$\frac{1}{N}\log\Pr\left\{\overline{X}_N > x\right\} \le -s_{X_1}(x) \tag{0.12}$$

0.8 Remark. For any non-constant random variable X, the function $p_X(\lambda)$ is strictly convex on any interval on which is it finite: Differentiating twice, we find

$$p'_X(\lambda) = \frac{\mathrm{E}(Xe^{\lambda X})}{\mathrm{E}(e^{\lambda X})} \qquad \text{and} \qquad p''_X(\lambda) = \frac{\mathrm{E}(X^2e^{\lambda X})}{\mathrm{E}(e^{\lambda X})} - \frac{(\mathrm{E}(Xe^{\lambda X}))^2}{(\mathrm{E}(e^{\lambda X}))^2} \ . \tag{0.13}$$

In particular, taking $\lambda = 0$, and assuming that $p_X(\lambda)$ is finite on an open interval about $\lambda = 0$,

$$p'_X(0) = \mathrm{E}(X) \qquad \text{and} \qquad p''_X(0) = \mathrm{Var}(X) \ . \tag{0.14}$$

To see that $p''_X(\lambda) > 0$ for other λ, note that, as a consequence of (0.14),

$$\mathrm{E}(e^{\lambda X})p''_X(\lambda) = \mathrm{E}\left(\left(X - \frac{e^{\lambda X}}{\mathrm{E}(e^{\lambda X})}\right)^2 e^{\lambda X}\right) > 0 \ .$$

The strict convexity proved in the previous remark is very helpful in computing the function $s_X(x) = \sup_{\lambda>0}\{\lambda x - p_X(\lambda)\}$ that we need to compute to apply Cramér's Theorem:

0.9 LEMMA. *Fix $x \in \mathbb{R}$, Suppose that there exists a value λ_0 in the interval on which $p_X(\lambda)$ is finite such that*

$$x = p'_X(\lambda_0) \ . \tag{0.15}$$

If $x > \mathrm{E}(X)$, then $\lambda_0 > 0$, and

$$s_X(x) = \sup_{\lambda>0}\{\lambda x - p_X(\lambda)\} = \lambda_0 x - p_X(\lambda_0) \ . \tag{0.16}$$

On the other hand, if $x < \mathrm{E}(X)$, then $\lambda_0 < 0$, and

$$s_X(x) = \sup_{\lambda>0}\{\lambda x - p_X(\lambda)\} = 0 \ . \tag{0.17}$$

Proof. For fixed x, define the function $\varphi(\lambda) = \lambda x - p_X(\lambda)$. Then $\varphi(\lambda)$ is strictly concave, and hence then ϕ' is strictly decreasing: If λ_0 is such that $\varphi'(\lambda_0) = 0$, then $\phi'(\lambda) > 0$ for $\lambda < \lambda_0$, and $\phi'(\lambda) < 0$ for $\lambda > \lambda_0$. Therefore, if $\lambda_0 > 0$, then

$$\sup_{\lambda>0}\{\lambda x - p_X(\lambda)\} = \lambda_0 x - p_X(\lambda_0) \ ,$$

while if $\lambda_0 < 0$, the supremum is attained by letting λ get as close as possible to λ_0; i.e., by taking $\lambda = 0$. Then since $p_X(0) = 0$,

$$\sup_{\lambda>0}\{\lambda x - p_X(\lambda)\} = 0.$$

Next, since $0 = \varphi'(\lambda_0) = x - p'_X(\lambda_0)$, we have $p'_X(\lambda_0) = x$, and we have seen above that $p'_X(0) = \mathrm{E}(X)$. Since $p_X(\lambda)$ is strictly convex, $p'_X(\lambda)$ is strictly increasing, and so $\lambda_0 > 0$ if and only if $x > \mathrm{E}(X)$. □

Lemma 0.9 tells us the following: For $x > \mathrm{E}(X)$, to compute $s_X(x)$, we should try to solve the equation $x = p_X(\lambda)$. If we find a solution λ_0, then $s(x) = \lambda_0 x - p_X(\lambda_9)$. It also tells us that if $x < \mathrm{E}(X)$, then $s_X(x) = 0$.

0.10 EXAMPLE. *The most basic example is already very interesting. Suppose that* $\{X_j\}_{j\in\mathbb{N}}$ *is an i.i.d. sequence of fair coin tossing variables. Then for each* j, X_j *takes values in* $\{-1,1\}$ *and*

$$\Pr\{X_1 = 1\} = \Pr\{X_1 = -1\} = 1/2 \ , \tag{0.18}$$

so that $\{X_j\}_{j\in\mathbb{N}}$ *models a sequence of fair coin tosses.*

We compute $\mathrm{E}e^{\lambda X_1} = \frac{1}{2}(e^{\lambda} + e^{-\lambda}) = \cosh(\lambda)$ *so that*

$$p_{X_1}(\lambda) = \log(\cosh(\lambda)) \ .$$

Therefore,

$$p'_{X_1}(\lambda) = \frac{\sinh(\lambda)}{\cosh(\lambda)} = \frac{e^{\lambda} - e^{-\lambda}}{e^{\lambda} + e^{-\lambda}} = \frac{1 - e^{-2\lambda}}{1 + e^{-2\lambda}} \ .$$

Now the equation we must solve, (0.15)*, becomes*

$$x = \frac{1 - e^{-2\lambda}}{1 + e^{-2\lambda}}$$

and the solution is

$$\lambda_0 = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right) \ .$$

Therefore, by Lemma 0.9, $s_{X_1}(x) = x\lambda_0 - p_{X_1}(\lambda_0)$*, and calculating*

$$\cosh(\lambda_0) = \frac{1}{2}\left(\left(\frac{1+x}{1-x}\right)^{1/2} + \left(\frac{1-x}{1+x}\right)^{1/2}\right) = \frac{1}{\sqrt{1-x^2}} \ .$$

Therefore,

$$s_{X_1}(x) = \begin{cases} \frac{1}{2}(1+x)\log(1+x) + \frac{1}{2}(1-x)\log(1-x) & 0 < x \le 1 \\ +\infty & x > 1 \ . \end{cases}$$

If Z *is a* $\{-1,1\}$ *valued random variable with* $\Pr\{Z=1\} = p$*, then* $\mathrm{E}Z = 2p-1$ *and if* $\{Z_j\}_{j\in\mathbb{N}}$ *is an i.i.d. sequences of such random variables, and* $\overline{Z}_N$ *denotes its empirical mean, then by the Law of Large Numbers,*

$$\lim_{N\to\infty} \overline{Z}_N = 2p - 1 \ .$$

Thus, one can measure p *by evaluating* $\overline{Z}_N(\omega)$ *for large* N*. What is the probability that we get this wrong, and how does this probability of error depend on* N*? Cramér's Theorem provides the answer. Suppose that we return to our i.i.d. sequence for fair coin tossing, and that we sample* N *times. What is the probability that* $\overline{X}_N \ge 2p - 1$ *for some* $p > 1/2$*?*

By Cramér's Theorem,

$$\lim_{N\to\infty} \frac{1}{N}\log(\Pr\{\overline{X}_N > 2p-1\}) = s_{X_1}(2p-1) = p\log(2p) + (1-p)\log(2(1-p)) \ ,$$

and for all N*,*

$$\Pr\{\overline{X}_N > 2p-1\} \le e^{-Ns_{X_1}(2p-1)}$$

For example, suppose we are testing a truly fair coin by making a sequence of tosses. What is the probability that after N tosses, it looks like a coin for which $p = \Pr\{X_1 = 1\} \geq 0.51$?

We compute

$$s_{X_1}(0.02) = 0.000200013333... \ .$$

Cramér's Theorem gives us the upper bounds

$$\Pr\{\overline{X}_{10^4} > 0.02\} \leq 0.136 \qquad \Pr\{\overline{X}_{10^5} > 0.02\} \leq 0.206{\times}10^{-8} \qquad \Pr\{\overline{X}_{10^6} > 0.02\} \leq 0.137{\times}10^{-87} \ .$$

The probability that a fair coin will appear to have a one percent bias after a million tosses is negligibly small.

0.11 EXAMPLE. *We now consider another example. Let $\{X_j\}$ be independent and identically distributed with $P(X_1 > x) = e^{-x}$. These random variables have mean 1, and variance 1. Since the variance is the same as it is for the coin-tossing variables in the previous example, the Central Limit Theorem gives the same prediction for the size of the fluctuations about the sample mean. However, for $N = 10^5$, noticeable fluctuations above the mean are much more likely in this example than in the coin tossing example, as we shall see. This is due to the long "tail" of the exponential distribution.*

We compute

$$p_{X_1}(\lambda) = \begin{cases} -\log(1-\lambda) & \lambda < 1 \\ \infty & \lambda \geq 1 \ . \end{cases} \ .$$

To maximize $\lambda x - p_{X_1}(\lambda)$, we solve (0.15), which reduces to

$$x = \frac{1}{1-\lambda} \ .$$

This leads to

$$s_{X_1}(x) = \begin{cases} x - 1 - \log x & x > 1 \\ 0 & x < 1 \end{cases} \ .$$

We compute

$$s_{X_1}(1.02) = 0.0001973727... \ .$$

This time we get the bound

$$P(\overline{X}_{10^5} \geq 1.01) \leq 0.268 \times 10^{-8}$$

which is small, but about 20% larger larger than what we found for the coin-tossing example. This is because of the long "tail" of the exponential distribution. Other examples will show even greater differences.

0.5 Proof of Cramér's Theorem

To prove Cramér's Theorem, we make use of the i.i.d. property to write

$$(\mathrm{E}e^{\lambda X_1})^N = \mathrm{E}\left(\prod_{j=1}^{N} e^{\lambda X_j}\right) = \mathrm{E}e^{N\lambda \overline{X}_N} \ .$$

Therefore,

$$\begin{aligned} p_{X_1}(\lambda) &= \frac{1}{N}\log\left(\mathrm{E}e^{N\lambda\overline{X}_N}\right) \\ &\geq \frac{1}{N}\log\left(e^{N\lambda x}\mathrm{Pr}\{e^{N\lambda\overline{X}_N} \geq e^{N\lambda x}\}\right) \\ &= \lambda x + \frac{1}{N}\log\left(\mathrm{Pr}\{\overline{X}_N \geq x\}\right) \ . \end{aligned}$$

In this last equality, we have used the positivity of λ: Since $\lambda > 0$, $e^{N\lambda y} \geq e^{N\lambda x}$ if and only if $y \geq x$. If λ we negative, we would have instead that $e^{N\lambda y} \geq e^{N\lambda x}$ if and only if $y \leq x$.

$$-\frac{1}{N}\log\left(\mathrm{Pr}\{\overline{X}_N \geq x\}\right) \geq \lambda x - p_{X_1}(\lambda) \ .$$

Since this is true for all $\lambda > 0$,

$$\frac{1}{N}\log\left(\mathrm{Pr}\{\overline{X}_N \geq x\}\right) \leq -p^*_{X_1}(x) \ .$$

It remains to show that $\lim_{N\to\infty} \frac{1}{N}\log\left(\mathrm{Pr}\{\overline{X}_N \geq x\}\right)$ and then that

$$\lim_{N\to\infty} \frac{1}{N}\log\left(\mathrm{Pr}\{\overline{X}_N \geq x\}\right) \geq -p^*_{X_1}(x) \ .$$

However, we have completed the proof of the upper bound that we used above.